



**HAL**  
open science

## Gestion des données manquantes

Vincent Audigier

► **To cite this version:**

Vincent Audigier. Gestion des données manquantes. 29èmes Rencontres de la Société Francophone de Classification, Société Francophone de Classification, Sep 2024, Marseille (CIRM, Centre International de Rencontres Mathématiques), France. hal-04696568

**HAL Id: hal-04696568**

**<https://hal.science/hal-04696568v1>**

Submitted on 13 Sep 2024

**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.

# Missing values

**V. Audigier**

*CNAM, CEDRIC-MSDMA, Paris*

29èmes Rencontres de la SFC, September 11th, 2024

# Missing values are everywhere

- unanswered questions in a survey
- lost data
- damaged plants
- machines that fail
- data integration
- ...

# Missing values

## Two kinds of missing values

- values which are not observed, but exist **Ex: measure of temperature is missing for one patient**
- values that are not observed because they would have no sense! **Ex: type of therapy followed by a healthy patient**

Here, we focus on values that are not observed, but exist.

# Example: Ozone data

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
0601	NA	15.6	18.5	18.4	4	4	8	NA	-1.7101	-0.6946	84
0602	82	17	18.4	17.7	5	5	7	NA	NA	NA	87
0603	92	NA	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82
0604	114	16.2	NA	NA	1	1	0	NA	NA	NA	92
0605	94	17.4	20.5	NA	8	8	7	-0.5	NA	-4.3301	114
0606	80	17.7	NA	18.3	NA	NA	NA	-5.6382	-5	-6	94
0607	NA	16.8	15.6	14.9	7	8	8	-4.3301	-1.8794	-3.7588	80
0610	79	14.9	17.5	18.9	5	5	4	0	-1.0419	-1.3892	NA
0611	101	NA	19.6	21.4	2	4	4	-0.766	NA	-2.2981	79
0612	NA	18.3	21.9	22.9	5	6	8	1.2856	-2.2981	-3.9392	101
0613	101	17.3	19.3	20.2	NA	NA	NA	-1.5	-1.5	-0.8682	NA
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0919	NA	14.8	16.3	15.9	7	7	7	-4.3301	-6.0522	-5.1962	42
0920	71	15.5	18	17.4	7	7	6	-3.9392	-3.0542	0	NA
0921	96	NA	NA	NA	3	3	3	NA	NA	NA	71
0922	98	NA	NA	NA	2	2	2	4	5	4.3301	96
0923	92	14.7	17.6	18.2	1	4	6	5.1962	5.1423	3.5	98
0924	NA	13.3	17.7	17.7	NA	NA	NA	-0.9397	-0.766	-0.5	92
0925	84	13.3	17.7	17.8	3	5	6	0	-1	-1.2856	NA
0927	NA	16.2	20.8	22.1	6	5	5	-0.6946	-2	-1.3681	71
0928	99	16.9	23	22.6	NA	4	7	1.5	0.8682	0.8682	NA
0929	NA	16.9	19.8	22.1	6	5	3	-4	-3.7588	-4	99
0930	70	15.7	18.6	20.7	NA	NA	NA	0	-1.0419	-4	NA

<http://www.airbreizh.asso.fr/>

# Problem

- Statistical models cannot be directly fitted on incomplete data
- Deleting incomplete observations (complete-case analysis) is generally irrelevant for many reasons
  - no complete-case
  - lost of data
  - bias
  - lost of power

Missing values cannot be avoided

# Outline

- 1 Introduction
- 2 Modelling with NA
  - Notations
  - Several mechanisms
  - Checking assumptions
- 3 Handling missing values by imputation
  - Univariate case
    - Single imputation
    - Multiple imputation
  - Multivariate case
  - Diagnostics
- 4 Beyond homogeneous data
- 5 Conclusion

## Notations and vocabulary (1)

- $n$ : number of individuals,  $p$ : number of variables
- $\mathbf{X}_{n \times p}$ : the full data matrix (partially unknown)
- $\mathbf{R}_{n \times p}$ : the missing data pattern  $\mathbf{R} = (r_{ij})$  with  $r_{ij} = 0$  if  $x_{ij}$  is missing and 1 otherwise

<b>X</b>	
-0.7	-0.7
0.5	-1.0
0.1	0.4

<b>R</b>	
1	1
1	0
1	0

- $\mathbf{x}_i^{obs}$  observed profile of the individual  $i$  et  $\mathbf{x}_i^{miss}$  the unobserved profile

$$\mathbf{x}_2 = \left( \underbrace{0.5}_{x_2^{obs}}, \underbrace{-1.0}_{x_2^{miss}} \right)$$



## Notations and vocabulary (2)

$\mathbf{X}_{n \times p}$ ,  $\mathbf{R}_{n \times p}$ ,  $\mathbf{x}_i^{obs}$  et  $\mathbf{x}_i^{miss}$  can be seen as realisations of random variables

- $X = (X_1, \dots, X_p)$ : random variables associated to  $\mathbf{X}_{n \times p}$
- $R = (R_1, \dots, R_p)$  random variables associated to  $\mathbf{R}_{n \times p}$
- $X^{obs}$  and  $X^{miss}$ : random variables associated to observed and unobserved parts of  $X$  so that  $X = (X^{obs}, X^{miss})$

$R$  is called the **missing data mechanism**

Handling missing values depends on the relationship between  
 $R$  and  $X$

## Several mechanisms

Three kinds of mechanisms (Rubin, 1976; Little, 1995):

- MCAR (missing completely at random)

$$f\left(R|X^{obs}, X^{miss}; \gamma\right) = f\left(R; \gamma\right)$$

→ lost data

- MAR (missing at random)

$$f\left(R|X^{obs}, X^{miss}; \gamma\right) = f\left(R|X^{obs}; \gamma\right)$$

→ A machine fails when the temperature is elevated

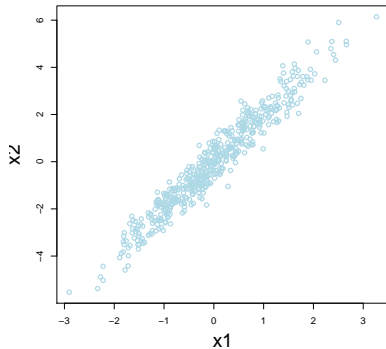
- MNAR (missing not at random)

$$f\left(R|X^{obs}, X^{miss}; \gamma\right) \neq f\left(R|X^{obs}; \gamma\right)$$

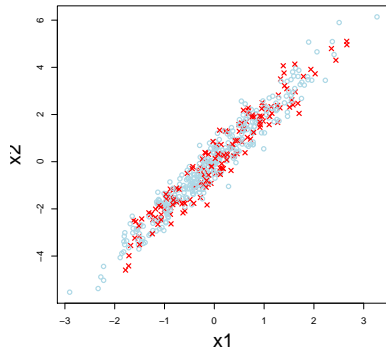
→ A thermometer fails when the temperature is elevated

# Examples

Full data



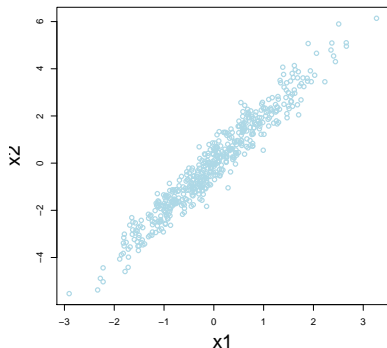
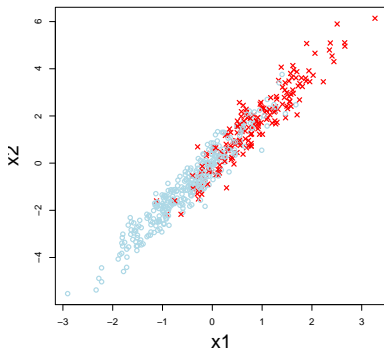
MCAR



$X_1$  always observed  
 $X_2$  incomplete

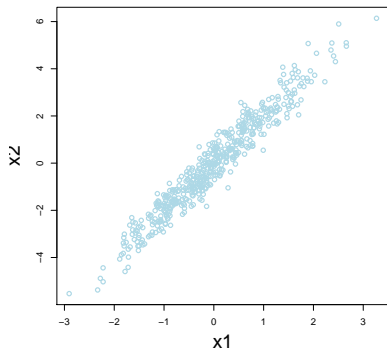
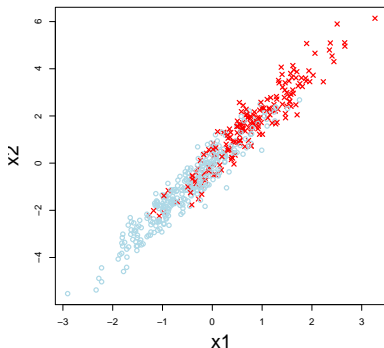
$$f(R_2 | X^{obs}, X^{miss}; \gamma) = \mathcal{B}(0.65)$$

# Examples

**Full data****MAR**

$$f(R_2 | X^{obs}, X^{miss}; \gamma) = \mathcal{B}(\Phi(-1.2 * x^{obs} + 0.5))$$

# Examples

**Full data****MNAR**

$$f(R_2 | X^{obs}, X^{miss}; \gamma) = \mathcal{B}(\Phi(-1.2 * x^{miss} + 0.5))$$

## Reasons for this typology

The type of the mechanism is important

- MCAR: complete individuals are representative of the data set, no bias with CCA
- MAR: complete individuals are not representative, inference could be biased, but valid inference can be obtained by modelling  $X$  only (as usual)
- MNAR: complete individuals are not representative. Valid inference requires modelling  $(X, R)$

# In detail

The full data distribution is  $f(X, R; \theta; \gamma)$ , but inference need to be based on the observed data distribution

$$\begin{aligned}
 f(X^{obs}, R; \theta, \gamma) &= \int f(X, R; \theta, \gamma) dX^{miss} \\
 &= \int f(R|X; \gamma) f(X; \theta) dX^{miss} \\
 &\stackrel{\text{MAR}}{=} \int f(R|X^{obs}; \gamma) f(X; \theta) dX^{miss} \\
 &= f(R|X^{obs}; \gamma) f(X^{obs}; \theta)
 \end{aligned}$$

Under MAR mechanisms, likelihood-based inferences about  $\theta$  can be done ignoring the missing data distribution!

## In practice ...

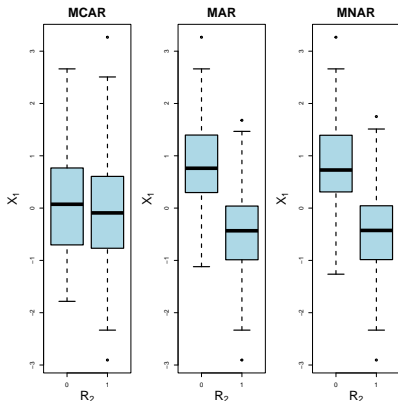
- Exploratory analysis can be used to suggest hypotheses:

Exploratory analysis of  $\mathbf{R}$       Exploratory analysis of  $\mathbf{R}, \mathbf{X}^{obs}$

- proportions per variable
  - Cramer's V
  - MCA
  - bivariate analysis based on CC
  - recode  $\mathbf{x}_i^{miss}$  as a new categories 'missing' and perform (bivariate) multivariate analysis.
- The relationship with  $X^{miss}$  is unknown
  - A knowledge of the data is required to valid the mechanism



# Example



In practice

- **MAR** assumption is often made **by default**
- A **sensitivity analysis** is performed

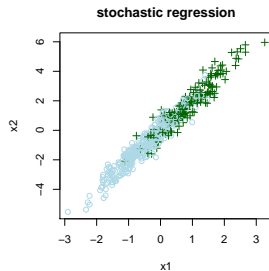
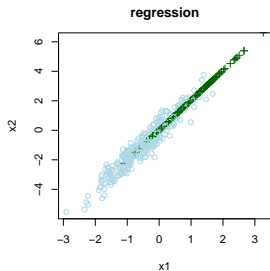
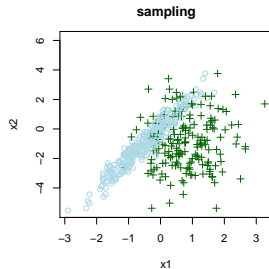
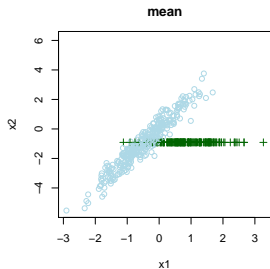
# Outline

- 1 Introduction
- 2 Modelling with NA
  - Notations
  - Several mechanisms
  - Checking assumptions
- 3 Handling missing values by imputation
  - Univariate case
    - Single imputation
    - Multiple imputation
  - Multivariate case
  - Diagnostics
- 4 Beyond homogeneous data
- 5 Conclusion

# Single imputation

- **Imputation** consists in replacing missing values by plausible values
- **Single imputation** consists in replacing by one unique value
- Assuming only one (continuous) variable is incomplete, several classical methods can be used:
  - mean
  - median
  - sampling observed data
  - regression
  - stochastic regression
  - PCA

# Examples

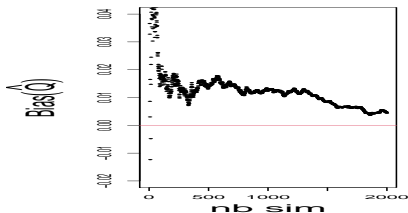
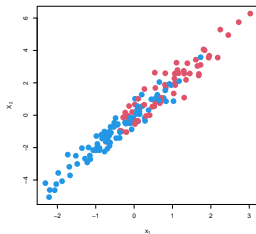


# Typology of single imputation methods

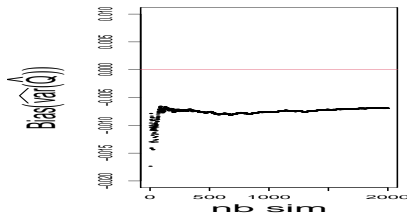
- parametric (ex: stochastic regression)
  - advantages: performs well on small datasets
  - drawbacks: sensitive to the model specification
- non-parametric (ex: knn, random forest)
  - advantages: preserves the nature of the variables
  - drawbacks: requires a large number of individuals
- semi-parametric (ex: predictive mean matching)
  - advantages: preserves the nature of the variables, more robust to model misspecification
  - drawbacks: requires a moderate number of individuals

# Single imputation is a limited approach

- $n = 150, p = 2$
- missing values on  $X_2$  (MAR)
- parameter:  $Q = \mathbb{E}[X_2]$



- The estimator of  $Q$  is unbiased

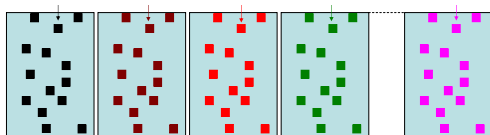


- The estimator of  $\text{Var}(\hat{Q})$  is downwardly biased

# Multiple imputation (Rubin, 1987)

- 1 Generate a set of  $M$  parameters  $(\theta_m)_{1 \leq m \leq M}$  of an **imputation model** to generate  $M$  plausible imputed data sets

$$f(X^{miss} | X^{obs}, \theta_1) \quad \dots \quad \dots \quad \dots \quad f(X^{miss} | X^{obs}, \theta_M)$$



- 2 Fit the **analysis model** on each imputed data set:  $\hat{Q}_m, \widehat{\text{Var}}(\hat{Q}_m)$

- 3 Combine the results:  $\hat{Q} = \frac{1}{M} \sum_{m=1}^M \hat{Q}_m$

$$\widehat{\text{Var}}(\hat{Q}) = \frac{1}{M} \sum_{m=1}^M \widehat{\text{Var}}(\hat{Q}_m) + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^M \left(\hat{Q}_m - \hat{Q}\right)^2$$

⇒ Provide estimation of the parameters and of their variability

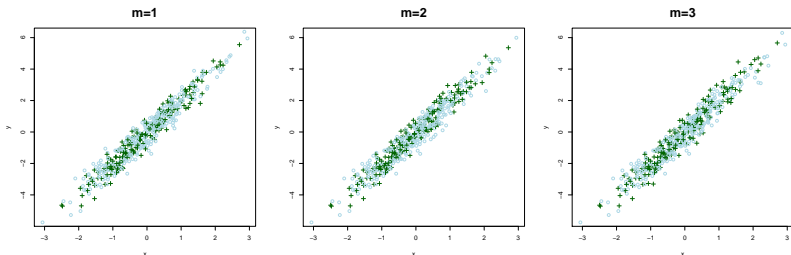
# Generation of $(\theta_m)_{1 \leq m \leq M}$

- Bayesian
  - Prior distribution  $p(\theta)$
  - Derive the posterior distribution  $p(\theta|X^{obs})$   
(Data-Augmentation)
  - Draw from  $p(\theta|X^{obs})$   $M$  times
- Non-parametric Bootstrap
  - Sampling observations with replacement  $M$  times
  - Estimate  $\theta_m$  from each one (EM)
- ...



# Illustration

## 1 Non-parametric Bootstrap + stochastic regression



```
1 library(mice)
2 res.mice <- mice(don, m = 3, method = "norm.boot")
```

## 2 Estimate $Q = E(X_2)$ and $Var(\hat{Q})$ from each imputed table

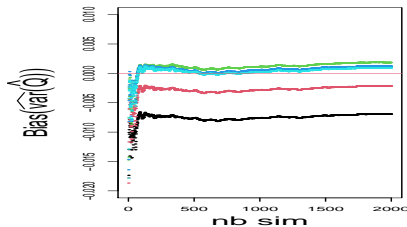
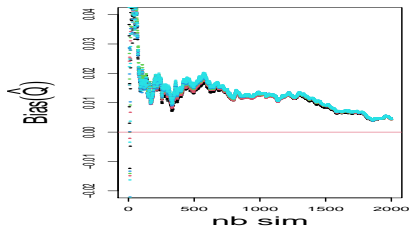
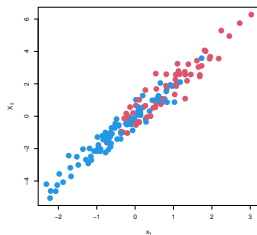
```
1 res.with <- with(res.mice, lm(X2~1))
```

## 3 Aggregate results

```
1 pool(res.with)
```

# Example

- $n = 150, p = 2$
- missing values on  $X_2$  (MAR)
- parameter:  $\mathbb{E}[X_2]$



— SI                      — MI improper  
— MI proper bayes    — MI proper boot    — MI proper asympt

# Outline

- 1 Introduction
- 2 Modelling with NA
  - Notations
  - Several mechanisms
  - Checking assumptions
- 3 Handling missing values by imputation
  - Univariate case
    - Single imputation
    - Multiple imputation
  - Multivariate case
  - Diagnostics
- 4 Beyond homogeneous data
- 5 Conclusion

# Joint modelling

When several variables are incomplete, a multivariate **imputation model**  $f(X; \theta)$  should be theoretically specified (joint modelling)

Some examples (Schafer, 1997):

- continuous data: multivariate gaussian distribution
- categorical data: log-linear model
- mixed data: general location model
- ...

In practice, joint models generally fit the data poorly

# Fully conditional specification

Instead of specifying one joint distribution  $f(X; \theta)$ , a conditional distribution is specified for each (incomplete) variable

$$f(X_j | X_{-j}; \theta_j)$$

$$\text{Ex: } f(X_j | X_{-j}; \theta_j) = \mathcal{N}(X_{-j}\beta, \sigma^2) \quad \theta_j = (\beta, \sigma)$$

To impute the  $m$ th data set

- initialize  $\mathbf{x}_i^{\text{miss}}$  for all  $i$
- for  $j$  in 1 ...  $p$ 
  - a generate  $\theta_j$  based on observed individuals on  $X_j$
  - b impute  $X_j^{\text{miss}}$  according to  $f(X_j | X_{-j}; \theta_j)$
- repeat until convergence

# FCS pros and cons

## Pros

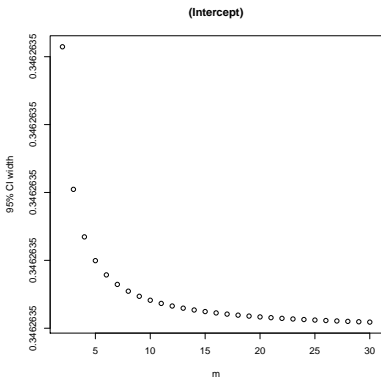
- sparsity
- accounting for interactions effects
- addressing outliers
- semi or non-parametric models
- handling categorical or mixed data

## Cons

- time consuming
- no theoretical guaranties (except in specific cases)
- checking convergence is not possible with a large number of variables

# How many imputed data sets?

- Confidence intervals are **valid** and inference **unbiased** for  $M \geq 2$
- Large value for  $M$  is more time consuming
- $M$  modifies the width of the confidence interval
- In practice,  $M$  between 5 and 100



```

1 library(mice)
2 res.mice <- mice.par(don, m = 30,
   method = "norm.boot", nnodes =
   8)
3 res.with <- with(res.mice, lm(y~1))
4 plot(res.with)

```

# Model fitting

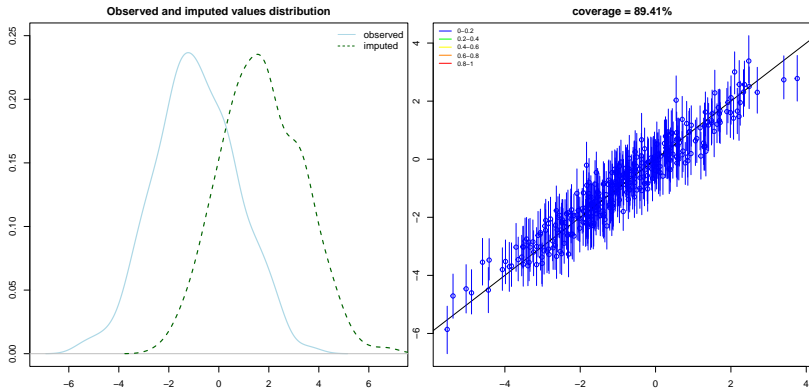


FIG.: Comparative distributions and overimputation

```
1 res.mice <- mice.par(don, m = 200, nnodes = 8)
2 overimpute(res.mice)
```



# Outline

- 1 Introduction
- 2 Modelling with NA
  - Notations
  - Several mechanisms
  - Checking assumptions
- 3 Handling missing values by imputation
  - Univariate case
    - Single imputation
    - Multiple imputation
  - Multivariate case
  - Diagnostics
- 4 Beyond homogeneous data
- 5 Conclusion

# Example: GREAT data

- 28 centres, 11685 patients
- 10 variables (patient characteristics and potential risk factors)
- sporadically and systematically missing data

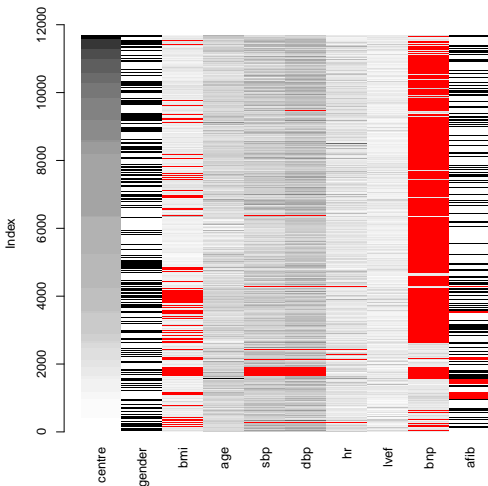


FIG.: Missing data pattern for GREAT

# Modelling

**Aim:** explain the relationship between biomarkers (BNP, AFIB,...) and the left ventricular ejection fraction (LVEF)

**Analysis model**

$$x_{ik}^{LVEF} = \beta^0 + \beta^1 x_{ik}^{BNP} + \beta^2 x_{ik}^{AFIB} + b_k^0 + b_k^1 x_{ik}^{BNP} + \varepsilon_{ik}$$
$$b_k \sim \mathcal{N}(0, \Psi) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$$

$\hat{\beta}$  and associated variability  $var(\hat{\beta})$

The **imputation model** (joint or conditional) needs to

- account for the **heterogeneity** between clusters
- account for the **types** of variables (continuous and binary)
- be identifiable with **sporadically and systematically** missing values

# FCS-GLM (Jolani, 2017)

## Conditional imputation models

$$y_{ik} = \mathbf{z}_{ik}\boldsymbol{\beta} + \mathbf{w}_{ik}\mathbf{b}_k + \varepsilon_{ik} \quad \mathbf{b}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$$

For each incomplete variable

- 1 generate  $\theta_m = (\boldsymbol{\beta}_m, \boldsymbol{\Psi}_m, \sigma^2_m)$   $1 \leq m \leq M$ 
  - prior: non-informative (Jeffreys)
  - posterior distribution

$$\sigma^2 | Y, \mathbf{b} \sim \text{Inv-}\Gamma\left(\frac{n-p}{2}, \frac{(n-p)\widehat{\sigma}^2}{2}\right)$$

$$\boldsymbol{\beta} | Y, \mathbf{b}, \sigma^2 \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}, \widehat{\text{var}}(\widehat{\boldsymbol{\beta}}))$$

$$\boldsymbol{\Psi}^{-1} | Y, \mathbf{b} \sim \mathcal{W}(K, \widehat{\mathbf{b}}\widehat{\mathbf{b}}^\top)$$

- 2 impute in each cluster  $k$  with **systematically missing data**
  - draw  $\mathbf{b}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_m)$
  - impute data according to the **imputation model**

# FCS-GLM (Jolani, 2017)

## Conditional imputation models

$$y_{ik} = \mathbf{z}_{ik}\boldsymbol{\beta} + \mathbf{w}_{ik}\mathbf{b}_k + \varepsilon_{ik} \quad \mathbf{b}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$$

For each incomplete variable

- generate  $\theta_m = (\boldsymbol{\beta}_m, \boldsymbol{\Psi}_m, \sigma^2_m)$   $1 \leq m \leq M$ 
  - prior: non-informative (Jeffreys)
  - posterior distribution

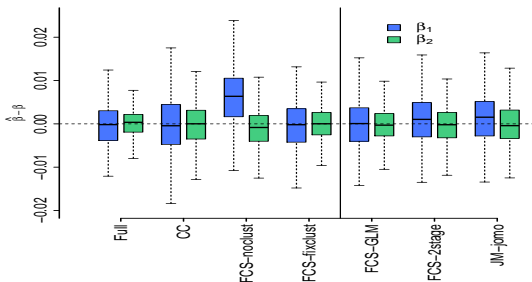
$$\sigma^2 | Y, \mathbf{b} \sim \text{Inv-}\Gamma\left(\frac{n-p}{2}, \frac{(n-p)\widehat{\sigma}^2}{2}\right)$$

$$\boldsymbol{\beta} | Y, \mathbf{b}, \sigma^2 \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}, \widehat{\text{var}}(\widehat{\boldsymbol{\beta}}))$$

$$\boldsymbol{\Psi}^{-1} | Y, \mathbf{b} \sim \mathcal{W}(K, \widehat{\mathbf{b}}\widehat{\mathbf{b}}^\top)$$

- impute in each cluster  $k$  with **sporadically missing data**
  - draw  $\mathbf{b}_k \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{b}_k|y_k}, \boldsymbol{\Psi}_{\mathbf{b}_k|y_k})$
  - impute data according to the **imputation model**

# Results for base-case configuration



Method	$\sqrt{\widehat{\text{var}}(\hat{\beta})}$		$\sqrt{\text{var}(\hat{\beta})}$		95% Cover		Time (min)
	$\beta_1$	$\beta_2$	$\beta_1$	$\beta_2$	$\beta_1$	$\beta_2$	
Full	0.0047	0.0029	0.0048	0.0030	93.8	94.2	
CC	0.0070	0.0053	0.0071	0.0053	92.2	94.4	
FCS-noclust	0.0041	0.0043	0.0067	0.0045	58.2	92.0	0.9
FCS-fixclust	0.0043	0.0043	0.0058	0.0042	87.0	94.6	1.1
FCS-GLM	0.0047	0.0046	0.0057	0.0043	89.7	95.8	103.3
FCS-2stage	0.0059	0.0049	0.0058	0.0044	95.0	96.2	0.9
JM-jomo	0.0066	0.0069	0.0056	0.0049	98.4	97.6	7.8

# Outline

- 1 Introduction
- 2 Modelling with NA
  - Notations
  - Several mechanisms
  - Checking assumptions
- 3 Handling missing values by imputation
  - Univariate case
    - Single imputation
    - Multiple imputation
  - Multivariate case
  - Diagnostics
- 4 Beyond homogeneous data
- 5 Conclusion

# R packages

## JM

norm  
Amelia  
cat  
mix  
missMDA  
jomo  
DPImputeCont (github)  
NPBayesImputeCat

## FCS

mice, micemd  
mi  
miceFast  
VIM  
clusterMI

More packages at <https://cran.r-project.org/web/views/MissingData.html>



## Take-home message

- **Imputation is not prediction** Multiple imputation aims to fit analysis models (taking into account the missing values uncertainty), not to predict missing values
- **Single imputation yields biased standard error**, but unbiased point estimates
- **MI  $\neq$  M single imputations**
- **Pool the analysis results** and never the imputed data sets
- **Use dedicated imputation model**
- **MI is one way to deal with missing values**

# References I

- D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.
- R. J. A. Little. Modelling the drop-out mechanism in repeated measures studies. *Journal of the American Statistical Association*, 90:1112–1121, 1995.
- D. B. Rubin. *Multiple Imputation for Non-Response in Survey*. Wiley, New-York, 1987.
- J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, London, 1997.
- S. Jolani. Hierarchical imputation of systematically and sporadically missing data: An approximate bayesian approach using chained equations. *Biometrical Journal*, 2017. doi: 10.1002/bimj.201600220.
- Anne Gégout-Petit, Myriam Maumy-Bertrand, Gilbert Saporta, and Christine Thomas-Agnan. *Données manquantes*. Editions Technip, June 2022. URL <https://hal-cnam.archives-ouvertes.fr/hal-03696270>.
- Global Research on Acute conditions Team (GREAT) Network. Managing Acute Heart Failure in the ED - Case Studies from the Acute Heart Failure Academy, 2013. <http://www.greatnetwork.org>.
- [http://factominer.free.fr/missMDA/appendix\\_These\\_Audigier.pdf](http://factominer.free.fr/missMDA/appendix_These_Audigier.pdf)
- <https://stefvanbuuren.name/fimd/>