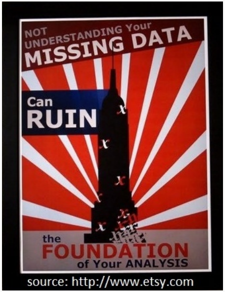
An overview of methods to handle missing values

Julie Josse
INRIA - Ecole Polytechnique

5 November 2020

Séminaire Science des Données Nantes



Introduction

Traumabase project: decision support for trauma patients.

- 20000 trauma patients
- 250 continuous and categorical variables: **heterogeneous**
- 11 hospitals: **multilevel data**
- 4000 new patients/ year

Center	Accident	Age	Sex	Lactactes	BP	Shock	Platelet	...
Beaujon	fall	54	m	NM	180	yes	292000	
Pitie	gun	26	m	NA	131	no	323000	
Beaujon	moto	63	m	3.9	NR	yes	318000	
Pitie	moto	30	w	Imp	107	no	211000	
HEGP	knife	16	m	2.5	118	no	184000	
⋮								⋱

[1] Doubly robust treatment effect estimation with incomplete confounders. Mayer, Wager, J. Annals Of Applied Statistics 2020.

Traumabase project: decision support for trauma patients.

- 20000 trauma patients
- 250 continuous and categorical variables: **heterogeneous**
- 11 hospitals: **multilevel data**
- 4000 new patients/ year

Center	Accident	Age	Sex	Lactactes	BP	Shock	Platelet	...
Beaujon	fall	54	m	NM	180	yes	292000	
Pitie	gun	26	m	NA	131	no	323000	
Beaujon	moto	63	m	3.9	NR	yes	318000	
Pitie	moto	30	w	Imp	107	no	211000	
HEGP	knife	16	m	2.5	118	no	184000	
$\vdots$								$\ddots$

$\Rightarrow$ **Estimate causal effect**: Administration of the **treatment** "tranexamic acid" on the **outcome** mortality for trauma brain patients.

Causal Inference (IPW) with covariates with missing values [1]

[1] Doubly robust treatment effect estimation with incomplete confounders. Mayer, Wager, J. Annals Of Applied Statistics 2020.

Traumabase project: decision support for trauma patients.

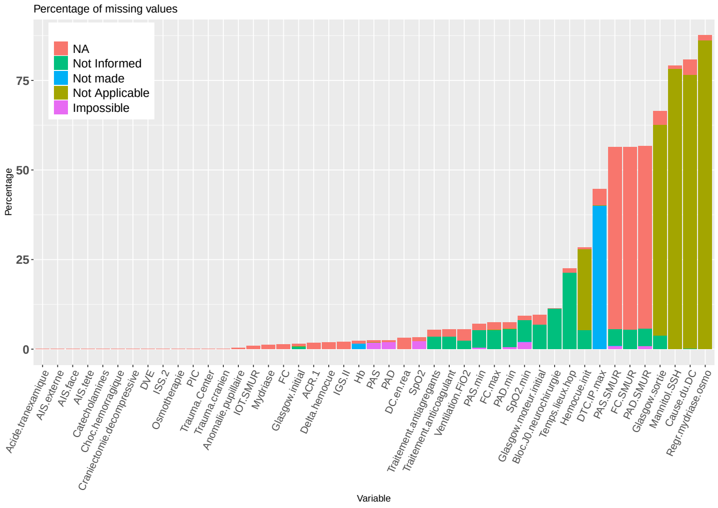
- 20000 trauma patients
- 250 continuous and categorical variables: **heterogeneous**
- 11 hospitals: **multilevel data**
- 4000 new patients/ year

Center	Accident	Age	Sex	Lactactes	BP	Shock	Platelet	...
Beaujon	fall	54	m	NM	180	yes	292000	
Pitie	gun	26	m	NA	131	no	323000	
Beaujon	moto	63	m	3.9	NR	yes	318000	
Pitie	moto	30	w	Imp	107	no	211000	
HEGP	knife	16	m	2.5	118	no	184000	
⋮								⋱

$\Rightarrow$ **Explain and Predict** platelet levels, hemorrhagic shock given pre-hospital features

Ex linear, logistic regression/ random forests with covariates with missing values

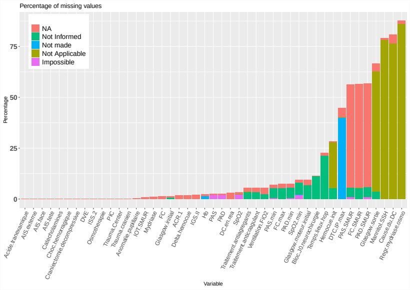
Missing values



Different pattern: sporadic & systematic (missing variable in one hospital)

Different types: MCAR, MAR, MNAR

Complete-case analysis



`?lm, ?glm, na.action = na.omit`

"One of the ironies of Big Data is that missing data play an ever more significant role" (R. Sameworth, 2019)

An $n \times p$ matrix, each entry is missing with probability 0.01

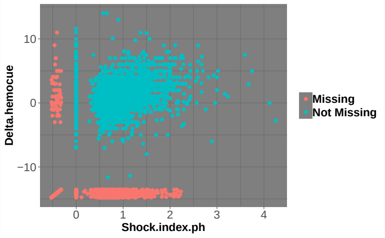
$p = 5 \implies \approx 95\%$ of rows kept

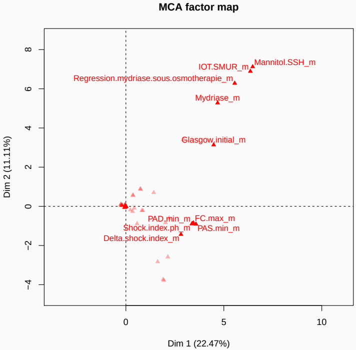
$p = 300 \implies \approx 5\%$ of rows kept

Visualization

The first thing to do with missing values (as for any analysis) is descriptive statistics: Visualization of patterns to get hints on how and why they occur

`VIM` (M. Templ), `naniar` (N. Tierney), `FactoMineR` (Husson *et al.*)





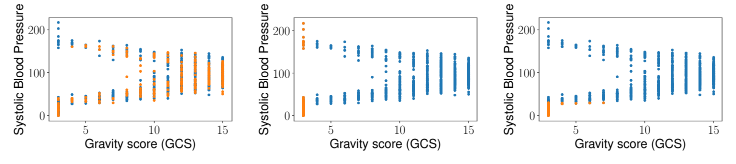
Right: *PAS_m* close to *PAD_m*: Often missing on both *PAS* & *PAD*

IOT: nested questions. Q1: yes/no, if yes Q2 - Q4, if no Q2 - Q4 "missing"

Note: Crucial **before** starting any treatment of missing values and **after**

Missing values mechanism

Rubin's taxonomy Rubin, 1976



MCAR - MAR - MNAR

Orange: missing values for Systolic Blood Pressure - Gravity index (GCS) is always observed

MCAR (completely at random): Proba to be missing does not depend on SBP neither on gravity
MAR: Proba depends on gravity (we do not measure for too severe patients)
MNAR (not at random): Proba depends on SBP (low SBP not measured)

Overview

1. Introduction

2. Inference with missing values/ Imputation

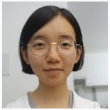
3. Supervised learning with missing values

    Random Forests with missing values

Inference with missing values/ Imputation

Collaborators on inference/imputation with missing values

- W. Jiang, A. Sportisse, PhD student at Polytechnique
- F. Husson, Professor Agronomy University. (package `missMDA`, `FactoMineR`)
- G. Bogdan, Professor Wroclaw. C. Boyer, Associate Professor Sorbonne
- Traumabase project: J.P. Nadal, T. Gauss, S. Hamada

Logistic Regression with Missing Covariates – Parameter Estimation, Model Selection and Prediction within a Joint-Modeling Framework. (2019). *CSDA*

Adaptive Bayesian SLOPE - High dimensional Model Selection with Missing Values. (2020) *In revision in JCGS.*

Estimation and imputation in Probabilistic Principal Component Analysis with Missing Not At Random data. *Neurips2020.*

Missing Data Imputation using Optimal Transport. *ICML2020*.

Debiasing Stochastic Gradient Descent to handle missing values. *Neurips2020.*

Solutions to handle missing values (M(C)AR)

Books: Schafer (2002), Little & Rubin (2019); Kim & Shao (2013); Carpenter & Kenward (2013); van Buuren (2018), etc.

Modify the estimation process to deal with missing values

Maximum likelihood: **EM algorithm** to obtain point estimates + Supplemented EM (Meng & Rubin, 1991) / Louis formulae for their variability

<u>Ex logistic regression</u>: EM to get $\hat{\beta}$ + Louis to get $\hat{V}(\hat{\beta})$

Aim: **Estimate parameters & their variance** from an incomplete data
$\Rightarrow$ Inferential framework

Solutions to handle missing values (M(C)AR)

Books: Schafer (2002), Little & Rubin (2019); Kim & Shao (2013); Carpenter & Kenward (2013); van Buuren (2018), etc.

Modify the estimation process to deal with missing values

Maximum likelihood: **EM algorithm** to obtain point estimates + Supplemented EM (Meng & Rubin, 1991) / Louis formulae for their variability

Ex logistic regression: EM to get $\hat{\beta}$ + Louis to get $\hat{V}(\hat{\beta})$

Cons: Difficult to establish - not many softwares even for simple models

One specific algorithm for each statistical method...

Aim: **Estimate parameters & their variance** from an incomplete data

$\Rightarrow$ Inferential framework

Solutions to handle missing values (M(C)AR)

Books: Schafer (2002), Little & Rubin (2019); Kim & Shao (2013); Carpenter & Kenward (2013); van Buuren (2018), etc.

Modify the estimation process to deal with missing values

Maximum likelihood: **EM algorithm** to obtain point estimates +
Supplemented EM (Meng & Rubin, 1991) / Louis formulae for their variability
Ex logistic regression: EM to get $\hat{\beta}$ + Louis to get $\hat{V}(\hat{\beta})$

Cons: Difficult to establish - not many softwares even for simple models
One specific algorithm for each statistical method...

Imputation (multiple) to get a complete data set

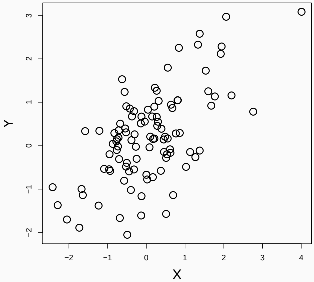
Any analysis can be performed
Ex logistic regression: Impute and apply logistic model to get $\hat{\beta}$, $\hat{V}(\hat{\beta})$

Aim: **Estimate parameters & their variance** from an incomplete data
$\Rightarrow$ Inferential framework

Mean imputation

- $(x_i, y_i) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_x, \mu_y), \Sigma_{xy})$

X	Y
-0.56	-1.93
-0.86	-1.50
.....	...
2.16	0.7
0.16	0.74



$\mu_y = 0$

$\sigma_y = 1$

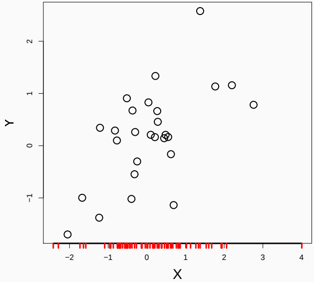
$\rho_{xy} = 0.6$

$\hat{\mu}_y = -0.01$
$\hat{\sigma}_y = 1.01$
$\hat{\rho} = 0.66$

Mean imputation

- $(x_i, y_i) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_x, \mu_y), \Sigma_{xy})$
- 70 % of missing entries completely at random on Y

X	Y
-0.56	NA
-0.86	NA
.....	...
2.16	0.7
0.16	NA



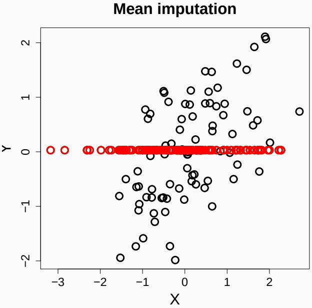
$\mu_y = 0$
$\sigma_y = 1$
$\rho_{xy} = 0.6$

$\hat{\mu}_y = 0.18$
$\hat{\sigma}_y = 0.9$
$\hat{\rho}_{xy} = 0.6$

Mean imputation

- $(x_i, y_i) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_x, \mu_y), \Sigma_{xy})$
- 70 % of missing entries completely at random on Y
- Estimate parameters on the mean imputed data

X	Y
-0.56	**0.01**
-0.86	**0.01**
.....	...
2.16	0.7
0.16	**0.01**

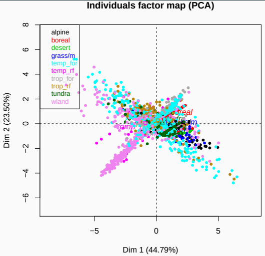


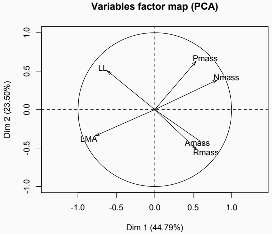
$$\mu_y = 0 \quad \sigma_y = 1 \quad \rho_{xy} = 0.6$$

$\hat{\mu}_y = 0.01$
$\hat{\sigma}_y = 0.5$
$\hat{\rho} = 0.30$

Mean imputation deforms joint and marginal distributions

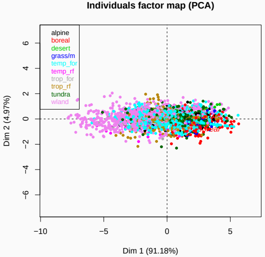
Mean imputation is bad for estimation

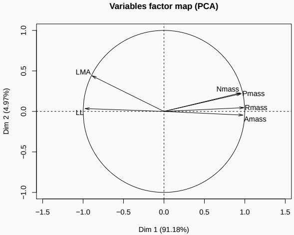




PCA with mean imputation

```
library(FactoMineR)
PCA(ecolo)
Warning message: Missing
are imputed by the mean
of the variable:
You should use imputePCA
from missMDA
```





EM-PCA

```
library(missMDA)
imp <- imputePCA(ecolo)
PCA(imp$comp)
```

J. (2016). missMDA: Handling Missing Values in Multivariate Data Analysis, JSS.

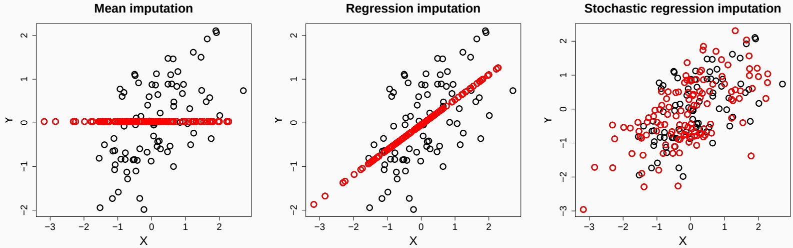
Ecological data: [2] $n = 69000$ species - 6 traits. Estimated correlation between Pmass & Rmass ≈ 0 (mean imputation) or ≈ 1 (EM PCA)

[2] Wright, I. et al. (2004). The worldwide leaf economics spectrum. *Nature*.

Imputation methods

- by regression takes into account the relationship: Estimate β - impute $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \Rightarrow$ variance underestimated and correlation overestimated
- by stochastic reg: Estimate β and σ - impute from the predictive $y_i \sim \mathcal{N}\left(x_i\hat{\beta}, \hat{\sigma}^2\right) \Rightarrow$ preserve distributions

Here $\hat{\beta}, \hat{\sigma}^2$ estimated with complete data, but MLE can be obtained with EM



	Mean imputation	Regression imputation	Stochastic regression imputation
$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho_{xy} = 0.6$	0.30	0.78	0.59

Imputation methods for multivariate data

Assuming a joint model

- Gaussian distribution: $x_{i.} \sim \mathcal{N}(\mu, \Sigma)$ (`Amelia` Honaker, King, Blackwell)
- low rank: $X_{n\times d} = \mu_{n\times d} + \varepsilon \ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with μ of low rank k
  (`softimpute` Hastie & Mazuder; missMDA J. & Husson, mimi [3])
- latent class - nonparametric Bayesian (`dpmpm` Reiter)
- deep learning using variational autoencoders (MIWAE, Mattei, 2018, VAEAC Ivanov et al., 2019), using GAN (GAIN, Yoon et al. 2018)

Using conditional models (joint implicitly defined)

- with logistic, multinomial, poisson regressions (`mice van Buuren`)
- iterative impute each variable by random forests (`missForest Stekhoven`)

Imputation for categorical, mixed, blocks/multilevel data [4], etc.

$\Rightarrow$ Rmistatic platform, more than 150 packages[5]

[3] J. et al. Main effects and interactions in mixed and incomplete data frames. (2018) *JASA*.

[4] J. et al. Imputation of mixed data with multilevel SVD. (2018). *JCGS*

[5] J., et al. `https://rmisstastic.netlify.com/`

Random forests versus PCA

	Feat1	Feat2	Feat3	Feat4	Feat5...
C1	1	1	1	1	1
C2	1	1	1	1	1
C3	2	2	2	2	2
C4	2	2	2	2	2
C5	3	3	3	3	3
C6	3	3	3	3	3
C7	4	4	4	4	4
C8	4	4	4	4	4
C9	5	5	5	5	5
C10	5	5	5	5	5
C11	6	6	6	6	6
C12	6	6	6	6	6
C13	7	7	7	7	7
C14	7	7	7	7	7
Igor	8	NA	NA	8	8
Frank	8	NA	NA	8	8
Bertrand	9	NA	NA	9	9
Alex	9	NA	NA	9	9
Yohann	10	NA	NA	10	10
Jean	10	NA	NA	10	10

Missing

Feat1	Feat2	Feat3	Feat4	Feat5
1	1.0	1.00	1	1
1	1.0	1.00	1	1
2	2.0	2.00	2	2
2	2.0	2.00	2	2
3	3.0	3.00	3	3
3	3.0	3.00	3	3
4	4.0	4.00	4	4
4	4.0	4.00	4	4
5	5.0	5.00	5	5
5	5.0	5.00	5	5
6	6.0	6.00	6	6
6	6.0	6.00	6	6
7	7.0	7.00	7	7
7	7.0	7.00	7	7
8	6.87	6.87	8	8
8	6.87	6.87	8	8
9	6.87	6.87	9	9
9	6.87	6.87	9	9
10	6.87	6.87	10	10
10	6.87	6.87	10	10

`missForest`

Feat1	Feat2	Feat3	Feat4	Feat5
1	1	1	1	1
1	1	1	1	1
2	2	2	2	2
2	2	2	2	2
3	3	3	3	3
3	3	3	3	3
4	4	4	4	4
4	4	4	4	4
5	5	5	5	5
5	5	5	5	5
6	6	6	6	6
6	6	6	6	6
7	7	7	7	7
7	7	7	7	7
8	8	8	8	8
8	8	8	8	8
9	9	9	9	9
9	9	9	9	9
10	10	10	10	10
10	10	10	10	10

`imputePCA`

$\Rightarrow$ Imputation inherits from the method: RF (computationaly costly) good for non linear relationships / PCA good for linear relationships

Single imputation: Underestimation of the variability

$\Rightarrow$ Incomplete Traumabase

X_1	X_2	X_3	...	Y
NA	20	10	...	shock
-6	45	NA	...	shock
0	NA	30	...	no shock
NA	32	35	...	shock
-2	NA	12	...	no shock
1	63	40	...	shock

Single imputation: Underestimation of the variability

⇒ Incomplete Traumabase

X_1	X_2	X_3	...	Y
NA	20	10	...	shock
-6	45	NA	...	shock
0	NA	30	...	no shock
NA	32	35	...	shock
-2	NA	12	...	no shock
1	63	40	...	shock

⇒ Completed Traumabase

X_1	X_2	X_3	...	Y
3	20	10	...	shock
-6	45	6	...	shock
0	4	30	...	no shock
-4	32	35	...	shock
-2	75	12	...	no shock
1	63	40	...	shock

Single imputation: Underestimation of the variability

$\Rightarrow$ Incomplete Traumabase

X_1	X_2	X_3	...	Y
NA	20	10	...	shock
-6	45	NA	...	shock
0	NA	30	...	no shock
NA	32	35	...	shock
-2	NA	12	...	no shock
1	63	40	...	shock

$\Rightarrow$ Completed Traumabase

X_1	X_2	X_3	...	Y
3	20	10	...	shock
-6	45	6	...	shock
0	4	30	...	no shock
-4	32	35	...	shock
-2	75	12	...	no shock
1	63	40	...	shock

A single value can't reflect the uncertainty of prediction

Multiple impute 1) Generate M plausible values for each missing value

X_1	X_2	X_3	Y
3	20	10	s
-6	45	6	s
0	4	30	no s
-4	32	35	s
-2	75	12	no s
1	63	40	s

X_1	X_2	X_3	Y
-7	20	10	s
-6	45	9	s
0	12	30	no s
13	32	35	s
-2	10	12	no s
1	63	40	s

X_1	X_2	X_3	Y
7	20	10	s
-6	45	12	s
0	-5	30	no s
2	32	35	s
-2	20	12	no s
1	63	40	s

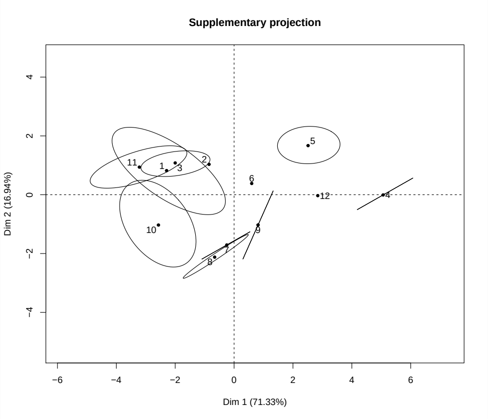
```
library(mice); mice(traumadata)
library(missMDA); MIPCA(traumadata)
```

Visualization of the imputed values

X_1	X_2	X_3	Y
3	20	10	s
-6	45	6	s
0	4	30	no s
-4	32	35	s
-2	15	12	no s
1	63	40	s

X_1	X_2	X_3	Y
-7	20	10	s
-6	45	9	s
0	12	30	no s
13	32	35	s
-2	10	12	no s
1	63	40	s

X_1	X_2	X_3	Y
7	20	10	s
-6	45	12	s
0	-5	30	no s
2	32	35	s
-2	20	12	no s
1	63	40	s




```
library(missMDA)
MIPCA(traumadata)
library(Amelia)
?compare.density
```

Percentage of NA?

Multiple imputation

1) Generate M plausible values for each missing value

X_1	X_2	X_3	Y
3	20	10	s
-6	45	6	s
0	4	30	no s
-4	32	35	s
1	63	40	s
-2	15	12	no s

X_1	X_2	X_3	Y
-7	20	10	s
-6	45	9	s
0	12	30	no s
13	32	35	s
1	63	40	s
-2	10	12	no s

X_1	X_2	X_3	Y
7	20	10	s
-6	45	12	s
0	-5	30	no s
2	32	35	s
1	63	40	s
-2	20	12	no s

2) Perform the analysis on each imputed data set: $\hat{\beta}_m, \widehat{Var}\left(\hat{\beta}_m\right)$

3) Combine the results (Rubin's rules):

$$\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m$$

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

```
imp.mice <- mice(traumadata)
lm.mice.out <- with(imp.mice, glm(Y ~ ., family = "binomial"))
```

$\Rightarrow$ Variability of missing values taken into account

Logistic regression with missing covariates: Parameter estimation, model selection and prediction (Jiang, J., Lavielle, Gauss, Hamada, 2018)

$x = (x_{ij})$ a $n \times d$ matrix of quantitative covariates
$y = (y_i)$ an n-vector of binary responses $\{0, 1\}$

Logistic regression model: $\mathbb{P}(y_i = 1 | x_i; \beta) = \frac{\exp(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij})}$
Covariables: $x_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}_d(\mu, \Sigma)$
Log-likelihood with $\theta = (\mu, \Sigma, \beta)$:
$\mathcal{LL}(\theta; x, y) = \sum_{i=1}^{n} \Big(\log(\mathrm{p}(y_i | x_i; \beta)) + \log(\mathrm{p}(x_i; \mu, \Sigma)) \Big).$

Logistic regression with missing covariates: Parameter estimation, model selection and prediction (Jiang, J., Lavielle, Gauss, Hamada, 2018)

$x = (x_{ij})$ a $n \times d$ matrix of quantitative covariates
$y = (y_i)$ an n-vector of binary responses $\{0, 1\}$

Logistic regression model: $\mathbb{P}(y_i = 1 | x_i; \beta) = \frac{\exp(\beta_0 + \sum_{j=1}^d \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^d \beta_j x_{ij})}$
Covariables: $x_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}_d(\mu, \Sigma)$
Log-likelihood with $\theta = (\mu, \Sigma, \beta)$:
$\mathcal{LL}(\theta; x, y) = \sum_{i=1}^n \Big(\log(\mathrm{p}(y_i | x_i; \beta)) + \log(\mathrm{p}(x_i; \mu, \Sigma)) \Big).$

X_1	X_2	X_3	...	Y
NA	20	10	...	shock
-6	45	NA	...	shock
0	NA	30	...	no shock
NA	32	35	...	shock
1	63	40	...	shock
-2	NA	12	...	no shock

Likelihood inference with Missing At Random values

$$\mathcal{LL}(\theta; x, y) = \sum_{i=1}^{n} \Big(\log(\mathrm{p}(y_i|x_i;\beta)) + \log(\mathrm{p}(x_i;\mu,\Sigma)) \Big)$$

X_1	X_2	X_3	...	M_1	M_2	M_3	...	Y
NA	20	10	...	1	0	0	...	shock
-6	45	NA	...	0	0	1	...	shock
0	NA	30	...	0	1	0	...	no shock
NA	32	35	...	1	0	0	...	shock

$m = (m_{ij})$ a $n \times d$ matrix $m_{ij} = 0$ if x_{ij} is observed and 1 otherwise

$(y_i, x_i, m_i) \underset{\text{i.i.d.}}{\sim} \{p_\theta(x,y) q_\phi(m \,|\, x, y)\}$ data & missing values mechanism

Likelihood inference with Missing At Random values

$$\mathcal{LL}(\theta; x, y) = \sum_{i=1}^{n} \Big(\log(\mathrm{p}(y_i|x_i;\beta)) + \log(\mathrm{p}(x_i;\mu,\Sigma)) \Big)$$

X_1	X_2	X_3	...	M_1	M_2	M_3	...	Y
NA	20	10	...	1	0	0	...	shock
-6	45	NA	...	0	0	1	...	shock
0	NA	30	...	0	1	0	...	no shock
NA	32	35	...	1	0	0	...	shock

$m = (m_{ij})$ a $n \times d$ matrix $m_{ij} = 0$ if x_{ij} is observed and 1 otherwise

$(y_i, x_i, m_i) \underset{\text{i.i.d.}}{\sim} \{p_\theta(x, y) q_\phi(m \,|\, x, y)\}$ data & missing values mechanism

Note $x_i = (x_{i,obs}, x_{i,mis})$ $\mathcal{L}(\theta, \phi) \triangleq \prod_{i=1}^{n} \int q_\phi(m_i|x_i, y_i) p_\theta(x_i, y_i) dx_{i,mis}$

MAR: $\forall\, \phi, \forall\, x'_{i,mis}$ such that $x'_i = (x_{i,obs}, x'_{i,mis})$, $q_\phi(m_i|x'_i) = q_\phi(m_i|x_i)$

Ignorable mechanism $\mathcal{L}(\theta, \phi) \triangleq \prod_{i=1}^{n} q_\phi(m_i|x_{i,obs}, y_i) \int p_\theta(x_i, y_i) dx_{i,mis}$

$$\mathcal{L}_{obs}(\theta) \triangleq \prod_{i=1}^{n} \int p_\theta(x_i, y_i) dx_{i,mis}$$

Stochastic Approximation EM - package misaem

$\arg\max \mathcal{LL}(\theta; x_{\text{obs}}, y) = \int \mathcal{LL}(\theta; x, y) dx_{\text{mis}}$

- **E-step:** Evaluate the quantity

$$\begin{aligned} Q_k(\theta) &= \mathbb{E}[\mathcal{LL}(\theta; x, y) | x_{\text{obs}}, y; \theta_{k-1}] \\ &= \int \mathcal{LL}(\theta; x, y) \mathrm{p}(x_{\text{mis}} | x_{\text{obs}}, y; \theta_{k-1}) dx_{\text{mis}} \end{aligned}$$

- **M-step:** $\theta_k = \arg\max_\theta Q_k(\theta)$

$\Rightarrow$ *Unfeasible computation of expectation*

MCEM (Wei & Tanner, 1990): Generate samples of missing data from $\mathrm{p}(x_{\text{mis}} | x_{\text{obs}}, y; \theta_{k-1})$ and replace the expectation by an empirical mean

$\Rightarrow$ *Require a huge number of samples*

SAEM (Lavielle, 2014) almost sure convergence to MLE (Metropolis Hasting - Variance estimation with Louis formulae).

Unbiased estimates: $\hat{\beta}_1, \ldots, \hat{\beta}_d$ - $\hat{V}(\hat{\beta}_1), \ldots, \hat{V}(\hat{\beta}_d)$ - good coverage

Low rank estimation with MNAR data

$Y \in \mathbb{R}^{n\times p}$ noisy realisation of a low-rank matrix $\Theta \in \mathbb{R}^{n\times p}$:

$$Y = \Theta + \epsilon, \text{where} \begin{cases} \Theta \text{ with rank } r < \min\{n, p\}, \\ \epsilon_i \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}(0_n, \sigma^2 I_{n\times n}), \forall i \in [1, n]. \end{cases}$$

$\dashrightarrow$ Access only to the missing-data matrix $Y \odot M$,

- How to estimate Θ ?
- How to impute the unknown entries of Y ?

Data distribution

$$p(y_{ij}; \Theta_{ij}) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2}\left(\frac{y_{ij} - \Theta_{ij}}{\sigma}\right)^2\right).$$

MNAR missing-data mechanism via a Logistic Model

$\forall i \in [1, n]$, $\phi_j = (\phi_{1j}, \phi_{2j})$ denoting a parameter vector:

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

$\rightsquigarrow$ self-masked MNAR : the lack only depends on the value itself.

EM algo with MNAR (self-mask logistic)[6]

MAR (ignorable): maximize the observed penalized log-likelihood

$$\hat{\Theta} \in \mathrm{argmin}_{\Theta} \|(Y - \Theta) \odot M\|_F^2 + \lambda \|\Theta\|_\star,$$

Classical: iterative soft-thresholding (ISTA) of SVD softimpute (Hastie), **its accelerated version: FISTA** (Beck & Teboulle)

MNAR (non ignorable) $\ell(\Theta, \phi; y_{\mathrm{obs}}, M) = \int p(y; \Theta) p(M|y; \phi) dy_{\mathrm{mis}}$.

- **E-step:**

$$Q(\Theta, \phi | \hat{\Theta}^{(t)}, \hat{\phi}^{(t)}) = -\mathbb{E}_{Y_{\mathrm{mis}}} \left[\ell(\Theta, \phi; y, \Omega) | Y_{\mathrm{obs}}, M; \Theta = \hat{\Theta}^{(t)}, \phi = \hat{\phi}^{(t)} \right]$$

- **M-step:**

$$\hat{\Theta}^{(t+1)}, \hat{\phi}^{(t+1)} \in \mathrm{argmin}_{\Theta, \phi} \, Q(\Theta, \phi | \hat{\Theta}^{(t)}, \hat{\phi}^{(t)}) + \lambda \|\Theta\|_\star$$

- E-step: Monte-Carlo approximation and SIR algorithm.
- M-step: Separability of Q:
  - Θ: `softImpute`, FISTA.
  - ϕ: Newton-Raphson algorithm.

$\Rightarrow$ Computationally costly, few variables with MNAR.

[6] Low-rank estimation with missing non at random data. (2018) *Statistics and Computing*

Take home message inference/imputation

- Few implementation of EM strategies

"**The idea of imputation is both seductive and dangerous**". *It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the imputed data have substantial biases." (Dempster & Rubin, 1983)*

- Single imputation aims at completing a dataset as best as possible
- **Multiple imputation** aims at estimating the parameters and their variability taking into account the uncertainty of the missing values
- Single imputation can be appropriate for point estimates
- Both % of NA & structure matter (5% of NA can be an issue)

Principal component methods powerful for single & multiple imputation of quanti & categorical data: Dimensionality reduction and capture similarities between observations and variables. `missMDA package`

- Still difficult to handle MNAR (Estim. & imput. in PPCA. *Neurips2020.*)

Supervised learning with missing values

Outline

1. Introduction
2. Inference with missing values/ Imputation
3. Supervised learning with missing values
   Random Forests with missing values

Collaborators on supervised learning with missing values

- M. Le Morvan, Postdoc at INRIA, Paris.
- E. Scornet, Associate Professor at Ecole Polytechnique, IP Paris.

Topic: random forests.

- G. Varoquaux, Senior researcher at INRIA, Paris.

Topic: machine learning. Creator of Scikitlearn in python.

⇒ **Random Forests with missing values**

1. *Consistency of supervised learning with missing values. (2019). Revis JMLR.*

⇒ **Linear regression with missing values - MultiLayer perceptron**

2. *Linear predictor on linearly-generated data with missing values: non consistency and solutions. AISTAT2020.*

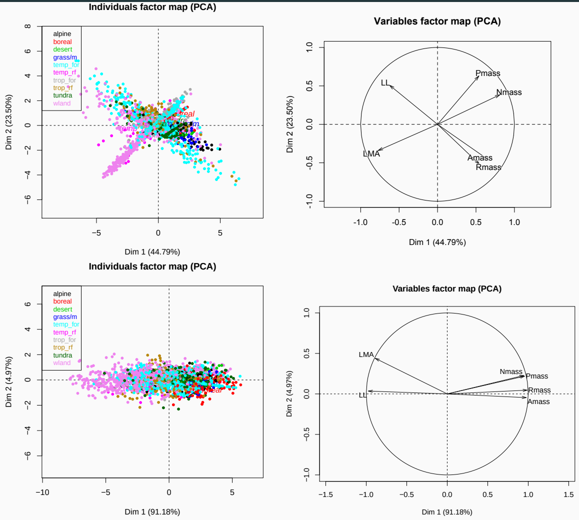
3. *Neumiss networks: differential programming for supervised learning with missing values. Neurips2020.*

Missing values in a predictive framework (not inferential)

- Aim: target an outcome Y (not estimate parameters and their variance)
- Specificities: train & test sets with missing values
- Methods : (in practice) imputation prior to prediction
  - Separate: impute train and test separately (with a different model)
  - Grouped/ semi-supervised: impute train and test simultaneously but the predictive model is learned only on the training imputed data set.
  - **Imputation train and test sets with the same model**
    Issue: methods (`missForest`) are "black-boxes" *i.e.* take as an input the incomplete data and output the completed data

    Easy for univariate imputation: **mean of each colum of the train**.

Mean imputation is bad for estimation



PCA with mean imputation

```
library(FactoMineR)
PCA(ecolo)
Warning message: Missing
are imputed by the mean
of the variable:
You should use imputePCA
from missMDA
```

EM-PCA

```
library(missMDA)
imp <- imputePCA(ecolo)
PCA(imp$comp)
```

J. (2016). missMDA: Handling Missing Values in Multivariate Data Analysis, JSS.

Ecological data: [7] $n = 69000$ species - 6 traits. Estimated correlation between Pmass & Rmass ≈ 0 (mean imputation) or ≈ 1 (EM PCA)

[7] Wright, I. et al. (2004). The worldwide leaf economics spectrum. *Nature*.

Constant (mean) imputation is consistent for prediction

$\tilde{X} = X \odot (1 - M) + \texttt{NA} \odot M$. New feature space is $\widetilde{\mathbb{R}}^d = (\mathbb{R} \cup \{\texttt{NA}\})^d$.

$$Y = \begin{pmatrix} 4.6 \\ 7.9 \\ 8.3 \\ 4.6 \end{pmatrix} \quad \tilde{X} = \begin{pmatrix} 9.1 & \texttt{NA} & 1 \\ 2.1 & \texttt{NA} & 3 \\ \texttt{NA} & 9.6 & 2 \\ \texttt{NA} & 5.5 & 6 \end{pmatrix} \quad X = \begin{pmatrix} 9.1 & 8.5 & 1 \\ 2.1 & 3.5 & 3 \\ 6.7 & 9.6 & 2 \\ 4.2 & 5.5 & 6 \end{pmatrix} \quad M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Find a prediction function that minimizes the risk.

$$\text{Bayes rule: } f^* \in \underset{f:\ \widetilde{\mathbb{R}}^d \to \mathbb{R}}{\arg\min}\ \mathbb{E}\left[\left(Y - f(\tilde{X})\right)^2\right].$$

$$f^*(\tilde{X}) = \mathbb{E}\left[Y \mid \tilde{X}\right] = \mathbb{E}\left[Y \mid X_{obs(M),M}\right]$$
$$= \sum_{m \in \{0,1\}^d} \mathbb{E}\left[Y | X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$$

$\Rightarrow$ One model per pattern (2^d) (Rubin, 1984, generalized propensity score)

Constant (mean) imputation is consistent for prediction

$\tilde{X} = X \odot (1 - M) + \texttt{NA} \odot M$. New feature space is $\widetilde{\mathbb{R}}^d = (\mathbb{R} \cup \{\texttt{NA}\})^d$.

$$Y = \begin{pmatrix} 4.6 \\ 7.9 \\ 8.3 \\ 4.6 \end{pmatrix} \quad \tilde{X} = \begin{pmatrix} 9.1 & \texttt{NA} & 1 \\ 2.1 & \texttt{NA} & 3 \\ \texttt{NA} & 9.6 & 2 \\ \texttt{NA} & 5.5 & 6 \end{pmatrix} \quad X = \begin{pmatrix} 9.1 & 8.5 & 1 \\ 2.1 & 3.5 & 3 \\ 6.7 & 9.6 & 2 \\ 4.2 & 5.5 & 6 \end{pmatrix} \quad M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Find a prediction function that minimizes the risk.

$$\text{Bayes rule: } f^* \in \underset{f:\ \widetilde{\mathbb{R}}^d \to \mathbb{R}}{\arg\min}\ \mathbb{E}\left[\left(Y - f(\tilde{X})\right)^2\right].$$

$$f^*(\tilde{X}) = \mathbb{E}\left[Y \mid \tilde{X}\right] = \mathbb{E}\left[Y \mid X_{obs(M),M}\right]$$

$$= \sum_{m \in \{0,1\}^d} \mathbb{E}\left[Y | X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$$

$\Rightarrow$ One model per pattern (2^d) (Rubin, 1984, generalized propensity score)

Constant (mean) imputation is consistent

Framework - assumptions

- $Y = f(X) + \varepsilon$
- $X = (X_1, \ldots, X_d)$ has a continuous density $g > 0$ on $[0, 1]^d$
- $\|f\|_\infty < \infty$
- Missing data MAR on X_1 with $M_1 \perp\!\!\!\perp X_1 | X_2, \ldots, X_d$.
- $(x_2, \ldots, x_d) \mapsto \mathbb{P}[M_1 = 1 | X_2 = x_2, \ldots, X_d = x_d]$ is continuous
- ε is a centered noise independent of (X, M_1)

(remains valid when missing values occur for several variables $X_1, \ldots, X_j$)

Constant (mean) imputation is consistent

Constant imputed entry $x' = (x_1', x_2, \dots, x_d)$: $x_1' = x_1 \mathbb{1}_{M_1=0} + \alpha \mathbb{1}_{M_1=1}$

Theorem. (J. et al. 2019)

$$
\begin{aligned}
f^\star_{impute}(x') =& \mathbb{E}[Y | X_2 = x_2, \dots, X_d = x_d, M_1 = 1] \\
& \qquad \mathbb{1}_{x_1'=\alpha]} \mathbb{1}_{\mathbb{P}[M_1=1 | X_2=x_2,\dots,X_d=x_d]>0} \\
& \qquad + \mathbb{E}[Y | X = x'] \mathbb{1}_{x_1'=\alpha} \mathbb{1}_{\mathbb{P}[M_1=1 | X_2=x_2,\dots,X_d=x_d]=0} \\
& \qquad + \mathbb{E}[Y | X_1 = x_1, X_2 = x_2, \dots, X_d = x_d, M_1 = 0] \mathbb{1}_{x_1' \neq \alpha}.
\end{aligned}
$$

Prediction with mean is equal to the Bayes function almost everywhere

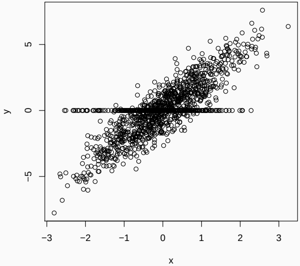
$$f^\star_{impute}(X') = f^\star(\tilde{X}) = \mathbb{E}[Y | \tilde{X} = \tilde{x}]$$

Rq: pointwise equality if using a constant out of range.

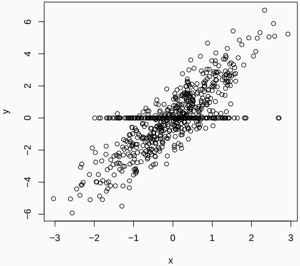
$\Rightarrow$ Learn on the mean-imputed training data, impute the test set with the same means and predict is optimal if the missing data are MAR and the **learning algorithm is universally consistent**

Consistency of supervised learning with NA: Rationale

- Specific value, systematic like a code for missing
- Need a lot of data (asymptotic result) and a super powerful learner
- The learner detects the code and recognizes it at the test time
- With categorical data, just code "Missing"
- With continuous data, any constant:



Train



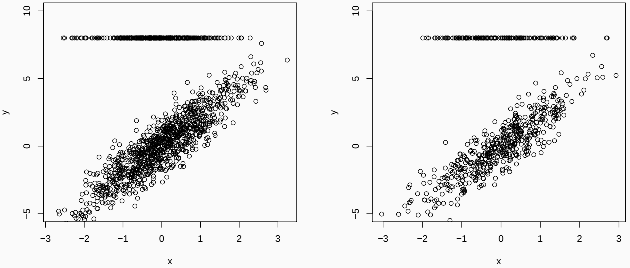
Test

<u>Mean imputation not bad for prediction; it is consistent; despite its drawbacks for estimation - Useful in practice!</u>

Empirically good results for MNAR

Consistency of supervised learning with NA: Rationale

- Specific value, systematic like a code for missing
- Need a lot of data (asymptotic result) and a super powerful learner
- The learner detects the code and recognizes it at the test time
- With categorical data, just code "Missing"
- With continuous data, any constant: out of range



Train Test

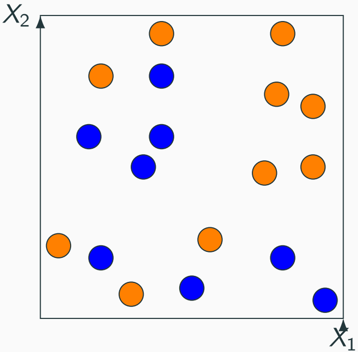
<u>Mean imputation not bad for prediction; it is consistent; despite its drawbacks for estimation - Useful in practice!</u>

Empirically good results for MNAR

CART (Breiman, 1984)

Built recursively by splitting the current cell into two children: Find the feature $j^\star$, the threshold $z^\star$ which minimises the (quadratic) loss

$$(j^\star, z^\star) \in \operatorname*{arg\,min}_{(j,z)\in\mathcal{S}} \mathbb{E}\Big[\big(Y - \mathbb{E}[Y|X_j \leq z]\big)^2 \cdot \mathbb{1}_{X_j \leq z} + \big(Y - \mathbb{E}[Y|X_j > z]\big)^2 \cdot \mathbb{1}_{X_j > z}\Big].$$

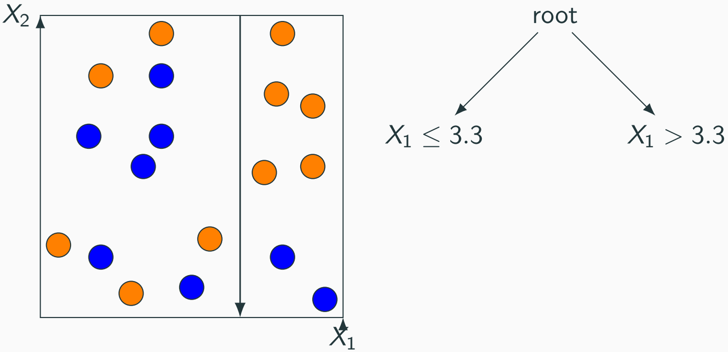


root

CART (Breiman, 1984)

Built recursively by splitting the current cell into two children: Find the feature $j^\star$, the threshold $z^\star$ which minimises the (quadratic) loss

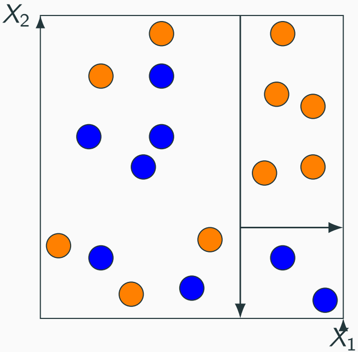
$$
\begin{aligned}
(j^\star, z^\star) \in \underset{(j,z)\in\mathcal{S}}{\arg\min}\ \mathbb{E}\Big[\big(Y - \mathbb{E}[Y|X_j \le z]\big)^2 \cdot \mathbb{1}_{X_j \le z} \\
+ \big(Y - \mathbb{E}[Y|X_j > z]\big)^2 \cdot \mathbb{1}_{X_j > z}\Big].
\end{aligned}
$$

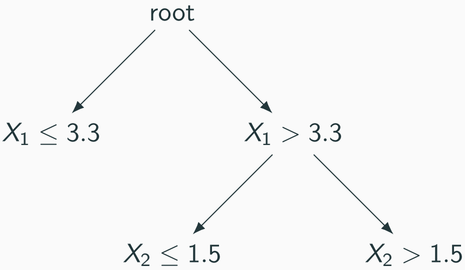


CART (Breiman, 1984)

Built recursively by splitting the current cell into two children: Find the feature $j^\star$, the threshold $z^\star$ which minimises the (quadratic) loss

$$(j^\star, z^\star) \in \underset{(j,z)\in\mathcal{S}}{\arg\min}\ \mathbb{E}\Big[\big(Y - \mathbb{E}[Y|X_j \leq z]\big)^2 \cdot \mathbb{1}_{X_j \leq z} + \big(Y - \mathbb{E}[Y|X_j > z]\big)^2 \cdot \mathbb{1}_{X_j > z}\Big].$$





CART with missing values

root

	X_1	X_2	Y
1			
2	NA		
3	NA		
4			

CART with missing values

$$\begin{tikzcd}[column sep=small] & \text{root} \arrow[dl] \arrow[dr] & \\ X_1 \leq s_1 & & X_1 > s_1 \end{tikzcd}$$

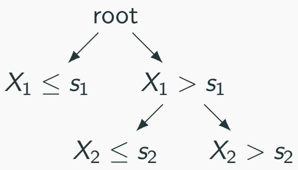
	X_1	X_2	Y
1			
~~2~~	~~NA~~		
~~3~~	~~NA~~		
4			

1) Select variable and threshold on observed values (1 & 4 for X_1)

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y|X_j \leq z, M_j = 0]\big)^2 \cdot \mathbb{1}_{X_j \leq z, M_j=0} + \big(Y - \mathbb{E}[Y|X_j > z, M_j = 0]\big)^2 \cdot \mathbb{1}_{X_j > z, M_j=0}\Big].$$

CART with missing values

	X_1	X_2	Y
1			
2	~~NA~~		
3	~~NA~~		
4			



1) Select variable and threshold on observed values (1 & 4 for X_1)

$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y|X_j \leq z, M_j = 0]\big)^2 \cdot \mathbb{1}_{X_j \leq z, M_j = 0} + \big(Y - \mathbb{E}[Y|X_j > z, M_j = 0]\big)^2 \cdot \mathbb{1}_{X_j > z, M_j = 0}\Big].$$

2) Propagate observations (2 & 3) with missing values?

- Probabilistic split: $\mathcal{B}ernoulli(\frac{\#L}{\#L+\#R})$ (`Rweeka`)
- Block: Send all to a side by minimizing the error (`xgboost`, `lightgbm`)
- Surrogate split: Search another variable that gives a close partition (rpart)

Missing incorporated in attribute (Twala et al. 2008)

One step: select the variable, the threshold and propagate missing values

1. $\{\widetilde{X}_j \leq z \text{ or } \widetilde{X}_j = \texttt{NA}\}$ vs $\{\widetilde{X}_j > z\}$
2. $\{\widetilde{X}_j \leq z\}$ vs $\{\widetilde{X}_j > z \text{ or } \widetilde{X}_j = \texttt{NA}\}$
3. $\{\widetilde{X}_j \neq \texttt{NA}\}$ vs $\{\widetilde{X}_j = \texttt{NA}\}$.

- The splitting location z depends on the missing values
- **Missing values treated like a category** (well to handle $\mathbb{R} \cup \texttt{NA}$)
- Good for informative pattern (M explains Y)

Targets one model per pattern:

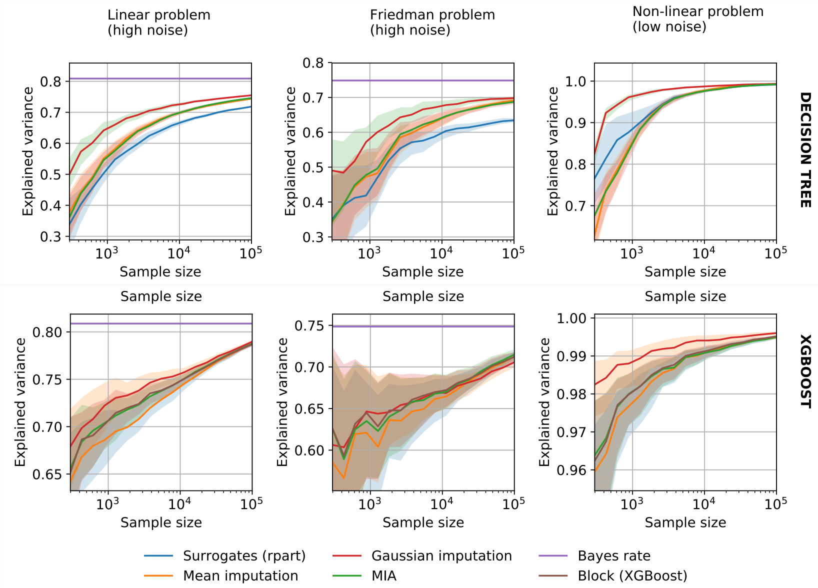
$$\mathbb{E}\left[Y \middle| \tilde{X}\right] = \sum_{m \in \{0,1\}^d} \mathbb{E}\left[Y | X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$$

- Implementation [8]: `grf package`, `scikit-learn`, `partykit`

⇒ Extremely **good performances** in practice **for any mechanism**.

[8] implementation trick, J. Tibshirani, duplicate the incomplete columns, and replace the missing entries once by $+\infty$ and once by $-\infty$

Consistency: 40% missing values MCAR



Take-home message. Supervised learning with missing values.

Supervised learning different from usual inferential probabilistic models. Solutions useful in practice robust to the missing-value mechanisms but needs powerful model.

Powerful learner with missing values

- Incomplete train and test → same imputation model
- Single constant imputation is consistent with a powerful learner
- Empirically, good imputation reduce sample complexity
- Tree-based models : Missing Incorporated in Attribute
- To be done: nonasymptotic results, uncertainty, distributional shift:

No NA in the test? Proofs in MNAR

Still an active area of research! Join this exciting field!

- New architecture for network with missing data: $\odot M$ nonlinearity.
- Supervised clustering with missing values
- Times series with missing values

Ressources

R-miss-tastic https://rmisstastic.netlify.com/R-miss-tastic

J., I. Mayer, N. Tierney & N. Vialaneix

Project funded by the R consortium (Infrastructure Steering Committee)[9]

Aim: a reference platform on the theme of missing data management

- list existing packages
- available literature
- tutorials
- analysis workflows on data
- main actors

$\Rightarrow$ Federate the community

$\Rightarrow$ Contribute!

[9] https://www.r-consortium.org/projects/call-for-proposals

Ressources

Examples:

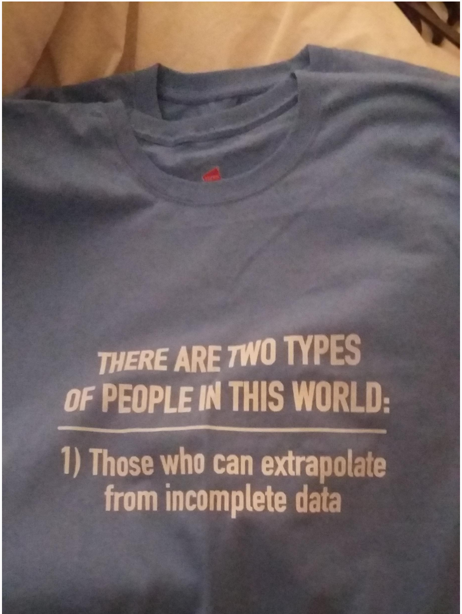
- Lecture [10] - General tutorial : Statistical Methods for Analysis with Missing Data (Mauricio Sadinle)
- Lecture - Multiple Imputation: `mice` by Nicole Erler [11]
- Longitudinal data, Time Series Imputation (Steffen Moritz - very active contributor of r-miss-tastic), Principal Component Methods[12]

[10] https://rmisstastic.netlify.com/lectures/

[11] https://rmisstastic.netlify.com/tutorials/erler_course_multipleimputation_2018/erler_practical_mice_2018

[12] https://rmisstastic.netlify.com/tutorials/Josse_slides_imputation_PCA_2018.pdf

Thank you


THERE ARE TWO TYPES
OF PEOPLE IN THIS WORLD:
1) Those who can extrapolate
from incomplete data