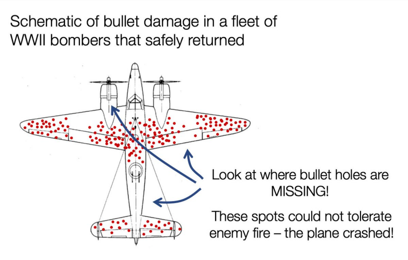
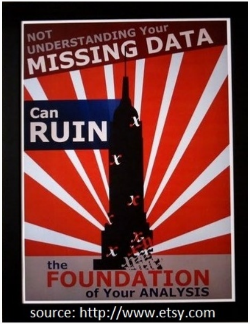
Statistical Learning with Missing Values

Julie Josse & Jeffrey Näf
Senior Researcher Inria, Head of Premedical Inria-Inserm team
Assistant Professor, Geneva School of Economics and Management (GSEM)

April 7th 2025

StatML & Bocconi Spring School 2025 â UK edition





Julie Josse: Research in Statistical Methodology

Academic background:

- Professor at **Ecole Polytechnique** (IP Paris) (2016 - 2020)
- Visiting researcher **Stanford Univ.** (2013-2016), **Google** (2019 - 2020)
- Senior Researcher at **Inria** Montpellier (Sept. 2020 -). Lead Inria-**Inserm** PreMeDICaL team: precision medicine by data integration & causal learning Composed of MD, researchers in ML, biostat, PhDs with medical degree

Research topics: Balance between theory and application

- Dimensionality reduction to visualize high dimensional **multi modal data**
- **Missing values**: max likelihood, matrix completion, supervised learning
- **Causal inference**: combining trials & observational data, optimal policy

⇒ **Transfert of research** through **software developments** (R foundation, packages, etc.)

Multidisciplinary and societal projects:

- Traumatrix: Clinical decision aid system to improve the triage & care of trauma patients ⇒ **Clinical trial** launching in 2025: real-time model implementation in ambulances via mobile data collection app.
- ICUBAM: Real time info gathering & vizualization on ICU beds availability

Presentation Jeffrey Näf

Academic background:

- PhD in Statistics from **ETH Zürich** (2018-2022)
- Postdoc in the PreMeDICaL **Inria** team, Montpellier (2023-2025)
- Assistant Professor in Business Analytics and Statistics at the Research Institute for Statistics and Information Science, **University of Geneva** (Feb. 2025 -)

Research topics: Distributional Estimation, Robust Estimation and Applications

- **Distributional Prediction:** Distributional Random Forest and Various extensions
- **Missing values**: Imputation, Imputation scoring
- **Robust Estimation**: Robust High-Dimensional Covariance Matrix estimation, MMD estimation

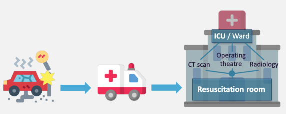
$\Rightarrow$ **Transfert of research** through **software developments**

Application in Marketing Research:

- CLVTools: Implementation of Probabilistic Modelling for Marketing Research

(Online) Decision support tool with quantified uncertainty

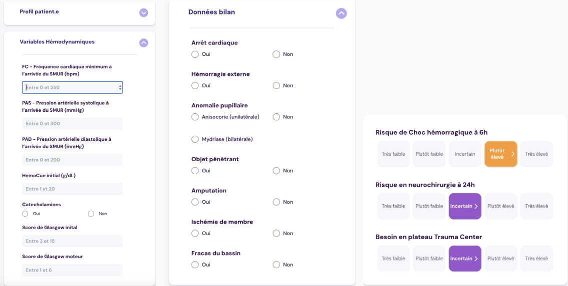
Ex: Traumatrix project[1]: Reducing under and over triage for improved resource allocation in trauma care



Major trauma: brain injuries, hemorrhagic shock from car accidents, falls, stab wounds

$\Rightarrow$ requires specialized care in "trauma centers"

Patients misdirected: human/ economical costs



Clinical trial launched in 2025: real-time implementation of Machine Learning models in ambulance dispatch via a mobile data collection application

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

Traumabase: an observational French registry on trauma[3]

- 40000 trauma patients
- 300 heterogeneous features from pre-hospital and in-hospital settings
- 40 trauma centers, 4000 new patients per year

Center	Accident	Age	Sex	Lactate	Blood Pres.	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	
⋮								⋱

$\Rightarrow$ **Explain and Predict** hemorrhagic shock, need for neurosurgery and need for a trauma center given pre-hospital features.
Ex: logistic regression/ random forests + **Quantify uncertainty**[2]

[2] Zaffran, **J.**, Dieuleveut, Romano. Conformal Prediction with Missing Values. *ICML 2023.*
[3] `www.traumabase.eu - https://www.traumatrix.fr/`

Personalization of treatment recommendation

Ex: Estimating treatment effect from the Traumabase data

- ▷ 40000 trauma patients
- ▷ 300 heterogeneous features from pre-hospital and in-hospital settings
- ▷ 40 trauma centers, 4000 new patients per year

Center	Accident	Age	Sex	Weight	Lactacte	Blood Press.	TXA.	Y
Beaujon	fall	54	m	85	NA	180	treated	0
Pitie	gun	26	m	NA	NA	131	untreated	1
Beaujon	moto	63	m	80	3.9	145	treated	1
Pitie	moto	30	w	NA	NA	107	untreated	0
HEGP	knife	16	m	98	2.5	118	treated	1
$\vdots$								$\ddots$

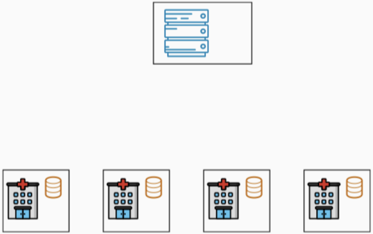
$\Rightarrow$ **Estimate causal effect** (with missing values[4]): Administration of the **treatment** *tranexamic acid (TXA)*, given within 3 hours of the accident, on the **outcome** (Y) *28 days in-hospital mortality* for trauma brain patients

[4] Mayer, I., Wager, S. & **J.**. (2020). Doubly robust treatment effect estimation with incomplete confounders. *Annals Of Applied Statistics. (implemented in package `grf`).*

Going beyond meta-analysis with federated causal inference[6]

⇒ **Difficulty to share individual-level data**: **data silos** & **regulations**

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \dots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \dots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

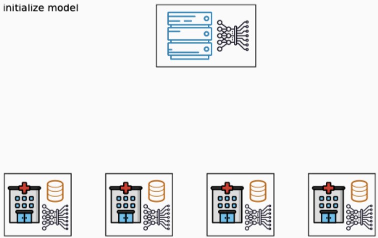
Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources** - Going beyond meta-analysis on individual data[5]

[5] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*
[6] Khellaf R, Bellet, A. & **J.**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Going beyond meta-analysis with federated causal inference[6]

⇒ **Difficulty to share individual-level data**: **data silos** & **regulations**

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \ldots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \ldots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

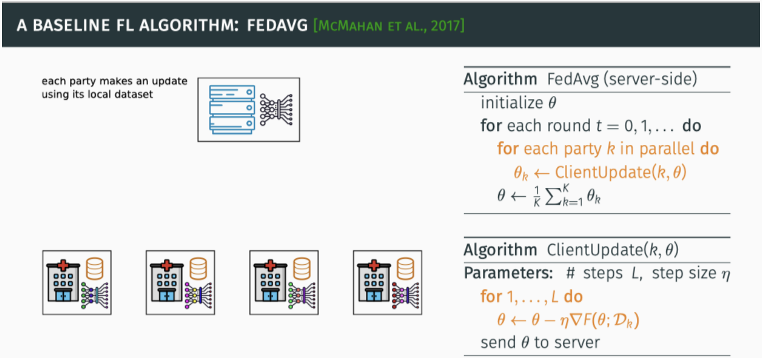
Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources** - Going beyond meta-analysis on individual data[5]

[5] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*

[6] Khellaf R, Bellet, A. & **J.**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Going beyond meta-analysis with federated causal inference[6]

⇒ **Difficulty to share individual-level data**: **data silos** & **regulations**



Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources** - Going beyond meta-analysis on individual data[5]

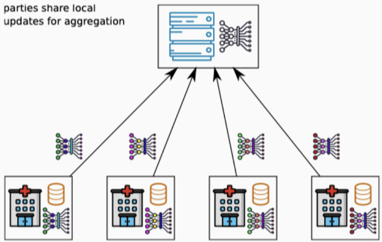
[5] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*

[6] Khellaf R, Bellet, A. & **J.**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Going beyond meta-analysis with federated causal inference[6]

⇒ **Difficulty to share individual-level data**: **data silos** & **regulations**

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \ldots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \ldots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources** - Going beyond meta-analysis on individual data[5]

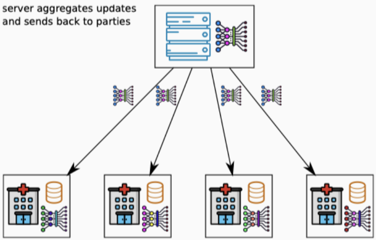
[5] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*

[6] Khellaf R, Bellet, A. & **J.**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Going beyond meta-analysis with federated causal inference[6]

⇒ **Difficulty to share individual-level data**: **data silos** & **regulations**

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ

for each round $t = 0, 1, \ldots$ **do**

  for each party k in parallel **do**

    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$

  $\theta \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η

for $1, \ldots, L$ **do**

  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$

send θ to server

Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources** - Going beyond meta-analysis on individual data[5]

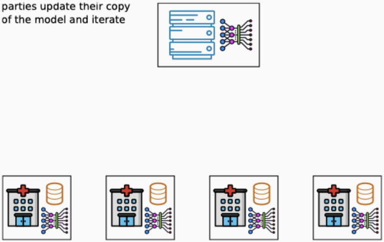
[5] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*

[6] Khellaf R, Bellet, A. & **J.**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Going beyond meta-analysis with federated causal inference[6]

⇒ **Difficulty to share individual-level data**: **data silos** & **regulations**

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \ldots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow$ ClientUpdate(k, θ)
  $\theta \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \ldots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources** - Going beyond meta-analysis on individual data[5]

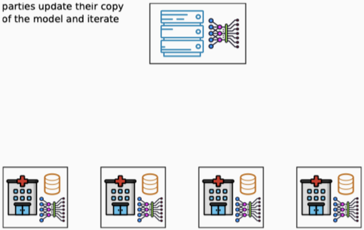
[5] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*

[6] Khellaf R, Bellet, A. & **J.**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Going beyond meta-analysis with federated causal inference[6]

$\Rightarrow$ **Difficulty to share individual-level data**: **data silos** & **regulations**

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \dots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \frac{1}{K}\sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \dots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

- Numerous extensions / improvements: fully decentralized (no server), dealing with highly heterogeneous data, privacy, fairness, compression... [Kairouz et al., 2021]

Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources** - Going beyond meta-analysis on individual data[5]

[5] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*

[6] Khellaf R, Bellet, A. & **J.**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Missing values[7], [8], [9]

are everywhere: unanswered questions in a survey, lost data, damaged plants, machines that fail...

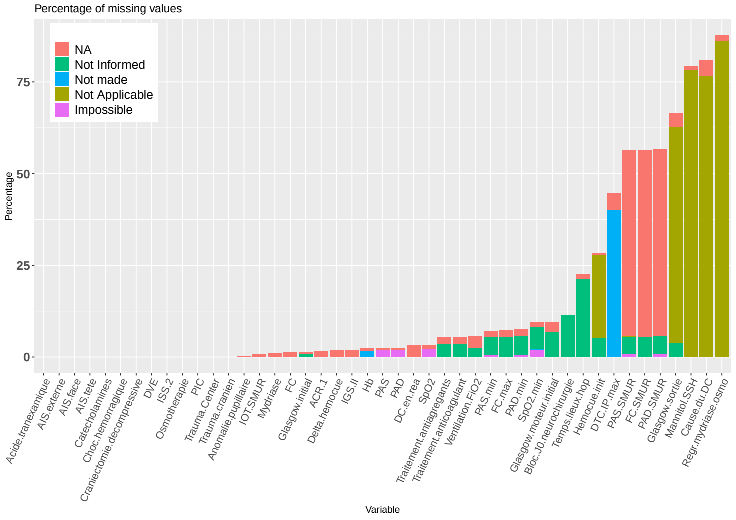
"The best thing to do with missing values is not to have any"

⇒ Still an issue in the "big data" area

Data integration: data from different sources

[7]Little & Rubin (2019). Statistical Analysis with Missing Data, Third Edition, Wiley.
[8]Van Buuren (2018). Flexible Imputation of Data. Second Edition, Chapman & Hall.
[9]Schafer (1997). Analysis of Incomplete Multivariate Data, Chapman & Hall.

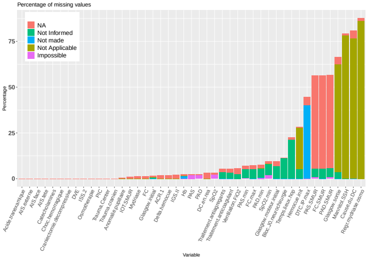
Missing data: important bottleneck in statistical practice



Different types of missing values

Multilevel data: Sporadic & Systematic missing values (feature/hospital)

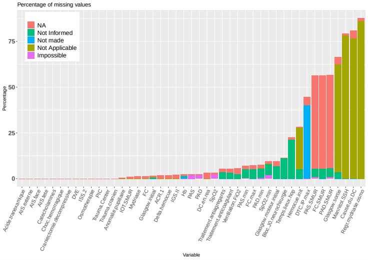
Missing data: important bottleneck in statistical practice



"One of the ironies of Big Data is that missing data play an ever more significant role"[10]

[10] 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"[10]

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

$$\begin{aligned} d = 5 \quad &\Longrightarrow \quad \approx 95\% \text{ of rows kept} \\ d = 300 \quad &\Longrightarrow \quad \approx \;\; 5\% \text{ of rows kept} \end{aligned}$$

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

Solutions to handle missing values in the covariates

Abundant literature: Creation of Rmistatic platform[11] ($>$ 150 packages)

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

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

[12] Jiang, **J.** et al. Logistic Regression with Missing Covariates *CSDA*. 2019. - `misaem package`

[13] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects & interac. in mixed data. *JASA*. 2019.

[14] **J.** et al. Consistency of supervised learning with missing values. *Stats papers*. 2018-2024.

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

Solutions to handle missing values in the covariates

Abundant literature: Creation of Rmistatic platform[11] (> 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[12]

[11] Mayer, **J.** et al. A unified platform for missing values methods and workflows. *R journal*. 2022.
[12] Jiang, **J.** et al. Logistic Regression with Missing Covariates *CSDA*. 2019. - `misaem package`
[13] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects & interac. in mixed data. *JASA*. 2019.
[14] **J.** et al. Consistency of supervised learning with missing values. *Stats papers*. 2018-2024.
[15] Le morvan, **J.** et al. What's a good imputation to predict with missing values? *Neurips2021*.

Solutions to handle missing values in the covariates

Abundant literature: Creation of Rmistatic platform[11] (> 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[12]

(Multiple) imputation to get a complete data set. Ex: (M)ICE

[11] Mayer, **J.** et al. A unified platform for missing values methods and workflows. *R journal*. 2022.
[12] Jiang, **J.** et al. Logistic Regression with Missing Covariates *CSDA*. 2019. - `misaem package`
[13] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects & interac. in mixed data. *JASA*. 2019.
[14] **J.** et al. Consistency of supervised learning with missing values. *Stats papers*. 2018-2024.
[15] Le morvan, **J.** et al. What's a good imputation to predict with missing values? *Neurips2021*.

Solutions to handle missing values in the covariates

Abundant literature: Creation of Rmistatic platform[11] ($>$ 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[12]

(Multiple) imputation to get a complete data set. Ex: (M)ICE

Matrix completion aim: **Predict the missing values** as well as possible. Solutions: using low rank matrix approximation[13]

Predictive aim: **Predict an outcome** with missing data in covariates[14][15]. Solutions: using deterministic (e.g. constant) imputation or Missing Incorporated in Attributes for trees based methods (`grf package`)

[11] Mayer, **J.** et al. A unified platform for missing values methods and workflows. *R journal*. 2022.
[12] Jiang, **J.** et al. Logistic Regression with Missing Covariates *CSDA*. 2019. - `misaem package`
[13] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects & interac. in mixed data. *JASA*. 2019.
[14] **J.** et al. Consistency of supervised learning with missing values. *Stats papers*. 2018-2024.
[15] Le morvan, **J.** et al. What's a good imputation to predict with missing values? *Neurips2021*.

Outline

- Monday
  - Introduction - Missing values mechanisms
  - Single imputation, Multiple imputation
  - Likelihood approaches
  - Practice
- Tuesday
  - PCA with missing values - Matrix completion
  - Supervised learning with missing values
  - Uncertainty quantification
  - Practice
- Wednesday
  - Causal inference with missing values
  - Advanced topics

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`[16]

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

[16] 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`[16]

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**

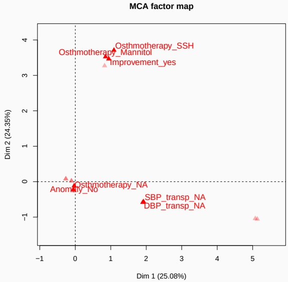
[16] Husson, **J.**, Le. FactoMineR: An R Package for Multivariate Analysis. *JSS*. (2008)

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[16]

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**

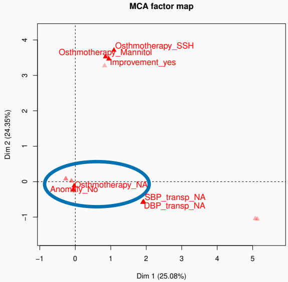
[16] Husson, **J.**, Le. FactoMineR: An R Package for Multivariate Analysis. *JSS*. (2008)

What is a 'true' missing value?

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

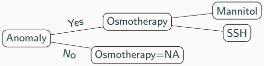
Visualize their patterns for clues as to how & why they occur FactoMineR[16]

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**

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

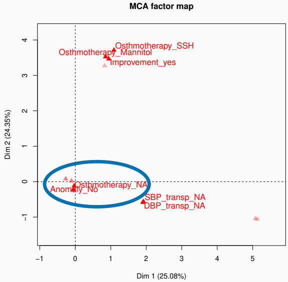
What is a 'true' missing value?

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

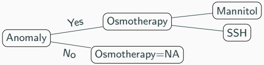
Visualize their patterns for clues as to how & why they occur FactoMineR[16]

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 nested variables:



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

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

⇒ Feedback on data collection/encoding process

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

Missing values mechanism

Missing values mechanism: Rubin's taxonomy[17,18]

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

- Realizations: For a pattern m, $o(x,m) = (x_j)_{j\in\{1,\dots,d\}:m_j=0}$ the observed elements of x and while $o^c(x,m) = (x_j)_{j\in\{1,\dots,d\}:m_j=1}$, the missing elements.

$$
\begin{aligned}
&x^\star = (1,2,3,8,5)\\
&x = (1,\mathrm{NA},3,8,\mathrm{NA})\\
&m = (0,1,0,0,1)\\
&o(x,m) = (1,3,8), \qquad o^c(x^\star,m) = (2,5)
\end{aligned}
$$

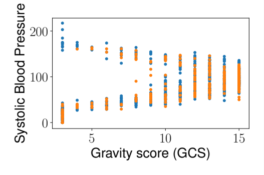
[17] Rubin. Inference and missing data. *Biometrika*. 1976.
[18] What Is Meant by "Missing at Random"? Seaman, et al. *Statistical Science*. 2013.

Missing values mechanism: Rubin's taxonomy[17],[18]

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

For a pattern m, $o(x,m) = (x_j)_{j\in\{1,\dots,d\}:m_j=0}$ the observed elements of x and while $o^c(x,m) = (x_j)_{j\in\{1,\dots,d\}:m_j=1}$, the missing elements.

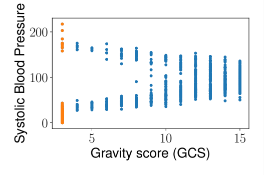
Ex: Simulated missing values according to the 3 mechanisms (Orange points will be missing) in Systolic Blood Pressure - GCS is always observed



Missing Completely at Random (MCAR)

$m \in \mathcal{M}, x \in \mathcal{X}$,

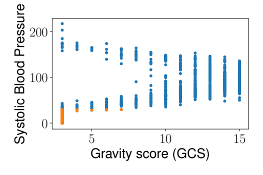
$\mathbb{P}(M = m|x) = \mathbb{P}(M = m)$



Missing at Random (MAR)

$\forall m \in \mathcal{M}, x \in \mathcal{X}$

$\mathbb{P}(M = m|x) = \mathbb{P}(M = m|o(x,m))$



Missing Not At Random (MNAR)

If not MAR: it is MNAR

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

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

Two views to model the joint distribution of (X, M)

Selection Model[19]**:** $p^*(M = m, x) = \mathbb{P}(M = m \mid x)p^*(x)$

Definition (SM-MAR)

$$\mathbb{P}(M = m|x) = \mathbb{P}(M = m|o(x, m)) \text{ for all } m \in \mathcal{M}, x \in \mathcal{X}.$$

The proba. of any m occurring only depends on the obs part of x.

Pattern Mixture Model[20]**:** $p^*(M = m, x) = p^*(x \mid M = m)\mathbb{P}(M = m)$

Definition (PMM-MAR)

$$p^*(o^c(x, m) \mid o(x, m), M = m) = p^*(o^c(x, m) \mid o(x, m)).$$

for all $m \in \mathcal{M}, x \in \mathcal{X}$. The conditional distrib. of missing given obs. in pattern m is equal to the unconditional one.[21]

[19] Heckman. Sample selection bias as a specification error. *Econometrica*. 1979

[20] Little. Pattern-mixture models for multivariate incomplete data. *JASA*. 1993

[21] Molenberghs et al. Every MNAR model has a MAR counterpart with equal fit. *JRSSB*. 2008

Two views to model the joint distribution of (X, M)

Selection Model[19]: $p^*(M = m, x) = \mathbb{P}(M = m \mid x)p^*(x)$

Definition (SM-MAR)

$$\mathbb{P}(M = m|x) = \mathbb{P}(M = m|o(x, m)) \text{ for all } m \in \mathcal{M}, x \in \mathcal{X}.$$

The proba. of any m occurring only depends on the obs part of x.

Pattern Mixture Model[20]: $p^*(M = m, x) = p^*(x \mid M = m)\mathbb{P}(M = m)$

Definition (PMM-MAR)

$$p^*(o^c(x, m) \mid o(x, m), M = m) = p^*(o^c(x, m) \mid o(x, m)).$$

for all $m \in \mathcal{M}, x \in \mathcal{X}$. The conditional distrib. of missing given obs. in pattern m is equal to the unconditional one.[21]

Proposition (SM-MAR is equivalent to PMM-MAR)

[19] Heckman. Sample selection bias as a specification error. *Econometrica*. 1979

[20] Little. Pattern-mixture models for multivariate incomplete data. *JASA*. 1993

[21] Molenberghs et al. Every MNAR model has a MAR counterpart with equal fit. *JRSSB*. 2008

MAR with shift in marginal distribution between patterns

- <u>Gaussian PMM</u>: $X^* \mid M = m \sim N(\mu_m \mid \Sigma_m)$. Ex: for two patterns $m_1 = (0, 0)$ and $m_2 = (1, 0)$ and **a shift**:

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} \\ NA & x_{2,2} \end{pmatrix}, \mathbf{M} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}.$$

MAR with shift in marginal distribution between patterns

- Gaussian PMM: $X^* \mid M = m \sim N(\mu_m \mid \Sigma_m)$. Ex: for two patterns $m_1 = (0, 0)$ and $m_2 = (1, 0)$ and **a shift**:

$$(X_1, X_2) \mid M = m_1 \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}2 & 1\\1 & 1\end{pmatrix}\right) (X_1, X_2) \mid M = m_2 \sim N\left(\begin{pmatrix}5\\5\end{pmatrix}, \begin{pmatrix}2 & 1\\1 & 1\end{pmatrix}\right).$$

MAR with shift in marginal distribution between patterns

- Gaussian PMM: $X^* \mid M = m \sim N(\mu_m \mid \Sigma_m)$. Ex: for two patterns $m_1 = (0,0)$ and $m_2 = (1,0)$ and **a shift**:

$$(X_1, X_2) \mid M = m_1 \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}2 & 1\\1 & 1\end{pmatrix}\right) (X_1, X_2) \mid M = m_2 \sim N\left(\begin{pmatrix}5\\5\end{pmatrix}, \begin{pmatrix}2 & 1\\1 & 1\end{pmatrix}\right).$$

- Not identifiable without restriction. How distributions can change?

$$= \underbrace{p^*(x_1 \mid x_2, M = m_2)}_{p^*(o^c(x,m_2) \mid o(x,m_2), M=m_2)} = N(x_2, 1)(x_1) = p^*(x_1 \mid x_2).$$

MAR with shift in marginal distribution between patterns

- Gaussian PMM: $X^* \mid M = m \sim N(\mu_m \mid \Sigma_m)$. Ex: for two patterns $m_1 = (0,0)$ and $m_2 = (1,0)$ and **a shift**:

$$(X_1, X_2) \mid M = m_1 \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}2 & 1\\1 & 1\end{pmatrix}\right) (X_1, X_2) \mid M = m_2 \sim N\left(\begin{pmatrix}5\\5\end{pmatrix}, \begin{pmatrix}2 & 1\\1 & 1\end{pmatrix}\right).$$

- Not identifiable without restriction. How distributions can change?

$$\underbrace{p^*(x_1 \mid x_2, M = m_1)}_{p^*(o^c(x,m_2) \mid o(x,m_2), M=m_1)} = \underbrace{p^*(x_1 \mid x_2, M = m_2)}_{p^*(o^c(x,m_2) \mid o(x,m_2), M=m_2)} = N(x_2, 1)(x_1) = p^*(x_1 \mid x_2).$$

Definition (Conditional indep. MAR - CIMAR)

$$p^*(o^c(x,m) \mid o(x,m), M = m') = p^*(o^c(x,m) \mid o(x,m)).$$

for all $m, m' \in \mathcal{M}, x \in \mathcal{X}$.equivalent to $o^c(X^*, m) \mid o(X^*, m) \perp\!\!\!\perp M$

MAR with shifts in conditional distribution between patterns

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ NA & x_{2,2} & x_{2,3} \\ NA & NA & x_{3,3} \end{pmatrix}, \mathbf{M} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix}$$

CIMAR

$p^*(x_1, x_2 \mid x_3, M = m_1) = p^*(x_1, x_2 \mid x_3, M = m_2) = p^*(x_1, x_2 \mid x_3, M = m_3) = p^*(x_1, x_2 \mid x_3)$

Distrib. of $X_1, X_2 \mid X_3$ is not allowed to change from one pattern to another, though the marginal distrib. of X_3 can change.

PMM-MAR

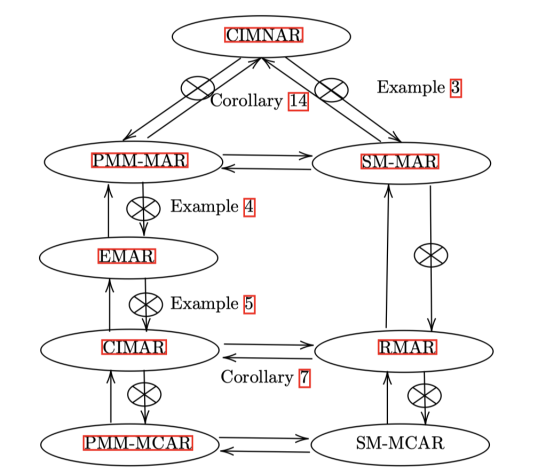
$p^*(x_1, x_2 \mid x_3, M = m_3) = p^*(x_1, x_2 \mid x_3)$

Both distrib. of observed variables and conditional ones can change from pattern to pattern.

MCAR: No change allowed.

$m \in \mathcal{M}, m' \in \mathcal{M}, x \in \mathcal{X},\ p^*(x) = p^*(x \mid M = m) = p^*(x \mid M = m')$

Relationships between the M(N)AR conditions[22]



[22]**Naf**, Scornet **J.**. (2024). What is a good imputation under MAR. *Submitted.*

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) [23] [24] [25]

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

Testing the missing values mechanism

- An obvious question is whether one can observe the missing value mechanism from the sample.
- The answer in general is no! (Unfortunately)
- However if we assume MAR is true we can test H_0 : MCAR vs H_A : MAR.
- A classical test is the Little test[26] that operates under the assumption of Gaussianity.
- One of the very few (if not only) useable nonparametric test is our PKLMTest[27]
- There is also interesting theoretical work[28]

[26] Little. *A Test of Missing Completely at Random for Multivariate Data with Missing Values*. 1988
[27] Michel, **Naf**, Spohn, Meinshausen. PKLM: a flexible MCAR test using classification, Psychometrika. 2025
[28] Berrett, Samworth. *Optimal nonparametric testing of missing completely at random and its connections to compatibility*, AoS. 2023

Hints on the missing values mechanism

- **Importance of contextual information:**
  - Important information is missing from datasets, which is often uncovered through collaborative discussions.
  - The context affects how data is coded and interpreted.
- **Examples:**
  - Distribution changes in gravity scores due to funding tied to patient severity.
  - Missing values due to team disagreements; Orientation depends of trust/reputation

Importance of communication with experts - Limits of AutoML?

Inference with missing values

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

Abundant literature: Creation of Rmistatic platform[29] (> 150 packages)

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

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

[30] 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[29] (> 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[30]
Need to design one specific algorithm for each statistical method

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

[30] 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[29] (> 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[30]
Need to design one specific algorithm for each statistical method

(Multiple) imputation to get a complete data set

Pros: Any analysis can be performed. Implementation: `mice R package`, IterativeImputer scikitlearn (option posterior equals true)
Cons: Generic

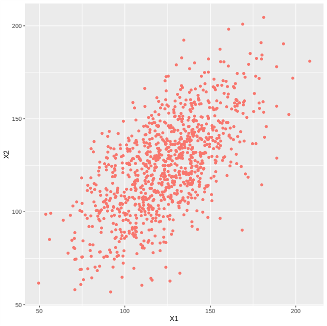
[29] Mayer, **J.** et al. A unified platform for missing values methods and workflows. *R journal*. 2022.
[30] Jiang, **J.** et al. Logistic Regression with Missing Covariates, Parameter Estimation, Model Selection and Prediction. *CSDA*. 2019. - Implementation in the `misaem package`

Single Imputation

Single imputation by the mean[31]

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

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



$\mu_{x_2} = 0$

$\sigma_{x_2} = 1$

$\rho = 0.6$

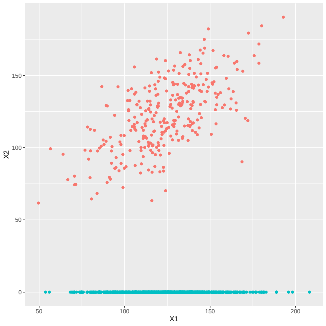
$\hat{\mu}_{x_2} = -0.01$
$\hat{\sigma}_{x_2} = 1.01$
$\hat{\rho} = 0.66$

[31] The code to reproduce the plots is available in Rmistastic

Single imputation by the mean[31]

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



$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.6$

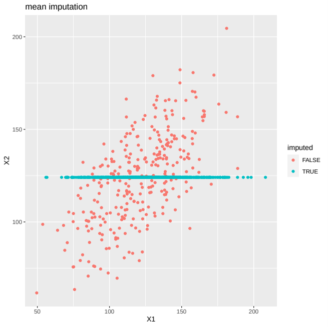
$\hat{\mu}_{x_2} = 0.18$
$\hat{\sigma}_{x_2} = 0.9$
$\hat{\rho} = 0.6$

[31] The code to reproduce the plots is available in Rmistastic

Single imputation by the mean[31]

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



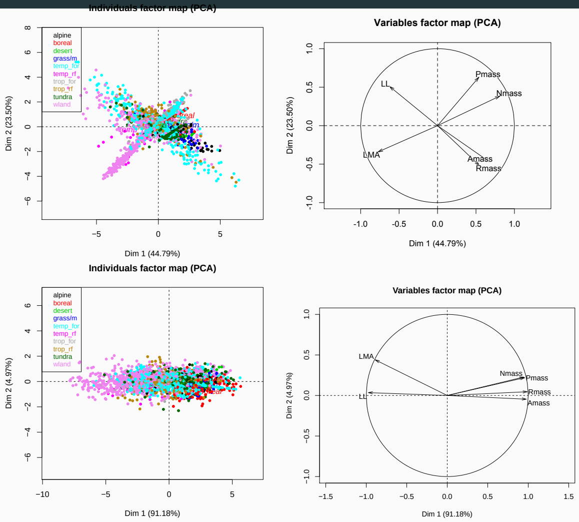
$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.6$

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

Mean imputation deforms joint and marginal distributions

[31] The code to reproduce the plots is available in Rmistastic

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)
```

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

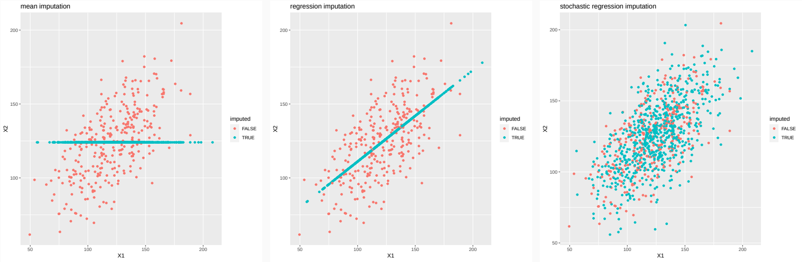
Ecological data: [32] $n = 69000$ species - 6 traits. Estimated correlation between Pmass & Rmass ≈ 0 (mean imputation) or ≈ 1 (EM PCA)

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

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



$\mu_{x_2} = 0$	0.01	0.01	0.01
$\sigma_{x_2} = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59

Impute while preserving distribution. Multivariate case

Assuming a joint distribution

- Gaussian model $x_i \sim \mathcal{N}(\mu, \Sigma)$
- Low rank : $X_{n\times d} = \mu_{n\times d} + \varepsilon \; \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with μ of low rank
  - $\Rightarrow$ Different regularization depending on noise regime [33]
  - $\Rightarrow$ Count data[34], ordinal data, categorical data, blocks/multilevel data
- Optimal transport [35], deep generative models: GAIN[36], MIWAE [37], etc. [38] [39]

[33] **J.** & Wager. Stable autoencoding for regularized low-rank matrix estimation. *JMLR*. 2016.
[34] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects & interac. in mixed data. *JASA*. 2019.
[35] Muzelec, Cuturi, Boyer, **J.** Missing Data Imputation using Optimal Transport. *ICML*. 2020.
[36] Yoon et al. GAIN: Missing data imputation using generative adversarial nets. *ICML*. 2018.
[37] Mattei & Frellsen. Miwae: Deep generative model. & imput. of incomplete data. *ICML*. 2018.
[38] Deng et al. Extended missing data imput. via gans. *Data Mining & Knowledge Discovery*. 2022.
[39] Fang Bao. Fragmgan gan for fragmentary data imputation. *Stat.theory & Related Fields*. 2023.
[40] van Buuren, S. Flexible Imputation of Missing Data. Chapman & Hall/CRC Press. 2018.
[41] Stekhoven & Bühlmann. MissForest–non-parametric imputation for mixed data. *Bioinfo. 2012.*

Impute while preserving distribution. Multivariate case

Assuming a joint distribution

- Gaussian model $x_i \sim \mathcal{N}(\mu, \Sigma)$
- Low rank : $X_{n\times d} = \mu_{n\times d} + \varepsilon \; \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with μ of low rank
  - $\Rightarrow$ Different regularization depending on noise regime [33]
  - $\Rightarrow$ Count data[34], ordinal data, categorical data, blocks/multilevel data
- Optimal transport [35], deep generative models: GAIN[36], MIWAE [37], etc. [38] [39]

Iterating conditional models (joint distribution implicitly defined)

- with parametric regression (M)ICE: (Multiple) Imput. by Chained Equations [40]
- iterative imputation of each variable by random forests [41]

---

[33] **J.** & Wager. Stable autoencoding for regularized low-rank matrix estimation. *JMLR*. 2016.
[34] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects & interac. in mixed data. *JASA*. 2019.
[35] Muzelec, Cuturi, Boyer, **J.** Missing Data Imputation using Optimal Transport. *ICML*. 2020.
[36] Yoon et al. GAIN: Missing data imputation using generative adversarial nets. *ICML*. 2018.
[37] Mattei & Frellsen. Miwae: Deep generative model. & imput. of incomplete data. *ICML*. 2018.
[38] Deng et al. Extended missing data imput. via gans. *Data Mining & Knowledge Discovery*. 2022.
[39] Fang Bao. Fragmgan gan for fragmentary data imputation. *Stat.theory & Related Fields*. 2023.
[40] van Buuren, S. Flexible Imputation of Missing Data. Chapman & Hall/CRC Press. 2018.
[41] Stekhoven & Bühlmann. MissForest–non-parametric imputation for mixed data. *Bioinfo. 2012.*

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

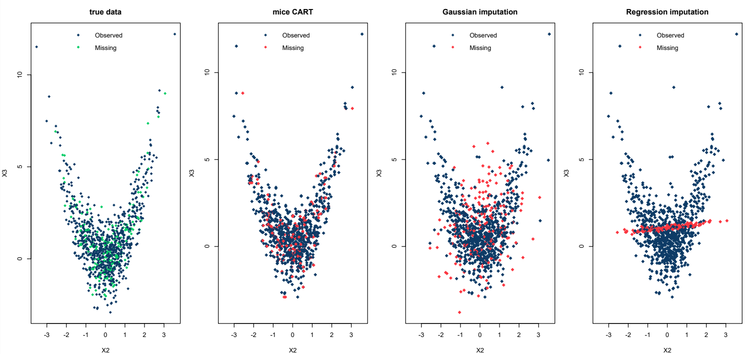
Feat1	Feat2	Feat3	Feat4	Feat5
1	1.0	1.00	1	1
1	1.0	1.00	1	1
2	2.0	2.00	2	2
2	2.0	2.00	2	2
3	3.0	3.00	3	3
3	3.0	3.00	3	3
4	4.0	4.00	4	4
4	4.0	4.00	4	4
5	5.0	5.00	5	5
5	5.0	5.00	5	5
6	6.0	6.00	6	6
6	6.0	6.00	6	6
7	7.0	7.00	7	7
7	7.0	7.00	7	7
8	6.87	6.87	8	8
8	6.87	6.87	8	8
9	6.87	6.87	9	9
9	6.87	6.87	9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

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

Missing `missForest` `imputePCA`

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

Imputation by forests versus regression imputation



Imputation with joint model with Gaussian distribution

$\Rightarrow$ Assumption joint gaussian model $x_i \sim \mathcal{N}(\mu, \Sigma)$

- Bivariate case with missing values on x_2: stochastic regression

▷ estimate β and σ

▷ impute from the predictive $\hat{x}_{i2} \sim \mathcal{N}\left(x_{i1}\hat{\beta}, \hat{\sigma}^2\right)$

- Extension to the multivariate case:

▷ Estimate μ and Σ from an incomplete data with EM

▷ Impute by drawing from the conditional distribution $X_{\text{mis}}|X_{\text{obs}} \sim \mathcal{N}(\mu_{\text{mis}|\text{obs}}, \Sigma_{\text{mis}|\text{obs}})$

$$\mu_{\text{mis}|\text{obs}} = \mathbb{E}[X_{\text{mis}}] + \Sigma_{\text{mis,obs}}\Sigma_{\text{obs,obs}}^{-1}\left(X_{\text{obs}} - \mathbb{E}[X_{\text{obs}}]\right).$$

$$\Sigma_{\text{mis}|\text{obs}} = \Sigma_{\text{mis}} - \Sigma_{\text{mis,obs}}\Sigma_{\text{obs,obs}}^{-1}\Sigma_{\text{obs,mis}}. \text{ Schur complement.}$$

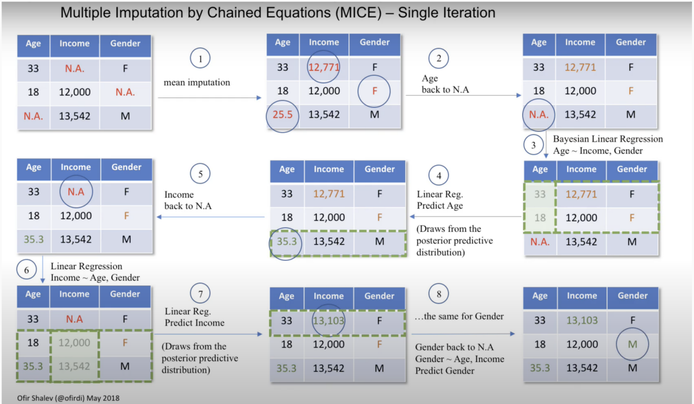
```
> library(norm)
> pre <- prelim.norm(as.matrix(don))
> thetahat <- em.norm(pre)
> imp <- imp.norm(pre, thetahat, don)
```

Fully conditional specification - FCS, (M)ICE

1. Fill NA with plausible values to get an initial completed dataset

2. For $j \in \{1, \ldots, d\}$, $t \geq 1$ use a univariate imputation to sample new imputed values $x_j^{(t+1)} \sim p^*(x_j \mid x_{-j}^{(t)})$, where $x_{-j}^{(t)} = \{x_l^{(t)}\}_{l \neq j}$ the imputed & observed values of other variables except j at the tth iteration.

3. Iterate until convergence



(Non) Identifiability under non-parametric MAR

Definition: Imputing with a mixture of distribution

$p^*(o^c(x,m) \mid o(x,m))$ **is identifiable** from $\mathcal{M}_0 \subset \mathcal{M}$ if there exists some weights $w_{m'}(o(x,m))$ (summing to 1) such that the mixture

$$h^*(o^c(x,m) \mid o(x,m)) = \sum_{m' \in \mathcal{M}_0} w_{m'}(o(x,m)) p^*(o^c(x,m) \mid o(x,m), M = m')$$

satisfies $p^*(o^c(x,m) \mid o(x,m)) = h^*(o^c(x,m) \mid o(x,m))$.

Proposition: Identifiability under PMM-MAR is not trivial[42]

Assume $|\mathcal{M}| > 3$. For any pattern $m \in \mathcal{M}$, $p^*(o^c(x,m) \mid o(x,m))$ is

- identifiable from any other pattern $m' \neq m$ under CIMAR,
- is not identifiable from any single pattern $m' \neq m$ under PMM-MAR.

If $\left|\sum_{j=1}^{d} m_j\right| > 1$, $p^*(o^c(x,m) \mid o(x,m))$ **is not identifiable from** L_m, **the set of patterns for which** $o^c(x,m)$ **is observed**.
$L_m = \{m' \in \mathcal{M} : m'_j = 0 \text{ for all } j \text{ such that } m_j = 1\}$.

[42] **Näf**, Scornet **J.**. (2024) What is a good imputation under MAR. *Submitted.*

Identifiability under MAR considering one variable at a time

- Consider the following mixture of distribution

$$h^*(x_j \mid x_{-j}) = \sum_{m \in L_j} \frac{\mathbb{P}(M=m)}{\sum_{m \in L_j} p^*(x_{-j} \mid M=m)\mathbb{P}(M=m)} p^*(x \mid M=m),$$

with $L_j = \{m \in \mathcal{M} : m_j = 0\}$, the patterns where x_j is observed

Theorem[43]: Identifiability of the right conditional distribution

Assume **PMM-MAR** holds,

$$h^*(x_j \mid x_{-j}) = \color{red}{p^*(x_j \mid x_{-j})}, \text{ for all } x_{-j} \text{ with } p^*(x_{-j}) > 0$$

At X_j, one can reduce the $|\mathcal{M}|$ patterns to two, one where X_j is missing, and one where it is observed. Though these two aggregated patterns are mixtures of several patterns $m \in \mathcal{M}$, MAR implies that both aggregated patterns have the same conditional distribution $X_j^* \mid X_{-j}^*$

[43] **Näf**, Scornet **J.**. (2024) What is a good imputation under MAR. *Submitted.*

Fully conditional specification - FCS, (M)ICE

1. Fill NA with plausible values to get an initial completed dataset
2. For $j \in \{1, \dots, d\}$, $t \geq 1$ use a univariate imputation to sample new imputed values $x_j^{(t+1)} \sim p^t(x_j \mid x_{-j}^{(t)})$, where $x_{-j}^{(t)} = \{x_l^{(t)}\}_{l \neq j}$ the imputed & observed values of other variables except j at the tth iteration.
3. Iterate until convergence

Theorem[44] shows that if we assume to have access to the true distribution $p^*(x_{-j})$ (assume x_{-j} is well imputed), we can impute according to the true distribution $p^*(x_j \mid x_{-j})$ by drawing from the conditional distrib. of $X_j \mid X_{-j}$ **learned from all patterns in which x_j is observed**

FCS approach can identify the right conditional distributions under PMM MAR

[44] **Näf**, Scornet **J.**. (2024) What is a good imputation under MAR. *Submitted*

What is a good imputation method under MAR?

- both conditional and marginal **distribution shifts** can occur for different patterns under MAR.
- conditional shifts are handled with FCS

An ideal imputation method should

- (1) be a distributional regression method,
- (2) be able to capture nonlinearities in the data,
- (3) be able to deal with distributional shifts in the observed variables,
- (4) be fast to fit,

1-3 are crucial for imputation under MAR
4 is only relevant to reduce the computational burden.

Rk: Block-wise FCS (multi-output methods to impute variables as blocks) should not be used: do not recover the correct distribution

What is a good imputation method?

(1) be a distributional regression method,
(2) be able to capture nonlinearities in the data,
(3) be able to deal with distributional shifts in the observed variables,

Method	(1)	(2)	(3)
missForest (Stekhoven & Bühlmann, 2011)		✓	
mice-cart (Burgette & Reiter, 2010)	✓	✓	
mice-RF (Doove et al., 2014)	✓	✓	
mice-DRF (Näf et al., 2024)	✓	✓	
mice-norm.nob (Gaussian)	✓		✓
mice-norm.predict (Regression)			✓

[45] Cevid, **Näf** et al., Distributional Random Forests. *JMLR*. 2022

What is a good imputation method?

(1) be a distributional regression method,
(2) be able to capture nonlinearities in the data,
(3) be able to deal with distributional shifts in the observed variables,

Method	(1)	(2)	(3)
missForest (Stekhoven & Bühlmann, 2011)		✓	
mice-cart (Burgette & Reiter, 2010)	✓	✓	
mice-RF (Doove et al., 2014)	✓	✓	
mice-DRF (Näf et al., 2024)	✓	✓	
mice-norm.nob (Gaussian)	✓		✓
mice-norm.predict (Regression)			✓

- mice-cart/RF estimate a tree, a forest, on observed data and then draw imputations from the leaves (approx conditional distribution) whereas distributional forest [45] is a distributional method

[45] Cevid, **Näf** et al., Distributional Random Forests. *JMLR*. 2022

Forests generalize poorly outside of the training set

Ex: Variables income & age with MAR missing values in income

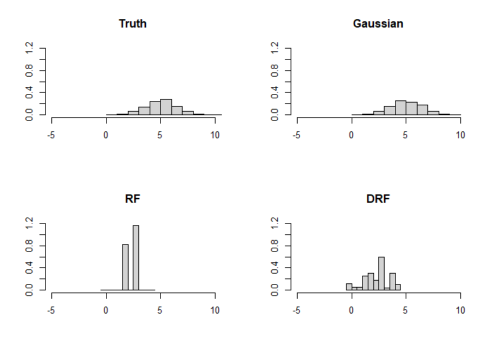
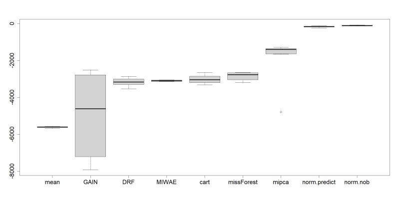


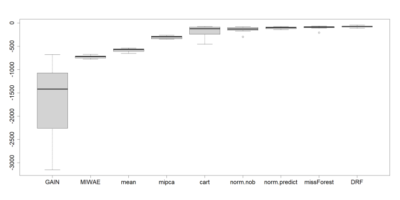
Figure 2: True distribution against a draw from different imputation methods.

DRF, a distributional method > mice-RF but fails to deal with the covariate shift (centering ≈ 2 instead of 5).

Finding an imputation method that meets (1) - (4) is still an open problem!

Empirical study: ranking with energy scores and not RMSE





Gaussian relation with shifts

Non linear relation with shifts

$$\text{Ex with } d = 6,\ n = 1500,\ 20\%\ \text{NA and CIMAR, } X_{O^c} = \mathbf{B}f(X_O) + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

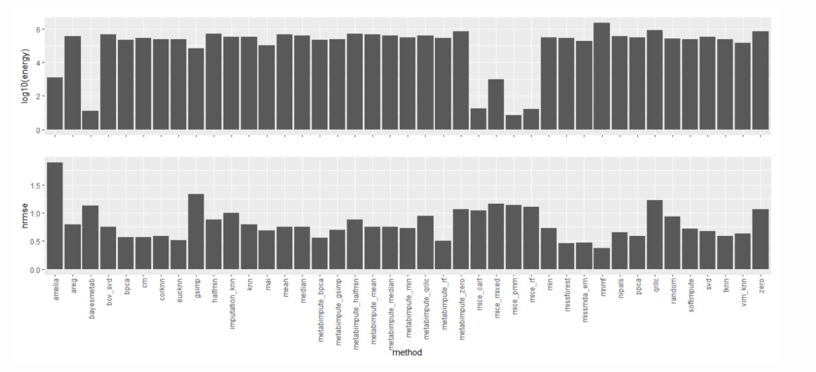
Energy distance[46] between imputed & real data

$$d(H, P^*) = 2\mathbb{E}[\|X - Y\|_{\mathbb{R}^d}] - \mathbb{E}[\|X - X'\|_{\mathbb{R}^d}] - \mathbb{E}[\|Y - Y'\|_{\mathbb{R}^d}],$$

where $\|\cdot\|_{\mathbb{R}^d}$ is the Euclidean metric on $\mathbb{R}^d$, $X \sim H$, $Y \sim P^*$ and X', Y' are independent copies of X and Y.

[46] Székely & Rizzo. Energy statistics *Journal of stat. planning & inference*. 2013

Empirical study: ranking with energy scores and not RMSE



credit: Krystyna Grzesiak, Michal Burdukiewicz[47] 230 scenarios (10 missing values patterns 23 different-size datasets)

---

[47] imputomics: web server and R package for missing values imputation in metabolomics data. *Bioinformatics 2024*.

What if the underlying values are not available?

- The question of how to evaluate imputation methods becomes much harder when the **true underlying values are not available**.
- The energy distance is directly related to the energy score[48]:

$$es(H, y) = \mathbb{E}_{X\sim H}[\|X - y\|_{\mathbb{R}^d}] - \frac{1}{2}\mathbb{E}_{X,X'\sim H}[\|X - X'\|_{\mathbb{R}^d}]$$

Theorem

In expectation, we score the true distribution lowest, i.e. :

$$S(P^*, H) := \mathbb{E}_{Y\sim P^*}[es(H, Y)] \geq \mathbb{E}_{Y\sim P^*}[es(P^*, Y)] := S(P^*, P^*)$$

[48] Gneiting, Raftery, Strictly Proper Scoring Rules, Prediction, and Estimation, *JASA*, 2007

General Idea of Scores

- The **energy** score can be used to score **distributional prediction**
- Assume we have learned a distribution H based on n samples, from which we can sample (for instance using DRF)
- We would like to test this distribution against a new test point y
- Can use the Energy score:

$$es(H, y) = \mathbb{E}_{X \sim H}[\|X - y\|_{\mathbb{R}^d}] - \frac{1}{2}\mathbb{E}_{X, X' \sim H}[\|X - X'\|_{\mathbb{R}^d}]$$

- **If we can sample from H, $es(H, y)$ can be easily approximated!**

Imputation Scores

- P refers to the distribution of X with missing values
- $P^* \in \mathcal{P}$ refers to the distribution of X^* without missing values.
- H refers to an imputation distribution.

Definition (Proper Imputation Score (I-Score))

A real-valued function $S_{NA}(H, P)$ is a proper I-Score iff

$$S_{NA}(H, P) \leq S_{NA}(P^*, P),$$

for any imputation distribution H.

Imputation Scores

- ▷ For this to work under the challenging MAR setting we need to have a set of variables O_j that is **observed whenever X_j is observed**:

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ NA & x_{2,2} & x_{2,3} & x_{2,4} \\ x_{3,1} & NA & x_{3,3} & x_{3,4} \\ x_{1,4} & NA & NA & x_{4,4} \end{pmatrix}$$

Figure 3: Illustration of O_j, for $j = 1, 2$. For X_2, $X_{O_j} = (X_3, X_4)$ in gray, while for X_1, $X_{O_j} = X_4$ in black.

$$S^j_{\mathrm{NA}}(H, P) =$$
$$\mathbb{E}_{X_{O_j} \sim P_{X_{O_j} \mid M \in L_j}} \left[\mathbb{E}_{\substack{X \sim H_{X_j \mid X_{O_j}} \\ Y \sim H^*_{X_j \mid X_{O_j}}}} [\|X - Y\|_2] - \frac{1}{2} \mathbb{E}_{\substack{X \sim H_{X_j \mid X_{O_j}} \\ X' \sim H_{X_j \mid X_{O_j}}}} [\|X - X'\|_2] \right], \quad (1)$$

$$S_{\mathrm{NA}}(H, P) = \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} S^j_{\mathrm{NA}}(H, P),$$

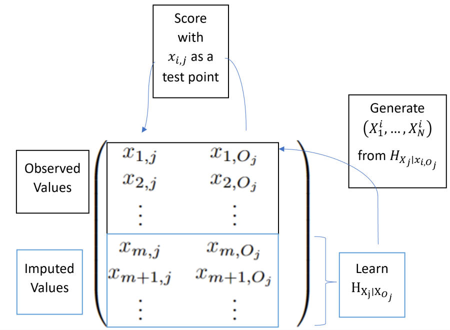


Figure 4: Illustration of the new scoring method. The PMM view shows that only certain conditional distributions can be compared under MAR. This is what we utilize here.

Theorem

Assume there exists $j \in \{1, \dots, d\}$ such that $O_j = \bigcap_{m \in L_j} \{l : m_l = 0\}$ is not empty and, for all k such that $O_k \neq \emptyset$, $X_k \perp\!\!\!\perp M_k \mid X_{O_k}$. Then the population version $S_{NA}^{es}(H, P)$ is a proper I-Score.

Propriety in Action

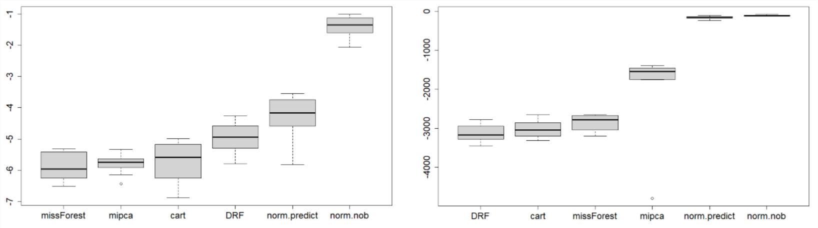


Figure 5: Left: Ordering of the I-score, Right: Ordering of the (negative) energy distance. The latter uses the true underlying values.

Conclusion on single imputation methods & FCS

- Non-parametric PMM view of missing (different environments) helps understand non-parametric imputation under MAR
- Identification result for FCS: the right conditional distributions are identifiable under MAR with no parametric assumption
- Identification under the weakest MAR assumption [49]. Beyond MAR. $\forall$j $\in \{1, \ldots, d\}$, $\forall$x $\in \mathcal{X}$, CIMNAR: $\mathbb{P}(M_j = 1|x) = \mathbb{P}(M_j = 1|x_{-j})$

---

[49] Deng et al., (2022) and Fang (2023) showed identifiability for GAN imputation under CIMAR
[50] Shen & Meinshausen (2024). Engression: extrapolation through the lens of distributional regression. *JRSS B.*

Conclusion on single imputation methods & FCS

- Non-parametric PMM view of missing (different environments) helps understand non-parametric imputation under MAR
- Identification result for FCS: the right conditional distributions are identifiable under MAR with no parametric assumption
- Identification under the weakest MAR assumption [49]. Beyond MAR. $\forall$j $\in \{1, \ldots, d\}$, $\forall$x $\in \mathcal{X}$, CIMNAR: $\mathbb{P}(M_j = 1|x) = \mathbb{P}(M_j = 1|x_{-j})$

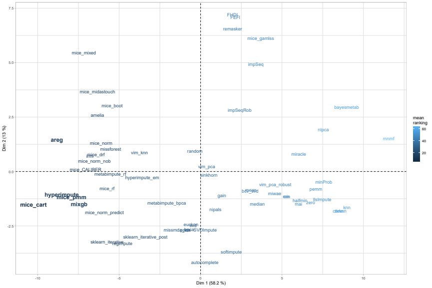
- The quest for an FCS imputation method meeting all 3 points is open
- mice-DRF promising (code available) - mice-Engression[50]
- Imputation scores with missing values that are proper under MAR: ranking imputation methods
- Simulations MAR for benchmarks

[49] Deng et al., (2022) and Fang (2023) showed identifiability for GAN imputation under CIMAR

[50] Shen & Meinshausen (2024). Engression: extrapolation through the lens of distributional regression. *JRSS B.*

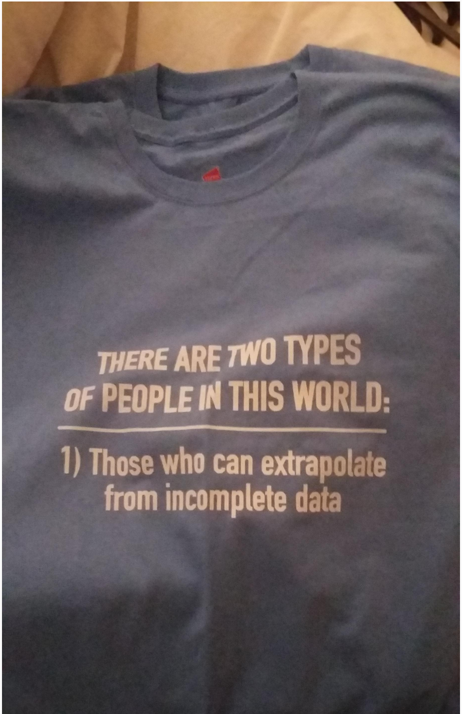
Benchmarking imputation methods

- 65 methods (R & Python)
- 14 datasets: 100-50000 observations and 3-400 features
- 10-30 % NA MCAR, MAR, Standardized energy distance



- Mice-cart[51], aregImpute (close to mice+splines+pmm)[52], Hyperimpute (mice + model selection RF, XGBoost, Logistic Reg., etc)[53], Mice mixed[54]

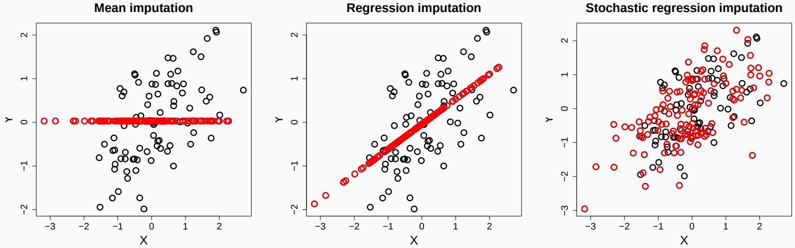
[51] Buuren & Groothuis-O. (2011). Multivariate imputation by chained equations in R. *JSS*.
[52] Harrell & Dupont (2018). Hmisc: Harrell miscellaneous. R package version 4.1-1. Stat. Comput.
[53] Jarrett et al. (2022). Hyperimpute: Generalized iterative imputation with automatic model selection. *ICML*.
[54] Varga (2020). missCompare: Intuitive Missing Data Imputation. R package version 1.0.3. Stat.


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

Multiple Imputation

Single imputation methods: Danger!

$$\left[\bar{y} - qt_{n-1}\frac{\hat{\sigma}_y}{\sqrt{n}}; \bar{y} - qt_{n-1}\frac{\hat{\sigma}_y}{\sqrt{n}}\right]$$



	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
$\rho = 0.6$	0.30	0.78	0.59
$CI\,\mu_y 95\%$	39.4	61.6	70.8

The idea of imputation is both seductive and dangerous *(Dempster and Rubin, 1983)*

Confidence interval for a mean

Let $Y = (Y_1, \ldots, Y_n)'$ be i.i.d. independent Gaussian random with expectation μ_y and variance $\sigma_y^2 > 0$.

- The empirical mean $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$
- $\bar{Y} \sim \mathcal{N}(\mu_y, \sigma_y^2/n)$
- A confidence interval for μ

$$\mathbb{P}\left(\bar{Y} - \frac{\sigma_y}{\sqrt{n}} \Phi^{-1}(1 - \alpha/2) \ \leq \mu \leq \bar{Y} + \frac{\sigma_y}{\sqrt{n}} \Phi^{-1}(1 - \alpha/2) \right) = 1 - \alpha$$

Confidence interval for a mean

Let $Y = (Y_1, \ldots, Y_n)'$ be i.i.d. independent Gaussian random with expectation μ_y and variance $\sigma_y^2 > 0$.

- The empirical mean $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$
- $\bar{Y} \sim \mathcal{N}(\mu_y, \sigma_y^2/n)$
- A confidence interval for μ

$$\mathbb{P}\left(\bar{Y} - \frac{\sigma_y}{\sqrt{n}} \Phi^{-1}(1-\alpha/2) \ \leq \mu \leq \bar{Y} + \frac{\sigma_y}{\sqrt{n}} \Phi^{-1}(1-\alpha/2)\right) = 1 - \alpha$$

Variance unknown:

$$\frac{\sqrt{n}}{\widehat{\sigma_y}} \left(\bar{Y} - \mu_y\right) \sim T(n-1)$$

$$\left[\bar{y} - \frac{\widehat{\sigma}_y}{\sqrt{n}} qt_{1-\alpha/2}(n-1) \ , \ \bar{y} + \frac{\hat{\sigma}_y}{\sqrt{n}} qt_{1-\alpha/2}(n-1)\right]$$

Simulation

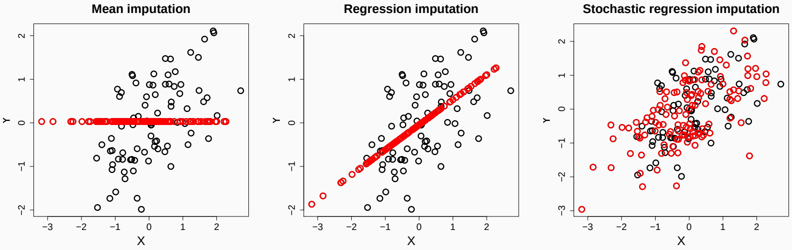
1. Generate bivariate Gaussian data ($\mu_y = 0, \sigma_y = 1, \rho = 0.6$)
2. Put missing values on y
3. Imput missing entries
4. Compute the confidence interval of μ_y - count if the true value $\mu_y = 0$ is in the confidence interval
5. Repeat the steps 1-4, 10000 times

$\Rightarrow$ Give the coverage

Code available on Rmistatic.

Single imputation methods: Danger!

$$\left[\bar{y} - qt_{n-1}\frac{\hat{\sigma}_y}{\sqrt{n}}; \bar{y} - qt_{n-1}\frac{\hat{\sigma}_y}{\sqrt{n}}\right]$$



	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
$\rho = 0.6$	0.30	0.78	0.59
$CI\,\mu_y 95\%$	39.4	61.6	70.8

The idea of imputation is both seductive and dangerous *(Dempster and Rubin, 1983)*

$\Rightarrow$ Standard errors of the parameters ($\hat{\sigma}_{\hat{\mu}_y}$) calculated from the imputed data set are underestimated

Underestimation of variance

Classical confidence interval for μ_y $\left[\bar{y} - qt_{n-1}\frac{\hat{\sigma}_y}{\sqrt{n}}; \bar{y} - qt_{n-1}\frac{\hat{\sigma}_y}{\sqrt{n}}\right]$

Asymptotic variance with MCAR values (Little & Rubin, 2019. p158):

$$\frac{\hat{\sigma}_y^2}{n_{obs}}\left(1 - \hat{\rho}^2\frac{n - n_{obs}}{n_{obs}}\right) = \frac{\hat{\sigma}_y^2}{n}\left(1 + \frac{n - n_{obs}}{n_{obs}}(1 - \hat{\rho}^2)\right)$$

$\Rightarrow$ When the $\rho = 1$, we trust the prediction and the coverage given by stochastic regression is OK.

$\Rightarrow$ Coverage of single imputation is too low: need to take into account the uncertainty associated to the predictions.

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

⇒ 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

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

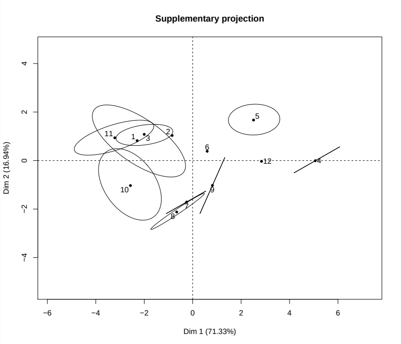
X_1	X_2	X_3	Y
7	20	10	s
-6	45	12	s
0	-5	30	no s
2	32	35	s
-2	20	12	no s
1	63	40	s

Visualization of the imputed values[55]

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)**

[55] 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):

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

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

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

$\Rightarrow$ Variability of missing values taken into account. **Metric: coverage.**

Multiple imputation: naive attempt

1. Generating M imputed data sets

   First idea: several stochastic regression
   for $m = 1, ..., M$, draw $\hat{y}_i$ from the predictive $\mathcal{N}(x_i\hat{\beta}, \hat{\sigma}^2)$

2. Performing the analysis on each imputed data set
3. Combining: variance = within + between imputation variance

	$M = 1$	$M = 50$
$\mu_y = 0$	0.01	0.01
$\sigma_y = 1$	0.99	0.99
$\rho = 0.6$	0.59	0.59
$CI\mu_y 95\%$	70.8	81.8

Multiple imputation: naive attempt

1. Generating M imputed data sets

   First idea: several stochastic regression
   for $m = 1, ..., M$, draw $\hat{y}_i$ from the predictive $\mathcal{N}(x_i\hat{\beta}, \hat{\sigma}^2)$

2. Performing the analysis on each imputed data set
3. Combining: variance = within + between imputation variance

	$M = 1$	$M = 50$
$\mu_y = 0$	0.01	0.01
$\sigma_y = 1$	0.99	0.99
$\rho = 0.6$	0.59	0.59
$CI\mu_y 95\%$	70.8	81.8

$\Rightarrow$ Variability of the parameters is missing: "improper" imputation

Multiple imputation: naive attempt

1. Generating M imputed data sets

   First idea: several stochastic regression
   for $m = 1, ..., M$, draw $\hat{y}_i$ from the predictive $\mathcal{N}(x_i\hat{\beta}, \hat{\sigma}^2)$

2. Performing the analysis on each imputed data set
3. Combining: variance = within + between imputation variance

	$M = 1$	$M = 50$
$\mu_y = 0$	0.01	0.01
$\sigma_y = 1$	0.99	0.99
$\rho = 0.6$	0.59	0.59
$CI \mu_y 95\%$	70.8	81.8

$\Rightarrow$ Variability of the parameters is missing: "improper" imputation
$\Rightarrow$ Prediction variance = estimation variance plus noise

Regression: variance of prediction

$$y_{n+1} = x'_{n+1}\beta + \varepsilon_{n+1}$$
$$\hat{y}_{n+1} = x'_{n+1}\hat{\beta}$$
$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\begin{aligned} V[\hat{y}_{n+1} - y_{n+1}] &= V[x'_{n+1}(\hat{\beta} - \beta) - \varepsilon_{n+1}] \\ &= x'_{n+1}V(\hat{\beta} - \beta)x_{n+1} + \sigma^2] \\ &= \hat{\sigma}^2\left(x'_{n+1}(X'X)^{-1}x_{n+1} + 1\right) \end{aligned}$$

CI for the prediction

$$\left[x'_{n+1}\hat{\beta} + -t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{\left(x'_{n+1}(X'X)^{-1}x_{n+1} + 1\right)}\right]$$

Multiple imputation: bivariate case [56]

$\Rightarrow$ Proper multiple imputation with $y_i = x_i\beta + \varepsilon_i$

1. Variability of the parameters, M plausible: $(\hat{\beta})^1, ..., (\hat{\beta})^M$

   $\Rightarrow$ Bootstrap
   $\Rightarrow$ Posterior distribution: Data Augmentation (Tanner & Wong, 1987)

2. Noise: for $m = 1, ..., M$, missing values $\hat{y}_i^m$ are imputed by drawing from the predictive distribution $\mathcal{N}(x_i\hat{\beta}^m, (\hat{\sigma}^2)^m)$

	Improper	Proper
$CI\mu_y 95\%$	0.818	0.935

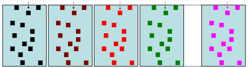
[56] Code available on Rmistatic.

Multiple imputation

$\Rightarrow$ Aim: provide estimation of the parameters and of their variability (taken into account the variability due to missing values)

Single imputation: a single value can't reflect the uncertainty of prediction $\Rightarrow$ underestimate the standard errors

1. Generating M imputed data sets: variance of prediction



2. Performing the analysis on each imputed data set[57], [58]
3. Combining: variance = within + between imputation variance

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

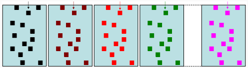
[57] The analysis model may be "in agreement" with the imputation model: congeniality.
[58] Little & Rubin. 2019. Statistical Analysis with Missing Data, 3rd Edition. Wiley

Multiple imputation

$\Rightarrow$ Aim: provide estimation of the parameters and of their variability (taken into account the variability due to missing values)

Single imputation: a single value can't reflect the uncertainty of prediction $\Rightarrow$ underestimate the standard errors

1. Generating M imputed data sets: variance of prediction



"1) Variance of estimation of the parameters + 2) Noise"

2. Performing the analysis on each imputed data set[57], [58]
3. Combining: variance = within + between imputation variance

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

[57] The analysis model may be "in agreement" with the imputation model: congeniality.

[58] Little & Rubin. 2019. Statistical Analysis with Missing Data, 3rd Edition. Wiley

Multiple Imputation with joint modeling

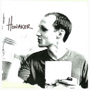
$\Rightarrow$ Hypothesis $x_i \sim \mathcal{N}(\mu, \Sigma)$

Algorithm Expectation Maximization Bootstrap:

1. Bootstrap rows: $X^1, \ldots, X^M$
   EM algorithm: $(\hat{\mu}^1, \hat{\Sigma}^1), \ldots, (\hat{\mu}^M, \hat{\Sigma}^M)$
2. Imputation: $\hat{x}^m_{i,miss}$ drawn from $\mathcal{N}\left(\hat{\mu}^m_{miss|obs}, \hat{\Sigma}^m_{miss|obs}\right)$

Easy to parallelized. Implemented in `Amelia` (website)

Amelia Earhart



James Honaker

Gary King

Matt Blackwell

Multiple imputation by chained equations or FCS[61]

- Impute variables 1 by 1 using all other variables as inputs (round-robin)
- One model/variable: flexible for categorical, ordinal variables
- Cycle through variables: iteratively refine the imputation

1. Initial imputation: mean imputation
2. For a variable j

   - Imputation of the missing values in variable j with a model of X_j on the other X_{-j}: stochastic regression imput. $\sim \mathcal{N}\left((x_{i,-j})'\hat{\beta}^{-j}, \hat{\sigma}^{-j}\right)$
3. Cycling through variables

$\Rightarrow$ Imputed values are draws from an (implicit) joint distribution
$\Rightarrow$ With continuous variables & regression/variable: gibbs $\mathcal{N}(\mu, \Sigma)$ [59], [60]

"*There is no clear-cut method for determining whether MICE has converged*"
Implemented in R package `mice` & IterativeImputer from `scikitlearn` (default iterative ridge regression)

Stef van Buuren

[59] Monte Carlo statistical methods (Robert, Casella, 2004) (p344),
[60] The EM algorithm and extensions (McLachlan, et al. 1998) (p243)
[61] van Buuren. 2018. Flexible Imputation of Missing Data. Second Edition. CRC Press

Multiple imputation by chained equations or FCS[61]

- Impute variables 1 by 1 using all other variables as inputs (round-robin)
- One model/variable: flexible for categorical, ordinal variables
- Cycle through variables: iteratively refine the imputation

1. Initial imputation: mean imputation
2. For a variable j
   - $(\hat{\beta}^{-j}, \hat{\sigma}^{-j})$ drawn from a Bootstrap: $(\hat{\beta}^{-j}, \hat{\sigma}^{-j})^1, ..., (\hat{\beta}^{-j}, \hat{\sigma}^{-j})^M$
   - Imputation of the missing values in variable j with a model of X_j on the other X_{-j}: stochastic regression imput. $\sim \mathcal{N}\left((x_{i,-j})'\hat{\beta}^{-j}, \hat{\sigma}^{-j}\right)$
3. Cycling through variables

⇒ Imputed values are draws from an (implicit) joint distribution

⇒ With continuous variables & regression/variable: gibbs $\mathcal{N}(\mu, \Sigma)$ [59], [60]

"*There is no clear-cut method for determining whether MICE has converged*"
Implemented in R package `mice` & IterativeImputer from `scikitlearn` (default iterative ridge regression)

Stef van Buuren

[59] Monte Carlo statistical methods (Robert, Casella, 2004) (p344),

[60] The EM algorithm and extensions (McLachlan, et al. 1998) (p243)

[61] van Buuren. 2018. Flexible Imputation of Missing Data. Second Edition. CRC Press

Joint versus Conditional modeling

$\Rightarrow$ Imputed values are both seen as draws from a joint distribution

Conditional modeling takes the lead?

- Flexible: one model/variable. Easy to deal with interactions and variables of different nature (binary, ordinal, categorical...)
- Many statistical models are conditional models
- Tailor to your data - Super powerful in practice

$\Rightarrow$ Drawbacks: one model/variable. **Computational costly**[62]

What to do with high correlation or when $n < p$

- JM shrinks the covariance $\Sigma + k\mathbb{I}$ (selection of k?)
- CM: ridge regression or predictors selection/variable

Challenges with multiple imputation

- MI in high dimension? Theory with small n, large p?
- Aggregating lasso regressions? clustering?

[62] Improvement on mice pmm for large sample size, see mice github repo - still costly for large d

Code Example

https://www.dropbox.com/scl/fo/8euubsr1l5tqhe1ksi8bk/ABd2NDfV2NR31KY7cLOY7h0?rlkey=2r6cfu614bvqk4hrn0xekxtyn&e=1&st=bexeahy2&dl=0

Expectation Maximization

A bit more notation

- P^*: the marginal distribution of the complete data variable X, which is assumed to be absolutely continuous with respect to Lebesgue's measure with density p^*.
- $P^*_{X,M}$: the joint distribution of (X^*, M) with joint density $p^*_{X,M}$.
- $\mathbb{P}_M$: the marginal distribution of the mask variable M, such that for every measurable set $A \subset \{0,1\}^d$, $\mathbb{P}_M(A) = \sum_{m \in A} \mathbb{P}[M = m]$.
- $\mathbb{P}_{M|X}$: the conditional distribution of the mask variable M given X, such that for every measurable set $A \subset \{0,1\}^d$, $\mathbb{P}_{M|X}(A) = \sum_{m \in A} \mathbb{P}(M = m \mid X)$.
- $P^*_{X|M}$: The conditional distribution of X^* given pattern M.

Now in addition we try to model P^* parametrically with a model $(P_\theta)_\theta/(p_\theta)_\theta$. If the model is correctly specified then there exists θ^* such that

$$P_{\theta^*} = P^*, \; p_{\theta^*} = p^*$$

Ignorable missing values mechanism

- Let us assume for the next two slide that there is also a parameter ϕ, such that $\mathbb{P}(M = m \mid x) = \mathbb{P}_{\phi^*}(M = m|x)$.
- Then with M(C)AR data, we get for all (θ, ϕ) the observed distribution:

$$
\begin{aligned}
p_{\theta,\phi}(o(x,m), m) &= \int p_\theta(x)\mathbb{P}_\phi(M = m|x) do^c(x,m) \\
&= \int p_\theta(x)\mathbb{P}_\phi(M = m|o(x,m)) do^c(x,m) \\
&= \mathbb{P}_\phi(M = m|o(x,m)) \int p_\theta(x) do^c(x,m) \\
&= \mathbb{P}_\phi(M = m|o(x,m)) p_\theta(o(x,m)).
\end{aligned}
$$

Ignorable missing values mechanism

- Thus the full likelihood problem becomes:

$(\theta_n^{full}, \phi_n^{full}) = \arg\max_{\theta,\phi} \sum_{i=1}^{n} \left\{ \log p_\theta^{(M_i)}\left(o(X_i, M_i)\right) + \log\left(\mathbb{P}_\phi(M = M_i | o(X_i, M_i))\right) \right\}$

- The likelihood ignoring the missing value mechanism is:

$$\theta_n^{\mathrm{ML}} = \arg\max_\theta \underbrace{\sum_{i=1}^{n} \log p_\theta^{(M_i)}\left(o(X_i, M_i)\right)}_{L_{obs}(\theta)}$$

- If the parameter space of (θ, ϕ) is given as the product of the space of θ and the one of ϕ [="Parameter Distinctness"]: $\theta_n^{full} = \theta_n^{\mathrm{ML}}$!!
  $\Rightarrow$ This was the main motivation to the practice of doing MLE while completely *ignoring* the missingness mechanism!

KL Divergence

- One cannot talking about the MLE without talking about KL Divergence.
- Let P_1, P_2 have densities p_1, p_2 with respect to the Lebesgue measure (though could be any measure dominating the two). Then

$$\mathsf{KL}\left(P_1 \| P_2\right) = \begin{cases} \int \log\left(\frac{p_1(x)}{p_2(x)}\right) p_1(x)\mathrm{d}x, & \text{if } P_2 \ll P_1 \\ \infty, & \text{else.} \end{cases} \tag{2}$$

- In fact, it can be shown that the MLE is effectively minimizing the KL Divergence between the proposed density and the true data distribution.

MLE + Missing Values

Informal variant of Theorems 1 and 2 in Golden et al 2019[63]

Under appropriate regularity conditions (including the existence of the involved quantities), θ_n^{ML} is strongly consistent and asymptotically normal:

$$\theta_n^{\mathrm{ML}} \xrightarrow[n\to+\infty]{\text{a.s.}} \theta_\infty^{\mathrm{ML}},$$

$$\sqrt{n}(\theta_n^{\mathrm{ML}} - \theta_\infty^{\mathrm{ML}}) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, A^{-1}VA^{-1}\right),$$

where

$$\theta_\infty^{\mathrm{ML}} = \arg\min_{\theta\in\Theta} \mathbb{E}_{M\sim\mathbb{P}_M}\left[\mathrm{KL}\left({P_{X|M}^*}^{(M)} \| P_\theta^{(M)}\right)\right],$$

$$A = \mathbb{E}_{(X^*,M)\sim P_{X,M}^*}\left[\nabla_{\theta,\theta}^2 \log p_{\theta_\infty^{\mathrm{ML}}}^{(M)}(o(X,M))\right],$$

and

$$V = \mathbb{E}_{(X,M)\sim P_{X,M}^*}\left[\nabla_\theta \log p_{\theta_\infty^{\mathrm{ML}}}^{(M)}(o(X,M)) \cdot \nabla_\theta \log p_{\theta_\infty^{\mathrm{ML}}}^{(M)}(o(X,M))^T\right].$$

[63] Golden, Henley, White, Kashner. *Consequences of Model Misspecification for Maximum Likelihood Estimation with Missing Data*. Econometrics. 2019

MLE + Missing Values

- Thus, under regularity conditions, θ_n^{ML} converges a.s. to the best approximation (in KL terms) to $P^*_{X|M}$, averaged over M.
- This is true under any missingness mechanism and under misspecification of the (complete) data distribution P_θ.
- Remarkably, it can be shown that if P_θ is correctly specified, $\mathbb{P}(M > 0)$, and MAR holds, $\theta_\infty^{\mathsf{ML}} = \theta^*$.
  $\Longrightarrow$**MLE is consistent under the MAR missingness, even without the assumption of Parameter Distinctness!**
- However, this consistency is intimately connected to the KL Divergence and not true for general M-Estimators.

Expectation - Maximization (Dempster et al., 1977)

Rationale to get ML estimates: max the observed data likelihood $L_{obs}(\theta)$ through max of $L_{comp}(\theta)$. Augment the data to simplify the problem.

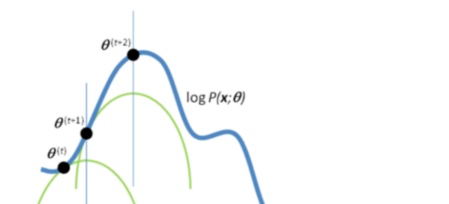
E step (conditional expectation):

$$Q(\theta, \theta^\ell) = \sum_{i=1}^{n} \mathbb{E}[\underbrace{\log(p_\theta(o(X_i, M_i), o^c(X_i^*, M_i)))}_{\text{Full Likelihood}} \mid o(X_i, M_i)]$$

M step (maximization):

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

Result: $L_{obs}(\theta) - L_{obs}(\theta^\ell) \geq Q(\theta, \theta^\ell) - Q(\theta^\ell, \theta^\ell)$. Thus if $\theta^{\ell+1} = \text{argmax}_\theta Q(\theta, \theta^\ell)$, $L_{obs}(\theta^{\ell+1}) \geq L_{obs}(\theta^\ell)$.



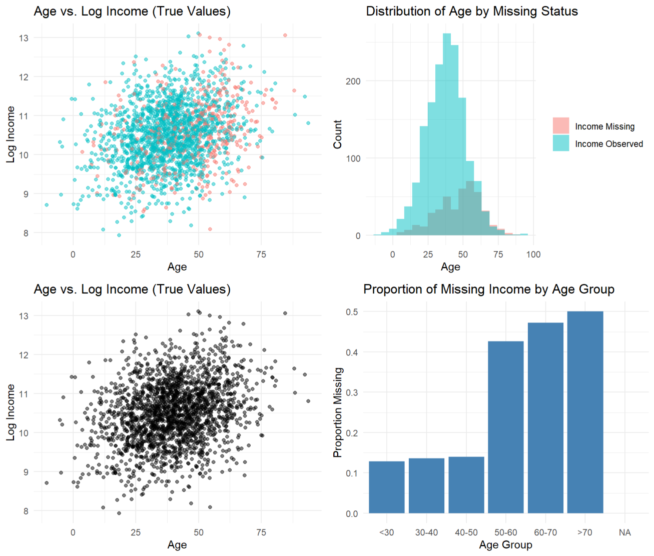
Income & Age Example

- Say X_1 is the logarithm of income of a person, X_2 is age.
- We assume joint normality: $(X_1^*, X_2^*) \sim P^* = N((\mu_1, \mu_2), \Sigma)$.
- Moreover, we assume that age is always observed, but income can be missing, leading to two patterns: $m_1 = (0, 0)$ and $m_2 = (1, 0)$.
- Income is missing randomly throughout the population, but there is a somewhat higher missingness for older people.
- In particular, we model this as:

$$\mathbb{P}(X_1 \text{ missing} \mid X = x) = \mathbb{P}(M = m_2 \mid X = x) = (1 - \varepsilon)\alpha + \varepsilon \mathbf{1}\{x_2 > 50\},$$

for $0 \leq \alpha < 0.5$, $0 \leq \varepsilon < 0.5$.

Example


Age vs. Log Income (True Values)
Log Income
Age
Distribution of Age by Missing Status
Count
Income Missing
Income Observed
Age vs. Log Income (True Values)
Log Income
Age
Proportion of Missing Income by Age Group
Proportion Missing
<30
30-40
40-50
50-60
60-70
>70
NA
Age Group

EM Algorithm Example

$$\log(p_\theta(o(X_i, M_i), o^c(X_i^*, M_i))) = \log(p_{\mu,\Sigma}(o(X_i, M_i), o^c(X_i^*, M_i)))$$
$$= -\frac{1}{2}\log(\det(\Sigma)) - \frac{1}{2}(X_i - \mu)^T\Sigma^{-1}(X_i - \mu)$$

Thus whenever $M_i = m_1 = (0, 0)$:

$$\mathbb{E}\left[\log(p_\theta(o(X_i, M_i), o^c(X_i^*, M_i))) \mid o(X_i, M_i)\right]$$
$$= -\frac{1}{2}\log(\det(\Sigma)) - \frac{1}{2}(X_i - \mu)^T\Sigma^{-1}(X_i - \mu)$$

and when $M_i = m_2 = (1, 0)$:

$$\mathbb{E}\left[\log(p_\theta(o(X_i, M_i), o^c(X_i^*, M_i))) \mid o(X_i, M_i)\right]$$
$$= -\tfrac{1}{2}\log(\det(\Sigma)) - \tfrac{1}{2}\mathbb{E}\left[\left(\begin{pmatrix} X_{i,1} \\ X_{2,2} \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\right)^T \begin{pmatrix} \sigma_{1,1} & \sigma_{2,1} \\ \sigma_{2,1} & \sigma_{2,2} \end{pmatrix}^{-1} \left(\begin{pmatrix} X_{i,1} \\ X_{2,2} \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\right) \mid X_{i,2}\right]$$

EM Algorithm Example

Thus for all i such that $M_i = m_2$,

$$\mathbb{E}\left[\log(p_\theta(o(X_i, M_i), o^c(X_i^*, M_i))) \mid o(X_i, M_i)\right]$$
$$= \frac{\sigma_{2,2}\mathbb{E}[(X_{i,1}-\mu_1)^2|X_{i,2}]-2\sigma_{2,1}\mathbb{E}[(X_{i,1}-\mu_1)|X_{i,2}](X_{i,2}-\mu_2)+\sigma_{1,1}(X_{i,2}-\mu_2)^2}{\sigma_{1,1}\sigma_{2,2}-\sigma_{2,1}^2}$$

Thus finding $Q(\theta, \theta^{(\ell)})$,

$$\begin{aligned} &Q(\theta, \theta^{(\ell)}) \\ &= \sum_{i:M_i=m_1} \mathbb{E}\left[\log(p_\theta(o(X_i, M_i), o^c(X_i^*, M_i))) \mid o(X_i, M_i)\right] \\ &+ \sum_{i:M_i=m_2} \mathbb{E}\left[\log(p_\theta(o(X_i, M_i), o^c(X_i^*, M_i))) \mid o(X_i, M_i)\right] \end{aligned}$$

boils down to finding $\mathbb{E}[(X_{i,1} - \mu_1)^2 \mid X_{i,2}]$, $\mathbb{E}[(X_{i,1} - \mu_1) \mid X_{i,2}]$.

Estimation of the mean and covariance matrix

Ex: Hypothesis $z_{i.} \sim \mathcal{N}(\mu, \Sigma)$

$\Rightarrow$ Point estimates with EM:

```
> library(norm)
> pre <- prelim.norm(as.matrix(don))
> thetahat <- em.norm(pre)
> getparam.norm(pre,thetahat)
```

Exercice: EM with bivariate data (2.1.1):
https://rmisstastic.netlify.app/tutorials/josse_bookdown_lecturenotesmissing_2020

Estimation of the mean and covariance matrix

Ex: Hypothesis $z_{i.} \sim \mathcal{N}(\mu, \Sigma)$

$\Rightarrow$ Point estimates with EM:

```
> library(norm)
> pre <- prelim.norm(as.matrix(don))
> thetahat <- em.norm(pre)
> getparam.norm(pre,thetahat)
```

Exercice: EM with bivariate data (2.1.1):
`https://rmisstastic.netlify.app/tutorials/josse_bookdown_lecturenotesmissing_2020`

$\Rightarrow$ Variances:

- ▷ Supplemented EM (Meng, 1991), Louis formulae
- ▷ Bootstrap approach:
  - ◇ Bootstrap rows: Z^1, ... , Z^B
  - ◇ EM algorithm: $(\hat{\mu}^1, \hat{\Sigma}^1)$, ... , $(\hat{\mu}^B, \hat{\Sigma}^B)$

Inference with EM algorithms

Jiang, **J.** et al. (2019). Logistic Regression with Missing Covariates, Parameter Estimation, Model Selection and Prediction. *CSDA*. 2019.

Bogdan, **J.** et al. (2020). Adaptive Bayesian SLOPE - High dimensional Model Selection with Missing Values. *JCGS.*

See slides in Mybox:slidesPhDdefenseJiangLogisticNASlopeNA

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***."*

- **Matrix completion** aims at completing data 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)

Challenges with heterogeneous sources and missing data

⇒ <u>What to do when you have both MCAR, MAR, **MNAR** in the data?</u>

⇒ <u>Federated learning with missing values</u>

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

Sporadic, systematic & missing modalities. Due to the pandemic, many patients did not complete their tests

Recap Day 1

- A true missing values mask an underlying values
- Different missing values mechanisms (MCAR, MAR, MNAR) to explain why values are missing

Recap Day 1

- A true missing values mask an underlying values
- Different missing values mechanisms (MCAR, MAR, MNAR) to explain why values are missing

Inference with missing values aim at estimating parameters (regression coefficient, causal effects) despite missing values

- Likelihood based methods: ignore the missing values mechanism to do inference: EM algorithm

Recap Day 1

- A true missing values mask an underlying values
- Different missing values mechanisms (MCAR, MAR, MNAR) to explain why values are missing

Inference with missing values aim at estimating parameters (regression coefficient, causal effects) despite missing values

- Likelihood based methods: ignore the missing values mechanism to do inference: EM algorithm
- Imputation: mean imputation should be avoided. Look for an imputation that preserve the joint distribution of the data
- Compare imputation methods with distributional metrics like energy distance, i-score with missing values
- Multiple imputation to get confidence intervals
- Proper multiple imputation to reflect the variance of prediction of missing values: variance of the parameters of the imputation model + noise

Matrix Completion: PCA imputation - Low rank approximation with missing values

PCA (complete)

Find the subspace that best represents the data

Figure 6: Camel or dromedary?

⇒ Best approximation when projecting the data

⇒ Best representation of the variability

⇒ Do not distort the distances between observations

PCA (complete)

Find the subspace that best represents the data

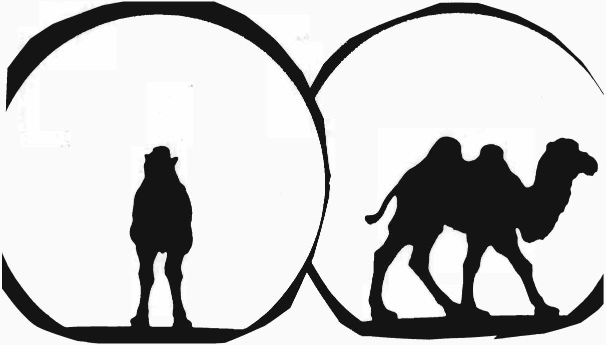


Figure 6: Camel or dromedary? source J.P. Fénelon

⇒ Best approximation when projecting the data
⇒ Best representation of the variability
⇒ Do not distort the distances between observations

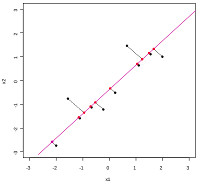
PCA reconstruction

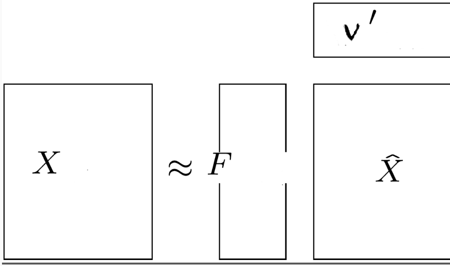
X

-2.00	-2.74
-1.56	-0.77
-1.11	-1.59
-0.67	-1.13
-0.22	-1.22
0.22	-0.52
0.67	1.46
1.11	0.63
1.56	1.10
2.00	1.00

$\hat{\mu}$

-2.16	-2.58
-0.96	-1.35
-1.15	-1.55
-0.70	-1.09
-0.53	-0.92
0.04	-0.34
1.24	0.89
1.05	0.69
1.50	1.15
1.67	1.33





$\Rightarrow$ Minimizes distance between observations and their projection

$\Rightarrow$ Approx $X_{n\times p}$ with a low rank matrix $S < p$ $\|A\|_2^2 = \text{tr}(AA^\top)$:

$$\arg\min_{\mu} \left\{ \|X - \mu\|_2^2 : \text{rank}\,(\mu) \leq S \right\}$$

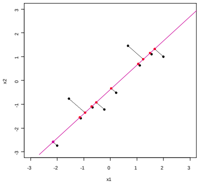
PCA reconstruction

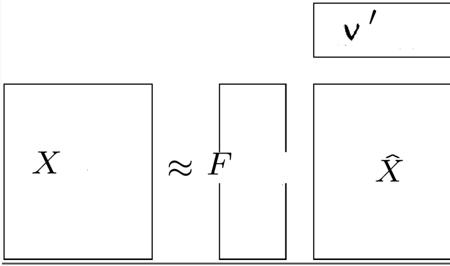
X

-2.00	-2.74
NA	-0.77
-1.11	-1.59
-0.67	-1.13
-0.22	NA
0.22	-0.52
0.67	1.46
NA	0.63
1.56	1.10
2.00	1.00

$\hat{\mu}$

-2.16	-2.58
-0.96	-1.35
-1.15	-1.55
-0.70	-1.09
-0.53	-0.92
0.04	-0.34
1.24	0.89
1.05	0.69
1.50	1.15
1.67	1.33





$\Rightarrow$ Minimizes distance between observations and their projection

$\Rightarrow$ Approx $X_{n\times p}$ with a low rank matrix $S < p$ $\|A\|_2^2 = \text{tr}(AA^\top)$:

$$\arg\min_{\mu} \left\{ \|X - \mu\|_2^2 : \text{rank}\,(\mu) \le S \right\}$$

SVD X: $\hat{\mu}^{\text{PCA}} = U_{n\times S}\Lambda^{\frac{1}{2}}_{S\times S}V'_{p\times S}$ $= F_{n\times S}V'_{p\times S}$

$F = U\Lambda^{\frac{1}{2}}$ PC - scores

V principal axes - loadings

Missing values in PCA

⇒ PCA: least squares

$$\arg\min_{\mu} \left\{ \|X_{n\times p} - \mu_{n\times p}\|_2^2 : \text{rank}(\mu) \leq S \right\}$$

⇒ PCA with missing values: weighted least squares

$$\arg\min_{\mu} \left\{ \|W_{n\times p} \odot (X - \mu)\|_2^2 : \text{rank}(\mu) \leq S \right\}$$

with $W_{ij} = 0$ if X_{ij} is missing, $W_{ij} = 1$ otherwise; $\odot$ elementwise multiplication

Many algorithms: weighted alternating least squares[64] ; iterative PCA[65].

See also Jan de Leeuw historical notes and NIPALS for 1 dim [66], [67].

[64] Gabriel, Zamir. 1979. Lower Rank Approximation of Matrices by Least Squares with Any Choize of Weights. Technometrics.

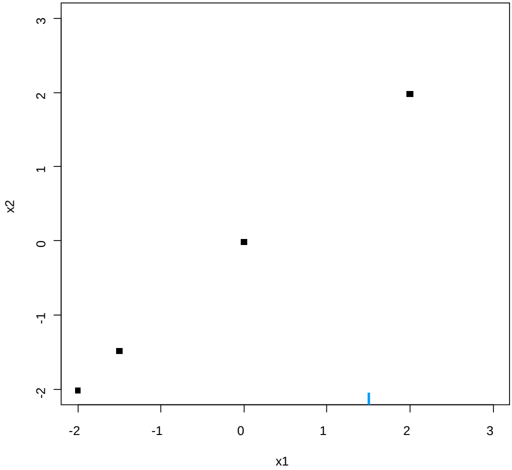
[65] Kiers, 1997. Weighted Least Squares Fitting Using Iterative OLS Algorithms. Psychometrika.

[66] Christofferson. 1969. The one-component model with incomplete data. PhD thesis, Uppsala University, Institute of statistics.

[67] Wold and Lyttkens. 1969. Nonlinear iterative partial least squares (nipals) estimation procedures. Bulletin. Int. Stat.

Iterative PCA

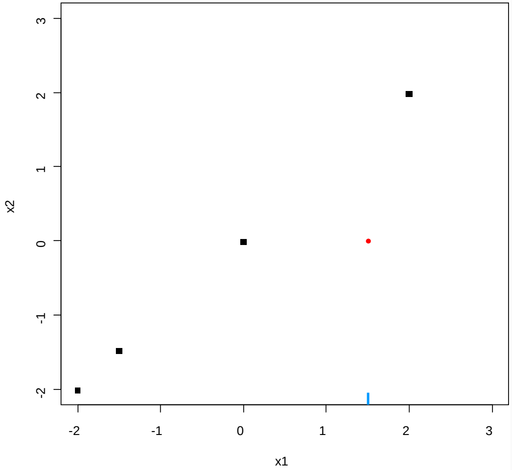
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98



Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	0.00
2.0	1.98



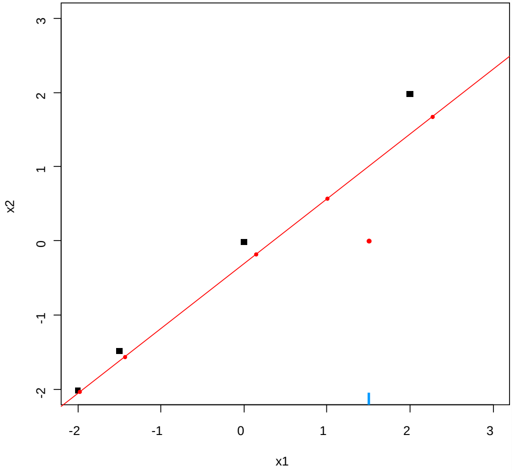
Initialization $\ell = 0$: X^0 (mean imputation)

Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.00**
2.0	1.98

$\widehat{x1}$	$\widehat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	0.57
2.27	1.67



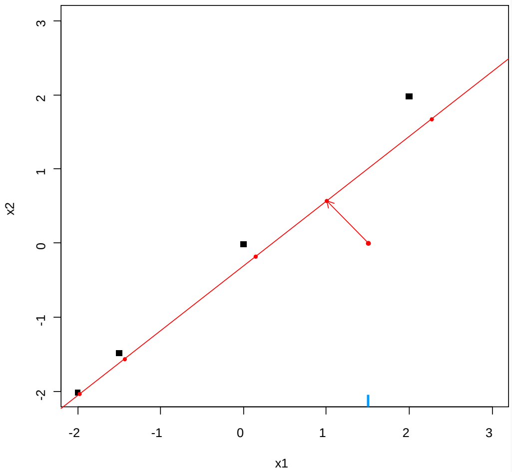
PCA on the completed data set $\rightarrow (U^\ell, \Lambda^\ell, V^\ell)$;

Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.00**
2.0	1.98

$\widehat{x1}$	$\widehat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	**0.57**
2.27	1.67



Missing values imputed with the fitted matrix $\hat{\mu}^\ell = U^\ell \Lambda^{1/2^\ell} V^{\ell\prime}$

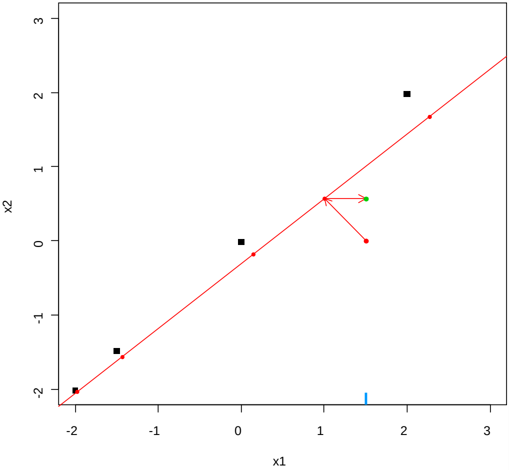
Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**NA**
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.00**
2.0	1.98

$\widehat{x1}$	$\widehat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	**0.57**
2.27	1.67

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



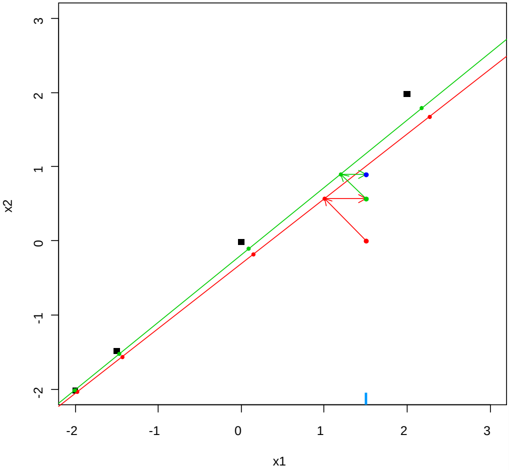
The new imputed dataset is $\hat{X}^{\ell} = W \odot X + (\mathbf{1} - W) \odot \hat{\mu}^{\ell}$

Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**NA**
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



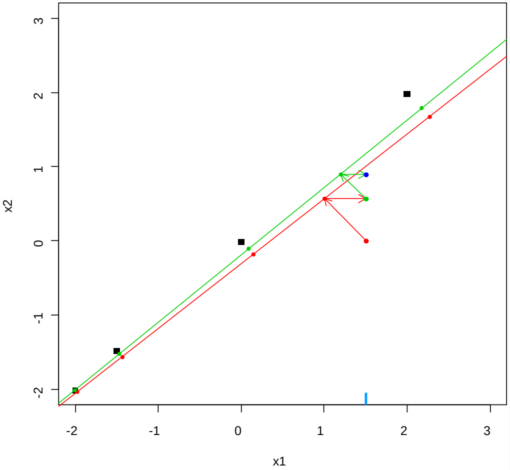
Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**NA**
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98

$\widehat{x1}$	$\widehat{x2}$
-2.00	-2.01
-1.47	-1.52
0.09	-0.11
1.20	**0.90**
2.18	1.78

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.90**
2.0	1.98



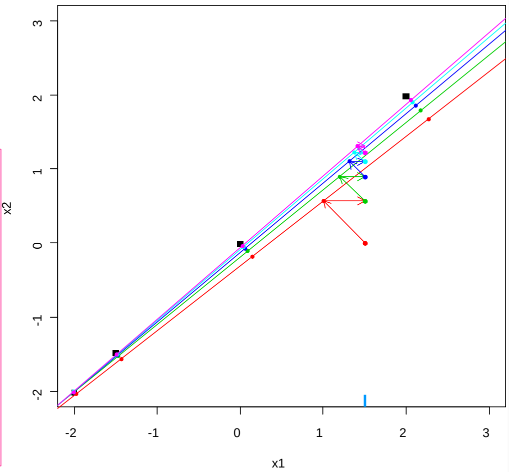
Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**NA**
2.0	1.98

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.00**
2.0	1.98

$\widehat{x1}$	$\widehat{x2}$
-1.98	-2.04
-1.44	-1.56
0.15	-0.18
1.00	0.57
2.27	1.67

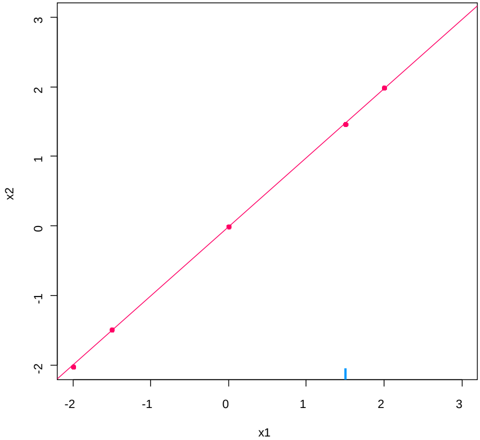
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



Steps are repeated until convergence

Iterative PCA

x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**NA**
2.0	1.98



x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**1.46**
2.0	1.98

PCA on the completed data set $\rightarrow (U^{\ell}, \Lambda^{\ell}, V^{\ell})$

Missing values imputed with the fitted matrix $\hat{\mu}^{\ell} = U^{\ell} {\Lambda^{1/2}}^{\ell} V^{\ell\prime}$

Iterative PCA

Iterative PCA/SVD algorithm

1. initialization $\ell = 0$: X^0 (mean imputation)
2. step ℓ:
   (a) PCA on the completed data $\rightarrow (U^\ell, \Lambda^\ell, V^\ell)$; S dim kept
   (b) missing values are imputed with $(\hat{\mu}^S)^\ell = U^\ell {\Lambda^{1/2}}^\ell V^{\ell\prime}$
   the new imputed data is $\hat{X}^\ell = W \odot X + (\mathbf{1} - W) \odot (\hat{\mu}^S)^\ell$
3. steps of estimation and imputation are repeated [68]

[68]In practice the means and variances are updated at each step to (re)center & (re)scale the data.
[69]**J.** & Husson, 2012. Selecting the number of components in PCA using cross-validation approximations. *CSDA*.

Iterative PCA

Iterative PCA/SVD algorithm

1. initialization $\ell = 0$: X^0 (mean imputation)
2. step ℓ:
   (a) PCA on the completed data $\rightarrow (U^\ell, \Lambda^\ell, V^\ell)$; S dim kept
   (b) missing values are imputed with $(\hat{\mu}^S)^\ell = U^\ell \Lambda^{1/2^\ell} V^{\ell\prime}$
   the new imputed data is $\hat{X}^\ell = W \odot X + (\mathbf{1} - W) \odot (\hat{\mu}^S)^\ell$
3. steps of estimation and imputation are repeated [68]

$\Rightarrow$ $\hat{\mu}$ from incomplete data: EM algo $X = \mu + \varepsilon$, $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$
with μ of low rank , $x_{ij} = \sum_{s=1}^{S} \sqrt{\tilde{\lambda}_s} \tilde{u}_{is} \tilde{v}_{js} + \varepsilon_{ij}$

$\Rightarrow$ Completed data: good imputation (matrix completion, Netflix)

[68] In practice the means and variances are updated at each step to (re)center & (re)scale the data.

[69] **J.** & Husson, 2012. Selecting the number of components in PCA using cross-validation approximations. *CSDA*.

Iterative PCA

Iterative PCA/SVD algorithm

1. initialization $\ell = 0$: X^0 (mean imputation)
2. step ℓ:
   (a) PCA on the completed data $\rightarrow (U^\ell, \Lambda^\ell, V^\ell)$; S dim kept
   (b) missing values are imputed with $(\hat{\mu}^S)^\ell = U^\ell \Lambda^{1/2^\ell} V^{\ell\prime}$
   the new imputed data is $\hat{X}^\ell = W \odot X + (\mathbf{1} - W) \odot (\hat{\mu}^S)^\ell$
3. steps of estimation and imputation are repeated [68]

$\Rightarrow$ $\hat{\mu}$ from incomplete data: EM algo $X = \mu + \varepsilon$, $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$
with μ of low rank , $x_{ij} = \sum_{s=1}^{S} \sqrt{\tilde{\lambda}_s} \tilde{u}_{is} \tilde{v}_{js} + \varepsilon_{ij}$

$\Rightarrow$ Completed data: good imputation (matrix completion, Netflix)

Reduction of variability (imputation by $U\Lambda^{1/2}V'$)

Selecting S (solution are not nested)? Generalized cross-validation[69]

---

[68] In practice the means and variances are updated at each step to (re)center & (re)scale the data.

[69] **J.** & Husson, 2012. Selecting the number of components in PCA using cross-validation approximations. *CSDA*.

Overfitting

Overfitting when:

- many parameters ($U_{n\times S}$, $V_{S\times p}$)/ the number of observed values: S large, many NA
- data are very noisy

$\Rightarrow$ "Trust too much the relationship between variables"

Remarks:

- missing values: special case of small data set
- iterative PCA: prediction method

Solution:
$\Rightarrow$ Regularization

Soft thresholding iterative SVD

$\Rightarrow$ Init - estimation - imputation steps:

The imputation step

$$\hat{\mu}_{ij}^{\text{PCA}} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js}$$

is replaced by [70]

$$\hat{\mu}_{ij}^{\text{Soft}} = \sum_{s=1}^{p} \left(\sqrt{\lambda_s} - \lambda\right)_{+} u_{is} v_{js}$$

$$X = \mu + \varepsilon \qquad \underset{\mu}{\arg\min} \left\{ \| W \odot (X - \mu) \|_2^2 + \lambda \| \mu \|_\star \right\},$$

with $\|\mu\|_\star$, the nuclear norm, *i.e.* the sum of its singular values.

Implemented in `softImpute`

[70] T. Hastie, R. Mazumber, 2015, Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. *JMLR*.

Regularized iterative PCA

The imputation step

$$\hat{\mu}_{ij}^{\text{PCA}} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js}$$

is replaced by [71],[72], [73] :

$$\hat{\mu}_{ij}^{\text{rPCA}} = \sum_{s=1}^{S} \left(\frac{\lambda_s - \hat{\sigma}^2}{\lambda_s} \right) \sqrt{\lambda_s} u_{is} v_{js} = \sum_{s=1}^{S} \left(\sqrt{\lambda_s} - \frac{\hat{\sigma}^2}{\sqrt{\lambda_s}} \right) u_{is} v_{js}$$

σ^2 small $\rightarrow$ regularized iterative PCA $\approx$ iterative PCA

σ^2 large $\rightarrow$ mean imputation

$$\hat{\sigma}^2 = \frac{RSS}{\text{df}} = \frac{n \sum_{s=S+1}^{p} \lambda_s}{np - p - nS - pS + S^2 + S} \qquad (X_{n \times p}; U_{n \times S}; V_{p \times S})$$

Implemented in `missMDA` (Youtube link)

[71] **J.**, Husson. 2012. Handling missing values in exploratory multivariate data analysis. *JSFDS*.

[72] Verbank, **J.**, Husson. 2013. Regularised PCA to denoise and visualise data *Stat & Computing*.

[73] Rationale: L2+L0 penalty, empirical bayes Efron Moris, 1979, PPCA

Properties of SVD based matrix completion

⇒ Powerful methods for matrix completion used in recommandation systems (ex Netflix prize: 99% missing)
⇒ Very good quality of imputation. Using similarities between observations and relationship between variables + reduction of dim

Model makes sense [74]: Data = structure of rank S + noise

⇒ Different noise regime [75], [76]

- ▷ low noise: iterative PCA (tuning S: CV - GCV)
- ▷ moderate: iterative regularized PCA (tuning S: CV - GCV, σ)
- ▷ high noise (SNR low, S large): soft thresholding (tuning λ: CV, σ)
  Implemented in `denoiseR`[77]

Imputed data should be analysed with caution by other methods

---

[74] Udell & Townsend. 2019. Why Are Big Data Matrices Approximately Low Rank? SIAM.
[75] **J.** & Sardy. 2015. Adaptive Shrinkage of singular values. *Stat & Computing*.
[76] **J.**& Wager. 2016. Stable Autoencoding: A Flexible Framework for Regularized Low-Rank Matrix Estimation. *JMLR*.
[77] **J.** Wager, Sardy. 2016: denoiseR: A Package for Low Rank Matrix Estimation.

Multiple imputation with Bootstrap PCA[80]

$$x_{ij} = \mu_{ij} + \varepsilon_{ij} = \sum_{s=1}^{S} \sqrt{\tilde{\lambda}_s} \tilde{u}_{is} \tilde{v}_{js} + \varepsilon_{ij} \ , \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

1. Variability of the parameters, M plausible: $(\hat{\mu}_{ij})^1, ..., (\hat{\mu}_{ij})^M$ [78]
2. Noise: for $m = 1, ..., M$, missing values x_{ij}^m drawn $\mathcal{N}(\hat{\mu}_{ij}^m, \hat{\sigma}^2)$

Implemented in missMDA (website)

François Husson

Revival with synthetic data generation! Avatar[79]: good performances in comparison to synthpop/CT-GAN, etc.

[78] Parametric bootstrap is used: noise resampled. Non parametric bootstrap implies different observations for each imputed data set. A trick consists in using tiny weights and not zero weights.

[79] Guillaudeux et al. (2023). Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *NPJ Digit Med.*

[80] **J.**, Pages. Husson. 2011. Multiple imputation in principal component analysis. *ADAC.*

Comparison of MICE, joint imputation and PCA imputation

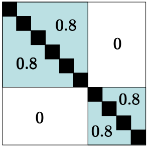
$\Rightarrow$ Good estimates of the parameters and their variance from an incomplete data (coverage close to 0.95)
The variability due to missing values is well taken into account

`Amelia` & `mice` can have difficulties with strong correlations or $n < p$
`missMDA` does not but requires a tuning parameter: number of dim.

`Amelia` & `missMDA` are based on linear relationships
`mice` is more flexible (one model per variable)

MI based on PCA works in a large range of configuration, $n < p$, $n > p$ strong or weak relationships, low or high percentage of missing values

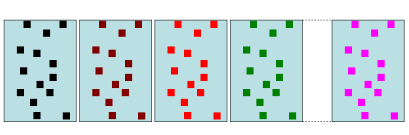
Simulations



The simulated data $\mathcal{N}(\mu, \Sigma)$

- ▷ vary number of obs. n, variables p, correlation ρ
- ▷ vary %NA, missing values mechanism (MCAR, MAR)

$\Rightarrow$ Multiple imputation $M = 100$ imputed tables with PCA, Joint Model, Conditional Model



$\Rightarrow$ Analysis model: estimate $\theta_1 = \mathbb{E}[Y], \theta_2 = \beta_1$ (regression coefficient)

$\Rightarrow$ Combine with Rubin's rule: $\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}_m$

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

Assess Bias, CI width & coverage - 1000 simulations

Questionnaire data[81]

```
region                      sex        age          year         edu          drunk        alcohol      glasses
Ile de France       :8120   F:29776   18_25: 6920   2005:27907   E1:12684   0    :44237   <1/m :12889   0  : 2812
Rhone Alpes         :5421   M:23165   26_34: 9401   2010:25034   E2:23521   1-2  : 4952   0    : 6133   0-2:37867
Provence Alpes      :4116             35_44:10899                E3:6563    10-19:  839   1-2/m: 7583   10+:  590
Nord Pas de Calais  :3819             45_54: 9505                E4:10100   20-29:  212   1-2/w: 9526   3-4: 9401
Pays de Loire       :3152             55_64: 9503                NA:73      3-5  : 1908   3-4/w: 6815   5-6: 1795
Bretagne            :3038             65_+ : 6713                           30+  :  404   5-6/w: 3402   7-9:  391
(Other)             :25275                                                  6-9  :  389   7/w  : 6593         NA: 85

binge                  Pbsleep                 Tabac
<2/m:10323             Never:20605             Frequent  : 9176
0   :34345             Often: 10172            Never     :39080
1/m : 6018             Rare :22134             Occasional: 4588
1/w : 1800             NA:  30                 NA:  97
7/w :  374
NA  :   81
```

- 'true' missing values: mask an underlying category among the available categories.
- not a missing values when it is a new category (keep a category NA).

Principal components method to explore categorical data: Multiple Correpondence Analysis[82]

---

[81] http://www.inpes.sante.fr

[82] M. Greenacre's books, MCA and related methods. 2006. Chapman and Hall/CRC.

Multiple Correspondence Analysis (MCA)

$X_{n\times m}$ m categorical variables coded with dummies in $A_{n\times C_j}$, with C_j the tot number of categories. For a category c, its frequency: $p_c = n_c/n$.

$$X = \left|\begin{array}{c|c|c} y & \dots & \textit{attack} \\ y & \dots & \textit{attack} \\ y & \dots & \textit{attack} \\ n & \dots & \textit{suicide} \\ & & \\ & & \\ n & \dots & \textit{accident} \\ n & \dots & \textit{suicide} \end{array}\right| \quad A = \left|\begin{array}{cc|c|ccc} 1 & 0 & \dots & 1 & 0 & 0 \\ 1 & 0 & \dots & 1 & 0 & 0 \\ 1 & 0 & \dots & 1 & 0 & 0 \\ 0 & 1 & \dots & 0 & 1 & 0 \\ & & & & & \\ & & & & & \\ 0 & 1 & \dots & 0 & 0 & 1 \\ 0 & 1 & \dots & 0 & 1 & 0 \end{array}\right| \quad D_p = \begin{bmatrix} p_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & p_J \end{bmatrix}$$

MCA: A SVD on weighted matrix: $Z = \frac{1}{\sqrt{mn}}(A - 1p^T)D_p^{-1/2} = U\Lambda V'$

The principal component ($F = U\Lambda^{1/2}$) satisfies:

$$\underset{F\in\mathbb{R}^n}{\arg\max} \quad \frac{1}{m}\sum_{j=1}^{m}\eta^2(F, X_j)$$

$$\eta^2(F, X_j) = \frac{\sum_{c=1}^{C_j} n_c(\bar{F}_c - \bar{F})^2}{\sum_{i=1}^{n}\sum_{c=1}^{C_j}(F_{ic} - \bar{F})^2} = \frac{\text{Between variance}}{\text{Total variance}}$$

Benzecri, 1973 :"*In data analysis the mathematical problems reduces to computing eigenvectors; all the science (the art) is in finding the right matrix to diagonalize*"

Regularized iterative MCA[83]

Iterative MCA algorithm:

	V1	V2	V3	...	V14
ind 1	a	NA	g	...	u
ind 2	NA	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	NA		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	NA	NA	1	0	...
ind 2	NA	NA	NA	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	NA	NA	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

`library(missMDA); ?imputeMCA`

[83] J. et al. 2012. Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis. *Journal of classification.*

Regularized iterative MCA[83]

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)

	V1	V2	V3	...	V14
ind 1	a	NA	g	...	u
ind 2	NA	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	NA		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	0.41	0.59	1	0	...
ind 2	0.20	0.30	0.50	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	0.27	0.78	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

```
library(missMDA); ?imputeMCA
```

[83] J. et al. 2012. Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis. *Journal of classification.*

Regularized iterative MCA[83]

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) estimation: MCA on the completed data $\rightarrow U, \Lambda, V$

	V1	V2	V3	...	V14
ind 1	a	NA	g	...	u
ind 2	NA	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	NA		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	0.41	0.59	1	0	...
ind 2	0.20	0.30	0.50	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	0.27	0.78	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

`library(missMDA); ?imputeMCA`

[83] J. et al. 2012. Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis. *Journal of classification.*

Regularized iterative MCA[83]

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) estimation: MCA on the completed data $\rightarrow U, \Lambda, V$
   (b) imputation with the fitted matrix $\hat{\mu} = U_S \Lambda_S^{1/2} V_S'$

	V1	V2	V3	...	V14
ind 1	a	NA	g	...	u
ind 2	NA	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	NA		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	0.65	0.35	1	0	...
ind 2	0.11	0.20	0.69	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	0.30	0.40	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

```
library(missMDA); ?imputeMCA
```

[83] J. et al. 2012. Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis. *Journal of classification.*

Regularized iterative MCA[83]

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) estimation: MCA on the completed data $\rightarrow U, \Lambda, V$
   (b) imputation with the fitted matrix $\hat{\mu} = U_S \Lambda_S^{1/2} V_S'$
   (c) column margins are updated

	V1	V2	V3	...	V14
ind 1	a	**NA**	g	...	u
ind 2	**NA**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**NA**		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	0.65	0.35	1	0	...
ind 2	0.11	0.20	0.69	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	0.30	0.40	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

`library(missMDA); ?imputeMCA`

[83] J. et al. 2012. Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis. *Journal of classification.*

Regularized iterative MCA[83]

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) estimation: MCA on the completed data $\rightarrow U, \Lambda, V$
   (b) imputation with the fitted matrix $\hat{\mu} = U_S \Lambda_S^{1/2} V_S'$
   (c) column margins are updated

	V1	V2	V3	...	V14
ind 1	a	**NA**	g	...	u
ind 2	**NA**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**NA**		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	**0.71**	**0.29**	1	0	...
ind 2	**0.12**	**0.29**	**0.59**	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	**0.37**	**0.63**	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

$\Rightarrow$ the imputed values can be seen as degree of membership

```
library(missMDA); ?imputeMCA
```

[83] J. et al. 2012. Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis. *Journal of classification.*

Regularized iterative MCA[83]

Iterative MCA algorithm:

1. initialization: imputation of the indicator matrix (proportion)
2. iterate until convergence
   (a) estimation: MCA on the completed data $\rightarrow U, \Lambda, V$
   (b) imputation with the fitted matrix $\hat{\mu} = U_S \Lambda_S^{1/2} V_S'$
   (c) column margins are updated

	V1	V2	V3	...	V14
ind 1	a	**e**	g	...	u
ind 2	**c**	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	**g**		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	**0.71**	**0.29**	1	0	...
ind 2	**0.12**	**0.29**	**0.59**	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	**0.37**	**0.63**	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

Two ways to obtain categories: majority or draw

```
library(missMDA); ?imputeMCA
```

[83] J. et al. 2012. Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis. *Journal of classification.*

Multiple imputation with MCA[84]

1. Variability of the parameters: M sets $(U_{n\times S}, \Lambda_{S\times S}, V^{\top}_{m\times S})$ using a non-parametric bootstrap

$\hat{X}_1$

1	0	...	1	0	0
1	0	...	1	0	0
1	0	...	0.01	0.80	0.19
			0	0	1
0.25	0.75				
0	1		0	0	1

$\hat{X}_2$

1	0	...	1	0	0
1	0	...	1	0	0
1	0	...	0.60	0.2	0.20
			0	0	1
0.26	0.74				
0	1		0	0	1

...

$\hat{X}_M$

1	0	...	1	0
1	0	...	1	0
1	0	...	0.11	0.74
			0	0
0.20	0.80			
0	1		0	0

2. Categories drawn from multinomial distribution using the values in $\left(\hat{X}_m\right)_{1\leq m\leq M}$

y	...	Attack
y	...	Attack
y	...	Suicide
	...	Accident
n		
n	...	S

y	...	Attack
y	...	Attack
y	...	Attack
	...	Accident
n		
n	...	B

...

y	...	Attack
y	...	Attack
y	...	Suicide
	...	Accident
n		
n	...	Suicide

```
library(missMDA); MIMCA()
```

[84] Audigier, Husson, **J.** MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis. (2017). *Statistics & Computing.*

Multiple imputation for categorical data

Joint modeling

- ▷ Log-linear model (Schafer, 1997) (cat): pb many levels
- ▷ Latent class models (Vermunt, 2014) - nonparametric Bayesian (Si & Reiter, 2014, Murray & Reiter, 2016) (MixedDataImpute, NPBayesImpute, NestedCategBayesImpute)

Conditional modeling

- ▷ logistic, multinomial logit, forests (mice)

⇒ MIMCA provides valid inference (ex. logistic reg with missing) applied to data of various size (many levels, rare levels)

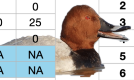
Time (seconds)	Titanic	Galetas	Income
rows-variables-levels	(2000 - 4 - 4)	(1000 - 4 -11)	(6000 - 14 - 9)
MIMCA	2.750	8.972	**58.729**
Loglinear	0.740	4.597	NA
Nonparametric bayes	10.854	17.414	143.652
Cond logistic	4.781	38.016	881.188
Cond forests	265.771	112.987	6329.514

Low-rank matrix completion for count data

- 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
Monitor the population and assess wetlands conservation policies.
$\Rightarrow$ 70% of missing values in contingency tables (drough, war, etc.) [85,86]

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

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

Missing values in multi-source heterogeneous 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	

- ▷ Mixed data (categorical & continuous): Imputation with Factorial Analysis for Mixed Data (FAMD)[87]. Good for rare categories.
- ▷ Multi-level data (groups of observations): imputation with random effects [88] - imputation with Multilevel SVD [89]. Close to meta-learning.
- ▷ Multi-block/modalities data: imputation with Multiple Factor Analysis[90]

[87] Audigier, Husson, **J.** (2016). A principal components method to impute mixed data. *ADAC.*
[88] Audigier et al. (2018). MI for multilevel data with continuous & binary variables. Stat. Science.
[89] Husson, **J.**, Narasimhan & Robin. (2019). Imputation of Mixed Data With Multilevel SVD.
[90] Husson, **J.** (2013). Handling missing values in MFA. *FQP.*

Low rank matrix completion for heterogeneous data

Works of Madeleine Udell:

- Mike et al. (2023). The Missing Indicator Method: From Low to High Dimensions. *SIGKDD Conference*.
- Zhao et al. (2022). Probabilistic Missing Value Imputation for Mixed Categorical and Ordered Data. *NeurIPS*.
- Zhao and Udell. (2020). Matrix Completion with Quantified Uncertainty through Low Rank Gaussian Copula. *NeurIPS*.
- Kallus et al. (2018). Causal Inference with Noisy and Missing Covariates via Matrix Factorization. *NeurIPS*.
- Software: gcimpute: imputation with the Gaussian copula - LowRankModels: low rank models for missing value imputation.

Take home message: estimation/imputation with low rank methods

- Principal component methods powerful for single & multiple imputation of quanti & categorical data (rare categories): dimensionality reduction & capture similarities between obs and variables.

  $\Rightarrow$ Correct inferences for analysis model based on relationships between pairs of variables

  $\Rightarrow$ Requires to choose the number of dimensions S

- SVD can be distributed/federated learning
- Handling missing values in PCA (quantitative), MCA (categorical), FAMD (mixed), MFA (groups/blocks), Correspondence analysis for contingency tables
- Preprocessing before clustering - clustering with missing values

Ressources implementation

Package missMDA:
http://factominer.free.fr/missMDA/index.html

Youtube: https://www.youtube.com/watch?v=OOM8_FH6_8o&list=PLnZgp6epRBbQzxFnQrcxg09kRt-PA66T_playlist

Article JSS: https://www.jstatsoft.org/article/view/v070i01

MOOC Exploratory Multivariate Data Analysis

Practice

Incomplete ozone data[91]

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
0601	87	15.6	18.5	18.4	4	4	8	NA	-1.7101	-0.6946	84
0602	82	NA	18.4	17.7	5	5	7	NA	NA	NA	87
0603	92	NA	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82
0604	114	16.2	NA	NA	1	1	0	NA	NA	NA	92
0605	94	17.4	20.5	NA	8	8	7	-0.5	NA	-4.3301	114
0606	80	17.7	NA	18.3	NA	NA	NA	-5.6382	-5	-6	94
0607	NA	16.8	15.6	14.9	7	8	8	-4.3301	-1.8794	-3.7588	80
0610	79	14.9	17.5	18.9	5	5	4	0	-1.0419	-1.3892	NA
0611	101	NA	19.6	21.4	2	4	4	-0.766	NA	-2.2981	79
0612	NA	18.3	21.9	22.9	5	6	8	1.2856	-2.2981	-3.9392	101
0613	101	17.3	19.3	20.2	NA	NA	NA	-1.5	-1.5	-0.8682	NA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
0919	NA	14.8	16.3	15.9	7	7	7	-4.3301	-6.0622	-5.1962	42
0920	71	15.5	18	17.4	7	7	6	-3.9392	-3.0642	0	NA
0921	96	NA	NA	NA	3	3	3	NA	NA	NA	71
0922	98	NA	NA	NA	2	2	2	4	5	4.3301	96
0923	92	14.7	17.6	18.2	1	4	6	5.1962	5.1423	3.5	98
0924	NA	13.3	17.7	17.7	NA	NA	NA	-0.9397	-0.766	-0.5	92
0925	84	13.3	17.7	17.8	3	5	6	0	-1	-1.2856	NA
0927	NA	16.2	20.8	22.1	6	5	5	-0.6946	-2	-1.3681	71
0928	99	16.9	23	22.6	NA	4	7	1.5	0.8682	0.8682	NA
0929	NA	16.9	19.8	22.1	6	5	3	-4	-3.7588	-4	99
0930	70	15.7	18.6	20.7	NA	NA	NA	0	-1.0419	-4	NA

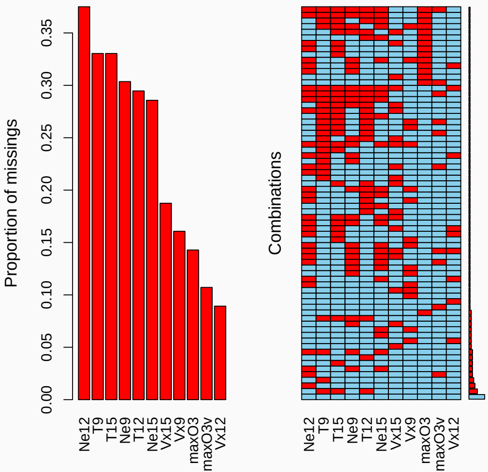
[91] Code and data availabkle on Rmistastic

Completed ozone data

```
         maxO3     T9    T12    T15   Ne9  Ne12  Ne15    Vx9   Vx12   Vx15  maxO3v
20010601  87.000 15.600 18.500 20.471 4.000 4.000 8.000  0.695 -1.710 -0.695  84.000
20010602  82.000 18.505 20.870 21.799 5.000 5.000 7.000 -4.330 -4.000 -3.000  87.000
20010603  92.000 15.300 17.600 19.500 2.000 3.984 3.812  2.954  1.951  0.521  82.000
20010604 114.000 16.200 19.700 24.693 1.000 1.000 0.000  2.044  0.347 -0.174  92.000
20010605  94.000 18.968 20.500 20.400 5.294 5.272 5.056 -0.500 -2.954 -4.330 114.000
20010606  80.000 17.700 19.800 18.300 6.000 7.020 7.000 -5.638 -5.000 -6.000  94.000
20010607  79.000 16.800 15.600 14.900 7.000 8.000 6.556 -4.330 -1.879 -3.759  80.000
20010610  79.000 14.900 17.500 18.900 5.000 5.000 5.016  0.000 -1.042 -1.389  99.000
20010611 101.000 16.100 19.600 21.400 2.000 4.691 4.000 -0.766 -1.026 -2.298  79.000
20010612 106.000 18.300 22.494 22.900 5.000 4.627 4.495  1.286 -2.298 -3.939 101.000
20010613 101.000 17.300 19.300 20.200 7.000 7.000 3.000 -1.500 -1.500 -0.868 106.000
.....

20010915  69.000 17.100 17.700 17.500 6.000 7.000 8.000 -5.196 -2.736 -1.042  71.000
20010916  71.000 15.400 18.091 16.600 4.000 5.000 5.000 -3.830  0.000  1.389  69.000
20010917  60.000 15.283 18.565 19.556 4.000 5.000 4.000  0.000  3.214  0.000  71.000
20010918  42.000 14.091 14.300 14.900 8.000 7.000 7.000 -2.500 -3.214 -2.500  60.000
20010919  65.000 14.800 16.425 15.900 7.000 7.982 7.000 -4.341 -6.062 -5.196  42.000
20010920  71.000 15.500 18.000 17.400 7.000 7.000 6.000 -3.939 -3.064  0.000  65.000
20010924  76.000 13.300 17.700 17.700 5.631 5.883 5.453 -0.940 -0.766 -0.500  65.139
20010925  75.573 13.300 18.434 17.800 3.000 5.000 5.001  0.000 -1.000 -1.286  76.000
20010927  77.000 16.200 20.800 20.499 5.368 5.495 5.177 -0.695 -2.000 -1.473  71.000
20010928  99.000 18.074 22.169 23.651 3.531 3.610 3.561  1.500  0.868  0.868  93.135
20010929  83.000 19.855 22.663 23.847 5.374 5.000 3.000 -4.000 -3.759 -4.000  99.000
20010930  70.000 15.700 18.600 20.700 7.000 6.405 7.000 -2.584 -1.042 -4.000  83.000
```

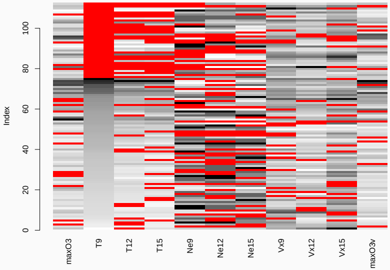
```
> library(missMDA)
> res.comp <- imputePCA(ozo[, 1:11])
> res.comp$comp
```

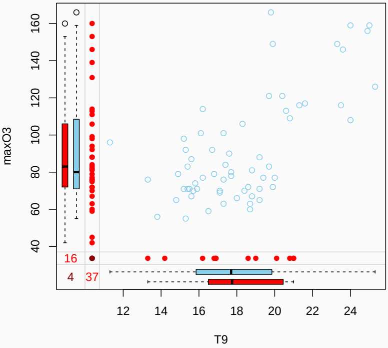
Visualization of the pattern of missing values




```
> library(VIM)
> aggr(don, sortVar = TRUE)
```

VVisualization of the pattern of missing values





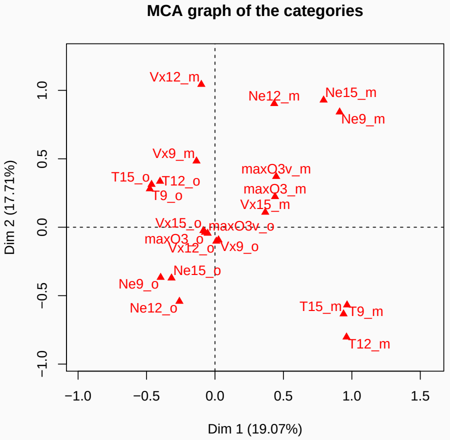

```
> library(VIM)
> matrixplot(don, sortby = 2)
> marginplot(don[ ,c("T9", "maxO3")])
```

Visualization with Multiple Correspondence Analysis

⇒ Create the missingness matrix

```
> mis.ind <- matrix("o", nrow = nrow(don), ncol = ncol(don))
> mis.ind[is.na(don)] = "m"
> dimnames(mis.ind) = dimnames(don)
> mis.ind

         maxO3 T9  T12 T15 Ne9 Ne12 Ne15 Vx9 Vx12 Vx15 maxO3v
20010601 "o"   "o" "o" "m" "o" "o"  "o"  "o" "o"  "o"  "o"
20010602 "o"   "m" "m" "m" "o" "o"  "o"  "o" "o"  "o"  "o"
20010603 "o"   "o" "o" "o" "o" "m"  "m"  "o" "m"  "o"  "o"
20010604 "o"   "o" "o" "m" "o" "o"  "o"  "m" "o"  "o"  "o"
20010605 "o"   "m" "o" "o" "m" "m"  "m"  "o" "o"  "o"  "o"
20010606 "o"   "o" "o" "o" "o" "m"  "o"  "o" "o"  "o"  "o"
20010607 "o"   "o" "o" "o" "o" "o"  "m"  "o" "o"  "o"  "o"
20010610 "o"   "o" "o" "o" "o" "o"  "m"  "o" "o"  "o"  "o"
```

Visualization with Multiple Correspondence Analysis

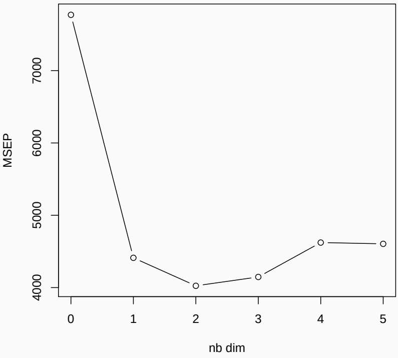



```
> library(FactoMineR)
> resMCA <- MCA(mis.ind)
> plot(resMCA, invis = "ind", title = "MCA graph of the categories")
```

Imputation with PCA in practice

⇒ Step 1: Estimation of the number of dimensions

```
> library(missMDA)
> nb <- estim_ncpPCA(don, method.cv = "Kfold")
> nb$ncp    #2
> plot(0:5, nb$criterion, xlab = "nb dim", ylab ="MSEP")
```

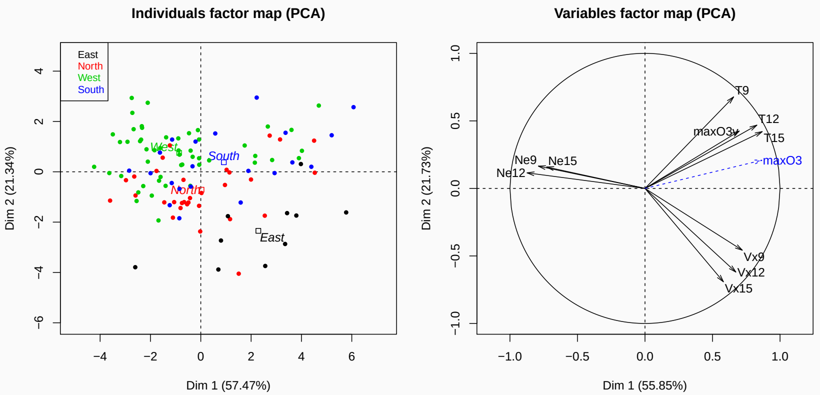


Imputation with PCA in practice

⇒ Step 2: Imputation of the missing values

```
> res.comp <- imputePCA(don, ncp = 2)
> res.comp$completeObs[1:3, ]
     maxO3    T9   T12   T15 Ne9 Ne12 Ne15   Vx9  Vx12  Vx15 maxO3v
0601    87 15.60 18.50 20.47   4 4.00 8.00  0.69 -1.71 -0.69     84
0602    82 18.51 20.88 21.81   5 5.00 7.00 -4.33 -4.00 -3.00     87
0603    92 15.30 17.60 19.50   2 3.98 3.81  2.95  1.97  0.52     82
```

Cherry on the cake: PCA on incomplete data!




```
> imp <- cbind.data.frame(res.comp$completeObs, ozo[, 12])
> res.pca <- PCA(imp, quanti.sup = 1, quali.sup = 12)
> plot(res.pca, hab = 12, lab = "quali"); plot(res.pca, choix = "var")
> res.pca$ind$coord #scores (principal components)
```

Matrix completion for continuous data

```
> library(softImpute)
> fit1 <- softImpute(XNA, rank = , lambda = )
> X.soft <- complete(XNA, fit1)

> library(denoiseR)
> adaNA <- imputeada(XNA,  gamma = 1) ## time consuming...
> X.ada <- adaNA$completeObs
```

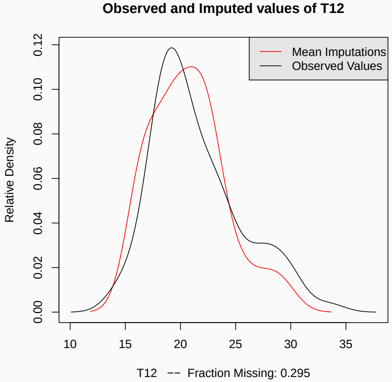
Multiple imputation in practice

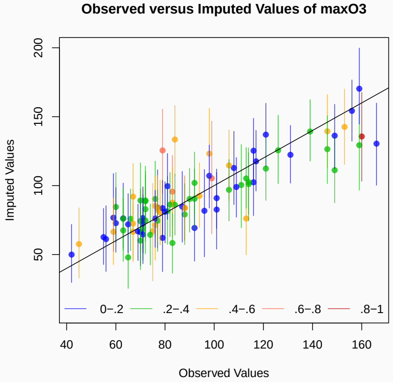
⇒ Step 1: Generate M imputed data sets

```
> library(Amelia)
> res.amelia <- amelia(don, m = 100)

> library(mice)
> res.mice <- mice(don, m = 100, defaultMethod = "norm.boot")

> library(missMDA)
>  res.MIPCA <- MIPCA(don, ncp = 2, nboot  = 100)
> res.MIPCA$res.MI
```

Multiple imputation in practice

⇒ Step 2: visualization



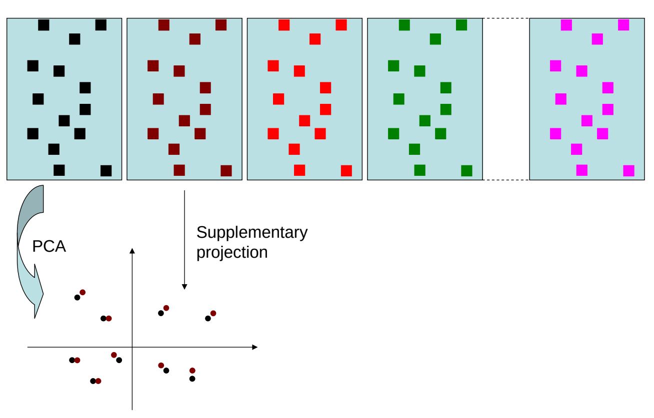



```
>  library(Amelia)
> res.amelia <- amelia(don, m = 100)
> compare.density(res.amelia, var = "T12")
> overimpute(res.amelia, var = "maxO3")

> library(missMDA)
res.over <- Overimpute(res.MIPCA)
```

Multiple imputation in practice

⇒ Step 2: visualization

⇒ Individuals position (and variables) with other predictions

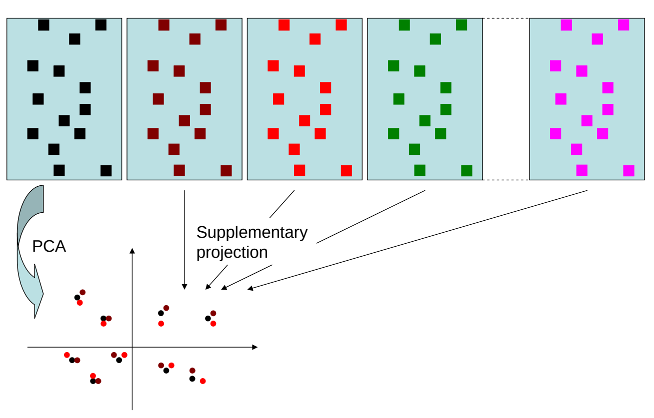


Regularized iterative PCA
⇒ reference configuration

Multiple imputation in practice

⇒ Step 2: visualization

⇒ Individuals position (and variables) with other predictions

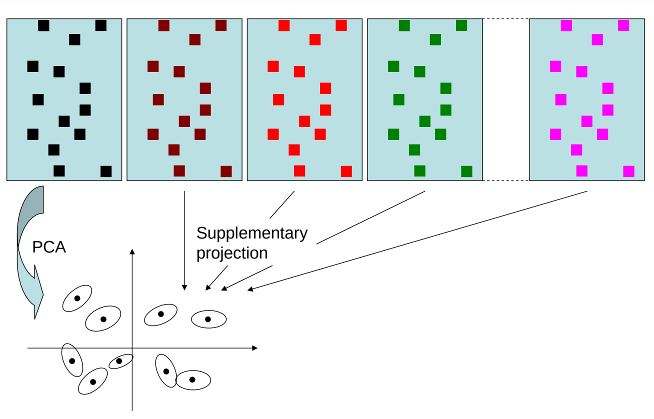


Regularized iterative PCA
⇒ reference configuration

Multiple imputation in practice

⇒ Step 2: visualization

⇒ Individuals position (and variables) with other predictions

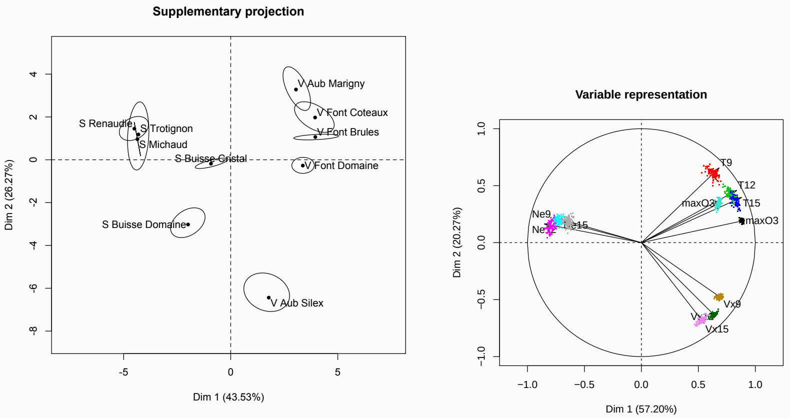


Regularized iterative PCA
⇒ reference configuration

Multiple imputation in practice

⇒ Step 2: visualization

```
> res.MIPCA <- MIPCA(don, ncp = 2)
> plot(res.MIPCA, choice = "ind.supp"); plot(res.MIPCA, choice = "var")
```



⇒ Percentage of NA?

Multiple imputation in practice

⇒ Step 3. Regression on each table and pool the results

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

```
> library(mice)
> res.mice <- mice(don, m = 100)
> imp.micerf <- mice(don, m = 100, defaultMethod = "rf")
> lm.mice.out <- with(res.mice, lm(maxO3 ~ T9+T12+T15+Ne9+...+Vx15+maxO3v))
> pool.mice <- pool(lm.mice.out)
> summary(pool.mice)
```

```
               est    se     t    df Pr(>|t|)  lo 95 hi 95 nmis  fmi lambda
(Intercept) 19.31 16.30  1.18 50.48     0.24 -13.43 52.05   NA 0.46   0.44
T9          -0.88  2.25 -0.39 26.43     0.70  -5.50  3.75   37 0.71   0.69
T12          3.29  2.38  1.38 27.54     0.18  -1.59  8.18   33 0.70   0.68
....
Vx15         0.23  1.33  0.17 39.00     0.87  -2.47  2.93   21 0.57   0.55
maxO3v       0.36  0.10  3.65 46.03     0.00   0.16  0.56   12 0.50   0.48
```

Categorical imputation with MCA in practice

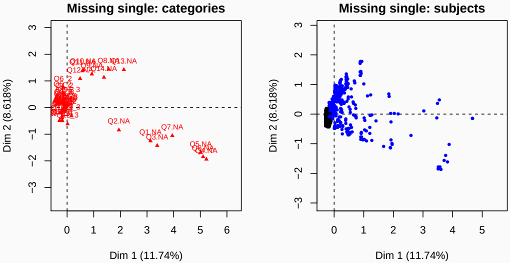
- 1232 respondents, 14 questions, 35 categories, 9% of missing values concerning 42% of respondents

In `missMDA` (Youtube)

```
data(vnf)
summary(vnf)
MCA(vnf)

#1) select the number of components
nb <- estim_ncpMCA(vnf, ncp.max = 5) #Time-consuming, nb = 4

#2) Impute the indicator matrix
res.impute <- imputeMCA(vnf, ncp = 4)
res.impute$tab.disj
res.impute$comp
summary(res.impute$comp)

# MCA on the incomplete data vnf
res.mca <- MCA(vnf, tab.disj = res.impute$tab.disj)
plot(res.mca, invisible=c("var"))
plot(res.mca,invisible=c("ind"),autoLab="yes", selectMod="cos2 5", cex  = 0.6)
```

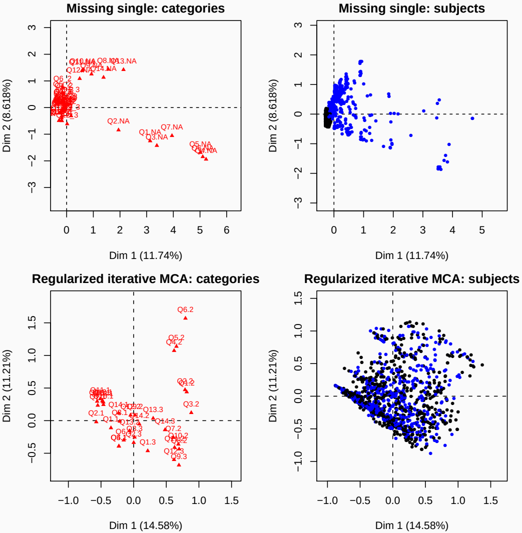
Categorical imputation with MCA in practice

- 1232 respondents, 14 questions, 35 categories, 9% of missing values concerning 42% of respondents



Categorical imputation with MCA in practice

- 1232 respondents, 14 questions, 35 categories, 9% of missing values concerning 42% of respondents



Mixed imputation in practice

```
> library(missMDA)
> res.ncp <- estim_ncpFAMD(ozo)
> res.famd <-imputeFAMD(ozo, ncp = 2)
> res.famd$completeObs

> library(missForest)
> res.rf <- missForest(ozo)
> res.rf$ximp
```

Ex of missing values per group of variables: Journal impact factors

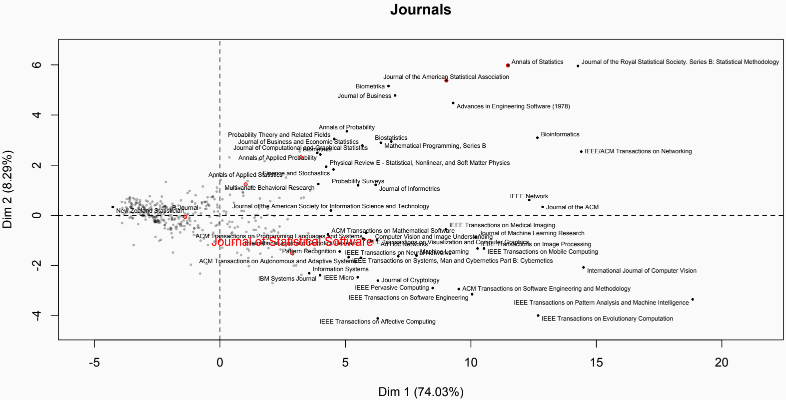
Data from journalmetrics.com

443 journals (Computer Science, Statistics, Probability and Mathematics),
15 years,
3 types of measures:

- IPP - Impact Per Publication: like the ISI impact factor but for 3 (rather than 2) years.
- SNIP - Source Normalized Impact Per Paper: Tries to weight by the number of citations per subject field to adjust for different citation cultures.
- SJR - SCImago Journal Rank: Tries to capture average prestige per publication.

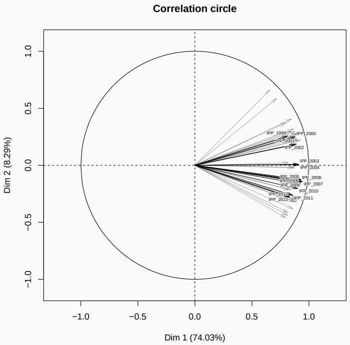
Many missing values per block of years.

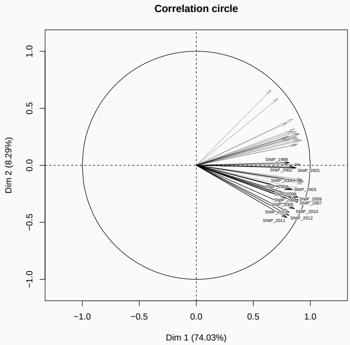
Multiple Factor Analysis (MFA) with missing values [92]



[92] Husson, J. 2013. Handling missing values in Multiple Factor Analysis. *FQP.*

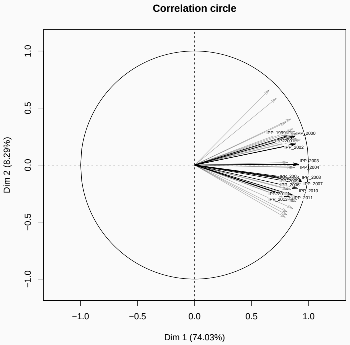
Multiple Factor Analysis (MFA) with missing values [92]

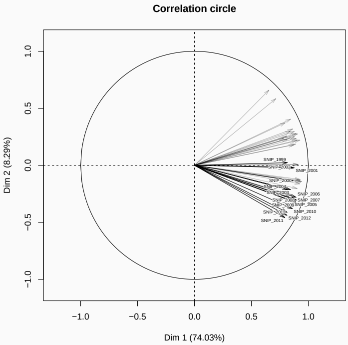




[92] Husson, J. 2013. Handling missing values in Multiple Factor Analysis. *FQP.*

Multiple Factor Analysis (MFA) with missing values [92]

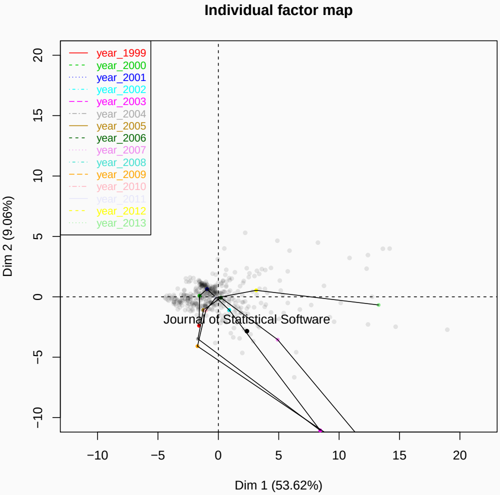




92 Husson, J. 2013. Handling missing values in Multiple Factor Analysis. *FQP.*

Multiple Factor Analysis (MFA) with missing values [92]

ACM Transactions on Networking trajectory



[92] Husson, J. 2013. Handling missing values in Multiple Factor Analysis. *FQP*.

MFA imputation in practice

```
> library(denoiseR)
> library(missMDA)
> data(impactfactor)
> year=NULL; for (i in 1: 15) year= c(year, seq(i,45,15))
> res.imp <- imputeMFA(impactfactor,  group = rep(3, 15),  type = rep("s", 15))

##
> res.mfa  <-MFA(res.imp$completeObs, group=rep(3,15),  type=rep("s",15),
name.group=paste("year", 1999:2013,sep="_"),graph=F)

plot(res.mfa, choix = "ind", select = "contrib 15", habillage = "group", cex = 0.7)
points(res.mfa$ind$coord[c("Journal of Statistical Software",
"Journal of the American Statistical Association", "Annals of Statistics"),
1:2], col=2, cex=0.6)
text(res.mfa$ind$coord[c("Journal of Statistical Software"), 1],
res.mfa$ind$coord[c("Journal of Statistical Software"), 2],cex=1,
labels=c("Journal of Statistical Software"),pos=3, col=2)

plot.MFA(res.mfa,choix="var", cex=0.5,shadow=TRUE, autoLab = "yes")

plot(res.mfa, select="IEEE/ACM Transactions on Networking",
partial="all",
habillage="group",unselect=0.9,chrono=TRUE)
```

Supervised Learning with Missing values

Collaborators on supervised learning with missing values

- M. Le Morvan, Researcher, INRIA, Paris.
- **E. Scornet**, Pr. Sorbonne. Topic: random forests, missing, causal.
- G. Varoquaux, Researcher, INRIA, Paris. Topic: machine learning/ Scikitlearn

⇒ **Random Forests with missing values**
Consistency of supervised learning with missing val. (2019-2024). Stat. papers.

⇒ **Linear regression with missing values - MultiLayer perceptron**
Linear predictor on linearly-generated data with missing values: non consistency and solutions. AISTAT2020.
Neumiss networks: differential programming for supervised learning with missing values. Neurips2020. Oral.

⇒ **Impute then regress**:
What's a good imputation to predict with missing val.? Neurips2021. Spotl.

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

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

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

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

$\Rightarrow$ One model per pattern m of missing values (2^d patterns)[93]

[93] 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)[93]

[93] 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

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

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

A new data $\mathcal{D}_{n,\mathrm{test}}$ 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: train & test sets with missing values. If not: 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: train & test sets with missing values. If not: distributional shift; data generating process (X, Y, M)

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

Imputation prior to learning: Impute then Regress

Common practice: use off-the-shelf methods 1) for imputation of missing values and 2) for supervised-learning on the completed data

- Separate imputat. Impute train & test separately (with a different model)
- Group imputation/ semi-supervised Impute train & test simultaneously but the predictive model is learned only on the training imputed data
- Imputation train & test with the same model. For instance, 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 the same means

Constant (mean) imputation is consistent for prediction[93]

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

[93] **J.** et al. (2019-2024.). Consistency of supervised learning with missing values. *Stat. papers.*

Constant (mean) imputation is consistent for prediction[93]

- 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)

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

Prediction with constant is equal to the Bayes function almost everywhere

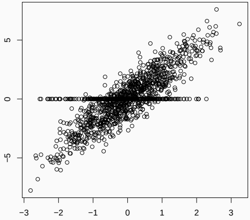
$$f^\star_{impute}(X') = f^\star(\tilde{X}) = \mathbb{E}[Y|\tilde{X}]$$

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

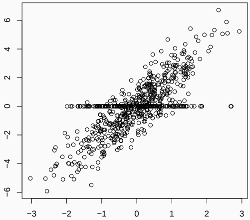
[93] **J.** et al. (2019-2024.). Consistency of supervised learning with missing values. *Stat. papers.*

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

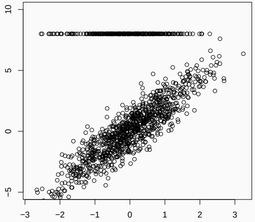


Test

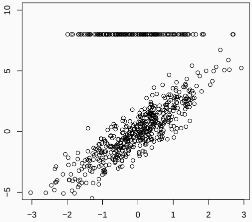
Imputing both train & test with the same constant and regress is consistent despite its drawbacks for estimation (useful in practice)

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



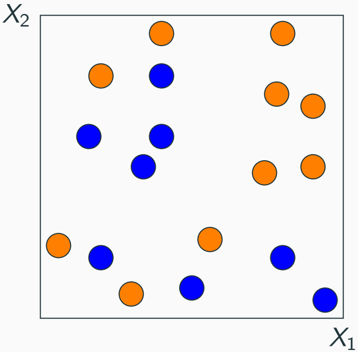
Test

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

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

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

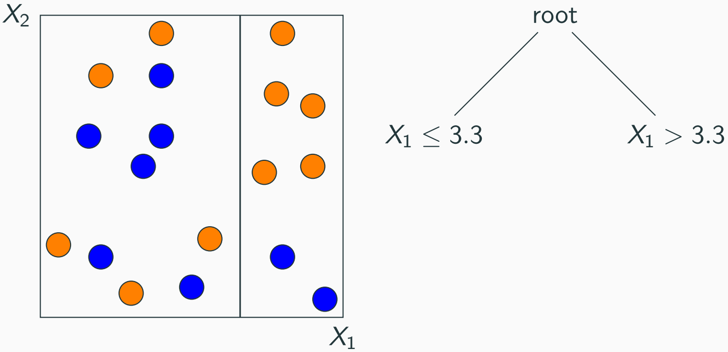


root

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

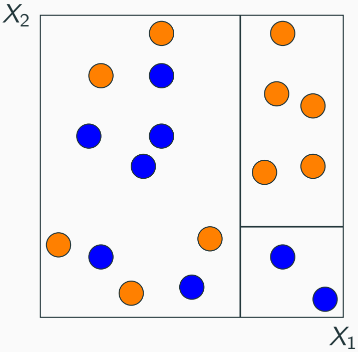
$$(j^\star, z^\star) \in \operatorname*{arg\,min}_{(j,z)\in\mathcal{S}} \mathbb{E}\Big[\big(Y - \mathbb{E}[Y|X_j \le z]\big)^2 \cdot \mathbb{1}_{X_j \le z} + \big(Y - \mathbb{E}[Y|X_j > z]\big)^2 \cdot \mathbb{1}_{X_j > z}\Big].$$

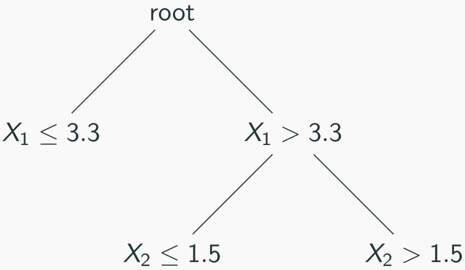


CART (Breiman, 1984)

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

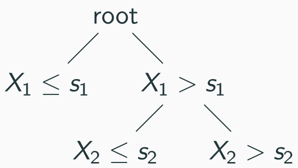
$X_1 \leq s_1$ $\quad$ $X_1 > s_1$

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 (MIA)[95]

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:

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

- ▷ Implem[94] `grf/partykit package`, `scikit HistGradientBoosting`

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

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

[95] Twala et al. (2008). Methods for coping with missing data in decision trees. *Pattern Recog.*

Bayes optimality of impute-n-regress[96]

- **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

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

- **Regression on imputed data:** $g^\star_\Phi \in \underset{g:\mathbb{R}^d \mapsto \mathbb{R}}{\operatorname{argmin}} \ \mathbb{E}\left[\left(Y - g \circ \Phi(\widetilde{X})\right)^2\right]$, minimizer of the risk on the imputed data

Theorem

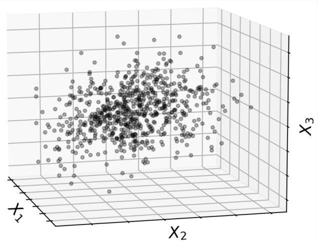
Assume that the response Y satisfies $Y = f^\star(X) + \varepsilon$
Then, for all missing data mechanisms & almost all imputation functions, $g^\star_\Phi \circ \Phi$ is Bayes optimal

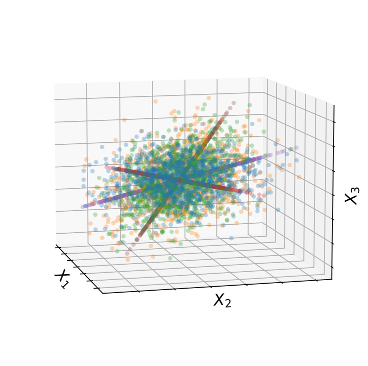
$\Rightarrow$ A universally consistent algorithm trained on the imputed data $\Phi(\widetilde{X})$ is Bayes consistent

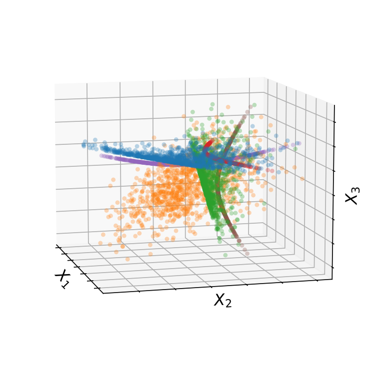
Asymptotically, imputing well is not needed to predict well

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

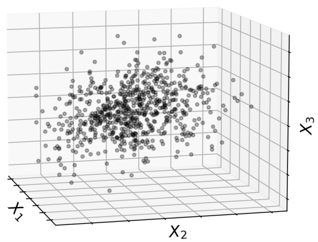
Rationale of proof: imputation creates manifolds



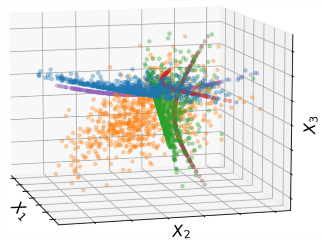




Bayes optimality of impute-n-regress (Le morvan et al. 2021)



Complete data



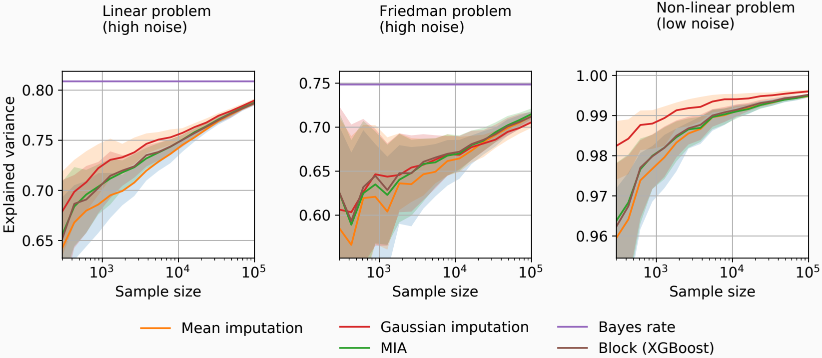
Imputed data (manifolds)

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)[97]
3. Given 2), we can build prediction functions, independent of m, that are Bayes optimal for all missing data patterns

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

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

⇒ Which imputation and predictor should one use?

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:

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

Assumptions on covariates and missing values (X, M)

1. Gaussian pattern mixture model, PMM: $X \mid (M = m) \sim \mathcal{N}(\mu_m, \Sigma_m)$

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

3. (Also for Gaussian assumption + MNAR self mask gaussian)

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 necessarily linear[98]

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.

[98] Le morvan, J. et al. Linear predictor on linearly-generated data with missing values: non consistency and solutions. *AISTAT2020*.

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

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

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

Neumiss Networks to approximate the covariance matrix

Order ℓ approx. of the Bayes predictor)

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

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

Neumiss Networks to approximate the covariance matrix

Order ℓ approx. of the Bayes predictor)

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

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**

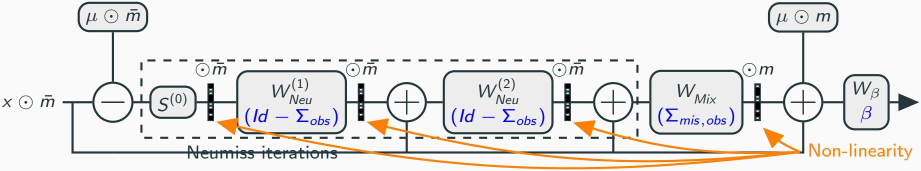
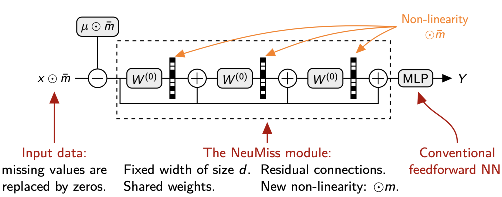


Figure 7: Depth of 3, $\bar{m} = 1 - m$. Each weight matrix $W^{(k)}$ corresponds to a simple transformation of the covariance matrix indicated in blue.

Networks with missing values: $\odot M$ nonlinearity [99]

- 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 $\odot M$



Couple Neumiss and MLP to jointly learn imputation and regression

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

Benchmark in supervised learning with missing values

- PhD Thesis Hava Chaptoukaev: Simulated data+ Real Data

Table 4.1: Test AUC (mean±std) over 10 random repetitions on the Breast Cancer, MIMIC III and UK Biobank datasets. **Bold** values denote the best prediction performances. Starred* values denote performances that significantly outperform the mean-imputation baseline. Values highlighted in blue denote the most reliable performances according to our guidelines.

	M1	M2	BC	UKB1	UKB2	UKB3	UKB4
Imp.-then-reg.							
Mean	79.7±2.1	72.1±4.6	64.9±4.5	**78.4±0.5**	76.3±0.5	86.0±0.3	**64.2±1.5**
KNN	79.6±2.7	72.2±3.7	64.2±4.6	62.3±0.7	76.1±0.5	85.4±0.3	63.1±1.5
MICE	79.6±1.9	73.4±4.2*	**67.6±3.3***	55.8±11.8	**76.5±0.4**	86.0±0.3	63.0±1.4
MIDA	**79.9±2.1**	72.0±4.0	66.1±2.4*	66.1±12.0	76.1±0.4	85.6±0.4	61.6±1.5
Imp.-and-reg.							
NeuMiss	75.4±2.5	73.5±4.4*	64.0±3.2	51.5±1.0	71.9±1.0	82.5±0.7	57.1±3.2
supMIWAE	76.8±3.0	71.6±4.0	66.7±5.7*	76.1±4.5	73.1±2.1	83.8±1.1	58.5±1.2
Imp.-free							
GBRT	**79.9±2.3**	**75.2±5.2***	65.7±3.1*	**78.4±0.5**	76.3±0.3	**86.2±0.3**	62.9±1.3

- Le Morvan, Varoquaux. (2025)[100]. Reg. models: MLP, Deep Tabular, XGBoost - Impute: Mean, MICE-Ridge, MissForest, Normal Cond.-Expect.

Good imputations matter less when using the mask[101]

$$\underbrace{\begin{array}{cc} X_1 & X_2 \\ \multicolumn{2}{c}{\begin{pmatrix} 1 & 2 \\ 3 & \mathrm{NA} \\ \mathrm{NA} & 4 \end{pmatrix}} \end{array}}_{\mu} \rightarrow \underbrace{\begin{array}{cccc} X_1 & X_2 & M_1 & M_2 \\ \multicolumn{4}{c}{\begin{pmatrix} 1 & 2 & 0 & 0 \\ 3 & \mathrm{NA} & 0 & 1 \\ \mathrm{NA} & 4 & 1 & 0 \end{pmatrix}} \end{array}}_{\Theta}$$

[100] Imputation for prediction: beware of diminishing returns. (ICLR 2025 spotlight)

[101] Mike et al. (2023). The Missing Indicator Method: From Low to High Dimensions. SIGKDD.

Benchmark in supervised learning with missing values

- Missing modalities - credit: PhD Thesis Hava Chaptoukaev

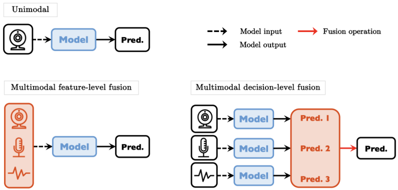


Table 4.2: Test F1-scores and accuracies (mean ± std) of state-of-the-art methods for the classification of stress. **Bold** values denote the best prediction performances. Starred* values denote performances that significantly outperform the reference feature-level fusion model. Values highlighted in blue denote the most reliable performances according to our guidelines.

		2-class	
	#tasks	F1-score (↑)	Accuracy (↑)
Multimodal (ref)			
Feature fusion	355	66.4 ± 4.3	61.2 ± 3.7
Decision fusion	355	72.9 ± 4.8	65.2 ± 4.9
Impute-then-regress			
Mean	711	**74.8±2.1***	73.7±2.7*
KNN	711	73.7±3.2*	72.8±2.9*
MICE	711	73.8±5.5*	73.4±5.5*
MIDA	711	74.3±3.1*	73.7±2.7*
Impute-and-regress			
NeuMiss	711	68.4±5.1*	58.0±4.7
supMIWAE	711	**74.8±4.0***	**73.8±3.9***
Imputation-free			
GBRT	711	73.9±2.8*	73.2±2.8*

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 powerful learner
- **Rethinking imputation: a good imputation is the one that makes the prediction easy**
- Close to conditional imputation but not CI
- Can even work in MNAR

MAR/MNAR settings are not tailored for prediction

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 powerful learner
- **Rethinking imputation: a good imputation is the one that makes the prediction easy**
- Close to conditional imputation but not CI
- Can even work in MNAR

Implicit and jointly learned Impute-then-Regress strategy

- Neumiss network: new architecture $\odot M$ nonlinearity
- Theoritically: differentiable approximation of the cond. expectation
- **Tree-based models: Missing Incorporated in Attribute**

MAR/MNAR settings are not tailored for prediction

Recent literature on supervised learning + resources

- Videos + slides in Mybox: Joint Imputation and Prediction, Linear+MLP

⇒ Erwan Scornet, Claire Boyer, Aymeric Dieuleveut.[102],[103],[104]
- Equivalence between imputing by zero in Linear Regression in high dimension and ridge regression.

⇒ K. A. Verchand, A. Montanari[105]
- Mean imputation + regularized logistic regression, in high dimension setting can reach Bayes risk

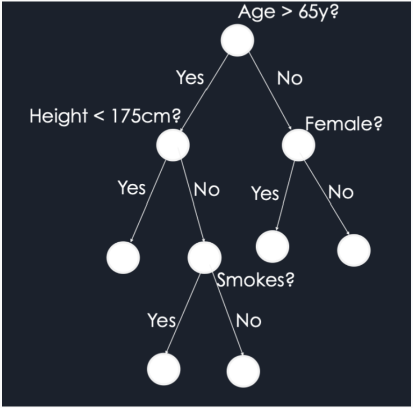
---

[102] Ayme et al. (2022) Near-optimal rate of consistency for linear models with missing val. *ICML*
[103] Ayme et al. (2023) Naive imputation implicitly regularizes high-dimensional linear models. *ICML*
[104] Ayme et al. (2024) Random features models to study the success of naive imputation. *ICML*
[105] High-dimensional logistic regression with missing data: Imputation, regularization, and universality

Distributional learning

Random Forest (RF) of Breimann (2001)

- Want to learn the conditional expectation of $Y \in \mathbb{R}$ given covariates $\mathbf{X} \in \mathbb{R}^p$ from i.i.d observations $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$
- Two steps:
  1. Construct a forest with N trees
  2. Predict for a test point $\mathbf{x}$

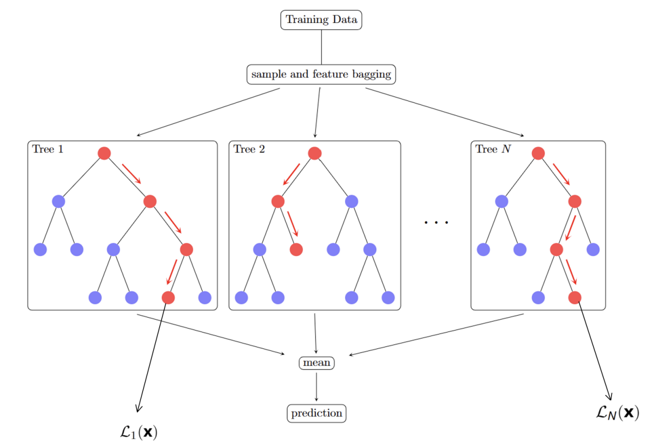


1. Forest construction

- Fit N trees
- Each tree splits the $Y_i's$ according to some rule depending on the covariates.
- Conventional RF uses the CART criterion, which compares the means of Y in the two child nodes.
- The split is taken where the squared difference in means is maximized.

2. Prediction

- ▷ Drop test point $\mathbf{x}$ in all trees $k = 1, \ldots, N$
- ▷ Let $\mathcal{L}_k(\mathbf{x})$ be the leaf where it falls.



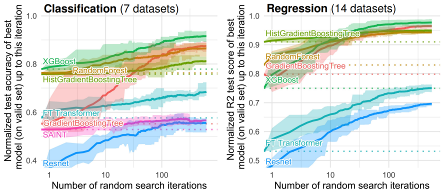
State-of-the-Art performance on Tabular Data

Why do tree-based models still outperform deep learning on tabular data?

Léo Grinsztajn
Soda, Inria Saclay
leo.grinsztajn@inria.fr

Edouard Oyallon
ISIR, CNRS, Sorbonne University

Gaël Varoquaux
Soda, Inria Saclay



Distributional Random Forest (DRF)

- ▷ Let's say we want to predict at $\mathbf{x}$
- ▷ RF implicitly also produces weights $w_i(\mathbf{x})$, $i = 1, \ldots, n$, indicating the importance of point i for this prediction:

$$w_i(\mathbf{x}) = \frac{1}{N} \sum_{k=1}^{N} \frac{\mathbf{1}\{\mathbf{X}_i \in \mathcal{L}_k(\mathbf{x})\}}{\#\mathcal{L}_k(\mathbf{x})}$$

- ▷ Can write the prediction as

$$\widehat{\mathbb{E}}[Y \mid \mathbf{X} = \mathbf{x}] = \sum_{i=1}^{n} w_i(\mathbf{x}) Y_i.$$

$\Longrightarrow$ RF is a nearest neighborhood method with a data-adaptive notion of neighborhood.

Distributional Random Forest (DRF)

- Can use the weights to approximate other things than conditional expectations
- Example: Conditional quantiles[106]
- However, doing this it might make sense to adapt the splitting criterion!
- Generalized Random Forest (GRF)[107]: Define an estimation target and adapt the splitting criterion by this target
- DRF: Define one splitting criterion that makes sense for many targets.

[106] Meinshausen. *Quantile regression forests*. JMLR, 2006
[107] Athey, Tibshirani, Wager. *Generalized random forests. AoS. 2019*

CART

CART criterion:

$$\min_{\text{splits}} \frac{1}{n_P} \left(\sum_{i \in C_L} (Y_i - \overline{Y}_L)^2 + \sum_{i \in C_R} (Y_i - \overline{Y}_R)^2 \right) \tag{3}$$

is equivalent to

$$\max_{\text{splits}} \frac{n_L n_R}{n_P^2} \left(\frac{1}{n_L} \sum_{i \in C_L} Y_i - \frac{1}{n_R} \sum_{i \in C_R} Y_i \right)^2 . \tag{4}$$

$\Longrightarrow$ Splits are chosen to make the means in the child nodes as different as possible.

Splitting Criteria

- RF:

$$\frac{n_L n_R}{n_P^2} \left(\bar{Y}_L - \bar{Y}_R \right)^2$$

- GRF:

$$\frac{n_L n_R}{n_P^2} \left(\hat{\tau}_L - \hat{\tau}_R \right)^2$$

Idea of DRF: Do CART but with means in a Reproducing Kernel Hilbert space (RKHS)!

MMD Criterion

Idea of DRF: Do CART but with means in a Reproducing Kernel Hilbert space (RKHS)!

- RKHS $\mathcal{H}$ is a Hilbert-space defined by a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$
- Any probability measure P can be mapped to $\mathcal{H}$, using the mapping Φ, where for all P,
$$\Phi(P) = \mathbb{E}_{\mathbf{Y} \sim P}[k(\mathbf{Y}, \cdot)] \in \mathcal{H}$$
- For certain choices of k learning this expectation is akin to learning the distribution!
- This is the idea of DRF: We use CART in $\mathcal{H}$ and estimate the conditional expectation in $\mathcal{H}$.

MMD Criterion

- ▷ Let Φ be the function that takes a probability measures and maps it into $\mathcal{H}$: $P \mapsto \Phi(P) := \mathbb{E}[k(\mathbf{Y}, \cdot)]$
- ▷ For the dirac measure $\Phi(\delta_{\mathbf{Y}_i}) = k(\mathbf{Y}_i, \cdot)$:

$$\max_{\text{split}} \frac{n_L n_R}{n_P^2} \left\| \Phi\left(\frac{1}{|n_L|} \sum_{i \in C_L} \delta_{\mathbf{Y}_i} \right) - \Phi\left(\frac{1}{|n_R|} \sum_{i \in C_R} \delta_{\mathbf{Y}_i} \right) \right\|_{\mathcal{H}}^2 =$$

$$\max_{\text{split}} \frac{n_L n_R}{n_P^2} \left\| \frac{1}{|n_L|} \sum_{i \in C_L} k(\mathbf{Y}_i, \cdot) - \frac{1}{|n_R|} \sum_{i \in C_R} k(\mathbf{Y}_i, \cdot) \right\|_{\mathcal{H}}^2$$

$\Longrightarrow$ Splits are chosen to make the means in the child nodes as different as possible, but now in the *Hilbert Space*.

DRF Estimator

- As a consequence, we get an estimate of the conditional mean embedding (CME)

$$\mu(\mathbf{x}) = \Phi(\mathbb{P}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}) = \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}]$$

- This has the form

$$\hat{\mu}_n(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) k(\mathbf{Y}_i, \cdot) \in \mathcal{H}$$

- This can easily be translated back into the empirical distribution:

$$\hat{\mathbb{P}}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} = \sum_{i=1}^{n} w_i(\mathbf{x}) \delta_{\mathbf{Y}_i}$$

- Access to $\hat{\mathbb{P}}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ is nice because a large range of targets can be calculated from it!

DRF Estimator: Summary

- i.i.d data $(\mathbf{Y}_1, \mathbf{X}_1), \ldots, (\mathbf{Y}_n, \mathbf{X}_n)$, $\mathbf{Y} \in \mathbb{R}^d$ and $\mathbf{X} \in \mathbb{R}^p$
- Random Forest (RF) is a powerful tool to estimate $\widehat{\mathbb{E}}[Y \mid \mathbf{X} = \mathbf{x}]$, for $d = 1$
- Idea of DRF: Use a RF in a Reproducing Kernel Hilbert space (RKHS) $\mathcal{H}$
- Learning the conditional expectation in this space
  $=$ Learning a representation of the conditional distribution $\mathbb{P}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$
- Resulting estimate can be conveniently written as

$$\hat{\mathbb{P}}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} = \sum_{i=1}^{n} w_i(\mathbf{x}) \delta_{\mathbf{Y}_i}$$

  with weights $w_i(\mathbf{x})$, $i = 1, \ldots, n$, indicating the importance of point i
- This also works when $\mathbf{Y}$ takes values in $\mathbb{R}^d$, for $d > 1$!

DRF Results

Theorem

Assume a list of conditions hold. Then there exists $\sigma_n \to 0$ such that,

$$\|\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\| = \mathcal{O}_p(\sigma_n) \tag{5}$$

$$\frac{1}{\sigma_n}\left(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\right) \xrightarrow{D} N(0, \mathbf{\Sigma}_{\mathbf{x}}) \tag{6}$$

holds.

Missing incorporated in attribute (MIA)[108]

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)

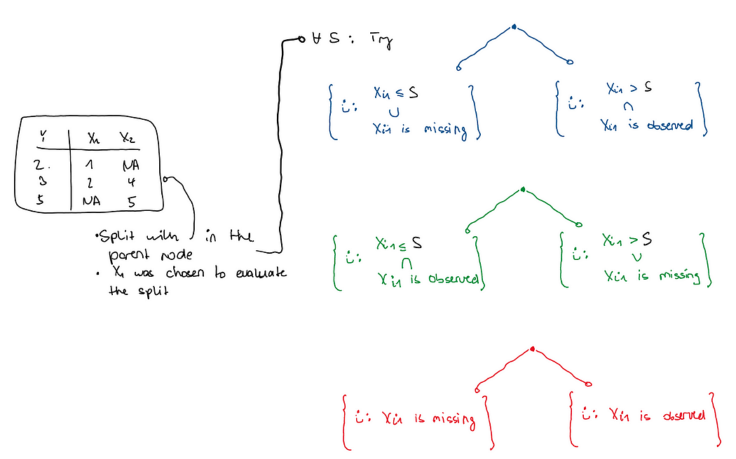
Targets one model per pattern:

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

$\Rightarrow$ Extremely **good performances** in practice **for any mechanism**

[108] Twala et al. (2008). Methods for coping with missing data in decision trees. *Pattern Recog.*

Missing incorporated in attribute (MIA)[109]



[109] Twala et al. (2008). Methods for coping with missing data in decision trees. *Pattern Recog.*

DRF & Missing Values

- DRF is based on the grf implementation and thus uses the same MIA implementation trick: duplicate the incomplete columns, and replace the missing entries once by $+\infty$ and once by $-\infty$ `grf/partykit package`, `scikit HistGradientBoosting`
- The good performance as well as theoretical results are likely to extend to DRF.
- This gives a way to obtain distributional prediction for $\mathbf{Y}$ when there are missing values only in $\mathbf{X}$!
- This allows for prediction intervals among other things.
- Examples:
  `https://1drv.ms/f/s!Ak6nJk4aN-80ndVCJwj8wCy8pDrufw?e=hw3BZk`
  `https://medium.com/data-science/random-forests-and-missing-values-3daaea103db0`

Other highlights and challenges with missing values

- Graphical models with missing values[110], works from R. Nabi, I. Shpitser.
- Weighting - doubly robust methods[111]; link with semi-supervised learning[112]
- SGD with MCAR values in linear models, see slides from A. Sportisse; Conformal Prediction with NA from M. Zaffran
- Researchers: R. Sameworth, T. Cannings, T. Berret, P. Ding, S. Seaman, F. Li, etc.

⇒ Some challenges

▷ Features importance with missing values
▷ Distributional shifts in the missing values
▷ SGD with NA under MAR and MNAR in logistic regression?[113]
▷ Times series with MNAR (predict intubation given online monitoring, features measured each 15 minutes/1 hour + clinical data - DTR
▷ Missing outcome/treatment/covariates?

[110] Mohan & Pearl. (2021). Graphical Models for Processing Missing Data. *JASA*
[111] Robins, Rotnitzky, Zhao. (1994). Estimation of regression coefficients when some regresors are not always observed. JASA.
[112] Sportisse et al. (2023). Are labels informative in semi-supervised learning? Estimating and leveraging the missing-data mechanism. *ICML*
[113] Sportisse, J. et al. Debiasing SGD to handle missing values. *Neurips2020*

Ressources

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

Project funded by the R consortium (Infrastructure Steering Committee)

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

- list existing packages
- available literature
- theoretical and practical tutorials
- analysis workflows on data (in R and in python)
- main actors
- popular datasets

$\Rightarrow$ Federate the community

$\Rightarrow$ Contribute!

Causal inference with missing values

Personalization of treatment recommendation

Ex: Estimating treatment effect from the Traumabase data

- 40000 trauma patients
- 300 heterogeneous features from pre-hospital and in-hospital settings
- 40 trauma centers, 4000 new patients per year

Center	Accident	Age	Sex	Weight	Lactacte	Blood Press.	TXA.	Y
Beaujon	fall	54	m	85	NA	180	treated	0
Pitie	gun	26	m	NA	NA	131	untreated	1
Beaujon	moto	63	m	80	3.9	145	treated	1
Pitie	moto	30	w	NA	NA	107	untreated	0
HEGP	knife	16	m	98	2.5	118	treated	1
$\vdots$								$\ddots$

$\Rightarrow$ **Estimate causal effect** (with missing values[114]): Administration of the **treatment** *tranexamic acid (TXA)*, given within 3 hours of the accident, on the **outcome** (Y) *28 days in-hospital mortality* for trauma brain patients

[114] Mayer, I., Wager, S. & **J.**. (2020). Doubly robust treatment effect estimation with incomplete confounders. *Annals Of Applied Statistics. (implemented in package `grf`).*

Causal inference questions in many fields

Assume a policy/intervention/**treatment** W causes an **outcome** Y
Aim: **estimate the effect** as acurately as possible (bias & variance)

- What is the effect of hydrochloroquine on mortality?
- Is there an effect of financial incentives on teacher performance (measured by teacher absences & class test scores)? (Duflo et al. 2012)
- Effect of reducing car traffic on air pollution
- What is the impact of the advertising campaign?
- What is the effect of social media on mental health?
- Does the students succeeded because of the new teacher?
  Had the students remained with the old teacher, they wouldn't have succeeded

Machine Learning VS Causal Inference

Machine learning: Powerful predictive models that rely on correlations. A central goal is to understand what usually happens in a given situation: Given today's weather, what's the chance tomorrow's air pollution levels will be dangerously high?

Causal inference: We want to predict what would happen if we change the system: How does the answer to the above question change if we reduce the number of cars on the road?

Concepts of causality are fundamental for having action levers, making recommendations and answering the questions "*what would happen if*"?

Human like AI: reasonable decisions in never experienced situations.

Long tradition in economics and epidemiology, public policies.

Causal inference[115]

Causal inference (simplest) question

Assume a policy/intervention/**treatment** W causes an **outcome** Y
Aim: **estimate the effect** as acurately as possible (bias & variance)

[115] Taskview to organize all packages on causal inference.

Causal inference[115]

Potential Outcome framework (Neyman, 1923; Rubin, 1974)

- n iid sample $(\underbrace{X_i}_{\text{covariates}}, \overbrace{W_i}^{\text{treatment}}, \underbrace{Y_i(1), Y_i(0)}_{\textit{potential outcomes}}) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R} \times \mathbb{R}$
- Individual **causal effect** of the binary treatment: $\Delta_i = Y_i(1) - Y_i(0)$

Problem: Δ_i never observed (only observe one outcome/indiv)

Covariates			Treatment	Outcome(s)	
X_1	X_2	X_3	W	Y(0)	Y(1)
1.1	20	F	1	**?**	200
-6	45	F	0	10	**?**
0	15	M	1	**?**	150
	...		...	...	...
-2	52	M	0	100	**?**

[115] Taskview to organize all packages on causal inference.

Causal inference[115]

Potential Outcome framework (Neyman, 1923; Rubin, 1974)

▷ n iid sample $(\underbrace{X_i}_{\text{covariates}}, \overbrace{W_i}^{\text{treatment}}, \underbrace{Y_i(1), Y_i(0)}_{\textit{potential outcomes}}) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R} \times \mathbb{R}$

▷ Individual **causal effect** of the binary treatment: $\Delta_i = Y_i(1) - Y_i(0)$

Problem: Δ_i never observed (only observe one outcome/indiv)

Covariates			Treatment	Outcome(s)	
X_1	X_2	X_3	W	Y(0)	Y(1)
1.1	20	F	1	**?**	200
-6	45	F	0	10	**?**
0	15	M	1	**?**	150
	...		...	...	...
-2	52	M	0	100	**?**

Average Treatment Effect (ATE): $\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$
The ATE is the difference of the average outcome had everyone gotten treated and the average outcome had nobody gotten treatment

[115] Taskview to organize all packages on causal inference.

Randomized Controlled Trial

Identifiability assumptions

- $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$ (consistency)
- $W_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\}$ **(random treatment assignment)**

Flip a coin to assign the treatment

$$\text{We can check that } \tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$$
$$= \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0]$$
$$= \mathbb{E}[Y_i|W_i = 1] - \mathbb{E}[Y_i|W_i = 0]$$

$\Rightarrow$ Although Δ_i never observe, τ is identifiable and can be estimated

Difference-in-means estimator

$$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{W_i=1} Y_i - \frac{1}{n_0} \sum_{W_i=0} Y_i, \text{ where } n_w = \sum_{i=1}^{n} 1_{W_i=w}$$

$\hat{\tau}_{DM}$ unbiased and $\sqrt{n}$-consistent $\sqrt{n}\,(\hat{\tau}_{DM} - \tau) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V_{DM})$

Randomized Controlled Trial

Identifiability assumptions

- $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$ (consistency)
- $W_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\}$ **(random treatment assignment)**

  Flip a coin to assign the treatment

$$\text{We can check that } \tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$$
$$= \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0]$$
$$= \mathbb{E}[Y_i|W_i = 1] - \mathbb{E}[Y_i|W_i = 0]$$

$\Rightarrow$ Although Δ_i never observe, τ is identifiable and can be estimated

Covariates			Treatment	Outcome(s)	
X_1	X_2	X_3	W	Y(0)	Y(1)
1.1	20	F	1	?	200
-6	45	F	0	10	?
0	15	M	1	?	150
	...		...	...	...
-2	52	M	0	100	?

$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{W_i=1} Y_i - \frac{1}{n_0} \sum_{W_i=0} Y_i;$ ATE = mean(red)-mean(blue)

Data sources and evidence to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions of treated and control groups
  $\Rightarrow$ High **internal** validity

- expensive, long, ethical limitations
- small sample size: restrictive inclusion criteria
  $\Rightarrow$ No personalized medicine
- **trial sample different from the population eligible for treatment**
  $\Rightarrow$ Low **external** validity

Data sources and evidence to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions of treated and control groups
  $\Rightarrow$ High **internal** validity

- expensive, long, ethical limitations
- small sample size: restrictive inclusion criteria
  $\Rightarrow$ No personalized medicine
- **trial sample different from the population eligible for treatment**
  $\Rightarrow$ Low **external** validity

Observational data

- low cost
- large amounts of data (registries, biobanks, EHR, claims)
  $\Rightarrow$ patient's heterogeneity
- **representative of the target populations**
  $\Rightarrow$ High **external** validity

Data sources and evidence to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions of treated and control groups
  $\Rightarrow$ High **internal** validity

- expensive, long, ethical limitations
- small sample size: restrictive inclusion criteria
  $\Rightarrow$ No personalized medicine
- **trial sample different from the population eligible for treatment**
  $\Rightarrow$ Low **external** validity

Observational data

- "big data": low quality
- lack of a controlled design opens the door to **confounding bias**
  $\Rightarrow$ Low **internal** validity

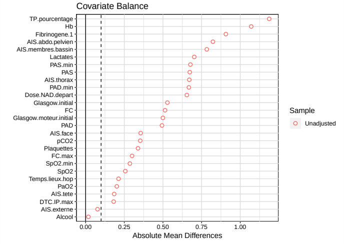
- low cost
- large amounts of data (registries, biobanks, EHR, claims)
  $\Rightarrow$ patient's heterogeneity
- **representative of the target populations**
  $\Rightarrow$ High **external** validity

Observational data: non random assignment

	survived	deceased	Pr(survived | treatment)	Pr(deceased | treatment)
TA not administered	6,238 (76%)	1,327 (16%)	0.82	0.18
TA administered	367 (4%)	316 (4%)	0.54	0.46

Mortality rate 20% - for treated 46% - not treated 18%: treatment kills?

Standardized mean differences between treated and control.



Severe patients (with higher risk of death) are more likely to be treated.

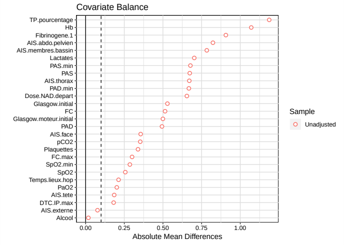
If the control group does not look like the treatment group, the difference in response may be **confounded** by the differences between the groups.

Observational data: non random assignment

	survived	deceased	Pr(survived | treatment)	Pr(deceased | treatment)
TA not administered	6,238 (76%)	1,327 (16%)	0.82	0.18
TA administered	367 (4%)	316 (4%)	0.54	0.46

Mortality rate 20% - for treated 46% - not treated 18%: treatment kills?

Standardized mean differences between treated and control.



Severe patients (with higher risk of death) are more likely to be treated.

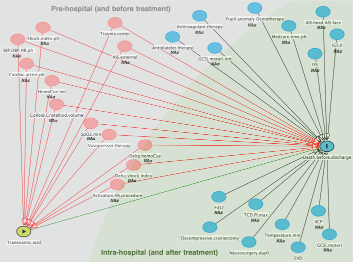
Treatment allocation W depends on covariates X, so the covariate distributions for treatment and control patients are different.

Assumption for ATE identifiability in observational data

Unconfoundedness

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$$

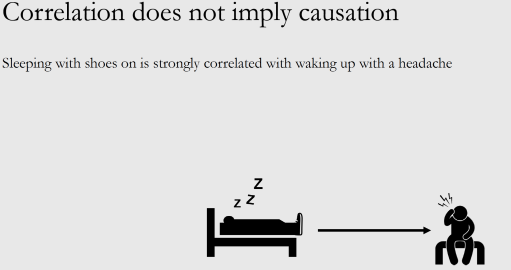
Measure all possible confounders: variables related to both treatment and outcome



http://www.dagitty.net/ - Obtained by a Delphi method: consensus group technique (do not ask for the complete graph)

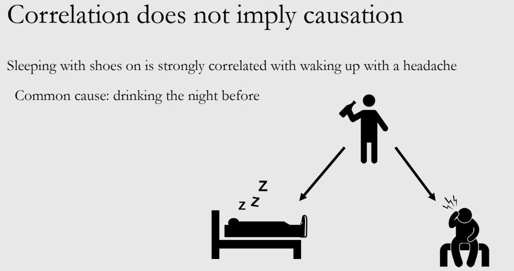
ATE not identifiable without it: it is not a sample size problem

Correlation versus Causation



credit: Brady Neal

Correlation versus Causation



credit: Brady Neal

Unobserved confounders make it impossible to separate correlation and causality.

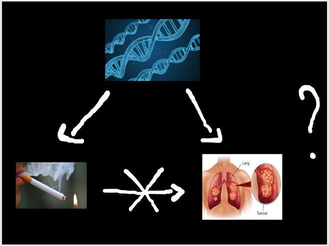
$\Rightarrow$ Many methods to mitigate this issue: sensitivity analysis, negative outcome control, instrumental variables, etc.

Assumption for ATE identifiability in observational data

Unconfoundedness

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$$

Measure all possible confounders



Unobserved confounders make it impossible to separate correlation and causality when correlated to both the outcome and the treatment.

$\Rightarrow$ Many methods to takle this issue: sensitivity analysis, negative outcome control, instrumental variables, etc.

Assumption for ATE identifiability in observational data

Overlap

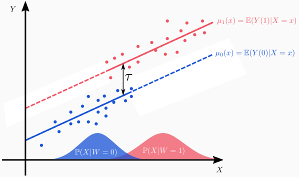
Propensity score: probability of treatment given observed covariates.

$$e(x) = \mathbb{P}(W_i = 1 \mid X_i = x) \quad \forall\, x \in \mathcal{X}.$$

We assume overlap, i.e. $\eta < e(x) < 1 - \eta, \quad \forall\, x \in \mathcal{X}$ and some $\eta > 0$

Regression adjustment

Outcome $\sim$ covariates: $\mu_{(w)}(x) = \mathbb{E}[Y_i(w) \mid X_i = x]$
OLS model: $w \in \{0, 1\}$ $Y_i(w) = c_{(w)} + X_i \beta_{(w)} + \varepsilon_i(w)$



$$
\begin{aligned}
\tau &= \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0)|X_i] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|W_i = 1, X_i = x] - \mathbb{E}[Y_i(0)|W_i = 0, |X_i = x]](\text{uncounfoud}) \\
&= \mathbb{E}[\mathbb{E}[Y_i|W_i = 1, X_i] - \mathbb{E}[Y_i|W_i = 0, X_i]](\text{consistency})
\end{aligned}
$$

Regression adjustment estimator (plug-in g-formula)

$$\hat{\tau}_g = \frac{1}{n}\sum_{i=1}^n (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)) = \frac{1}{n}\sum_{i=1}^n \left((\hat{c}_{(1)} + X_i\hat{\beta}_{(1)}) - (\hat{c}_{(0)} + X_i\hat{\beta}_{(0)})\right)$$

$\Rightarrow$ Consistent if $\hat{\mu}_{(w)}$ consistent

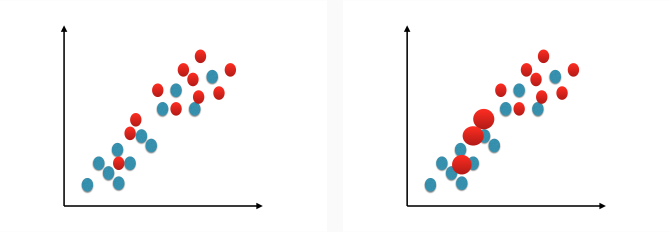
Inverse-propensity weighting estimator

Average treatment effect (ATE): $\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$
Propensity score (proba treated|covariates): $e(x) = \mathbb{P}(W_i = 1 \mid X_i = x)$

IPW estimator

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

$\Rightarrow$ Balance the differences between the two groups
$\Rightarrow$ High variance (divide by probability)



$\Rightarrow$ Consistent estimator of τ when $\hat{e}(\cdot)$ consistent (logistic regression)

Doubly robust estimator

Define $\mu_{(w)}(x) = \mathbb{E}[Y_i \mid X_i = x, W_i = w]$ and $e(x) = \mathbb{P}(W_i = 1 \mid X_i = x)$.

Augmented IPW - Double Robust (DR)

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

is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent.

- $\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right)$: Treatment assignment $\sim$ covariates
- $\hat{\tau}_{OLS} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$: Outcome $\sim$ covariates

$\Rightarrow$ Both sensitive to misspecification. DR: combine ols + ipw of residuals

Doubly robust estimator

Define $\mu_{(w)}(x) = \mathbb{E}[Y_i \mid X_i = x, W_i = w]$ and $e(x) = \mathbb{P}(W_i = 1 \mid X_i = x)$.

Augmented IPW - Double Robust (DR)

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

is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent.

- $\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right)$: Treatment assignment $\sim$ covariates
- $\hat{\tau}_{OLS} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$: Outcome $\sim$ covariates

$\Rightarrow$ Both sensitive to misspecification. DR: combine ols + ipw of residuals

Rationale: makes group similar before extrapolation

$$\sum_{i:W_i=1}(\widetilde{\hat{\mu}}_{(0)}(X_i) - \mu_{(0)}(X_i)) = \underbrace{(\overline{X}_1 - \hat{\gamma}^T\overline{X}_0)}_{\textit{covariate balancing}}\underbrace{(\hat{\beta}^{(0)} - \beta^{(0)})}_{\textit{extrapolation}} + \text{noise term}$$

where $\hat{\gamma} = (1 - \hat{e}(X_j))^{-1}$

Doubly robust ATE estimation

Model Treatment on Covariates $e(x) = \mathbb{P}(W_i = 1 \mid X_i = x)$
Model Outcome on Covariates $\mu_{(w)}(x) = \mathbb{E}[Y_i(w) \mid X_i = x]$

Augmented IPW - Double Robust (DR)

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

is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent

Possibility to use **any (machine learning) procedure** such as **random forests**, deep nets, etc. to estimate $\hat{e}(x)$ and $\hat{\mu}_{(w)}(x)$ without harming the interpretability of the causal effect estimation

Properties - Double Machine Learning (chernozhukov, et al. 2018)

If $\hat{e}(x)$ and $\hat{\mu}_{(w)}(x)$ converge at the rate $n^{1/4}$ then
$\sqrt{n}\,(\hat{\tau}_{DR} - \tau) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V^*)$, V^* semiparametric efficient variance.
In practice: random forests (+ outcome related variables for precision?)

Missing attributes alter causal analyses

Coupled causal and missing assumptions

1. Classical unconfoundedness + classical missing values mechanisms[116]
2. Unconfoundedness with missing + (no) missing values mechanisms[117]
3. Latent unconfoundedness + MCAR[118]

- New proposals to handle missing values in causal inference
- Implemented in the grf R package

[116] Seaman and White. IPW with missing predictors of treatment assignment, *Communications in Statistics, Theory & Methods*. 2014.

[117] Mayer, Wager, J. Doubly robust estimation with incomplete confounders. *AOAS*. 2020.

[118] Kallus et al. Causal inf. with noisy & missing covariates via matrix factorization. *Neurips*. 2018.

Popular multiple imputation for estimating treatment effect?

X_1^*	X_2^*	X_3^*	...	W	Y(0)	Y(1)
NA	20	10	...	1	?	200
-6	45	NA	...	1	10	?
0	NA	30	...	0	?	150
NA	32	35	...	0	?	100
-2	NA	12	...	0	20	?

1) Generate M plausible values for each missing value

X_1	X_2	X_3	...	W	Y	X_1	X_2	X_3	...	W	Y	X_1	X_2	X_3	...	W	Y
3	20	10	...	1	200	-7	20	10	...	1	200	7	20	10	...	1	200
-6	45	6	...	1	10	-6	45	9	...	1	10	-6	45	12	...	1	10
0	4	30	...	0	150	0	12	30	...	0	150	0	-5	30	...	0	150
-4	32	35	...	0	100	13	32	35	...	0	100	2	32	35	...	0	100
-2	15	12	...	0	20	-2	10	12	...	0	20	-2	20	12	...	0	20

2) Estimate Average Treatment Effect on each imputed data set with IPW: $\hat{\tau}_m$

3) Combine the results (Rubin's rules): $\hat{\tau} = \frac{1}{M}\sum_{m=1}^{M}\hat{\tau}_m$

Consistency of multiple imputation with IPW [119]

Assume: **MAR** $\mathbb{P}(M = m \mid X, Y, W) = \mathbb{P}(M = m \mid X_{obs(m)}, Y, W)$,

Classical **unconfoundedness** $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$,

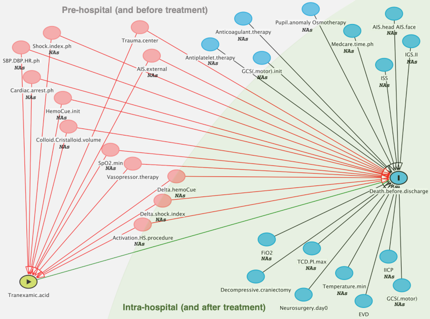
Propensity Score and model for $(X \mid Y, W)$ correctly specified,

$\Rightarrow$ Multiple imputation (Mice using (X^*, W, Y)) with IPW is consistent

[119] Seaman and White. 2014. IPW with missing predictors of treatment assignment, *Communications in Statistics, Theory & Methods*.

Causal identifiability assumptions adapted to missing values

http://www.dagitty.net/



Covariates			Treatment	Outcome(s)	
X_1^*	X_2^*	X_3^*	W	Y(0)	Y(1)
NA	20	F	1	?	200
-6	45	NA	0	10	?
0	NA	M	1	?	150
NA	32	F	1	?	100
1	63	M	1	15	?
-2	NA	M	0	20	?

Unconfoundedness: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \mid X$

$\Rightarrow$ Doctors give us the DAG (do not ask for the complete graph only for adjustment set), obtained by a Delphi method

Unconfoundedness with missing values: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \mid X^*$

$X^* = (1 - M) \odot X + M \odot NA$, $M_{ij} = 1$ if X_{ij} missing, 0 otherwise; $(\mathbb{R} \cup \{\mathtt{NA}\})^d$

$\Rightarrow$ Doctors decide to treat a patient based on what they observe/record. We have access to the same information as the doctors

AIPW under unconfoundeness with missing values [121]

Augmented IPW [120] with missing values

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

Generalized propensity score

$$e^*(x^*) = \mathbb{P}(W = 1 \,|\, X^* = x^*)$$

One model per pattern: $\sum_{m \in \{0,1\}^d} \mathbb{E}\left[W | X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$

In practice: combine two non-parametric estimations (imputation + forests or forest with MIA)

Properties

$\hat{\tau}_{AIPW^*}$ is $\sqrt{n}$-consistent, asympt. normal with semi parametric variance given: $\mathbb{E}\left[\left(\hat{e}^*(X_i^*)^{(-i)} - e^*(X_i^*))\right)^2\right]^{\frac{1}{2}} \times \mathbb{E}\left[\left(\hat{\mu}^*_{(W)}(X_i^*)^{(-i)} - \mu^*_{(W)}(X_i^*)\right)^2\right]^{\frac{1}{2}} = o\left(\frac{1}{\sqrt{n}}\right)$

[120] Robins, Rotnitzky, Zhao. (1994). Estimation of regression coefficients when some regressors are not always observed. JASA.

[121] Mayer, Wager, J. (2020). Doubly robust treat. effect estim. with incomplete confounders *AOAS*.

Methods to do causal inference with missing values

	Covariates		Missingness		Unconfoundedness			Models for (W, Y)	
	multivariate normal	general	M(C)AR	general	Missing	Latent	Classical	logistic-linear	non-param.
1. *(SA)EM* [122]	✓	✗	✓	✗	✓	✗	✗	✓	✗
1. Mean.GRF	✓	✓	✓	(✓)	✓	✗	✗	✓	✓
1. MIA.GRF	✓	✓	✓	(✓)	✓	✗	✗	✓	✓
2. Mult. Imp.	✓	✓	✓	✗	(✗)	✗	✓	✓	(✗)
3. MatrixFact.	✓	✗	✓	✗	✗	✓	✗	✓	(✗)
3. MissDeep-Causal	✓	✓	✓	✗	✗	✓	✗	✓	✓

Methods & assumptions on data generating process: models for covariates, missing values mechanism, identifiability conditions, models for treatment/outcome.

✓: can be handled ✗: not applicable in theory

(✓): empirical results and ongoing work on theoretical guarantees

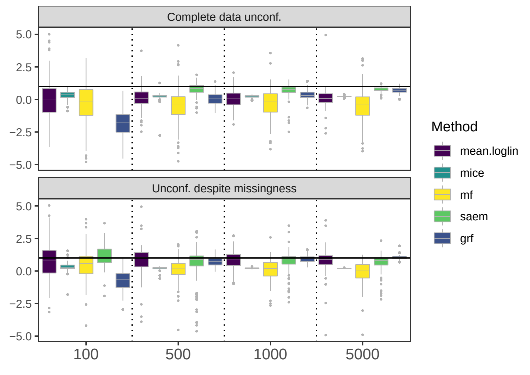
(✗): no theoretical guarantees but heuristics.

[122] Use of EM algorithms for logistic regression with missing values. Jiang, et al. 2019

Simulations: no overall best performing method.

- 10 covariates generated with Gaussian mixture model $X_i \sim \mathcal{N}_d(\mu_{(c_i)}, \Sigma_{(c_i)}) | C_i = c_i$, C from a multinomial distribution with three categories.
- Unconfoundedness on complete/observed covariates, 30% NA
- Logistic-linear for $(W,\ Y)$, $logit(e(X_{i.})) = \alpha^T X_{i.}$, $Y_i \sim \mathcal{N}(\beta^T X_{i.} + \tau W_i, \sigma^2)$

Figure 8: Estimated with AIPW and true ATE $\tau = 1$



→ grf-MIA is asymptotically unbiased under unconfoundedness despite missingness.
→ Multiple imputation requires many imputations to remove bias.

Simulations: importance of unconfoundedness assumption and choice of estimator

Setup

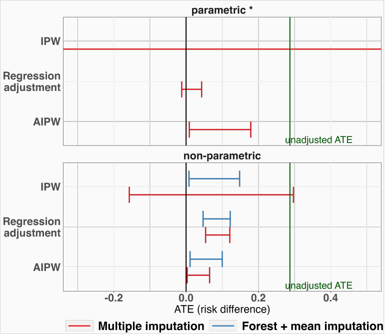
- Different data generating models (linear, nonlinear, latent, etc.)
- Different missingness mechanisms

Results

- AIPW estimators outperform their IPW counterparts.
- For $\hat{\tau}_{mia}$, the *unconfoundedness despite missingness* is indeed necessary.
- $\hat{\tau}_{mia}$ unbiased for all missingness mechanisms, especially for MNAR.
- Multiple imputation (mice) only requires standard unconfoundedness, but needs MAR

ATE estimations: effect of tranexamic acid on in-ICU mortality

- 40 covariates, 18 confounders (categorical and quantitative). 8248 patients
- Multiple imputation assumes **MAR & classical unconfoundeness** while other **unconfoundeness with missing & (no) assumptions on missing mechanism**

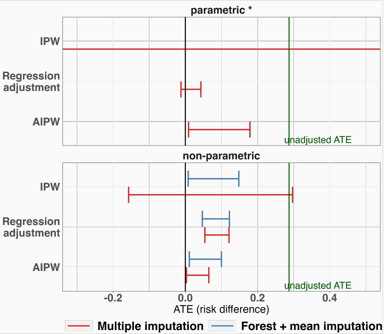


x-axis: Estimat. of the ATE ($\times 100$), bootstrap CI, y-axis: Methods with logistic regression or forests for nuisances. Missing values handled with multiple imputation or MIA [123]

[123] Other estimators (latent confounding, Kallus 2018 or parametric models with EM algorithms Jiang, J. 2019) are available bur not displayed for clarity (all tend to a slightly detrimental effect)

ATE estimations: effect of tranexamic acid on in-ICU mortality

- 40 covariates, 18 confounders (categorical and quantitative). 8248 patients
- Multiple imputation assumes **MAR & classical unconfoundeness** while other **unconfoundeness with missing & (no) assumptions on missing mechanism**



x-axis: Estimat. of the ATE ($\times100$), bootstrap CI, y-axis: Methods with logistic regression or forests for nuisances. Missing values handled with multiple imputation or MIA [123]

$\Rightarrow$ Do we need to include outcome related variables to improve precision?

Compromise for final sample size with non parametric methods

[123] Other estimators (latent confounding, Kallus 2018 or parametric models with EM algorithms Jiang, J. 2019) are available bur not displayed for clarity (all tend to a slightly detrimental effect)