

# How to handle missing values?

Model-based approaches

Aude Sportisse

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

December 8, 2022

# Overview

1. Summary of the first lesson
2. Code the EM algorithm
3. Other methods to impute missing values

# Basics on missing data analysis

- Deleting missing values **is not** a solution: loss of information and bias in the estimate if the sub-population is not representative of the overall population.
- Two key objects:
  - the **missing-data pattern**  $M$  which indicates where are the missing values,
  - the **missing-data mechanism**  $p(M|X)$  which describes the causal relationships between the missing-data pattern and the data (MCAR, MAR, MNAR)
- Model of the joint distribution  $(X, M)$  but the missing-data mechanism may be **ignored** in some cases.
- Example of a method for parameter estimation (and imputation): EM algorithm.

# Missing-data mechanism (Rubin, 1976)

## Missing Completely At Random (MCAR)

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

## Missing At Random (MAR)

$X^{\text{obs}}$ : 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.

- Example 1: the measuring instrument does not work because the battery is empty.
- Example 2: doctors do not take the time to report the heart rate in the data table when the situation is too critical (patient in a serious situation).
- Example 3: survey with three variables (Heart rate, Weight, Trauma center). The hospital Pitié measures both the heart rate and the weight of the patient. But the hospital Beaujon does not measure the weight.

# Overview

1. Summary of the first lesson
2. Code the EM algorithm
3. Other methods to impute missing values

# 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  $X_{12}, \dots, X_{r2}$  are missing, with  $0 < r < n$ .

# 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.

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

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

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

# EM algorithm in a toy example

- **E-step:** computation of

$$s_1 = \sum_{i=1}^n x_{i1},$$

$$s_{11} = \sum_{i=1}^n x_{i1}^2$$

$$s_2 = \sum_{i=m+1}^n x_{i2} + \sum_{i=1}^m \left( \mu_2^r + \frac{\sigma_{21}^r}{\sigma_{11}^r} (x_{i1} - \mu_1^r) \right)$$

$$s_{22} = \sum_{i=m+1}^n x_{i2}^2 + \sum_{i=1}^m \left( \left( \mu_2^r + \frac{\sigma_{21}^r}{\sigma_{11}^r} (x_{i1} - \mu_1^r) \right)^2 + \sigma_{22}^r - \frac{(\sigma_{21}^r)^2}{\sigma_{11}^r} \right)$$

$$s_{12} = \sum_{i=m+1}^n x_{i1} x_{i2} + \sum_{i=1}^m x_{i1} \left( \mu_2^r + \frac{\sigma_{21}^r}{\sigma_{11}^r} (x_{i1} - \mu_1^r) \right)$$

- **M-step:** update the parameters:  $\mu_1^{r+1} = \frac{s_1}{n}$ ,  $\mu_2^{r+1} = \frac{s_2}{n}$ ,  $\sigma_{11}^{r+1} = \frac{s_{11}}{n} - (\mu_1^{r+1})^2$ ,  $\sigma_{22}^{r+1} = \frac{s_{22}}{n} - (\mu_2^{r+1})^2$  and  $\sigma_{12}^{r+1} = \frac{s_{12}}{n} - (\mu_1^{r+1} \mu_2^{r+1})$ .



# EM algorithm in a toy example

We have seen that the EM algorithm can be used to **estimate the parameters** of the underlying data distribution. **Question:** Can we impute missing values?

## Imputation of the missing values using EM algorithm

We can use the conditional expectation.

$\forall i \in \{1, \dots, n\}$  such that  $M_{ij} = 1$ ,

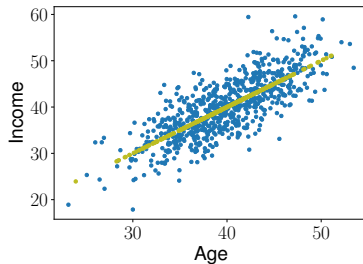
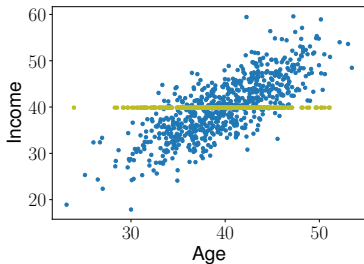
$$X_{i1}^{\text{imp}} = \mathbb{E}[X_{i2}|X_{i1}] = \mu_2 + \frac{\sigma_{21}}{\sigma_{11}}(X_{i1} - \mu_1)$$

# Overview

1. Summary of the first lesson
2. Code the EM algorithm
3. Other methods to impute missing values

# Naive imputation

Mean imputation, performing regression.



**X** bias in the estimates, correlation between the variables overestimated.

# Low rank models

## Definition: low rank matrix

$\Theta \in \mathbb{R}^{n \times d}$  has a *low rank*, if its rank  $r \geq 1$ , referred to as the dimension of the vector space generated by its columns, is small compared to the dimensions  $n$  and  $d$ , i.e. if  $r \ll \min\{n, d\}$ , where  $\ll$  can be interpreted as  $\exists r_{\max} \geq 1, r < r_{\max} < \min\{n, d\}$ .

Low rank models: the dataset  $X$  is a **noisy** realisation of a low rank matrix  $\Theta \in \mathbb{R}^{n \times d}$

$$X = \Theta + \epsilon.$$

- $X$  contain MCAR missing values.
- The goal is to estimate  $\Theta$ .
- Low rank approximation is often relevant: individual profiles can be summarized into a limited number of general profiles, or dependencies between variables can be established.

# Low rank models

Classical methods to handle missing values solve the following optimization problem:

$$\hat{\Theta} \in \operatorname{argmin}_{\Theta} \underbrace{\|(\mathbf{1}_{n \times d} - M) \odot (X - \Theta)\|_F^2}_{\text{to fit the data at best}} + \lambda \underbrace{\|\Theta\|_{\star}}_{\text{to satisfy the low rank constraint}},$$

with  $\lambda > 0$  a regularization term,  $\odot$  the Hadamard product (by convention  $0 \times \text{NA} = 0$ ) and  $\mathbf{1}_{n \times d} \in \mathbb{R}^{n \times d}$  with each of its entry equal to 1.

# softImpute, Hastie et al. (2015)

Iterative algorithm: starting from an initial point  $\Theta^0$ ,

- **Estimation-step:** perform the threshold SVD of the complete matrix

$$X^t = (\mathbf{1}_{n \times d} - M) \odot X + M \odot \Theta^t,$$

which leads to

$$\text{SVD}_\lambda(X^t) = U^t D_\lambda^t V^t,$$

where  $U^t \in \mathbb{R}^{n \times r}$ ,  $V^t \in \mathbb{R}^{r \times d}$  are orthonormal matrices containing the singular vectors of  $X^t$  and  $D_\lambda^t \in \mathbb{R}^{r \times r}$  is a diagonal matrix such that its diagonal terms are  $(D_\lambda^t)_{ii} = \max((\sigma_i - \lambda), 0)$ ,  $i \in \{1, \dots, r\}$ , with  $\sigma_i$  the singular values of  $X^t$ .

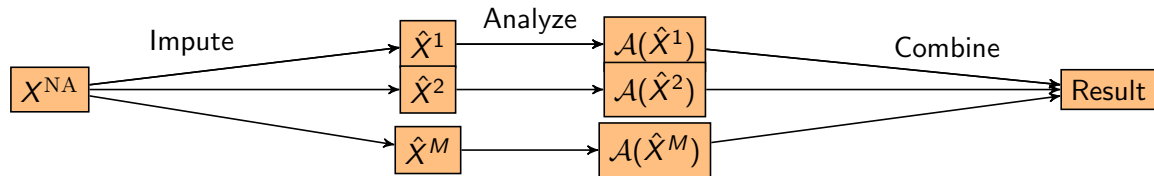
- **Imputation-step:** the entries of  $\Theta^t$  corresponding to missing values in  $X$  are replaced by the values of  $\text{SVD}_\lambda(X^t)$ ,

$$\Theta^{t+1} \odot M = \text{SVD}_\lambda(X^t) \odot M.$$

# Multiple imputation

✗ Single imputation does not reflect the variability of imputation.

- Generating  $M$  plausible values for each missing values:  $M$  complete datasets,  $\hat{X}^1, \dots, \hat{X}^M$ .
- Analysis performed on each imputed data set
- Results are combined.




`mice` (Buuren et al., 2010): use chained equations (iterative conditional distributions assuming a Bayesian framework).

# Summary

Method	Simple to implement	Imputation	Confidence intervals	Main drawbacks
Single imputation	✓	single	✗	biased estimates if too simple imputation
Multiple imputation	✓	multiple	✓	combining results can be delicate
EM	✗	not directly	can be obtained	specific algorithm for each statistical model



# References

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