# Model-based approaches to handle **missing values**

## MSc2 DSAI : Model-based statistical learning

## Aude Sportisse & <u>Vincent Vandewalle</u>

Professor in Applied Mathematics
Université Côte d'Azur
vincent.vandewalle@univ-cotedazur.fr

November 4, 2024

# Overview

# Your viewpoint

Missing values

# 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
> $\Rightarrow$ `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

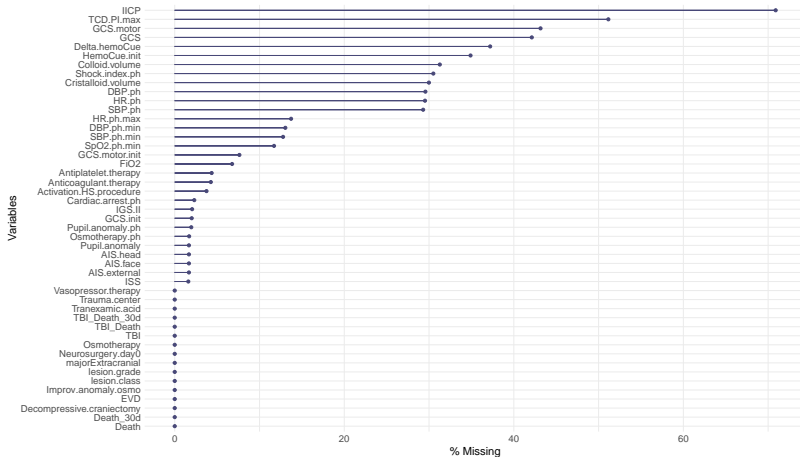| Trauma.center | Heart rate | Death | Anticoagulant. therapy | Glascow score | . . . |
|---|---|---|---|---|---|
| Pitie-Salpêtrière | 88 | 0 | No | 3 | |
| Beaujon | 103 | 0 | NA | 5 | |
| Bicêtre | NA | 0 | Yes | 6 | |
| Bicêtre | NA | 0 | No | NA | |
| Lille | 62 | 0 | Yes | 6 | |
| Lille | NA |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

$$
\begin{pmatrix}
\text{Pitie-Salpêtrière} & 88 & 0 & \text{No} & 3 \\
\text{Beaujon} & 103 & 0 & \text{NA} & 5 \\
\text{Bicêtre} & \text{NA} & 0 & \text{Yes} & 6 \\
\text{Bicêtre} & \text{NA} & 0 & \text{No} & \text{NA} \\
\text{Lille} & 62 & 0 & \text{Yes} & 6 \\
\text{Lille} & \text{NA} & 0 & \text{No} & \text{NA}
\end{pmatrix}
$$

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

- Loss of information .

  Traumabase$^{\circledR}$: only 5% of the rows are kept.

- Bias in the analysis .

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

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 $\mu_2$ with the empirical mean in the complete case.
- see Rmarkdown!
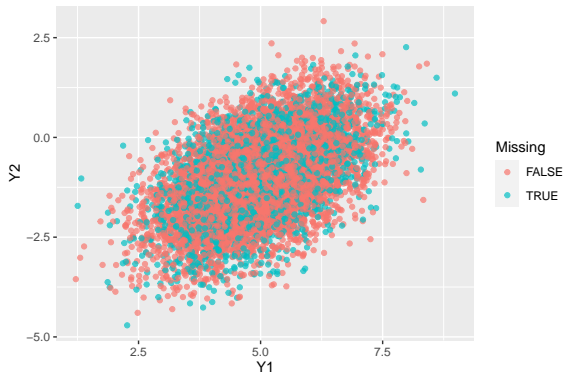
# What we should not do: ~~discard individuals~~



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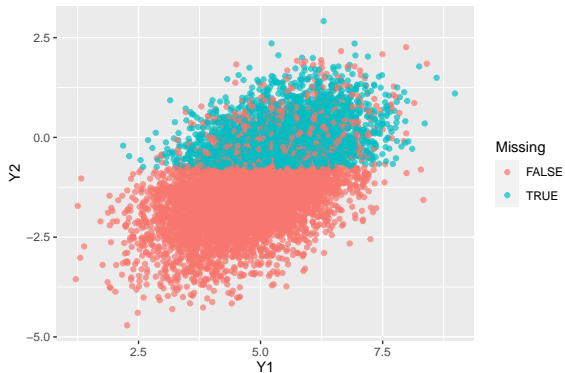
Figure: The sub-population is **not** representative of the overall population.

# 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
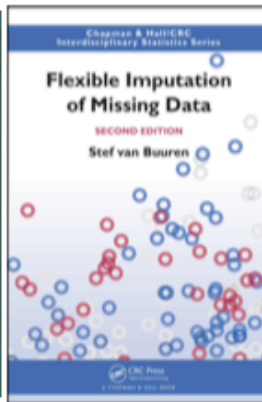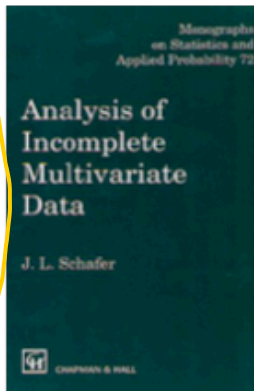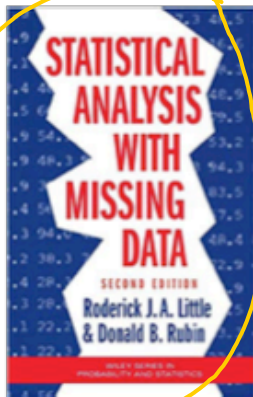
# Goal of this course[1]

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),
- Classical Imputation methods

---

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

# Overview

$$X = \begin{pmatrix} 30 & 100 & 61 \\ 85 & 31 & 50 \end{pmatrix} \qquad X^{\mathrm{NA}} = \begin{pmatrix} 30 & \mathrm{NA} & 61 \\ \mathrm{NA} & \mathrm{NA} & 50 \end{pmatrix}$$

$\underbrace{\hphantom{XXXXXXXXXXXX}}_{\text{not observed}}$ $\underbrace{\hphantom{XXXXXXXXXXXX}}_{\text{observed}}$

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

**Definition:** missing-data pattern (mask)

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

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

$$X = \begin{pmatrix} 30 & 100 & 61 \\ 85 & 31 & 50 \end{pmatrix} \qquad X^{\mathrm{NA}} = \begin{pmatrix} 30 & \mathrm{NA} & 61 \\ \mathrm{NA} & \mathrm{NA} & 50 \end{pmatrix} \qquad M = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

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

**Question:** What to model?

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

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. $(X_1, M_1), \ldots, (X_n, M_n)$ 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)**

$X^{\mathrm{obs}}$: observed component of $X$.

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

**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.

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

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

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

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

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

# 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. ✓  MAR
- MNAR: one should consider the mechanism.

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

$\rightarrow$ Similar trick for logistic regression.

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

# Overview

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

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

- 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 $\theta^0$, 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.

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

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

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

# 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 $r < i \leq n$ (i.e. $r$ values of $X_{.2}$ are observed and $n - r$ values are missing)

**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 $X_{i2}|X_{i1} \sim \mathcal{N}(\mathbb{E}[X_{i2}|X_{i1}], \mathrm{Var}(X_{i2}|X_{i1}))$ with

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

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

# 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.

$$\ell(X_{.1}, X_{.2}^{\mathrm{obs}}; \mu, \Sigma) = -\frac{n}{2} \log(\sigma_{11}) - \frac{1}{2} \sum_{i=1}^{n} \frac{(X_{i1} - \mu_1)^2}{\sigma_{11}}$$

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

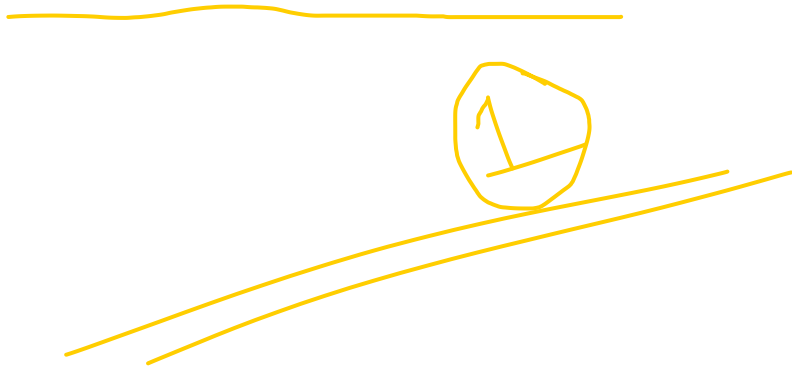More fun: let us derive the EM algorithm!

# 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.

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

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

**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 $\theta$.

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

# Summary: 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 $\theta$.

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

# Summary: EM algorithm in a toy example

- **E-step:** computation of

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

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

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

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

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

- **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 $\sigma_{12}^{r+1} = \frac{s_{12}}{n} - (\mu_1^{r+1} \mu_2^{r+1})$.

# Summary: 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.
$\forall i \in \{1, \ldots, n\}$ such that $M_{ij} = 1$,

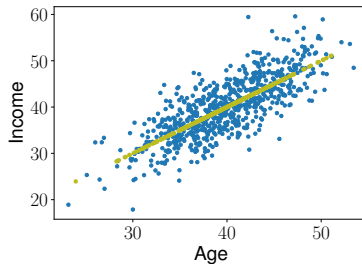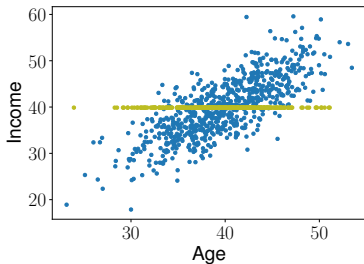$$X_{i2}^{\mathrm{imp}} = \mathbb{E}[X_{i2}|X_{i1}] = \mu_2 + \frac{\sigma_{21}}{\sigma_{11}}(X_{i1} - \mu_1)$$

# Overview

# Naive imputation

Mean imputation, performing regression.



✗ 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 $r \geq 1$, refereed 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:

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

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

# R package `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

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

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 $(D_\lambda^t)_{ii} = \max((\sigma_i - \lambda), 0), i \in \{1, \ldots, r\}$, 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 $\mathrm{SVD}_\lambda(X^t)$,

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

**Iterative Random Forests imputation**

- Initial imputation: mean imputation and sort the variables according to the amount of missing values
- Repeat until convergence:
  - **fit a random forest** with $X_j^{\mathrm{obs}}$ on $X_{-j}^{\mathrm{obs}}$ (all the observed variables except variable $j$) and then predict $X_j^{\mathrm{mis}}$
  - Cycling through variables

# Multiple imputation

✗ Single imputation does not reflect the variability of imputation.

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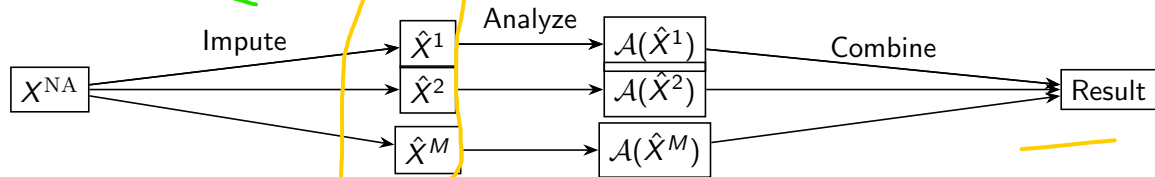


$\Leftrightarrow$ `mice` (Buuren et al., 2010). use chained equations (iterative conditional distributions assuming a Bayesian framework).

# 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.