

How to handle missing values?

Model-based approaches

Aude Sportisse

Postdoctoral researcher
Institut 3iA Côte d'Azur & Maasai, Inria
aude.sportisse@inria.fr

November 29, 2022

Overview

1. Introduction
2. Statistical framework in missing-data literature
 - Missing-data pattern
 - Missing-data mechanism
3. EM algorithm for handling missing values

Your viewpoint



Missing values are **everywhere!**

- unanswered questions in a survey,
- lost data,
- sensing machines that fail,
- aggregation of dataset, ...

Take-home message

Growing masses of data + Multiplication of sources

⇒ Not available values, NA

The more data we have, the more missing data we have!

The Traumabase dataset


Trauma.center	Heart rate	Death	Anticoagulant. therapy	Glascow score	...
Pitie-Salpêtrière	88	0	No	3	
Beaujon	103	0	NA	5	
Bicêtre	NA	0	Yes	6	
Bicêtre	NA	0	No	NA	
Lille	62	0	Yes	6	
Lille	NA	0	No	NA	
⋮	⋮	⋮	⋮	⋮	

250 clinical
variables
(heterogeneous)

1 patient; in total: 30 000 patients

The Traumabase dataset

Trauma.center	Heart rate	Death	Anticoagulant. therapy	Glascow score	...
Pitie-Salpêtrière	88	0	No	3	
Beaujon	103	0	NA	5	
Bicêtre	NA	0	Yes	6	
Bicêtre	NA	0	No	NA	
Lille	62	0	Yes	6	
Lille	NA	0	No	NA	
⋮	⋮	⋮	⋮	⋮	



**23 different
hospitals**

The Traumabase dataset

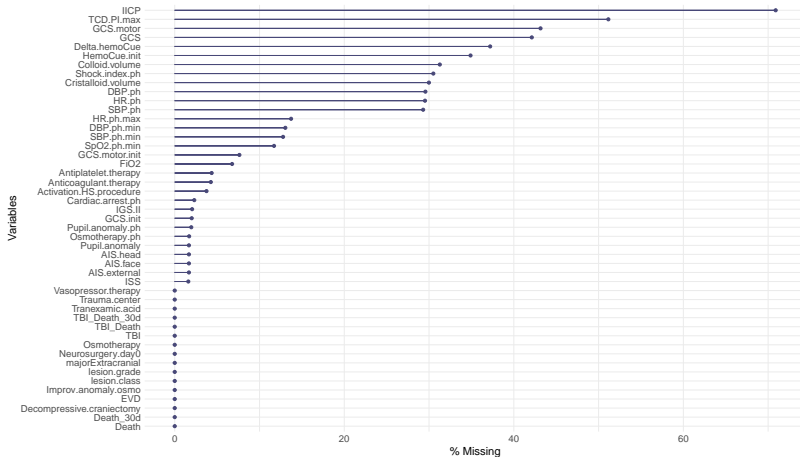


Figure: Percentage of missing values for 40 variables.

The Traumabase dataset

Traumabase[®] dataset

- now **30 000** patients (in 2018: 10 000).
 - **250** heterogeneous variables: continuous, categorical, ordinal,...
 - **23** different hospitals
 - **missing** values everywhere (1% to 90% NA in each variable).
-
- **Imputation:** provide a **complete dataset** to the doctors.
 - **Estimation:** explain the level of platelet with pre-hospital characteristics.
 - **Prediction:** predict the administration or not of the tranexomic acid.
 - **Clustering:** identify relevant groups of patients sharing similarities.

Question: How to deal with missing values? A first naive idea?

What we should not do

Pitie-Salpêtrière	88	0	No	3
Beaujon	103	0	NA	5
Bicêtre	NA	0	Yes	6
Bicêtre	NA	0	No	NA
Lille	62	0	Yes	6
Lille	NA	0	No	NA

What we should not do: ~~discard individuals~~

Discarding individuals with missing values **is not** a solution

- Loss of information .

Traumabase[®]: only 5% of the rows are kept.

- Bias in the analysis .

Kept observations: sub-population **not necessarily representative** of the overall population.

What we should not do: ~~discard individuals~~

Example:

- We consider a bivariate Gaussian variable. $X \sim \mathcal{N}(\mu, \Sigma)$, with

$$\mu = \begin{pmatrix} 5 \\ -1 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

- X_2 is missing.
- We estimate μ_2 with the empirical mean in the complete case.
- see Rmarkdown!

What we should not do: ~~discard individuals~~

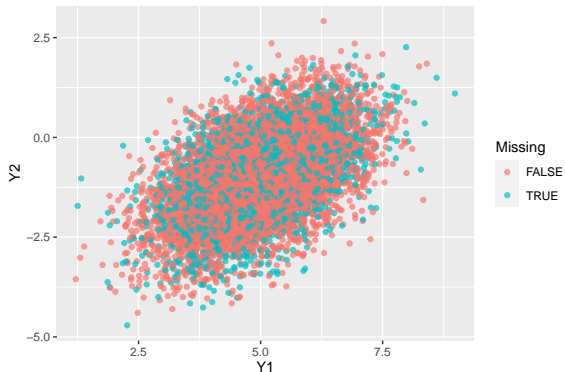


Figure: The sub-population is representative of the overall population.

What we should not do: ~~discard individuals~~

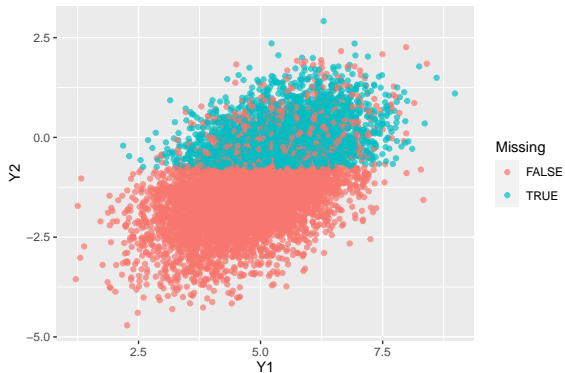


Figure: The sub-population is **not** representative of the overall population.

Need for assumption

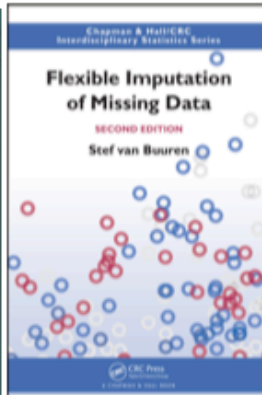
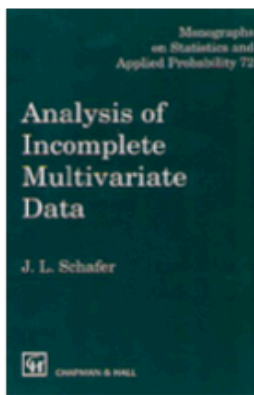
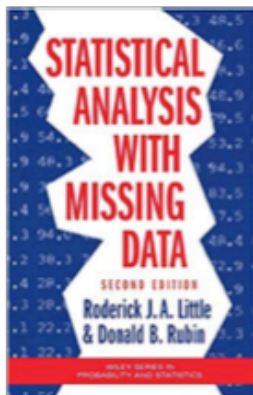
Example: survey with two variables, Income and Age, with missing values only on Income.

- Poor and rich respondents would be less incline to reveal their income.
- There are missing values for the smallest and highest values of Income.
- Even though Age and Income are related, the process that causes the missing data is not fully explained by Age.
- Knowing the value of Age is not enough to retrieve the value of Income.

Take-home message

- Knowing **why** the data is missing is an important issue.
- The process that causes the missing data should be modeled in some situations.

Main references



Goal of this course¹

This is only an **introduction** to missing data.

- Dangers of naive methods in the analysis,
- Importance of the missing-data mechanism (*type of missing data*),
- EM algorithm for handling missing data (+ R code session),
- **To be continued next week** (*other methods, comparison of the methods on a real dataset*).

¹Inspired by the courses of Pierre-Alexandre Mattei (2019-2020) and Julie Josse (2020) on missing values.

1. Introduction

2. Statistical framework in missing-data literature

- Missing-data pattern

- Missing-data mechanism

3. EM algorithm for handling missing values

A statistical framework for incomplete data

$$X = \underbrace{\begin{pmatrix} 30 & 100 & 61 \\ 85 & 31 & 50 \end{pmatrix}}_{\text{not observed}} \quad X^{\text{NA}} = \underbrace{\begin{pmatrix} 30 & \text{NA} & 61 \\ \text{NA} & \text{NA} & 50 \end{pmatrix}}_{\text{observed}}$$

We observe also where are the missing values in X^{NA} .

Definition: missing-data pattern (mask)

$M \in \{0, 1\}^{n \times d}$: indicates where are the missing values in X^{NA} .

$$\forall i, j, \quad M_{ij} = \begin{cases} 1 & \text{if } X_{ij}^{\text{NA}} \text{ is missing,} \\ 0 & \text{otherwise.} \end{cases}$$

A statistical framework for incomplete data

$$X = \underbrace{\begin{pmatrix} 30 & 100 & 61 \\ 85 & 31 & 50 \end{pmatrix}}_{\text{not observed}} \quad X^{\text{NA}} = \underbrace{\begin{pmatrix} 30 & \text{NA} & 61 \\ \text{NA} & \text{NA} & 50 \end{pmatrix}}_{\text{observed}} \quad M = \underbrace{\begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}}_{\text{observed}}$$

Question: What to model?

- model $p(X^{\text{NA}})$: too difficult because the entries $X_{ij}^{\text{NA}} \in \mathbb{R} \cup \{\text{NA}\}$ (semi-discrete set).
- model $p(X, M)$: entries are in a well-behaved mathematical set $\mathbb{R}^{n \times d} \cup \{1, 0\}^{n \times d}$

Model the joint distribution (X, M)

We want to model the **joint** distribution of the data X and the missing-data pattern M .

The observations are assumed to be i.i.d., i.e. $(X_1, M_1), \dots, (X_n, M_n)$ have the same distribution and are independent

$$p(X, M) = \prod_{i=1}^n p(X_i, M_i).$$

Model the joint distribution (X, M)

We want to model the **joint** distribution of the data X and the missing-data pattern M .

Selection model factorization

$$p(X, M) = p(X)p(M|X)$$

where

- $p(X)$: distribution of the data,
- $p(M|X)$: conditional distribution of the missing-data pattern given the data, it is the **missing-data mechanism**.

Parametric approach:

$$p(X, M; \theta, \phi) = p(X; \theta)p(M|X; \phi)$$

where $\theta \in \Omega_\theta$ and $\phi \in \Omega_\phi$.

Missing-data mechanism (Rubin, 1976)

Missing Completely At Random (MCAR)

$$p(M|X; \phi) = p(M; \phi)$$

Missing At Random (MAR)

X^{obs} : observed component of X .

$$p(M|X; \phi) = p(M|X^{\text{obs}}; \phi)$$

Missing Not At Random (MNAR)

The MAR assumption does not hold.
The missingness can depend on the missing data value itself.

Question: Which mechanism is realistic? How to choose the right mechanism for real data?

Example of models

$$p(X, M; \theta, \phi) = p(X; \theta)p(M|X; \phi)$$

- For $p(X)$: models seen in the rest of the course, e.g. mixture model, single Gaussian, variational autoencoder, ...
- For $p(M|X)$: typically Logit or Probit distribution.

$$p(M_{ij}|\mathbf{X}_{ij}; \phi) = [(1 + e^{-\phi_{1j}(\mathbf{X}_{ij} - \phi_{2j})})^{-1}]^{M_{ij}} [1 - (1 + e^{-\phi_{1j}(\mathbf{X}_{ij} - \phi_{2j})})^{-1}]^{(1-M_{ij})}.$$

But it is a **strong assumption**. We will see that in some situations, the missing-data mechanism can be *ignored* (not modelled).

Likelihood approach with incomplete data

- Goal of the **parametric estimation**: model the joint distribution (X, M) parametrized by $\theta, \phi \in \Omega_\theta \times \Omega_\phi$.
- Likelihood-approach **without missing data**: maximizing the **full likelihood**

$$L_{\text{full}}(\theta, \phi; X, M) = p(X; \theta)p(M|X; \phi)$$

- Split X into two components X^{obs} (observed features), X^{mis} (missing features).
- Likelihood-approach **with missing data**: maximizing **the full observed likelihood**

$$L_{\text{full,obs}}(\theta, \phi; X^{\text{obs}}, M) = \int L_{\text{full}}(\theta, \phi; X, M) dX^{\text{mis}}$$

Ignorable mechanisms

Question: How can we ignore the missing-data mechanism?

Ignorable mechanisms

For MCAR and MAR data, we can **ignore** the missing-data mechanism:

$$L_{\text{full,obs}}(\theta, \phi; X^{\text{obs}}, M) \propto L_{\text{ign}}(\theta; X^{\text{obs}}) = \int p(X; \theta) dX^{\text{mis}} = p(X^{\text{obs}}; \theta)$$

Take-home message

- M(C)AR: one can ignore the mechanism.
- MNAR: one should consider the mechanism.

Link with the logistic regression

Ignorability in missing-data analysis: to model (X, M) , we can in some cases ignore the mechanism $(M|X)$, by treating ϕ as a nuisance parameter.

→ Similar trick for logistic regression.

- $p(x, y) = p(y|x; \theta)p(x)$ with $p(x)$ which does not involve θ .
- Likelihood written as $L_{\text{full}}(\theta; x, y) = p(y|x; \theta)p(x)$.
- Goal: estimate θ .
- We do not model $p(x)$ because $\hat{\theta} \in \operatorname{argmax}_{\theta} L_{\text{full}}(\theta; x, y) = \operatorname{argmax}_{\theta} p(y|x; \theta)$

Overview

1. Introduction
2. Statistical framework in missing-data literature
 - Missing-data pattern
 - Missing-data mechanism
3. EM algorithm for handling missing values

- Goal: estimate $\theta \in \Omega_\theta$, when X contain **MCAR or MAR** values.
- We can maximize the fully observed **log**-likelihood (logarithm more convenient):

$$\hat{\theta} = \operatorname{argmax}_{\theta} \ell_{\text{ign}}(\theta; X^{\text{obs}}) = \log(p(X^{\text{obs}}; \theta))$$

- When it has no closed form, a solution can be to use the EM algorithm.
Idea: consider the missing variables as latent variables.

Expectation Maximization algorithm (Dempster et al., 1977)

Starting from an initial point θ^0 , the EM algorithm proceeds two steps **iteratively**:

- **E-step:** computation of the expected full log-likelihood knowing the observed data and a current value of the parameters.

$$Q(\theta; \theta^r) = \mathbb{E}[\ell_{\text{full}}(X; \theta) | X^{\text{mis}}, \theta^r]$$

- **M-step:** maximization of $Q(\theta; \theta^r)$ over θ .

$$\theta^{r+1} = \operatorname{argmax}_{\theta} Q(\theta; \theta^r)$$

EM algorithm in a toy example

Consider a Gaussian bivariate variable $X = (X_{.1}^T, X_{.2}^T) \in \mathbb{R}^{n \times 2}$.

$$X \sim \mathcal{N}(\mu, \Sigma),$$

with $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$.

$X_{.2}$ contain some **M(C)AR missing values**. Without loss of generality, assume that X_{i2} is missing, with $r < i \leq n$.

Question: First, we want to know if it is possible to maximize the observed log-likelihood directly. Write the observed log-likelihood.

EM algorithm in a toy example

Question: Write the observed log-likelihood.

Tip: use the classical formula $X_{i2}|X_{i1} \sim \mathcal{N}(\mathbb{E}[X_{i2}|X_{i1}], \text{Var}(X_{i2}|X_{i1}))$ with

$$\mathbb{E}[X_{i2}|X_{i1}] = \mu_2 + \frac{\sigma_{21}}{\sigma_{11}}(X_{i1} - \mu_1)$$

$$\text{Var}(X_{i2}|X_{i1}) = \sigma_{22} - \frac{\sigma_{21}^2}{\sigma_{11}}$$

EM algorithm in a toy example

EM algorithm in a toy example

Question: Write the observed log-likelihood.

In this simple setting, directly maximizing the log-likelihood is possible.

$$\begin{aligned}\ell(X_{.1}, X_{.2}^{\text{obs}}; \mu, \Sigma) = & -\frac{n}{2} \log(\sigma_{11}^2) - \frac{1}{2} \sum_{i=1}^n \frac{(X_{i1} - \mu_1)^2}{\sigma_{11}^2} \\ & - \frac{r}{2} \log \left(\sigma_{22} - \frac{\sigma_{21}^2}{\sigma_{11}} \right)^2 - \frac{1}{2} \sum_{i=1}^r \frac{(X_{i2} - \mu_2 + \frac{\sigma_{21}}{\sigma_{11}}(X_{i1} - \mu_1))^2}{\left(\sigma_{22} - \frac{\sigma_{21}^2}{\sigma_{11}} \right)^2}\end{aligned}$$

More fun: let us derive the EM algorithm!

EM algorithm in a toy example

E-step: computation of the expected full log-likelihood knowing the observed data and a current value of the parameters.

$$Q(\theta; \theta^r) = \mathbb{E}[\ell_{\text{full}}(X; \theta) | X^{\text{mis}}, \theta^r]$$

Question: Write the full log-likelihood (easy question).

EM algorithm in a toy example


Question: Write $Q(\theta; \theta^r)$. What quantities should be computed in the E-step?

EM algorithm in a toy example

M-step: maximization of $Q(\theta; \theta^r)$ over θ .

$$\theta^{r+1} = \operatorname{argmax}_{\theta} Q(\theta; \theta^r)$$

References

-  Little, Roderick JA and Rubin, Donald B (2019)
Statistical analysis with missing data
[John Wiley & Sons.](#)