Supervised learning with missing values

Julie Josse
Head of the Inria-Inserm team **PreMeDICaL**:
"**Pre**cision **Me**dicine by **D**ata **I**ntegration & **Ca**usal **L**earning"

February 9, 2023

Presentation Julie Josse: Computational statistics

Academic background:

- Assistant Professor at Institut **Agro Rennes**-Angers (2011-15)
- Visiting Scholar at **Stanford University** (2013-2015, 18 months)
- Professor at **École Polytechnique** (IP Paris) (2016 - 20)
- Visiting Researcher at **Google Brain** Paris (2019 - 2020). 2 days/week
- Senior Researcher at **Inria** Montpellier (Sept. 2020 -)
- Visiting Researcher at **Apple** Paris (2023 -). 1 day/week

Research topics:

- Dimensionality reduction to visualize high dimensional heterogeneous data
- Missing values: EM alg., matrix completion, MNAR, supervised learning
- Causal inference: combining RCT & observational data, optimal policy
- Collaborations: medical (hospitals, SANOFI, etc.), energy (EDF), ecology

Software:

- R community: book R for Statistics, R foundation, R Forwards (widen the participation of minorities), R packages, R Task Views (missing, causal inf.)
- Website on missing values (R-miss-tastic), mobile application (ICUBAM)

Traumabase project: decision support for trauma patients

- 30 000 French trauma patients[1]
- 250 features from the accident site to the hospital discharge
- 30 hospitals
- 4000 new patients/ year

Center	Accident	Age	Sex	Lactactes	BP	Shock	Platelet	...
Beaujon	fall	54	m	NM	180	yes	292 000	
Pitie	gun	26	m	NA	131	no	323 000	
Beaujon	moto	63	m	3.9	NR	yes	318 000	
Pitie	moto	30	w	Imp	107	no	211 000	
⋮								⋱

[1] www.traumabase.eu - https://www.traumatrix.fr/

Traumabase project: decision support for trauma patients

- ▷ 30 000 French trauma patients[1]
- ▷ 250 features from the accident site to the hospital discharge
- ▷ 30 hospitals
- ▷ 4000 new patients/ year

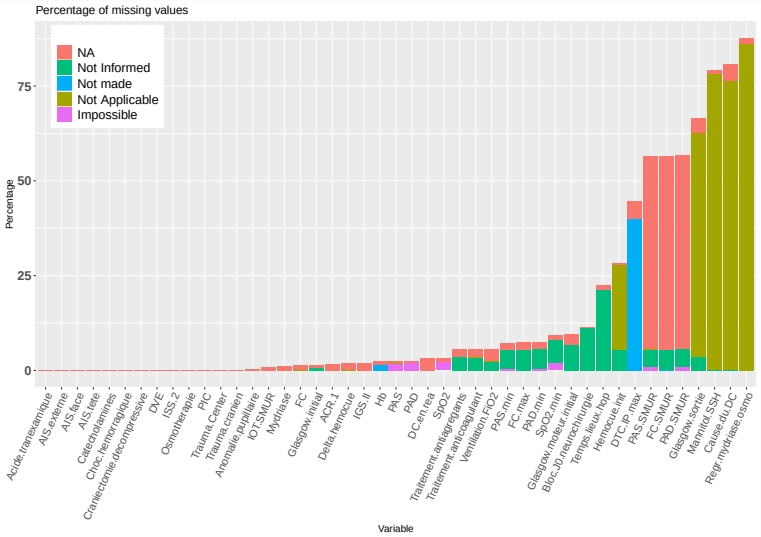
Center	Accident	Age	Sex	Lactactes	BP	Shock	Platelet	...
Beaujon	fall	54	m	NM	180	yes	292 000	
Pitie	gun	26	m	NA	131	no	323 000	
Beaujon	moto	63	m	3.9	NR	yes	318 000	
Pitie	moto	30	w	Imp	107	no	211 000	
⋮								⋱

⇒ **Explain and predict** hemorrhagic shock given pre-hospital features.
Ex: logistic regression/ random forests with missing values in covariates

Prospective study: real-time testing of models in the ambulance via a mobile data collection application (ShockMatrix playstore)

[1] www.traumabase.eu - https://www.traumatrix.fr/

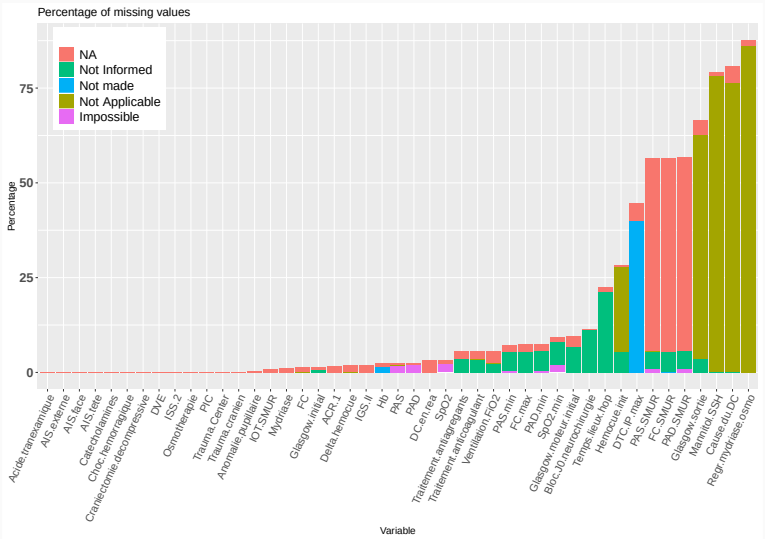
Missing data: important bottleneck in statistical practice



"One of the ironies of Big Data is that missing data play an ever more significant role"[2]

[2] Zhu, Wang, Samworth. High-dimensional PCA with heterogeneous missingness. *JRSSB*. 2022.

Missing data: important bottleneck in statistical practice



"One of the ironies of Big Data is that missing data play an ever more significant role"[2]

Complete case analysis: delete incomplete samples

- **Bias**: Resulting sample not representative of the target population
- **Information loss**: Take a matrix with d features where each entry is missing with probability $1/100$, remove a row (of length d) when one entry is missing

$$d = 5 \implies \approx 95\% \text{ of rows kept}$$
$$d = 300 \implies \approx 5\% \text{ of rows kept}$$

[2] Zhu, Wang, Samworth. High-dimensional PCA with heterogeneous missingness. *JRSSB*. 2022.

Inference with missing values

What is a 'true' missing value?

First analysis to perform with missing data (and any data): descriptive study

Visualize their patterns for clues as to how & why they occur `FactoMineR`[3]

Anomaly	**Osthmot.**	Improv.	SBP	DBP
No	NA	NA	150	100
Yes	Mannitol	Yes	99	41
No	NA	NA	110	76
Yes	SSH	NA	114	50
No	NA	NA	116	NA

[3] Husson, **J.**, Le. FactoMineR: An R Package for Multivariate Analysis. *JSS*. (2008)

What is a 'true' missing value?

First analysis to perform with missing data (and any data): descriptive study

Visualize their patterns for clues as to how & why they occur `FactoMineR`[3]

Anomaly	**Osthmot.**	Improv.	SBP	DBP
No	NA	NA	Obs	Obs
Yes	Mannitol	Yes	Obs	Obs
No	NA	NA	Obs	Obs
Yes	SSH	NA	Obs	Obs
No	NA	NA	Obs	NA

Multiple Correspondence Analysis with numeric values coded as **Obs** & missing as **NA**

[3] Husson, **J.**, Le. FactoMineR: An R Package for Multivariate Analysis. *JSS*. (2008)

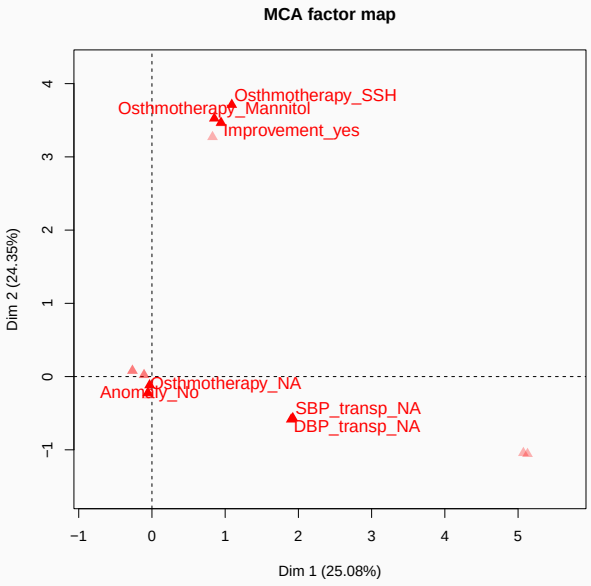
What is a 'true' missing value?

First analysis to perform with missing data (and any data): descriptive study

Visualize their patterns for clues as to how & why they occur `FactoMineR`[3]

Anomaly	**Osthmot.**	Improv.	SBP	DBP
No	NA	NA	Obs	Obs
Yes	Mannitol	Yes	Obs	Obs
No	NA	NA	Obs	Obs
Yes	SSH	NA	Obs	Obs
No	NA	NA	Obs	NA

Multiple Correspondence Analysis with numeric values coded as **Obs** & missing as **NA**



[3] Husson, **J.**, Le. FactoMineR: An R Package for Multivariate Analysis. *JSS*. (2008)

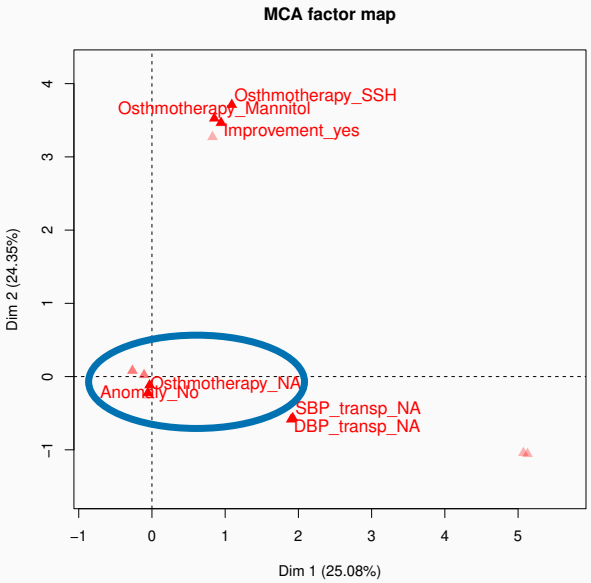
What is a 'true' missing value?

First analysis to perform with missing data (and any data): descriptive study

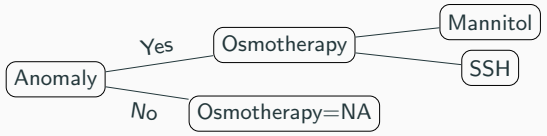
Visualize their patterns for clues as to how & why they occur `FactoMineR`[3]

Anomaly	**Osthmot.**	Improv.	SBP	DBP
No	NA	NA	Obs	Obs
Yes	Mannitol	Yes	Obs	Obs
No	NA	NA	Obs	Obs
Yes	SSH	NA	Obs	Obs
No	NA	NA	Obs	NA

Multiple Correspondence Analysis with numeric values coded as **Obs** & missing as **NA**



- Detect <u>nested variables</u>:



⇒ Not a 'true' missing value, does not **mask an underlying value**

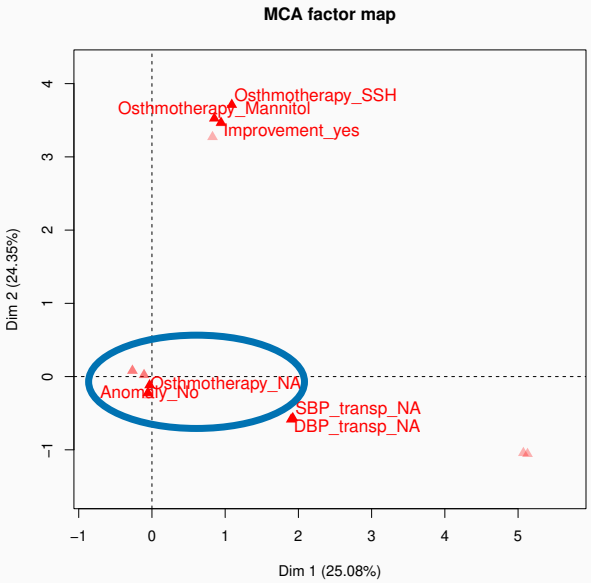
[3] Husson, **J.**, Le. FactoMineR: An R Package for Multivariate Analysis. *JSS*. (2008)

What is a 'true' missing value?

First analysis to perform with missing data (and any data): descriptive study

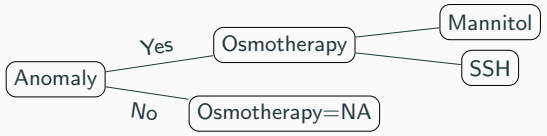
Visualize their patterns for clues as to how & why they occur `FactoMineR`[3]

Anomaly	**Osthmot.**	Improv.	SBP	DBP	**Anomaly-Osthmot.**
No	NA	NA	Obs	Obs	No
Yes	Mannitol	Yes	Obs	Obs	Yes Mannitol
No	NA	NA	Obs	Obs	No
Yes	SSH	NA	Obs	Obs	Yes SSH
No	NA	NA	Obs	NA	No



Multiple Correspondence Analysis with numeric values coded as **Obs** & missing as **NA**

- Detect <u>nested variables</u>:



⇒ Not a 'true' missing value, does not **mask an underlying value**

⇒ Solution: <u>recode</u> with a 3-level variable 'Yes Mannitol', 'Yes SSH', 'no'

⇒ Feedback on data collection/encoding process

[3] Husson, **J.**, Le. FactoMineR: An R Package for Multivariate Analysis. *JSS*. (2008)

Missing values mechanism: Rubin's taxonomy[4,5]

- Random Variables:
  - ▷ $X \in \mathbb{R}^d$: complete unavailable data
  - ▷ $M \in \{0,1\}^d$: missing pattern, or mask, $M_j = 1$ if and only if X_j is missing

  For a pattern m, obs(m) indices of observed entries, $X_{\text{obs}(m)}$ the vector of observed components

[4] Rubin. Inference and missing data. *Biometrika.* 1976.

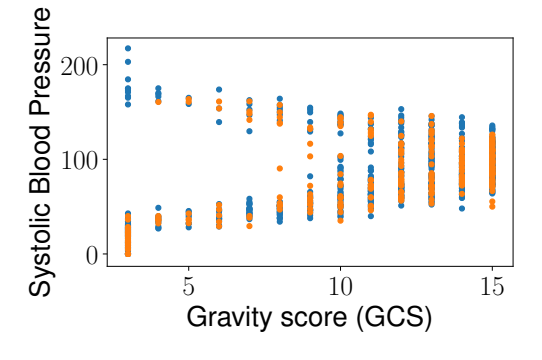
[5] What Is Meant by "Missing at Random"? Seaman, et al. Statistical Science. 2013.

Missing values mechanism: Rubin's taxonomy[4,5]

- Random Variables:
  - ▷ $X \in \mathbb{R}^d$: complete unavailable data
  - ▷ $M \in \{0,1\}^d$: missing pattern, or mask, $M_j = 1$ if and only if X_j is missing

  For a pattern m, obs(m) indices of observed entries, $X_{\text{obs}(m)}$ the vector of observed components

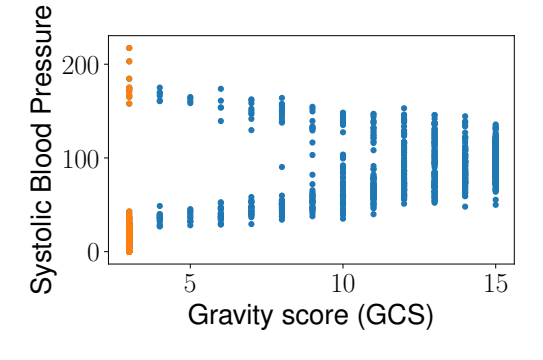
Ex: Simulated missing values according to the 3 mechanisms (Orange points will be missing) in Systolic Blood Pressure - GCS is always observed



Missing Completely at Random (MCAR)

$\forall m \in \{0,1\}^d$,

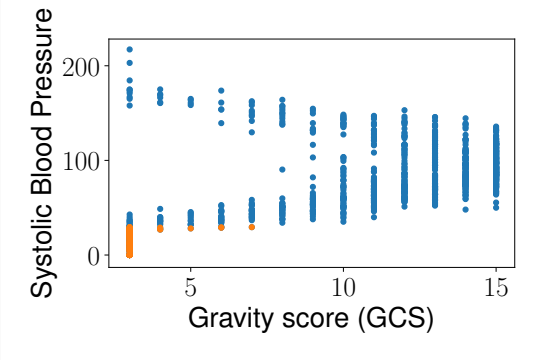
$\mathbb{P}(M = m|X) = \mathbb{P}(M = m)$



Missing at Random (MAR)

$\forall m \in \{0,1\}^d$,

$\mathbb{P}(M = m|X) = \mathbb{P}\left(M = m|X_{\text{obs}(m)}\right)$



Missing Not At Random (MNAR)

If not MAR: it is MNAR

[4] Rubin. Inference and missing data. *Biometrika.* 1976.

[5] What Is Meant by "Missing at Random"? Seaman, et al. Statistical Science. 2013.

Solutions to handle M(C)AR values (in the covariates)

Abundant literature: Creation of Rmistatic platform[6] ($>$ 150 packages)

Inferential aim: **Estimate parameters & their variance, i.e.** $\hat{\beta}$, $\hat{V}(\hat{\beta})$ to get confidence intervals with the appropriate coverage

[6] Mayer, **J.** et al. A unified platform for missing values methods and workflows. *R journal*. 2022.
[7] Jiang, **J.** et al. Logistic Regression with Missing Covariates, Parameter Estimation, Model Selection and Prediction. *CSDA*. 2019. - Implementation in the `misaem package`

Solutions to handle M(C)AR values (in the covariates)

Abundant literature: Creation of Rmistatic platform[6] ($>$ 150 packages)

Inferential aim: **Estimate parameters & their variance, i.e.** $\hat{\beta}$, $\hat{V}(\hat{\beta})$ to get confidence intervals with the appropriate coverage

Modify the estimation process to deal with missing values

Maximum likelihood inference: Expectation Maximization algorithms

Pros: Tailored toward a specific problem

Cons: Few softwares even for simple models. Ex: logistic regression[7]

Need to design one specific algorithm for each statistical method

[6] Mayer, **J.** et al. A unified platform for missing values methods and workflows. *R journal*. 2022.

[7] Jiang, **J.** et al. Logistic Regression with Missing Covariates, Parameter Estimation, Model Selection and Prediction. *CSDA*. 2019. - Implementation in the `misaem package`

Solutions to handle M(C)AR values (in the covariates)

Abundant literature: Creation of Rmistatic platform[6] ($>$ 150 packages)

Inferential aim: **Estimate parameters & their variance, i.e.** $\hat{\beta}$, $\hat{V}(\hat{\beta})$ to get confidence intervals with the appropriate coverage

Modify the estimation process to deal with missing values

Maximum likelihood inference: Expectation Maximization algorithms

Pros: Tailored toward a specific problem

Cons: Few softwares even for simple models. Ex: logistic regression[7]
Need to design one specific algorithm for each statistical method

(Multiple) imputation to get a complete data set

Pros: Any analysis can be performed, `mice` R package

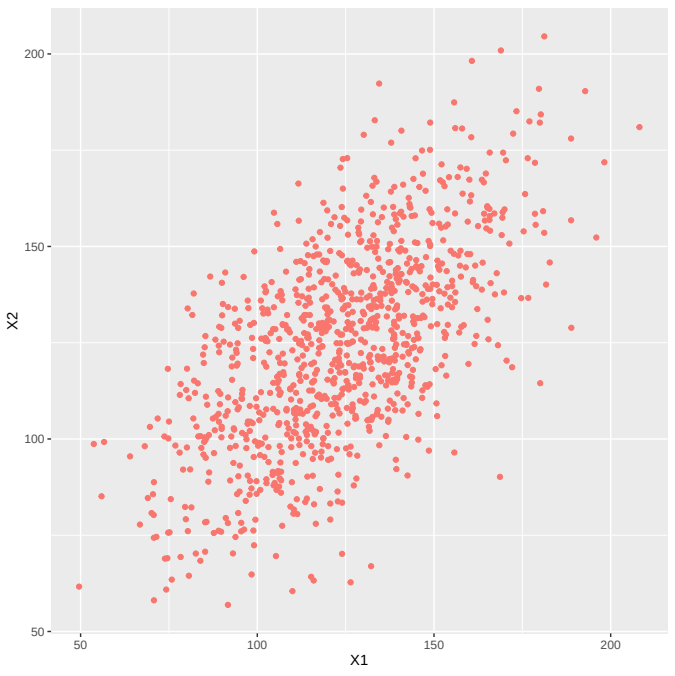
Cons: Generic

[6] Mayer, **J.** et al. A unified platform for missing values methods and workflows. *R journal*. 2022.
[7] Jiang, **J.** et al. Logistic Regression with Missing Covariates, Parameter Estimation, Model Selection and Prediction. *CSDA*. 2019. - Implementation in the `misaem package`

Single imputation by the mean

▷ $(x_{i1}, x_{i2}) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_{x_1}, \mu_{x_2}), \Sigma_{x_1 x_2})$

$\mathbf{X}_1$	$\mathbf{X}_2$
-0.56	-1.93
-0.86	-1.50
.....	...
2.16	0.7
0.16	0.74



$\mu_{x_2} = 0$

$\sigma_{x_2} = 1$

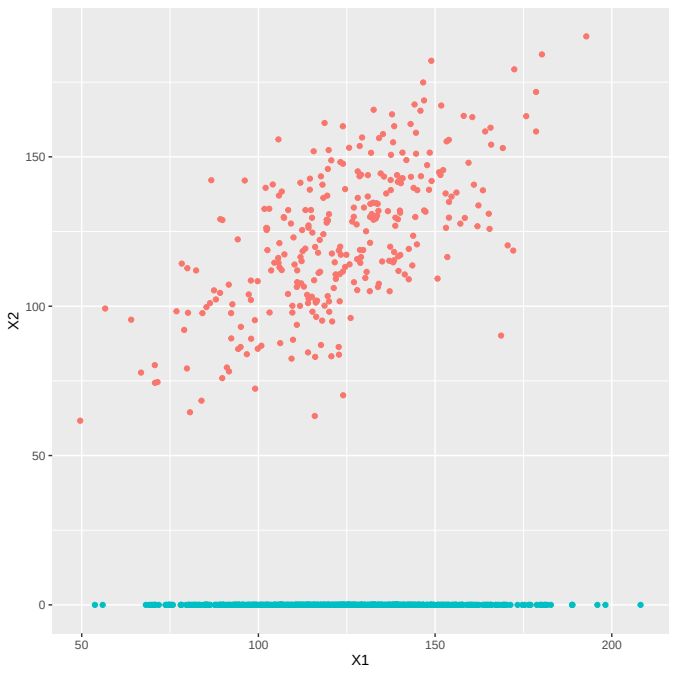
$\rho = 0.6$

$\hat{\mu}_{x_2} = -0.01$
$\hat{\sigma}_{x_2} = 1.01$
$\hat{\rho} = 0.66$

Single imputation by the mean

- $(x_{i1}, x_{i2}) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_{x_1}, \mu_{x_2}), \Sigma_{x_1 x_2})$
- 70 % of missing entries completely at random on X_2

$\mathbf{X}_1$	$\mathbf{X}_2$
-0.56	NA
-0.86	NA
.....	...
2.16	0.7
0.16	NA



$\mu_{x_2} = 0$

$\sigma_{x_2} = 1$

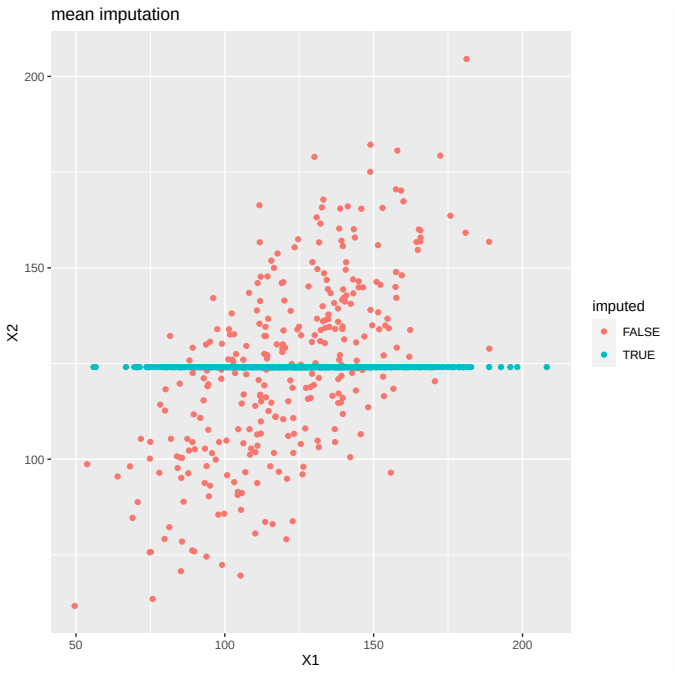
$\rho = 0.6$

$\hat{\mu}_{x_2} = 0.18$
$\hat{\sigma}_{x_2} = 0.9$
$\hat{\rho} = 0.6$

Single imputation by the mean

- $(x_{i1}, x_{i2}) \underset{\text{i.i.d.}}{\sim} \mathcal{N}_2((\mu_{x_1}, \mu_{x_2}), \Sigma_{x_1 x_2})$
- 70 % of missing entries completely at random on X_2
- Estimate parameters on the mean imputed data

$\mathbf{X}_1$	$\mathbf{X}_2$
-0.56	**0.01**
-0.86	**0.01**
.....	...
2.16	0.7
0.16	**0.01**



$\mu_{x_2} = 0$

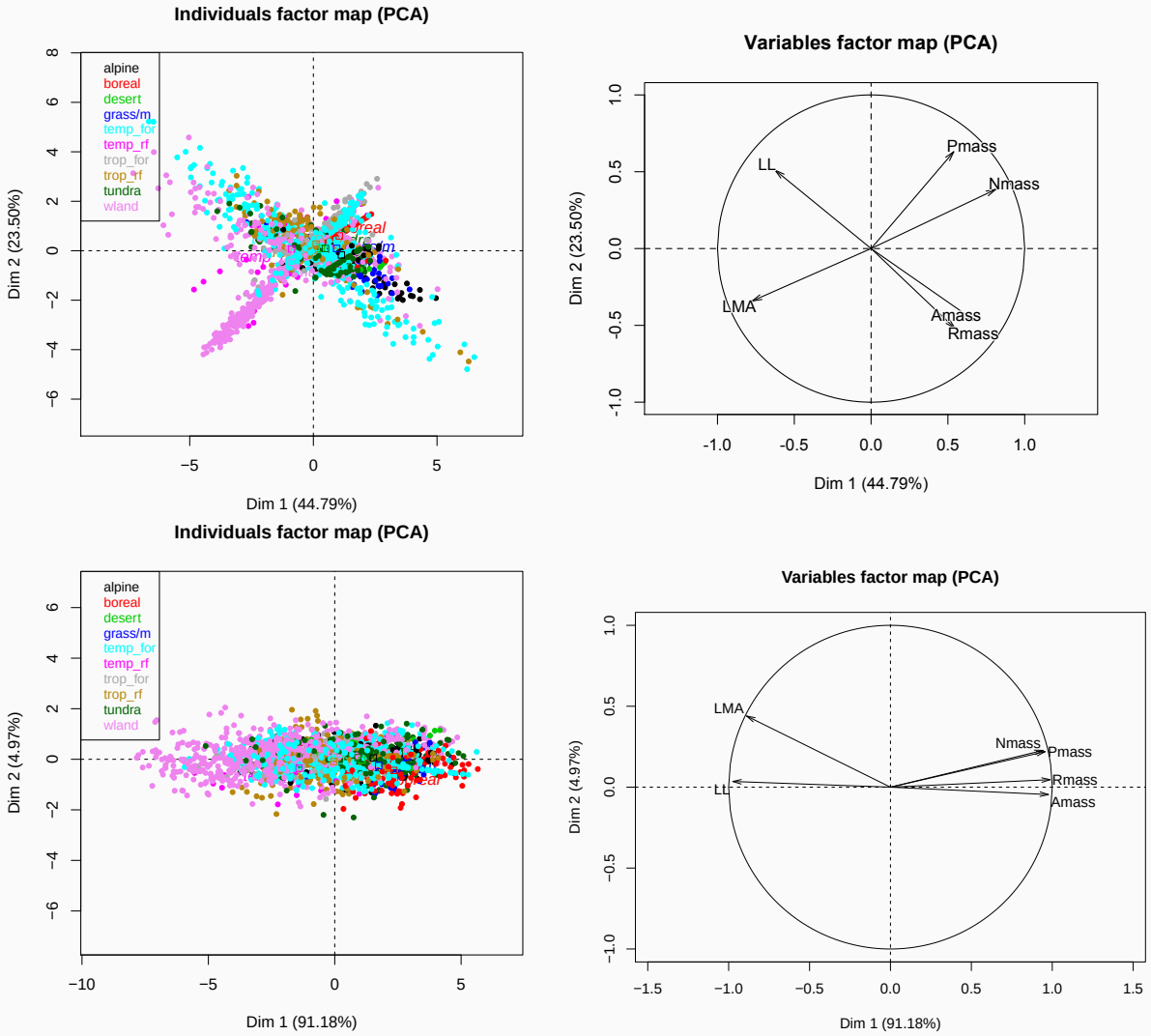
$\sigma_{x_2} = 1$

$\rho = 0.6$

$\hat{\mu}_{x_2} = 0.01$
$\hat{\sigma}_{x_2} = 0.5$
$\hat{\rho} = 0.30$

Mean imputation deforms joint and marginal distributions

Mean imputation should be avoided 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)
```

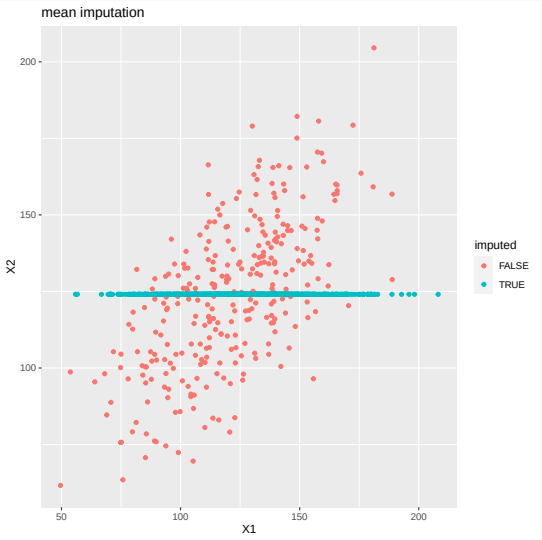
Ecological data: $n = 69000$ species - 6 traits. Estimated correlation between Pmass & Rmass ≈ 0 (mean imputation) or ≈ 1 (EM PCA)[8]

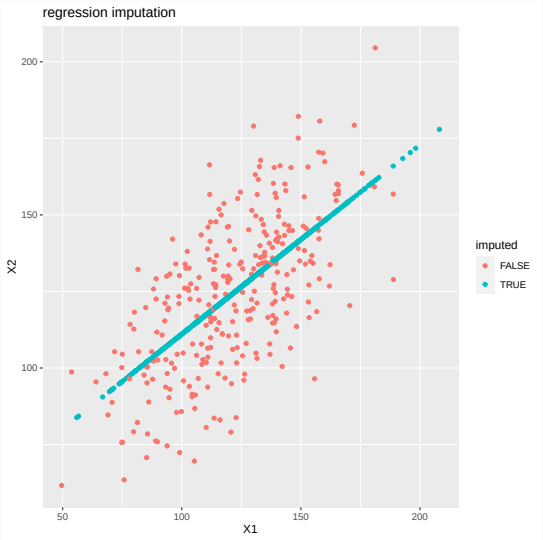
[8] **J.** & Husson. missMDA: Handling Missing Values in Multivariate Data Analysis, *JSS*. 2016.

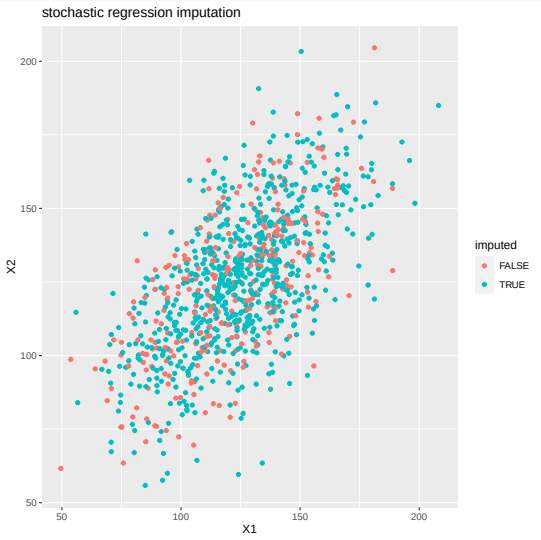
Objective: to impute while preserving distribution

Assuming a bivariate gaussian distribution $x_{i2} = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

- ▷ Regression imputation: Estimate β (here with complete data) and impute $\hat{x}_{i2} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} \Rightarrow$ variance underestimated and correlation overestimated
- ▷ Stochastic reg. imputation: Estimate β and σ - impute from the predictive $\hat{x}_{i2} \sim \mathcal{N}\left(\beta_0 + \hat{\beta}_1 x_{i1}, \hat{\sigma}^2\right) \Rightarrow$ preserve distributions







$\mu_{x_2} = 0$

$\sigma_{x_2} = 1$

$\rho = 0.6$

0.01
0.5
0.30

0.01
0.72
0.78

0.01
0.99
0.59

Impute while preserving distribution. Multivariate case

Assuming a joint distribution

- Gaussian model $x_i \sim \mathcal{N}(\mu, \Sigma)$
- <u>Low rank</u> : $X_{n\times d} = \mu_{n\times d} + \varepsilon\ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\, \sigma^2\right)$ with μ of low rank
  - $\Rightarrow$ Powerful in recommendation system: Netflix prize 90% of missing
  - $\Rightarrow$ Use similarities between rows & links between variables + reduct. of dim.
  - $\Rightarrow$ Different regularization depending on noise regime[9,10]
  - $\Rightarrow$ Count data,[11] ordinal data, categorical data, blocks/multilevel data[12]
- Using optimal transport,[13] deep generative models

[9] **J.** & Sardy. Adaptive Shrinkage of singular values. *Stat & Computing.* 2015.
[10] **J.** & Wager. Stable autoencoding for regularized low-rank matrix estimation. *JMLR.* 2016.
[11] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects and interactions in mixed and incomplete data frames. 2019. *JASA*.
[12] **J.** et al. Imputation of mixed data with multilevel SVD. *JCGS*. 2018.
[13] Muzelec, Cuturi, Boyer, **J.** Missing Data Imputation using Optimal Transport. *ICML*. 2020.

Impute while preserving distribution. Multivariate case

Assuming a joint distribution

- Gaussian model $x_i \sim \mathcal{N}(\mu, \Sigma)$
- Low rank : $X_{n\times d} = \mu_{n\times d} + \varepsilon\ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\, \sigma^2\right)$ with μ of low rank
  - $\Rightarrow$ Powerful in recommendation system: Netflix prize 90% of missing
  - $\Rightarrow$ Use similarities between rows & links between variables + reduct. of dim.
  - $\Rightarrow$ Different regularization depending on noise regime[9,10]
  - $\Rightarrow$ Count data,[11] ordinal data, categorical data, blocks/multilevel data[12]
- Using optimal transport,[13] deep generative models

Iterating conditional models (joint distribution implicitly defined)

- with multinomial, Poisson regression (ICE: Imputation by Chained Equations)
- iterative imputation of each variable by random forests

---

[9] **J.** & Sardy. Adaptive Shrinkage of singular values. *Stat & Computing.* 2015.
[10] **J.** & Wager. Stable autoencoding for regularized low-rank matrix estimation. *JMLR.* 2016.
[11] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects and interactions in mixed and incomplete data frames. 2019. *JASA.*
[12] **J.** et al. Imputation of mixed data with multilevel SVD. *JCGS.* 2018.
[13] Muzelec, Cuturi, Boyer, **J.** Missing Data Imputation using Optimal Transport. *ICML.* 2020.

Single imputation is not enough: Underestimates 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 is not enough: Underestimates 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

Single imputation is not enough: Underestimates 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

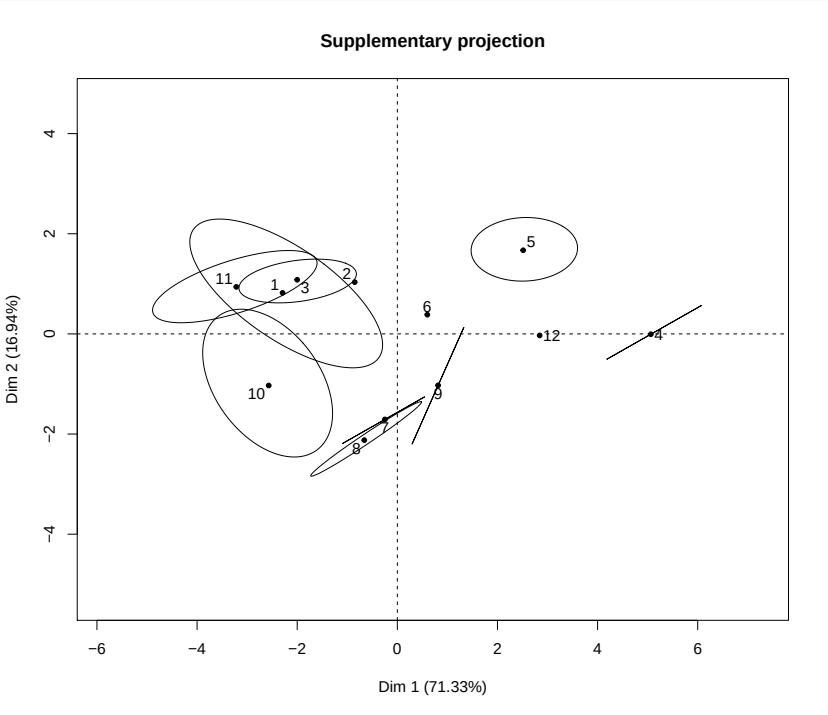
X_1	X_2	X_3	Y
7	20	10	s
-6	45	12	s
0	-5	30	no s
2	32	35	s
-2	20	12	no s
1	63	40	s

Visualization of the imputed values[14]

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)
```

Projection of the M imputed data on a 'compromise' subspace (PCA with missing values)

Is it possible to handle 30% of missing values? 50%?, etc. **Both % of missing values & signal matter (5% of NA can be an issue)**

[14] **J.** et al. Multiple imputation in principal component analysis. *ADAC*. 2011.

Multiple imputation: standard errors are not underestimated

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):

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

$$T = \underbrace{\frac{1}{M}\sum_{m=1}^{M}\widehat{Var}\left(\hat{\beta}_m\right)}_{\textbf{Within}\text{-imputation variance}} + \underbrace{\left(1+\tfrac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_m-\hat{\beta}\right)^2}_{\textbf{Between}\text{-imputation variance}}$$

$\Rightarrow$ Variability of missing values taken into account.

Take home message on inference & imputation

- **Methods used in practice are the one implemented in a sustainable way**: few implementations of EM strategies

- **"Imputation is both seductive & dangerous"** (Dempster & Rubin, 1983).

Seductive: *"can lull the user into the pleasant state of* **believing that the data are complete**

Dangerous: *"it lumps together situations where the problem is minor enough to be handled in this way & situations where estimators applied to the imputed data have* **substantial biases***."*

Take home message on inference & imputation

- **Methods used in practice are the one implemented in a sustainable way**: few implementations of EM strategies

- **"Imputation is both seductive & dangerous"** (Dempster & Rubin, 1983).
Seductive: *"can lull the user into the pleasant state of* **believing that the data are complete**

Dangerous: *"it lumps together situations where the problem is minor enough to be handled in this way & situations where estimators applied to the imputed data have* **substantial biases**.*"*

- **Single imputation** aims at completing data as best as possible
  $\Rightarrow$ **low rank approaches** are powerful for heterogeneous data

- **Multiple imputation** aims at estimating the parameters and their variability taking into account **the uncertainty of the missing values**

[15] Bogdan, **J.** et al. Adaptive Bayesian SLOPE - High dimensional Model Selection with Missing Values. *Journal of Computational and Graphical Statistics.* 2020.

Take home message on inference & imputation

- **Methods used in practice are the one implemented in a sustainable way**: few implementations of EM strategies

- **"Imputation is both seductive & dangerous"** (Dempster & Rubin, 1983).
Seductive: *"can lull the user into the pleasant state of* **believing that the data are complete**

Dangerous: *"it lumps together situations where the problem is minor enough to be handled in this way & situations where estimators applied to the imputed data have* **substantial biases**.*"*

- **Single imputation** aims at completing data as best as possible
  $\Rightarrow$ **low rank approaches** are powerful for heterogeneous data

- **Multiple imputation** aims at estimating the parameters and their variability taking into account **the uncertainty of the missing values**
  - How to aggregate lasso regressions? Alternatives EM[15]

[15] Bogdan, **J.** et al. Adaptive Bayesian SLOPE - High dimensional Model Selection with Missing Values. *Journal of Computational and Graphical Statistics.* 2020.

Challenges and on-going work with heterogeneous data sources and missing data

		Clinical Data					Biological Data				Questionnaire on lifestyle		
		X_1	....	X_p	W	Y	Z_1	.....	Z_q	....	C_1	...	C_r
Obs Hospital 1	1		NA							....			
			NA	NA						...			
			NA							...			
	n_1	NA	NA							...			
Obs Hospital 2	1				NA	NA				...		NA	NA
		NA		NA	NA	NA	NA	NA	NA	...			
					NA	NA				...	NA	NA	NA
	n_2				NA	NA				...			
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Obs Hospital K	1	NA	NA	NA						...		NA	
		NA								...		NA	
		NA								....		NA	
	n_K	NA								....		NA	

Sporadic & systematic (missing variable in one hospital). Due to the pandemic, many patients did not complete their tests

- What to do when you have both MCAR, MAR, **MNAR** in the data?

Supervised learning with missing values

Prediction with missing values

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

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

Find a regression function that minimizes the expected risk

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

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

$\Rightarrow$ One model per pattern m of missing values (2^d patterns)[16]

[16] Rosenbaum & Rubin. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *JASA*.

Prediction with missing values

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

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

Find a regression function that minimizes the expected 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 m of missing values (2^d patterns)[16]

[16] Rosenbaum & Rubin. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *JASA*.

Prediction with missing values

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

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

Find a regression function that minimizes the expected risk

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

A learner estimates the regression function from a train set minimizing the empirical risk: $\hat{f}_{\mathcal{D}_{n,\text{train}}} \in \underset{f\colon \widetilde{\mathbb{R}}^d \to \mathbb{R}}{\arg\min} \left(\frac{1}{n}\sum_{i=1}^n \ell\left(f(\tilde{X}_i), Y_i\right)\right)$

A new data $\mathcal{D}_{n,\text{test}}$ to estimate the generalization error rate

- Bayes consistent: $\mathbb{E}[\ell(\hat{f}_n(\tilde{X}), Y)] \xrightarrow[n\to\infty]{} \mathbb{E}[\ell(f^\star(\tilde{X}), Y)]$

Supervised learning with missing values

Differences with classical litterature

Aim: predict an outcome Y (not estimate parameters & their variance)

Specificities: both train & test sets with missing values;
Otherwise, distributional shift (data generating process (X, Y, M))

$\Rightarrow$ Is it possible to use previous approaches (EM - impute), consistent?
$\Rightarrow$ Do we need to design new ones?

Supervised learning with missing values

Differences with classical litterature

Aim: predict an outcome Y (not estimate parameters & their variance)

Specificities: both train & test sets with missing values;
Otherwise, distributional shift (data generating process (X, Y, M))

$\Rightarrow$ Is it possible to use previous approaches (EM - impute), consistent?
$\Rightarrow$ Do we need to design new ones?

Imputation prior to learning: Impute then Regress

Common practice: use off-the-shelf methods for

1) imputation of missing values
2) supervised-learning on the completed data

Impute train & test sets with the same model. Easy with univariate imputation: compute the means on the observed data $(\hat{\mu}_1, ..., \hat{\mu}_d)$ of each column of the train set & impute the test set with such means

Constant (mean) imputation is consistent for prediction[16]

Framework - assumptions

- Regression model: $Y = f^\star(X) + \varepsilon$
  $f^\star : \mathbb{R}^d \to \mathbb{R}$ a continuous function of the complete data X
  $\varepsilon \in \mathbb{R}$ is a centered random noise variable independent of (X, M_1)
  $X = (X_1, \ldots, X_d)$ has a continuous density $g > 0$ on $[0,1]^d$
  $\|f^\star\|_\infty = \sup_{x \in \mathbb{R}^d} |f^\star(x)| < \infty$

- 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

[16] **J.** et al. Consistency of supervised learning with missing values. 2019.

Constant (mean) imputation is consistent for prediction[16]

- Constant imputation $x' = (x'_1, x_2, \ldots, x_d)$: $x'_1 = x_1 \mathbb{1}_{M_1=0} + \alpha \mathbb{1}_{M_1=1}$
- Use a **universally consistent algorithm** (for all distribution) to approach the regression function $f^\star_{impute}(x') = \mathbb{E}[Y|X = x']$

Theorem. (J. et al. 2019)

$$\begin{aligned} f^\star_{impute}(x') =& \mathbb{E}[Y|X_2 = x_2, \ldots, X_d = x_d, M_1 = 1] \\ & \mathbb{1}_{x'_1=\alpha} \mathbb{1}_{\mathbb{P}[M_1=1|X_2=x_2,\ldots,X_d=x_d]>0} \\ & + \mathbb{E}[Y|X = x'] \mathbb{1}_{x'_1=\alpha} \mathbb{1}_{\mathbb{P}[M_1=1|X_2=x_2,\ldots,X_d=x_d]=0} \\ & + \mathbb{E}[Y|X = x', M_1 = 0] \mathbb{1}_{x'_1 \neq \alpha}. \end{aligned}$$

Prediction with constant is equal to the Bayes function almost everywhere

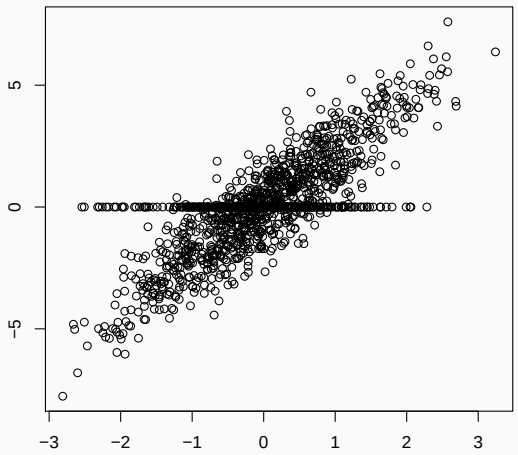
$$f^\star_{impute}(X') = f^\star(\tilde{X}) = \mathbb{E}[Y|\tilde{X}]$$

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

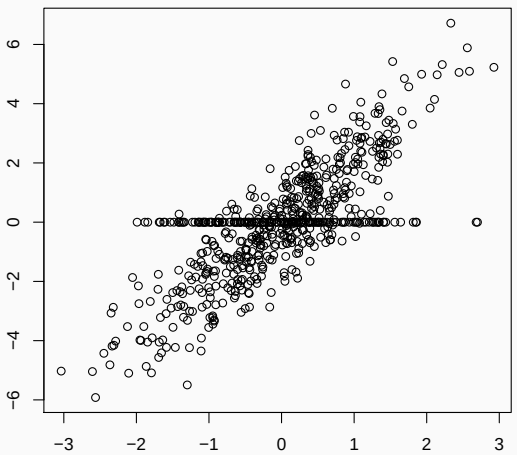
[16] **J.** et al. Consistency of supervised learning with missing values. 2019.

Consistency of constant imputation: Rationale

- Specific value, systematic like a code for missing
- The learner detects the code and recognizes it at the test time (the imputed data distribution shouldn't differ between train and test)
- With categorical data, just code "Missing"
- With continuous data, any constant:
- **De-identified/imputed missing data**: recovers from which pattern it comes
- Need a lot of data (asymptotic result) and a universally consistent learner



Train

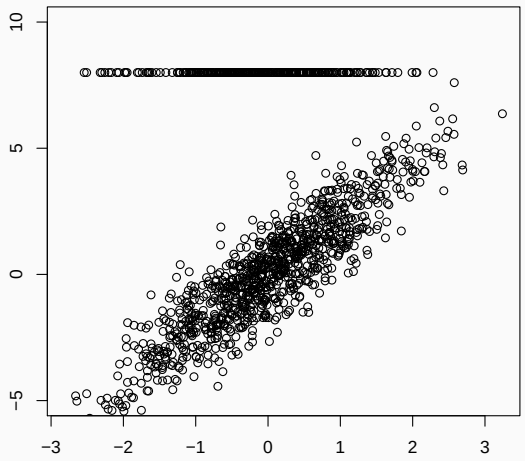


Test

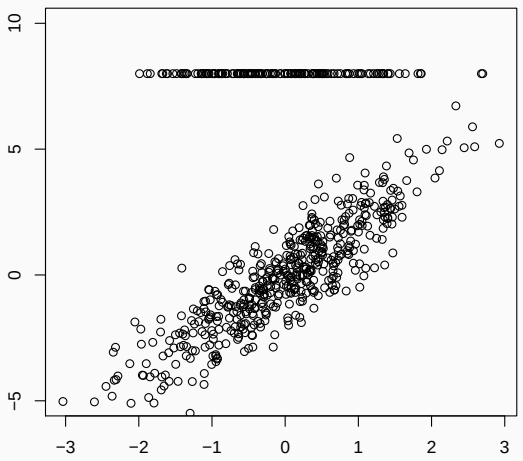
Imputing both train & test with the same constant and regress is consistent despite its drawbacks for estimation (useful in practice)

Consistency of constant imputation: Rationale

- Specific value, systematic like a code for missing
- The learner detects the code and recognizes it at the test time (the imputed data distribution shouldn't differ between train and test)
- With categorical data, just code "Missing"
- With continuous data, any constant: out of range
- **De-identified/imputed missing data**: recovers from which pattern it comes
- Need a lot of data (asymptotic result) and a universally consistent learner



Train



Test

Imputing both train & test with the same constant and regress is consistent despite its drawbacks for estimation (useful in practice)

Bayes optimality of impute-n-regress[17]

- **Imputation function**: $\forall m \in \{0,1\}^d$, let $\phi^{(m)} \in \mathcal{C}_\infty$: $\mathbb{R}^{|obs(m)|} \to \mathbb{R}^{|mis(m)|}$

which outputs values for the missing entries based on the observed ones

$$\Phi : (\mathbb{R} \cup \{\texttt{NA}\})^d \to \mathbb{R}^d : \forall j \in [\![1, d]\!] \,,\ \Phi_j(\widetilde{X}) = \begin{cases} X_j & \text{if } M_j = 0 \\ \phi_j^{(M)}(X_{obs(M)}) & \text{if } M_j = 1 \end{cases}$$

- **Regression on imputed data:** $g_\Phi^\star \in \underset{g:\mathbb{R}^d \mapsto \mathbb{R}}{\text{argmin}} \quad \mathbb{E}\left[\left(Y - g \circ \Phi(\widetilde{X})\right)^2\right]$,

minimizer of the risk on the imputed data

Theorem

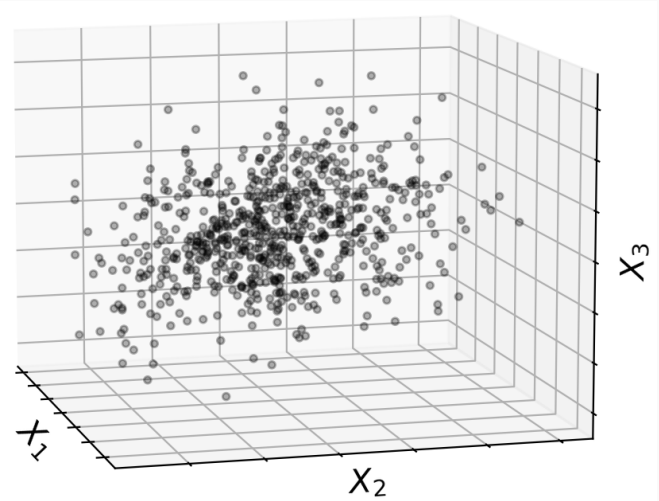
Assume that the response Y satisfies $Y = f^\star(X) + \varepsilon$
Then, for all missing data mechanisms & almost all imputation functions,
$g_\Phi^\star \circ \Phi$ is Bayes optimal

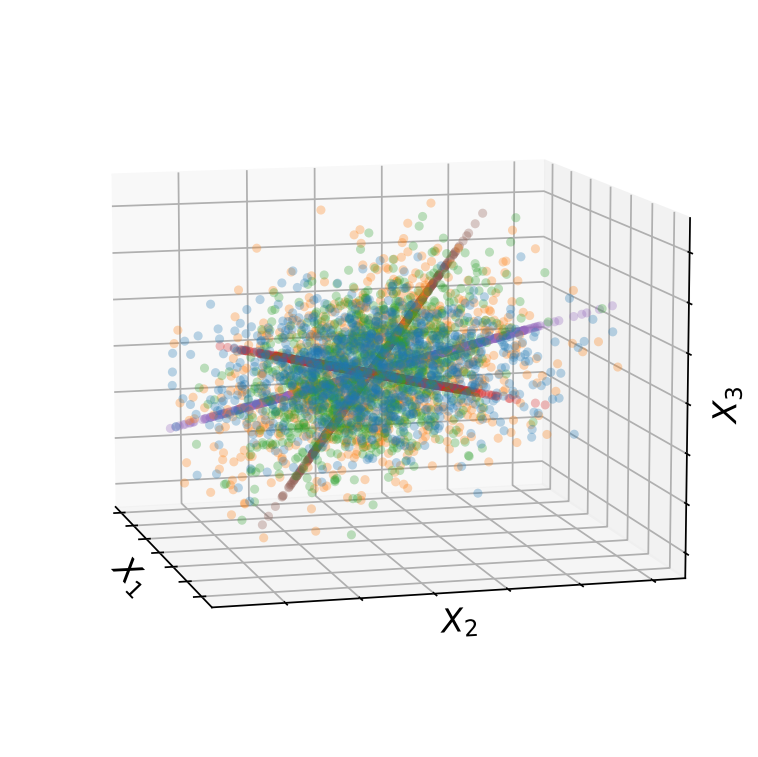
$\Rightarrow$ A universally consistent algorithm trained on the imputed data $\Phi(\widetilde{X})$ is Bayes consistent

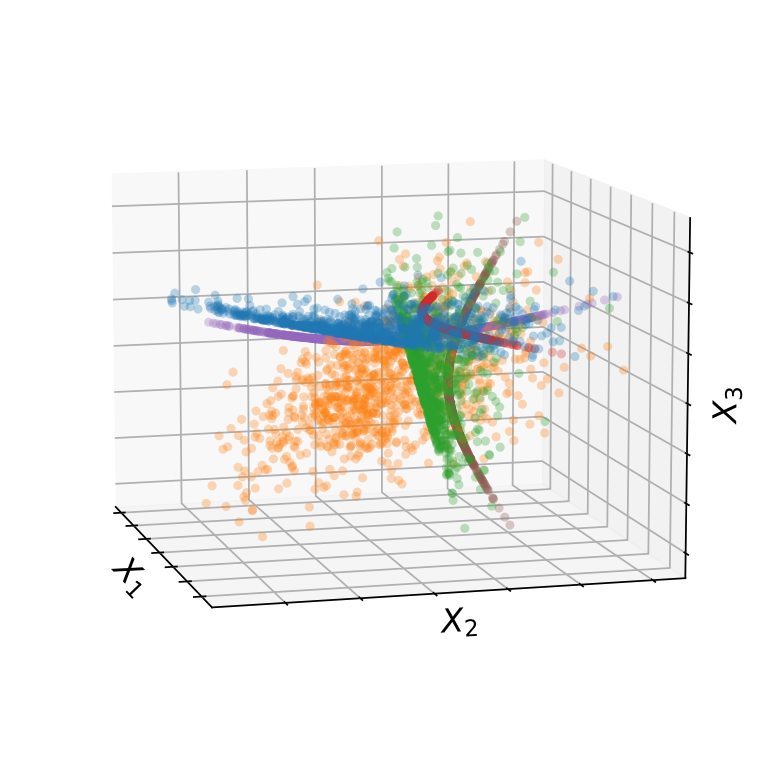
Asymptotically, imputing well is not needed to predict well

[17] Le Morvan, **J.** et al. What's a good imputation to predict with missing values? Neurips2021

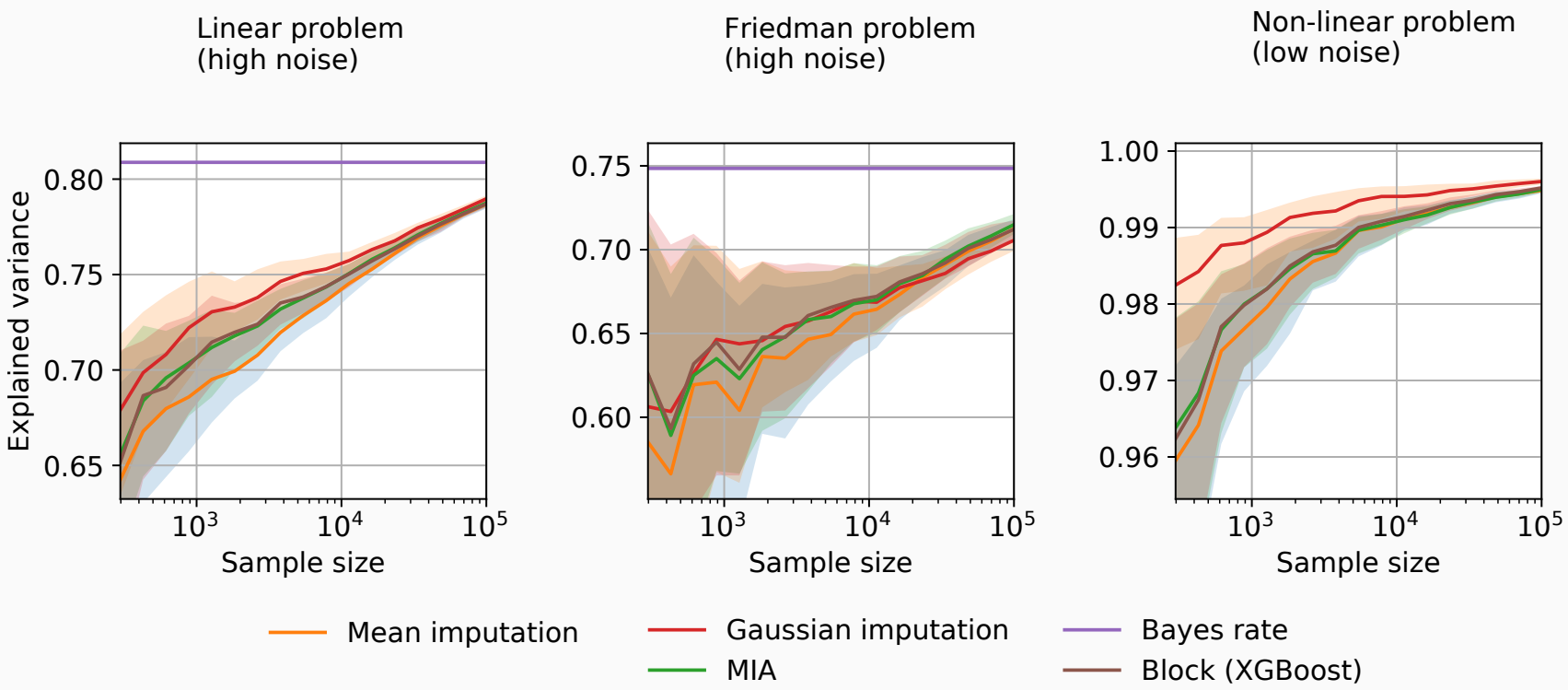
Rationale of proof: imputation creates manifolds







Which imputation function should one choose?



Consistency of impute-then-regress. Ex: 3 regression models, 40% of MCAR in covariates, different imputation methods, then regress with random forests.

- A "better" imputation could create an easier learning problem
- Constant imputation is consistent but introduces strong discontinuities

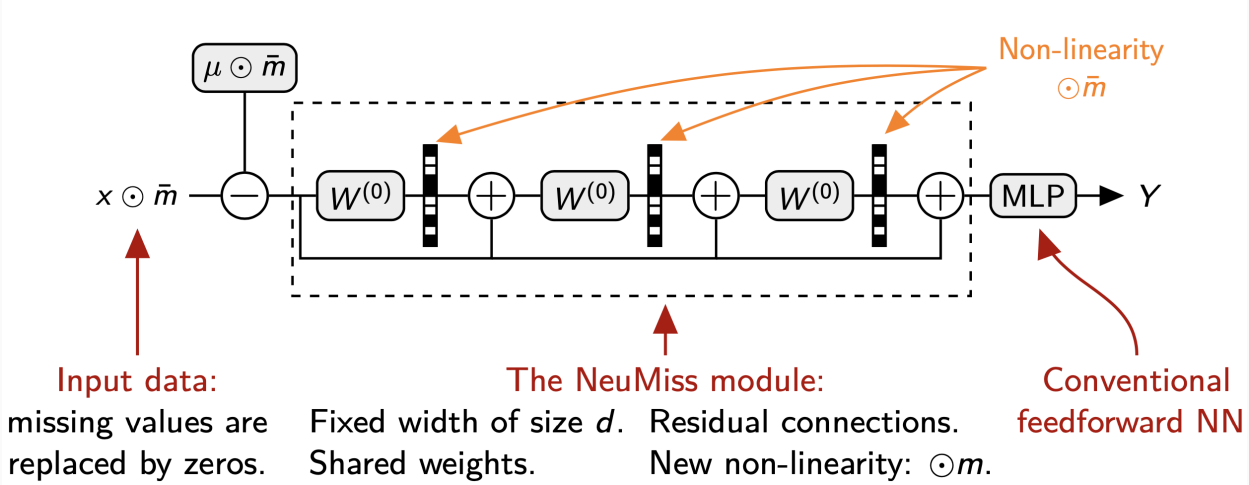
$\Rightarrow$ Which imputation and predictor should one use?

Best imputation is joint learn with regression[18]

- Neumiss network:

▷ Classic network with multiplications by the mask nonlinearities $\odot M$
▷ Motivated by linear regression with missing values in the covariates
▷ Theoritically grounded: approximation of the Bayes predictor (truncated neumiss series to approximate inverses of covariance matrices)

- Couple Neumiss and MLP to jointly learn imputation and regression



[18] Le morvan, **J.** et al. Neumiss networks: differential programming for supervised learning with missing values. *Neurips2020 (Oral)*.

Take home message in supervised learning with missing values

Supervised learning different from inferential aim

Bayes optimality of Impute then Regress

- Single constant imputation is consistent with a universally consistent learner
- Can even work in MNAR
- Rethinking imputation: a good imputation is the one that makes the prediction easy

Implicit and jointly learned Impute-then-Regress strategy

- Neumiss network: new architecture $\odot M$ nonlinearity
- **Tree-based models: Missing Incorporated in Attribute (MIA)** (implemented in generalized random forest package `grf`, `partykit`)

Challenges and on-going works with missing values

- Uncertainty quantification with conformal prediction[19] (with Y. Romano & A. Dieuleveut)
  Predictive intervals for any predictive algorithm (neural nets, random forests), in finite samples with no assumption on the data distribution except for the exchangeability

[19] Vovk, et al. Algorithmic Learning in a Random World. Springer US, 2005.

Challenges and on-going works with missing values

- Uncertainty quantification with conformal prediction[19] (with Y. Romano & A. Dieuleveut)
  Predictive intervals for any predictive algorithm (neural nets, random forests), in finite samples with no assumption on the data distribution except for the exchangeability

- Times series with MNAR (prediction with covariates measured regularly over time & static covariates)

- Federated learning with missing values

[19] Vovk, et al. Algorithmic Learning in a Random World. Springer US, 2005.
[20] Wager, **J.**, Doubly robust estimation with incomplete confounders. *Ann. Appl. Stat.* 2020.
[21] Colnet, **J.** et al. Generalization: sensitivity analysis & missing data. *Journal of causal inf.* 2022.
[22] Mayer & **J.** Generalizing treatment effects with incomplete covariates. 2022.
[23] Colnet, **J.** et al. Reweighting RCT for generalization: finite sample analysis & var. select. 2022.

Challenges and on-going works with missing values

- Uncertainty quantification with conformal prediction[19] (with Y. Romano & A. Dieuleveut)
  Predictive intervals for any predictive algorithm (neural nets, random forests), in finite samples with no assumption on the data distribution except for the exchangeability

- Times series with MNAR (prediction with covariates measured regularly over time & static covariates)

- Federated learning with missing values

- **Causal inference** from incomplete heterogeneous sources,[20],[21],[22],[23] (treatment estimation, generalisation of randomized control trial, etc.)

[19] Vovk, et al. Algorithmic Learning in a Random World. Springer US, 2005.
[20] Wager, **J.**, Doubly robust estimation with incomplete confounders. *Ann. Appl. Stat.* 2020.
[21] Colnet, **J.** et al. Generalization: sensitivity analysis & missing data. *Journal of causal inf.* 2022.
[22] Mayer & **J.** Generalizing treatment effects with incomplete covariates. 2022.
[23] Colnet, **J.** et al. Reweighting RCT for generalization: finite sample analysis & var. select. 2022.

Collaborators on missing values

- F. Husson, Prof. Agronomy University. (package `missMDA`, `FactoMineR`)
- Gosia Bogdan, Prof. Wroclaw. High dimensional regression
- Claire Boyer, Assoc. Prof. Sorbonne. Signal processing, missing values
- Aymeric Dieuleveut, Asso. Prof. Ecole Polytechnique, Paris. Optimization
- Imke Mayer, Postdoc Charité Institute, Berlin. Causal inference
- Aude Sportisse, Postdoc Inria Nice. Missing values
- Marine Le Morvan, Junior researcher at INRIA, Paris. Supervised learning
- Erwan Scornet, Asso. Prof. Ecole Polytechnique, Paris. Random forests
- Gael Varoquaux, Senior researcher at INRIA, Paris. ML, Scikit-learn
- Margaux Zaffran, PhD student, EDF. Conformal prediction

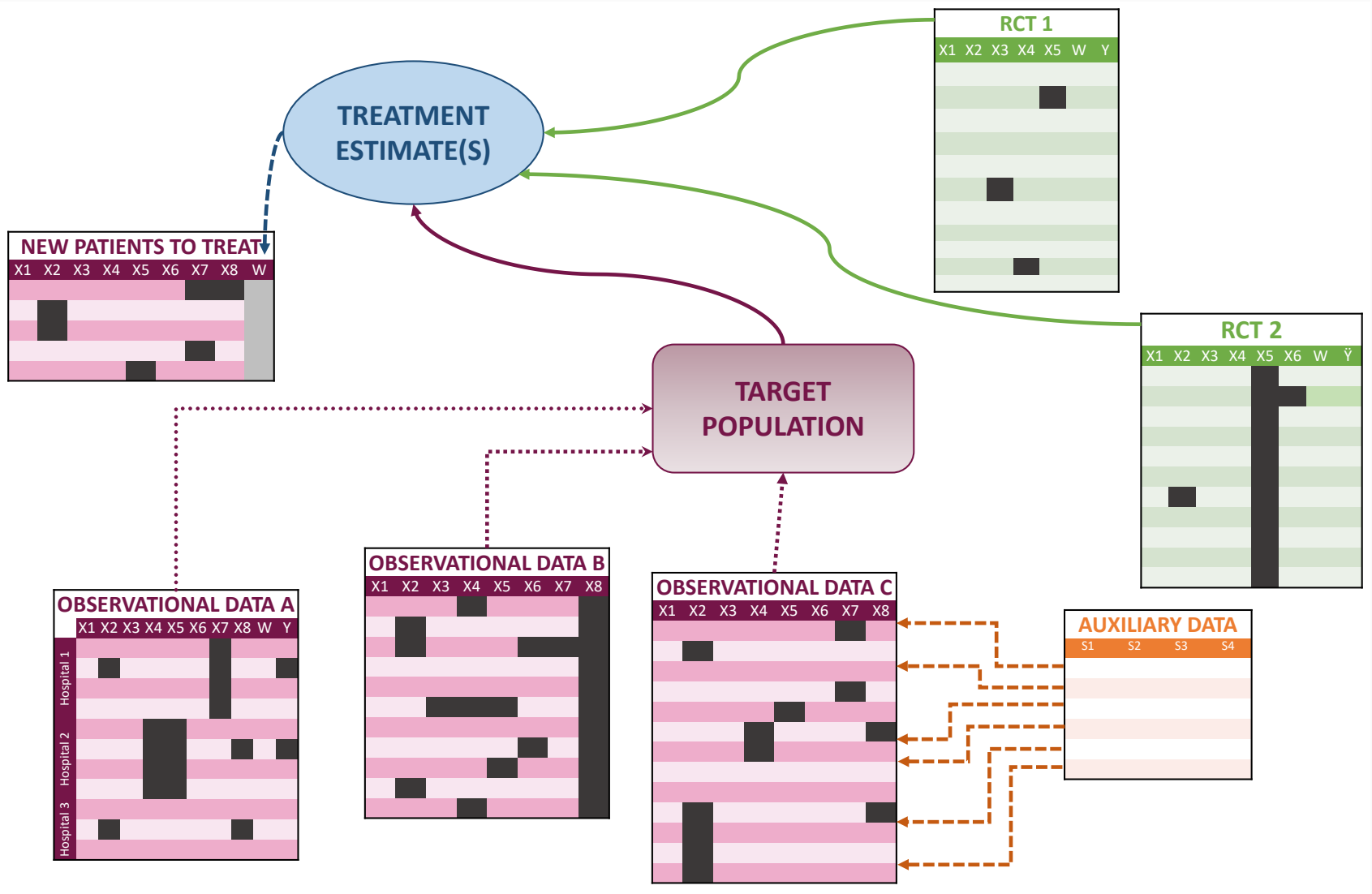









Challenges with heterogeneous sources and missing data



Monitor population & assess wetlands conservation policies

- National agency for wildlife and hunting management (ONCFS) data
- Contingency tables: Water (785 wetland sites) - bird (23 species) count data, from 1990-2016 in 5 countries in North Africa
- Side information (17 variables) on sites & years: meteo, altitude, etc.

Common pochard (canard milouin)

Site	2008	2009	2010
1	NA	0	0
2	4	50	25
3	NA	0	0
4	NA	NA	NA
5	NA	NA	NA
6	0	0	0
7	5	75	870
8	9	34	0
9	10	8	30

Site	Year	Rain	Eco	Country	Agri
1	2008	163.7	0.8	Algeria	16.2
2	2008	60.7	0.8	Algeria	16.2
3	2008	227.9	0.8	Algeria	16.2
[illegible]	2008	174.8	0.8	Algeria	16.2
5	2008	163.7	0.8	Algeria	16.2
6	2008	230.7	0.8	Algeria	16.2
7	2008	243.5	0.8	Algeria	16.2
8	2008	262.6	0.8	Algeria	16.2
9	2008	197.3	0.8	Algeria	16.2
10	2008	227.9	0.8	Algeria	16.2

$\Rightarrow$ Aims: Assess the effect of time on species abundances

$\Rightarrow$ 70% of missing values in contingency tables (drough, war, etc.)[24,25]

[24] Robin, **J.**, Moulines Sardy. 2019. Low-rank model with covariates for count data with missing values. *Journal of Multivariate Analysis.*

[25] Robin, Klopp, **J.**, Moulines Tibshirani. Main effects and interactions in mixed and incomplete data frames. 2019. *JASA.*

Iterative imputation by random forests versus by low rank (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

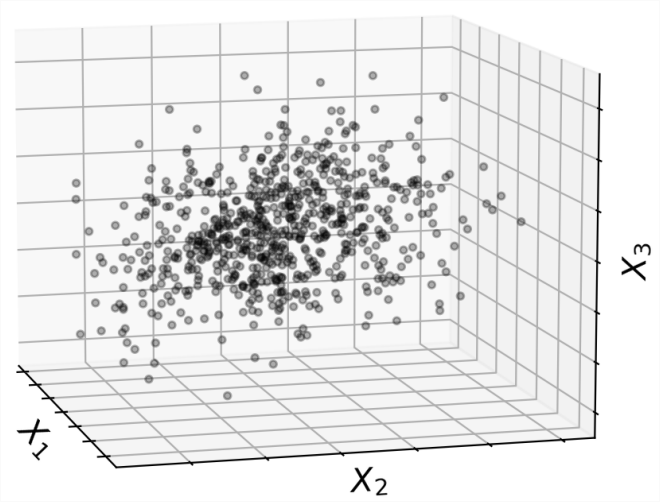
`missForest`

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

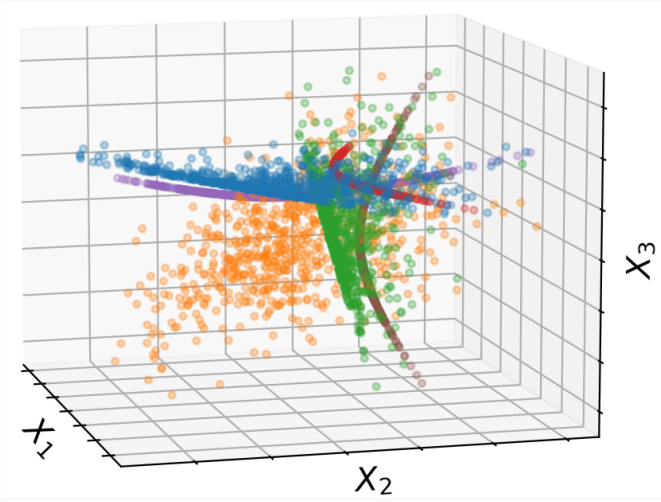
`imputePCA`

$\Rightarrow$ Imputation inherits from the method: Random forests (computationaly costly) handles non linear relationships/ PCA linear ones

Bayes optimality of impute-n-regress (Le morvan et al. 2021)



Complete data



Imputed data (manifolds)

Rationale: Imputation create manifolds to which the learner adapts

1. All data points with a missing data pattern m are mapped to a manifold $\mathcal{M}^{(m)}$ of dimension $|obs(m)|$ (Preimage Theorem)
2. The missing data patterns of imputed data points can almost surely be de-identified (Thom transversality Theorem)[26]
3. Given 2), we can build prediction functions, independent of m, that are Bayes optimal for all missing data patterns

[26] Non transverse: the manifolds on which the data with either x1 missing or x2 missing are projected are exactly the same (the same line)

Jointly learn imputation & prediction: Neumiss

Linear regression with missing values

Linear model:

$Y = \beta_0 + \langle X, \beta \rangle + \varepsilon, \quad X \in \mathbb{R}^d, \ \varepsilon$ Gaussian

Bayes predictor for the linear model:

$$\begin{aligned} f^\star(\tilde{X}) &= \mathbb{E}[Y|\tilde{X}] = \mathbb{E}[\beta_0 + \beta^\mathsf{T} X \mid \ M, X_{obs(M)}] \\ &= \beta_0 + \beta_{obs(M)}^\mathsf{T} X_{obs(M)} + \beta_{mis(M)}^\mathsf{T} \ \mathbb{E}[X_{mis(M)} \mid M, X_{obs(M)}] \\ &= \sum_{m \in \{0,1\}^d} \beta_0 + \beta_{obs(m)}^\mathsf{T} X_{obs(m)} + \beta_{mis(m)}^\mathsf{T} \ \mathbb{E}[X_{mis(m)} \mid M = m, X_{obs(m)}] \end{aligned}$$

Assumptions on covariates and missing values (X, M)

1. Gaussian assumption $X \sim \mathcal{N}(\mu, \Sigma)$ + MCAR and MAR

Under Assump. the Bayes predictor is linear per pattern

$$f^\star(X_{obs}, M) = \beta_0 + \langle \beta_{obs}, X_{obs} \rangle + \langle \beta_{mis}, \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}) \rangle$$

use of obs instead of $obs(M)$ for lighter notations - Expression for 2.

Linear model with missing values not necessarely linear

Example

Let $Y = X_1 + X_2 + \varepsilon$, where $X_2 = \exp(X_1) + \varepsilon_1$. Now, assume that only X_1 is observed. Then, the model can be rewritten as

$$Y = X_1 + \exp(X_1) + \varepsilon + \varepsilon_1,$$

where $f(X_1) = X_1 + \exp(X_1)$ is the Bayes predictor. In this example, the submodel for which only X_1 is observed is not linear.

$\Rightarrow$ There exists a large variety of submodels for a same linear model. Depend on the structure of X and on the missing-value mechanism.

Neumiss Networks to approximate the covariance matrix

Bayes predictor requires inverting many covariance matrices

$$f^\star(X_{obs}, M) = \beta_0^+ \langle \beta_{obs}, X_{obs} \rangle + \langle \beta_{mis}, \mu_{mis} + \Sigma_{mis,obs} (\Sigma_{obs})^{-1} (X_{obs} - \mu_{obs}) \rangle$$

Order-ℓ approx of $(\Sigma_{obs(m)}^{-1})$ for any m defined recursively:

$$S_{obs(m)}^{(\ell)} = (Id - \Sigma_{obs(m)}) S_{obs(m)}^{(\ell-1)} + Id.$$

Neuman Series, $S^{(0)} = Id$, $\ell = \infty$: $(\Sigma_{obs(m)})^{-1} = \sum_{k=0}^{\infty} (Id - \Sigma_{obs(m)})^k$

Neumiss Networks to approximate the covariance matrix

Order ℓ approx. of the Bayes predictor)

$$f_\ell^\star(X_{obs}, M) = \langle \beta_{obs}, X_{obs} \rangle + \langle \beta_{mis}, \mu_{mis} + \Sigma_{mis,obs} S_{obs(m)}^{(\ell)} (X_{obs} - \mu_{obs}) \rangle$$

Order-ℓ approx of $(\Sigma_{obs(m)}^{-1})$ for any m defined recursively:

$$S_{obs(m)}^{(\ell)} = (Id - \Sigma_{obs(m)}) S_{obs(m)}^{(\ell-1)} + Id.$$

Neuman Series, $S^{(0)} = Id$, $\ell = \infty$: $(\Sigma_{obs(m)})^{-1} = \sum_{k=0}^{\infty} (Id - \Sigma_{obs(m)})^k$

Neumiss Networks to approximate the covariance matrix

Order ℓ approx. of the Bayes predictor)

$$f_\ell^\star(X_{obs}, M) = \langle \beta_{obs}, X_{obs} \rangle + \langle \beta_{mis}, \mu_{mis} + \Sigma_{mis,obs} S^{(\ell)}_{obs(m)} (X_{obs} - \mu_{obs}) \rangle$$

Order-ℓ approx of $(\Sigma^{-1}_{obs(m)})$ for any m defined recursively:

$$S^{(\ell)}_{obs(m)} = (Id - \Sigma_{obs(m)}) S^{(\ell-1)}_{obs(m)} + Id.$$

Neuman Series, $S^{(0)} = Id$, $\ell = \infty$: $(\Sigma_{obs(m)})^{-1} = \sum_{k=0}^{\infty} (Id - \Sigma_{obs(m)})^k$

$\Rightarrow$ **Neural network architecture to approximate the Bayes predictor**

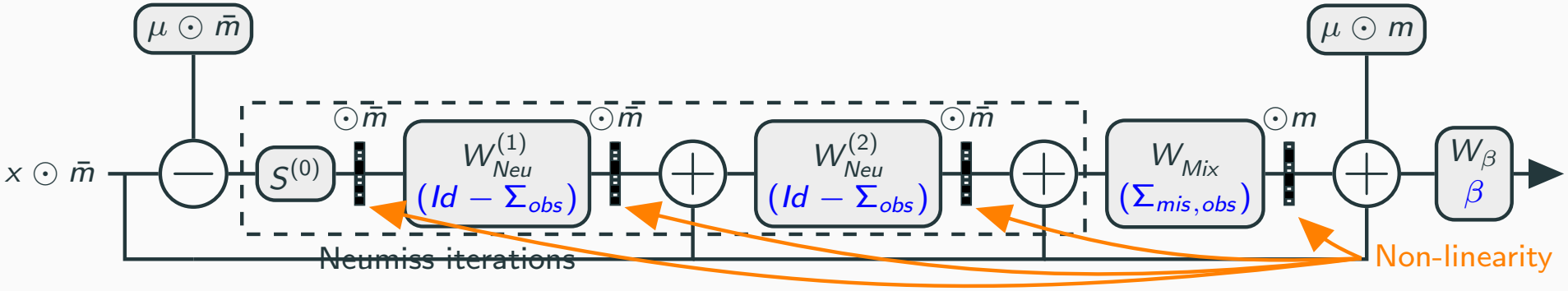


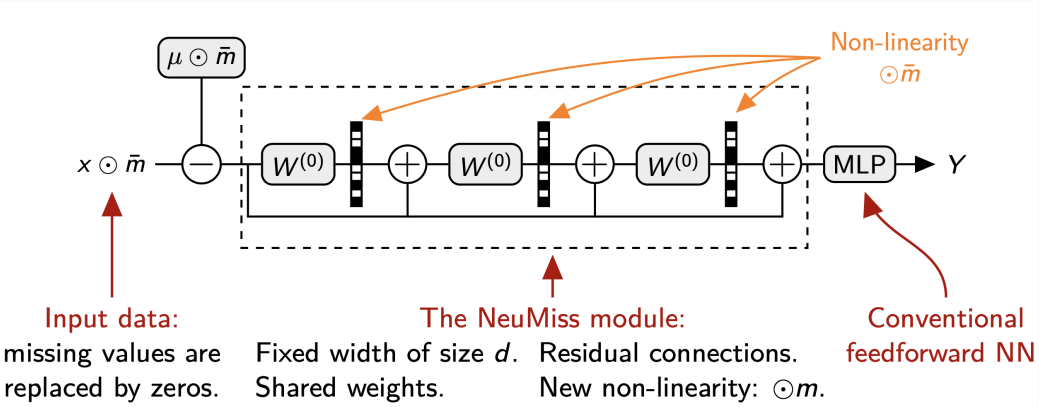
Figure 1: Depth of 3, $\bar{m} = 1 - m$. Each weight matrix $W^{(k)}$ corresponds to a simple transformation of the covariance matrix indicated in blue.

Networks with missing values: $\odot M$ nonlinearity[27]

- Implementing a network with the matrix **weights** $W^{(k)} = (I - \Sigma_{obs(m)})$ **masked differently for each sample** can be challenging
- Masked weights is **equivalent to masking input & output vector.**

Let v a vector, $\bar{m} = 1 - m$. $(W \odot \bar{m}\bar{m}^\top)v = (W(v \odot \bar{m})) \odot \bar{m}$

Classic network with multiplications by the mask nonlinearities $\odot M$



Jointly learn imputation and regression

[27] Le morvan, **J.** et al. Neumiss networks: differential programming for supervised learning with missing values. *Neurips2020 (Oral)*.

Which imputation function and predictor should one choose?

- Oracles regression $f^\star$ & conditional imputation $\mathbb{E}[X_{mis}|X_{obs}, M]$: $f^\star \circ \Phi^{CI}$

Proposition (excess of risk)

Assum PSD matrices $\bar{H}^+$ & $\bar{H}^-$ s.t. for all $X \in \mathcal{S}, \bar{H}^- \leq H(X) \leq \bar{H}^+$, $H(X)$ the Hessian of f^* at X (min. & max. curvatures of f^* in any direction are uniformly bounded over the entire space)

$$\mathcal{R}\left(f^\star \circ \Phi^{CI}\right) - \mathcal{R}^\star \leq \frac{1}{4}\mathbb{E}_M[\max\left(\mathrm{tr}\left(\bar{H}^-_{mis,mis}\Sigma_{mis|obs,M}\right)^2, \mathrm{tr}\left(\bar{H}^+_{mis,mis}\Sigma_{mis|obs,M}\right)^2\right)]$$

High excess risk if both 1) the curvature of $f^\star$ is high and 2) the variance of the missing data given the observed one is high (linear regression consistent)

Which imputation function and predictor should one choose?

- Oracles regression $f^\star$ & conditional imputation $\mathbb{E}[X_{mis}|X_{obs}, M]$: $f^\star \circ \Phi^{CI}$

Proposition (excess of risk)

Assum PSD matrices $\bar{H}^+$ & $\bar{H}^-$ s.t. for all $X \in \mathcal{S}, \bar{H}^- \leq H(X) \leq \bar{H}^+$, $H(X)$ the Hessian of f^* at X (min. & max. curvatures of f^* in any direction are uniformly bounded over the entire space)

$$\mathcal{R}\left(f^\star \circ \Phi^{CI}\right) - \mathcal{R}^\star \leq \frac{1}{4}\mathbb{E}_M[\max\left(\mathrm{tr}\left(\bar{H}^-_{mis,mis}\Sigma_{mis|obs,M}\right)^2, \mathrm{tr}\left(\bar{H}^+_{mis,mis}\Sigma_{mis|obs,M}\right)^2\right)]$$

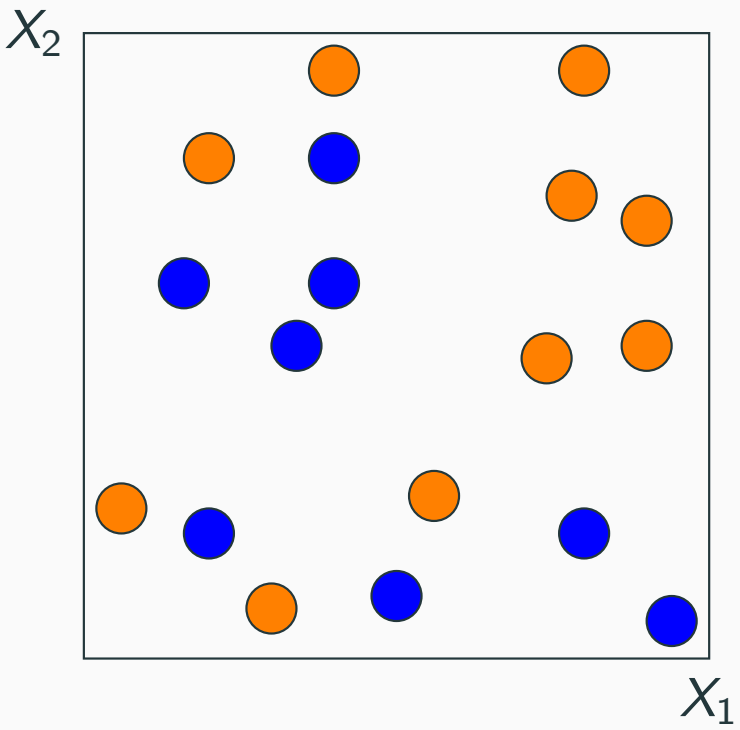
High excess risk if both 1) the curvature of $f^\star$ is high and 2) the variance of the missing data given the observed one is high (linear regression consistent)

- Is there a <u>continuous</u> function g, s.t. $g \circ \Phi^{CI}$ is Bayes optimal? *No.*

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

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

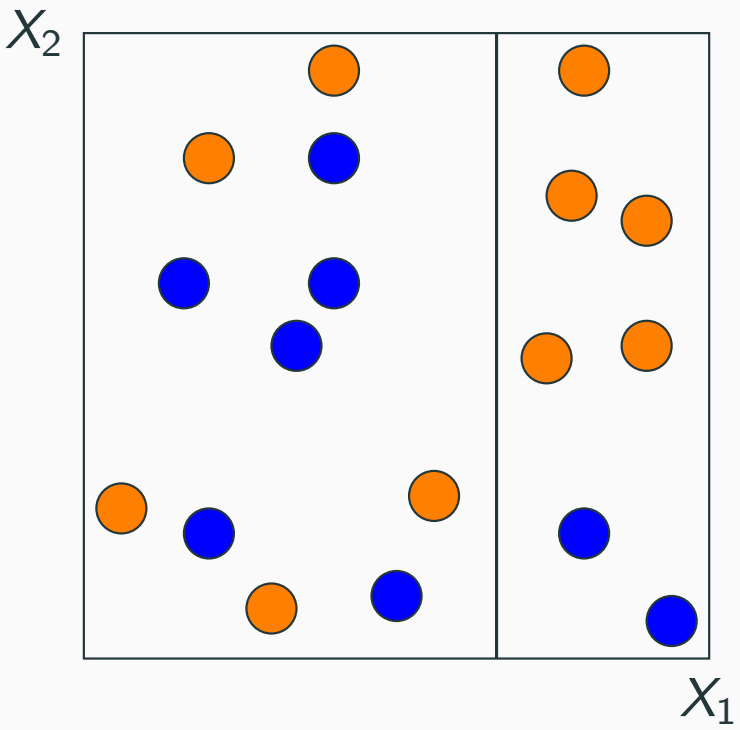


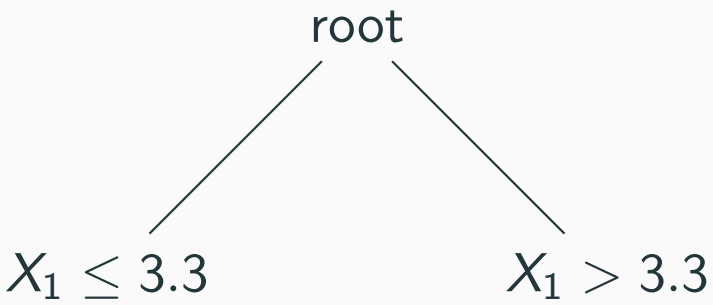
root

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

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

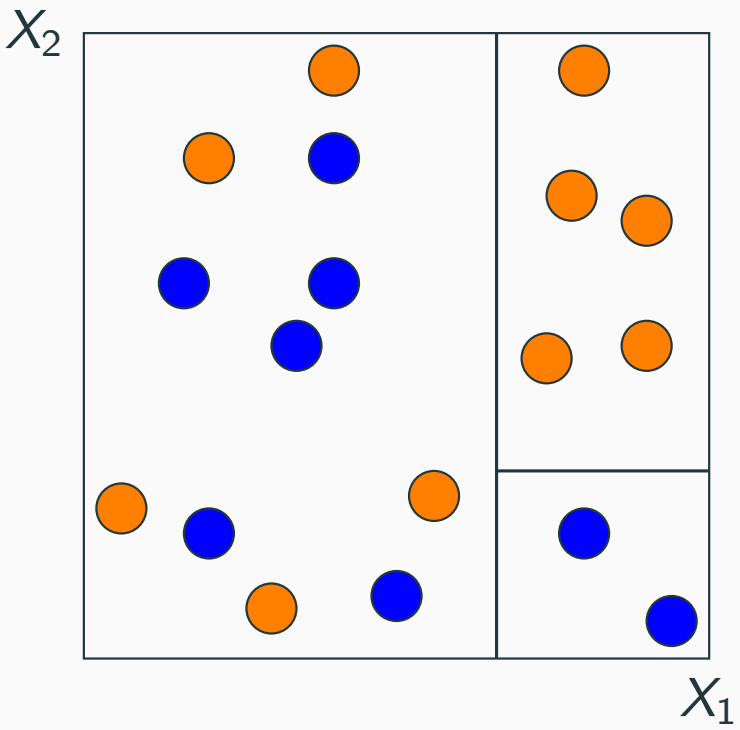


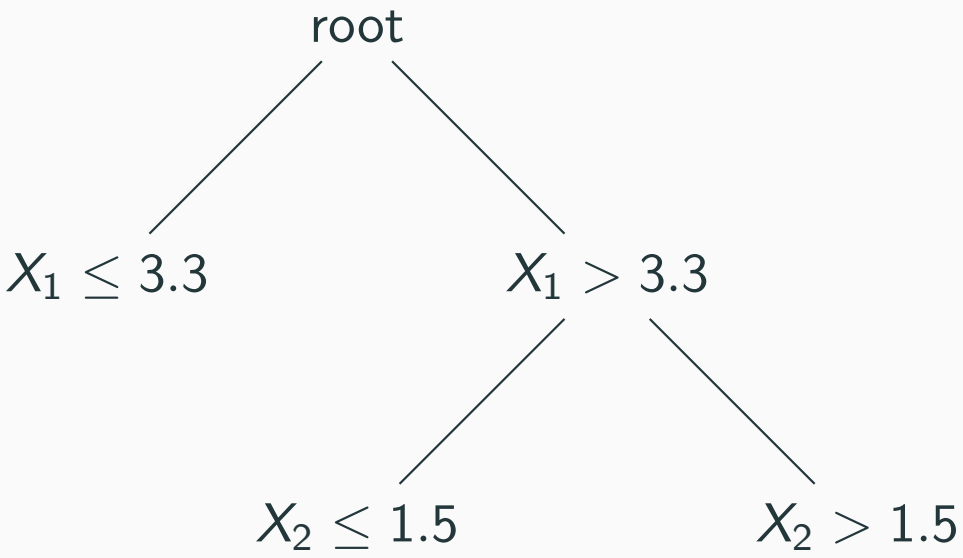


CART (Breiman, 1984)

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

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





CART with missing values

root

	X_1	X_2	Y
1			
2	NA		
3	NA		
4			

CART with missing values

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

root

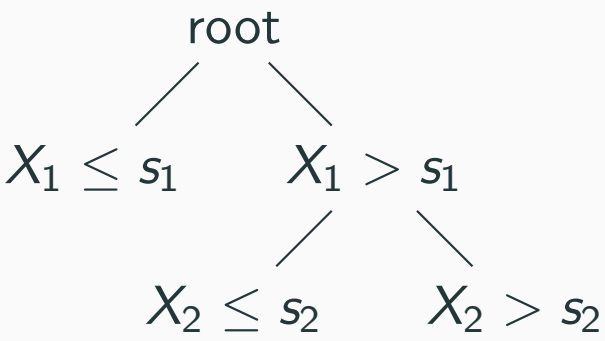
$X_1 \leq s_1$ $\quad$ $X_1 > s_1$

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[28]: `grf package`, `scikit-learn`, `partykit`

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

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

Causal inference from (incomplete) heterogeneous sources

- **Estimate causal effect:** Example on trauma brain patients[29]

"tranexamic acid" (**treatment**) impact on "28 days mortality"(**outcome**)

Estimator: Augmented Inverse Propensity Weighting uses non-parametric regression models Treatment $\sim$ covariates & Outcome $\sim$ covariates

$\Rightarrow$ Extended to handle missing values (implemented in `grf` package)

[29] Wager, **J.**, Doubly robust estimation with incomplete confounders. *Ann. Appl. Stat.* 2020.

Causal inference from (incomplete) heterogeneous sources

- **Estimate causal effect:** Example on trauma brain patients[29]

"tranexamic acid" (**treatment**) impact on "28 days mortality" (**outcome**)

Estimator: Augmented Inverse Propensity Weighting uses non-parametric regression models Treatment $\sim$ covariates & Outcome $\sim$ covariates

$\Rightarrow$ Extended to handle missing values (implemented in `grf` package)

- **Generalization of randomized control trial**'s findings toward a target pop., with distributional shift (data fusion, recovery from selection biais)[30,31,32,33]

			Covariates			Treat	Outcomes
	Set	S	X_1	X_2	X_3	W	Y
1	$\mathcal{RCT}$	1	1.1	20	NA	1	24.1
	$\mathcal{RCT}$	1	-6	NA	NA	0	26.3
n	$\mathcal{RCT}$	1	0	15	NA	1	23.5
$n+1$	Target	?	NA	35	7.1		
$n+2$	*Target*	?	-2	52	2.4		
	Target	?		...			
$n+m$	*Target*	?	-2	22	NA		

[29] Wager, **J.**, Doubly robust estimation with incomplete confounders. *Ann. Appl. Stat.* 2020.
[30] Colnet, **J.** et al. (2021). Causal inference for combining RCT & obs. studies: a review.
[31] Mayer & **J.** Generalizing treatment effects with incomplete covariates. 2022.
[32] Colnet, **J.** et al. Generalization: sensitivity analysis & missing data. *Journal of causal inf.* 2022.
[33] Colnet, **J.** et al. Reweighting RCT for generalization: finite sample analysis & var. select. 2022.

Traumabase project: decision support for trauma patients

- ▷ 30 000 French trauma patients[34]
- ▷ 250 features from the accident site to the hospital discharge
- ▷ 30 hospitals
- ▷ 4000 new patients/ year

Center	Accident	Age	Sex	Weight	Lactactes	BP	TXA.	Y
Beaujon	fall	54	m	85	NM	180	treated	0
Pitie	gun	26	m	NR	NA	131	untreated	1
Beaujon	moto	63	m	80	3.9	145	treated	1
Pitie	moto	30	w	NR	Imp	107	untreated	0
HEGP	knife	16	m	98	2.5	118	treated	1
⋮								⋱

⇒ **Estimate causal effect**: Administration of the **treatment** "tranexamic acid (TXA)" given within 3 hours of the accident, on the **outcome** *28 days intra hospital mortality (Y)* for trauma brain patients.[35]

[34] `www.traumabase.eu - https://www.traumatrix.fr/`

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

Effect of tranexamic acid on in-ICU mortality

Model Treatment on Covariates $e(x) = \mathbb{P}(W_i = 1 \,|\, X_i = x)$

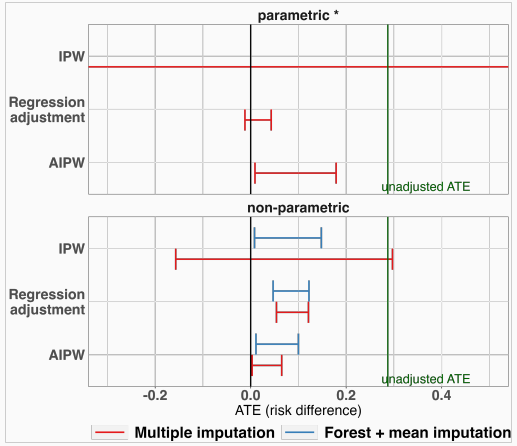
Model Outcome on Covariates $\mu_{(w)}(x) = \mathbb{E}[Y_i(w) \,|\, X_i = x]$

Augmented Inverse Propensity Weighting - double robust

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + W_i\frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1 - W_i)\frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1 - \hat{e}(X_i)}\right)$$

$\hat{\tau}_{AIPW}$ is $\sqrt{n}$-consistent, asympt. normal with semi parametric variance given:

$$\mathbb{E}\left[\left(\hat{e}(X_i)^{(-i)} - e(X_i)\right)^2\right]^{\frac{1}{2}} \times \mathbb{E}\left[\left(\hat{\mu}_{(W)}(X_i)^{(-i)} - \mu_{(W)}(X_i)\right)^2\right]^{\frac{1}{2}} = o\left(\frac{1}{\sqrt{n}}\right)$$



x-axis: Estimat. of the Average Treatment Effect ($\times 100$), bootstrap CI

AIPW with missing implemented in generalized random forest package `grf`

Generalization of trial's findings toward a target population[39]

$\Rightarrow$ RCT gold standart to estimate treatment effect
Trial sample can be different from the population eligible for treatment

$\Rightarrow$ Leveraging RCT and covariates from target population to transport the treatment effect estimated from the RCT to another population with a distributional shift (data fusion, recovery from selection biais),[36],[37] [38]

$\rightarrow$ Reduce drug approval times and costs for patients who could benefit

$\rightarrow$ Prices depend on efficiency

			Covariates			Treat	Outcomes
	Set	S	X_1	X_2	X_3	W	Y
1	$\mathcal{RCT}$	1	1.1	20	NA	1	24.1
	$\mathcal{RCT}$	1	-6	45	NA	0	26.3
n	$\mathcal{RCT}$	1	0	15	NA	1	23.5
$n+1$	*Obs*	?	-1	35	7.1		
$n+2$	*Obs*	?	-2	52	2.4		
	$\mathcal{O}\lfloor\int$	?		. . .			
$n+m$	*Obs*	?	-2	22	3.4		

[36] Mayer & **J.** Generalizing treatment effects with incomplete covariates. Archiv. 2022.
[37] Colnet, **J.** et al. Generalizing a causal effect: sensitivity analysis and missing covariates. *Journal of causal inference*. 2022.
[38] Colnet, **J.** et al. Reweighting RCT for generalization: finite sample analysis & var. select. 2022.
[39] Colnet, **J.** et al. (2021). Causal inference for combining RCT & obs. studies: a review.

MNAR data: identifiability issues, few solutions in practice

Before estimation, we should prove the identifiability of the parameters

Example: Credit: Ilya Shpitser $X^{\mathrm{NA}} = [1, \mathrm{NA}, 0, 1, \mathrm{NA}, 0]$

- **Case 1:** X missing only if $X = 1$.

$$X = [1, 1, 0, 1, 1, 0],\ \mathbb{P}(X = 1) = 2/3$$

- **Case 2:** X missing only if $X = 0$.

$$X = [1, 0, 0, 1, 0, 0],\ \mathbb{P}(X = 1) = 1/3$$

$\Rightarrow$ Start from 2 equal observed distribution. It leads to different parameters of the data distribution $\mathbb{P}(X = 1)$

Identifiability: the parameters of (X, M) are uniquely determined from available information $(X, M = 0)$

Estimation: restrictive setting (few variables, only missing values on the outcome, simple models)[40 41 42]

[40] Ibrahim, et al. Missing covariates in glm when the mechanism is non-ignorable. *JRSSB*. 1999.
[41] Tang. Statistical inference for nonignorable missing-data. *Statistic. theory & rel. fields.* 2018.
[42] Mohan, Thoemmes, Pearl. Estimation with incomplete data: The linear case. *IJCAI*. 2018.

Notations

- Random Variables:

▷ $X \in \mathbb{R}^d$: the complete unvailable data
▷ $\widetilde{X} \in \{\mathbb{R} \cup \{\mathrm{NA}\}\}^d$: incomplete data (observed), NA: Not Available
▷ $M \in \{0,1\}^d$: the missing-data pattern, the mask

$obs(M)$ (resp. $\mathrm{mis}(M)$) indices of the observed (resp. missing) entries.

- Realizations:

$$x = (1.1, 2.3, 3.1, 8, 5.27)$$
$$\widetilde{x} = (1.1, \mathrm{NA}, -3.1, 8, \mathrm{NA})$$
$$m = (0, 1, 0, 0, 1)$$
$$x_{\mathrm{obs(m)}} = (1.1, 3.1, 8), \qquad x_{\mathrm{mis}(m)} = (2.3, 5.27)$$

MCAR[43]: For all $m \in \{0,1\}^d, P(M = m \mid X) = P(M = m)$
MAR[44]: For all $m \in \{0,1\}^d, P(M = m \mid X) = P\left(M = m \mid X_{obs(m)}\right)$

[43] Michel, Naf, Spohn, Â¨ Meinshausen. 2021. PKLM: a flexible mcar test using classification.
[44] What Is Meant by "Missing at Random"? Seaman, et al. Statistical Science. 2013.