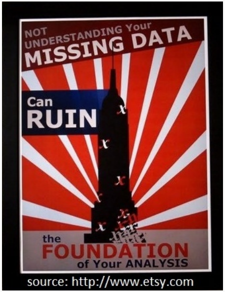
An overview of methods to deal with missing values

Julie Josse

CUSO doctoral school in Statistics and Applied Probability

Les Diableret, February 8-9 2022



Presentation Julie Josse: Stat and ML for bio-sciences

Academic background:

- Engineer and Assistant Professor in Agronomy University (2007-2015)
- Visiting Researcher + Teaching at Stanford University (18 months)
- Professor at Ecole Polytechnique (IP Paris) (2016-2020). Still Teaching
- Visiting Researcher at Google Brain Paris (2019-2020). Still Collaborating
- Senior Researcher at Inria Montpellier (Sept. 2020-)

Research topics:

- Dimensionality reduction to visualize high dimensional heterogeneous data
- Missing values: supervised learning, inference, matrix completion, MNAR
- Causal inference: estimating treatment effect, combining RCT and observational data, personalized recommendation
- Medical collaborations: Traumabase, IGR, CHU Nancy, Curie, etc.

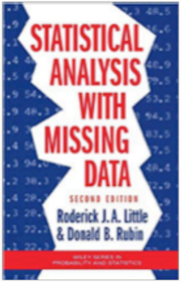
Implementations - transfert:

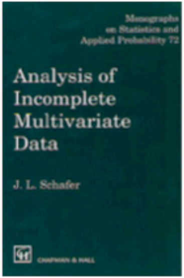
- R community: book R for Statistics, R foundation, R Forwards (widen the participation of minorities), R packages
- Rmisstastic https://rmisstastic.netlify.app/

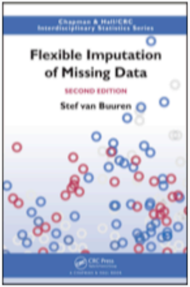
Outline

- Lecture 1: Introduction
    - Single imputation, Multiple imputation
    - Likelihood approaches
- Lecture 2: Low rank methods
    - PCA with missing values - (Multiple) Imputation with PCA
    - Practice
    - MNAR data
    - Heterogeneous data
- Lecture 3:
    - Supervised learning with missing values
        - Random Forest with missing values
        - Linear regression with missing values
    - Causal inference with missing values

References


STATISTICAL ANALYSIS WITH MISSING DATA
SECOND EDITION
Roderick J.A. Little
& Donald B. Rubin


Monographs on Statistics and Applied Probability 72
Analysis of Incomplete Multivariate Data
J. L. Schafer
CHAPMAN & HALL


Chapman & Hall/CRC Interdisciplinary Statistics Series
Flexible Imputation of Missing Data
SECOND EDITION
Stef van Buuren
CRC Press

Outline

Missing values

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

Traumabase

- 30000 patients
- 250 continuous and categorical variables: **heterogeneous**
- 20 hospitals
- 4000 new patients/ year

Center	Accident	Age	Sex	Weight	Lactactes	BP	shock	...
Beaujon	fall	54	m	85	NM	180	yes	
Pitie	gun	26	m	NR	NA	131	no	
Beaujon	moto	63	m	80	3.9	145	yes	
Pitie	moto	30	w	NR	Imp	107	no	
HEGP	knife	16	m	98	2.5	118	no	
⋮								⋱

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

Traumabase

- 30000 patients
- 250 continuous and categorical variables: **heterogeneous**
- 20 hospitals
- 4000 new patients/ year

Center	Accident	Age	Sex	Weight	Lactactes	BP	shock	...
Beaujon	fall	54	m	85	NM	180	yes	
Pitie	gun	26	m	NR	NA	131	no	
Beaujon	moto	63	m	80	3.9	145	yes	
Pitie	moto	30	w	NR	Imp	107	no	
HEGP	knife	16	m	98	2.5	118	no	
⋮								⋱

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

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

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

Traumabase

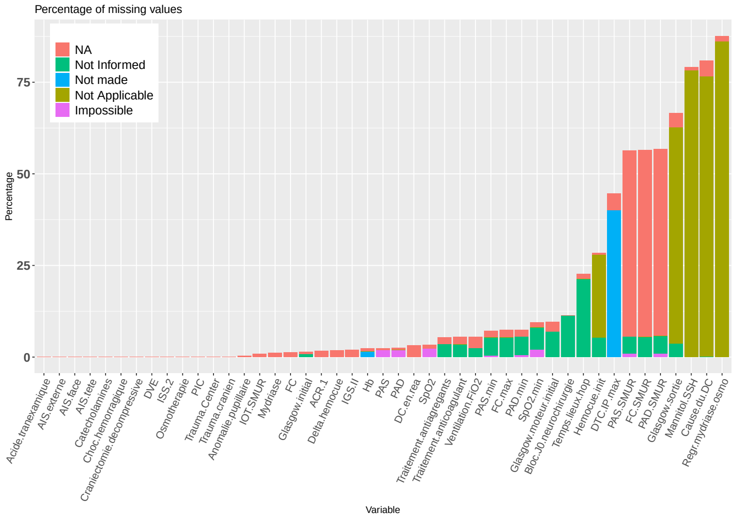
- 30000 patients
- 250 continuous and categorical variables: **heterogeneous**
- 20 hospitals
- 4000 new patients/ year

Center	Accident	Age	Sex	Weight	Lactactes	BP	shock	...
Beaujon	fall	54	m	85	NM	180	yes	
Pitie	gun	26	m	NR	NA	131	no	
Beaujon	moto	63	m	80	3.9	145	yes	
Pitie	moto	30	w	NR	Imp	107	no	
HEGP	knife	16	m	98	2.5	118	no	
⋮								⋱

⇒ **Explain and Predict** platelet levels, hemorrhagic shock given pre-hospital features

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

Missing values

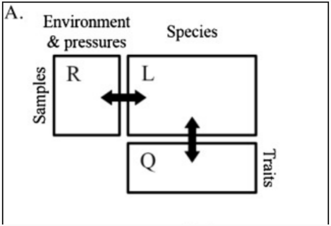


Different types of missing values

Multilevel data/ data integration: Systematic missing variable in one hospital

Contingency tables with side information

- National agency for wildlife and hunting management (ONCFS) data
- Contingency tables: Water (722 wetland sites) - bird (species) count data, from 1990-2016 in 5 countries in North Africa
- Additional sites & years info: meteo, geographical (altitude, etc.)

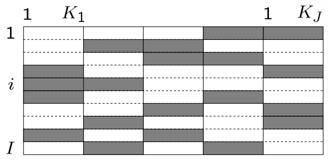


$\Rightarrow$ Aims: Assess the effect of time on species abundances;
Monitor the population and assess wetlands conservation policies.
$\Rightarrow$ 70% of missing values in contingency tables [2] [3]

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

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

Multi-blocks data set



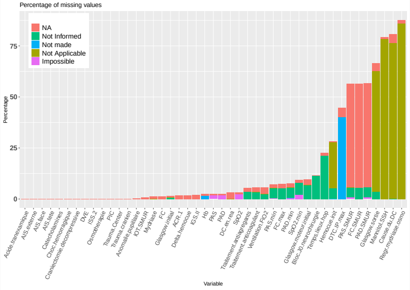
L'OREAL data: 100 000 women in many countries - 300 questions in groups:

- Self-assessment questionnaire: life style, skin and hair characteristics, care and consumer habits
- Clinical assessments by a dermatologist: facial skin complexion, wrinkles, scalp dryness, greasiness
- Hair assessments by a hair dresser: abundance, volume, breakage, curly
- Skin and Hair photographs and measurements: sebum quantity, etc.

⇒ Aim: Clustering women for marketing targeting
⇒ Missing values structured by group [4] [5]

[4] Handling missing values in exploratory multivariate data analysis. J., Husson. *JSFDS* 2012.
[5] Handling missing values in Multiple Factor Analysis. J., Husson. FQP 2013.

Complete-case analysis



Deleting rows with missing values?

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

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

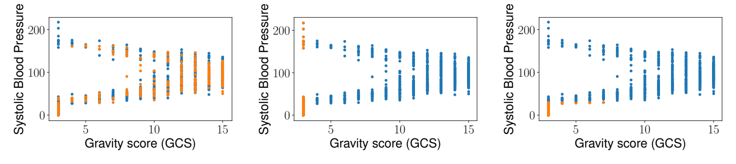
An $n \times p$ matrix, each entry is missing with probability 0.01

$p = 5 \quad \Longrightarrow \approx 95\%$ of rows kept

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

Distribution of missing values

Missing values mechanisms taxonomy Rubin, 1976



MCAR - MAR - MNAR

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

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

Missing values mechanisms

- Random Variables:
  - $X \in \mathbb{R}^d$: the complete unvailable data
  - $\widetilde{X} \in \{\mathbb{R} \cup \{\mathrm{NA}\}\}^p$: incomplete data (available)
  - $M \in \{0,1\}^d$: the missing-data pattern

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

- Realizations:

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

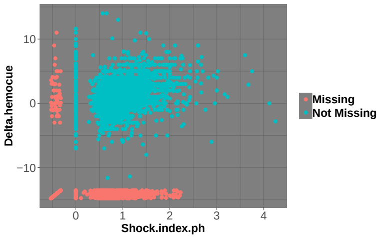
MCAR: For all $m \in \{0,1\}^d, P(M = m \mid X) = P(M = m)$
MAR[6]: For all $m \in \{0,1\}^d, P(M = m \mid X) = P\left(M = m \mid X_{obs(m)}\right)$

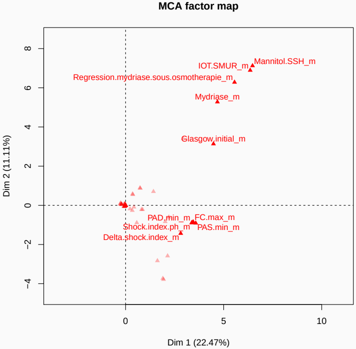
[6] What Is Meant by "Missing at Random"? Seaman, et al. Statistical Science. 2013.

Visualization

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

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





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

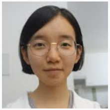
IOT: nested questions. Q1: yes/no, if yes Q2 - Q4, if no Q2 - Q4 "missing"

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

Outline

Collaborators on inference/imputation with missing values

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

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

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

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

Missing Data Imputation using Optimal Transport. *ICML2020*.

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

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

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

Modify the estimation process to deal with missing values

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

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

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

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

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

Modify the estimation process to deal with missing values

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

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

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

One specific algorithm for each statistical method...

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

$\Rightarrow$ Inferential framework

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

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

Modify the estimation process to deal with missing values

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

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

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

Imputation (multiple) to get a complete data set

Any analysis can be performed

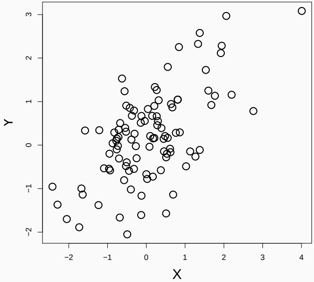
Ex logistic regression: Impute and apply logistic model to get $\hat{\beta}$, $\hat{V}(\hat{\beta})$

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

Mean imputation

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

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



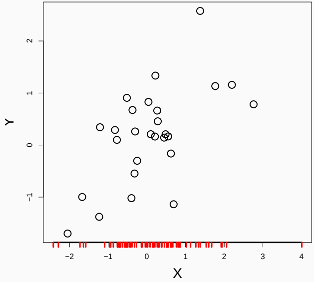
$\mu_y = 0$
$\sigma_y = 1$
$\rho_{xy} = 0.6$

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

Mean imputation

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

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



$\mu_y = 0$

$\sigma_y = 1$

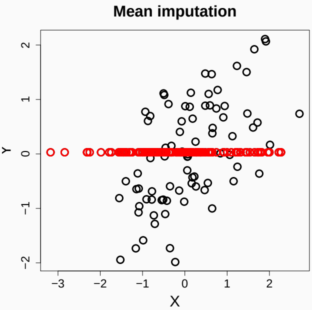
$\rho_{xy} = 0.6$

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

Mean imputation

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

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

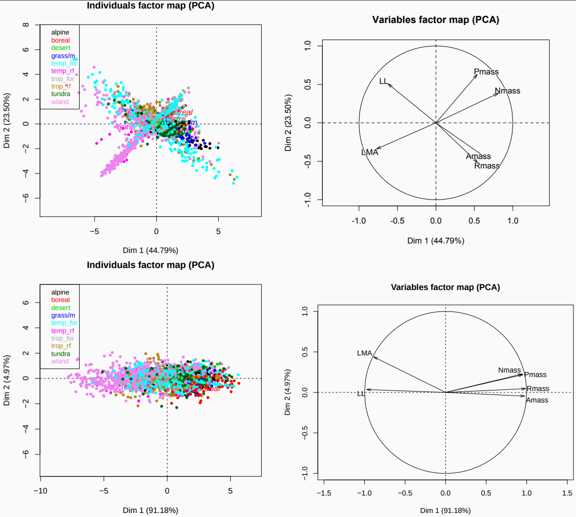


$\mu_y = 0$
$\sigma_y = 1$
$\rho_{xy} = 0.6$

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

Mean imputation deforms joint and marginal distributions

Mean imputation is bad for estimation



PCA with mean imputation

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

EM-PCA

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

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

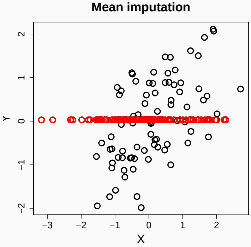
Ecological data: [7] $n = 69000$ species - 6 traits. Estimated correlation between Pmass & Rmass ≈ 0 (mean imputation) or ≈ 1 (EM PCA)

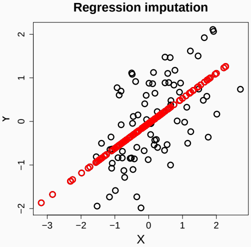
[7] Wright, I. et al. (2004). The worldwide leaf economics spectrum. *Nature*.

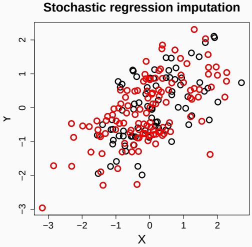
Imputation methods

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

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







$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho_{xy} = 0.6$	0.30	0.78	0.59

Imputation with joint model with gaussian distribution

$\Rightarrow$ Assumption joint gaussian model $z_i = (x_i, y_i)$, $z_{i.} \sim \mathcal{N}(\mu, \Sigma)$

- Bivariate case with missing values on y (stochastic regression):
  - estimate β and σ
  - impute from the predictive $\hat{y}_i \sim \mathcal{N}\left(x_i\hat{\beta}, \hat{\sigma}^2\right)$

- Extension to the multivariate case:
  - Estimate μ and Σ from an incomplete data with EM
  - Impute by drawing from the conditional distribution $Z_{\text{mis}}|Z_{\text{obs}} \sim \mathcal{N}(\mu_{\text{mis}|\text{obs}}, \Sigma_{\text{mis}|\text{obs}})$

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

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

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

Imputation methods for multivariate data

Assuming a joint model

- Gaussian distribution: $z_{i.} \sim \mathcal{N}(\mu, \Sigma)$ (`Amelia` Honaker, King, Blackwell)
- low rank: $Z_{n \times d} = \mu_{n \times d} + \varepsilon \; \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with μ of low rank k
  (`softimpute` Hastie & Mazuder; missMDA J. & Husson, mimi [8])
- latent class - nonparametric Bayesian (dpmpm Reiter)
- deep learning using variational autoencoders (MIWAE, Mattei, 2018, VAEAC Ivanov et al., 2019), using GAN (GAIN, Yoon et al. 2018)

Using conditional models (joint implicitly defined)

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

Imputation for categorical, mixed, blocks/multilevel data [9], etc.

$\Rightarrow$ Rmistatic platform, more than 150 packages[10]

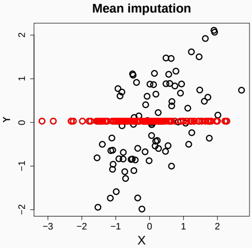
[8] J. et al. Main effects and interactions in mixed and incomplete data frames. (2018) *JASA*.

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

[10] J., et al. `https://cran.r-project.org/web/views/MissingData.html`

Outline

Single imputation methods: Danger!



$\mu_y = 0$	0.01
$\sigma_y = 1$	0.5
$\rho = 0.6$	0.30
$CI\mu_y 95\%$	

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

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

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

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

Variance unknown:

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

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

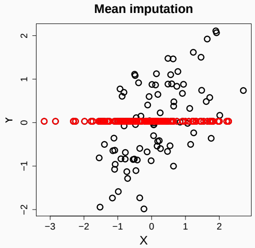
Simulation

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

Single imputation methods: Danger!

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

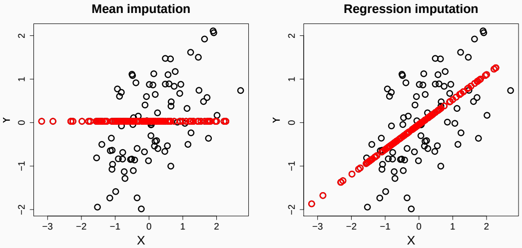


$\mu_y = 0$	0.01
$\sigma_y = 1$	0.5
$\rho = 0.6$	0.30
$CI\mu_y 95\%$	39.4

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

Single imputation methods: Danger!

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

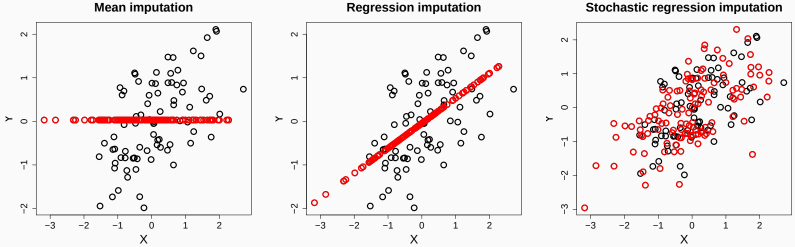


$\mu_y = 0$	0.01	0.01
$\sigma_y = 1$	0.5	0.72
$\rho = 0.6$	0.30	0.78
$CI\mu_y 95\%$	39.4	61.6

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

Single imputation methods: Danger!

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

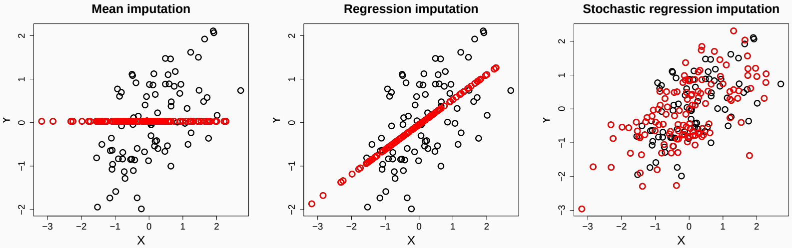


$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59
$CI\mu_y 95\%$	39.4	61.6	70.8

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

Single imputation methods: Danger!

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



$\mu_y = 0$	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: Underestimation of the variability

$\Rightarrow$ Incomplete Traumabase

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

Single imputation: Underestimation of the variability

$\Rightarrow$ 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: Underestimation of the variability

⇒ Incomplete Traumabase

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

⇒ Completed Traumabase

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

A single value can't reflect the uncertainty of prediction

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

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

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

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

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

Multiple imputation

1) Generate M plausible values for each missing value

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

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

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

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

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

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

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

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

⇒ Variability of missing values taken into account

Multiple imputation

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

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

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 continuous data: bivariate case

$\Rightarrow$ 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 $\mathcal{N}(x_i\hat{\beta}^m, (\hat{\sigma}^2)^m)$

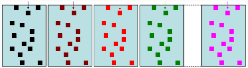
	Improper	Proper
$CI\mu_y 95\%$	0.818	0.935

Multiple imputation

⇒ 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 ⇒ underestimate the standard errors

❶ Generating M imputed data sets: variance of prediction



❷ Performing the analysis on each imputed data set[11], [12]

❸ Combining: variance = within + between imputation variance
$\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m \;\; T = \frac{1}{M}\sum \widehat{Var}\left(\hat{\beta}_m\right) + \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum\left(\hat{\beta}_m - \hat{\beta}\right)^2$

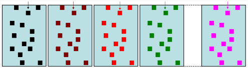
[11] The analysis model may be "in agreement" with the imputation model: congeniality.
[12] Little & Rubin. 2019. Statistical Analysis with Missing Data, 3rd Edition. Wiley

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[11], [12]

3. Combining: variance = within + between imputation variance
$\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m \;\; T = \frac{1}{M}\sum \widehat{Var}\left(\hat{\beta}_m\right) + \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum\left(\hat{\beta}_m - \hat{\beta}\right)^2$

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

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

Joint modeling

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

Algorithm Expectation Maximization Bootstrap:

1. Bootstrap rows: $Z^1, \ldots, Z^M$
   EM algorithm: $(\hat{\mu}^1, \hat{\Sigma}^1), \ldots, (\hat{\mu}^M, \hat{\Sigma}^M)$
2. Imputation: $\hat{z}^m_{imiss}$ drawn from $\mathcal{N}\left(\hat{\mu}^m, \hat{\Sigma}^m\right)$

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

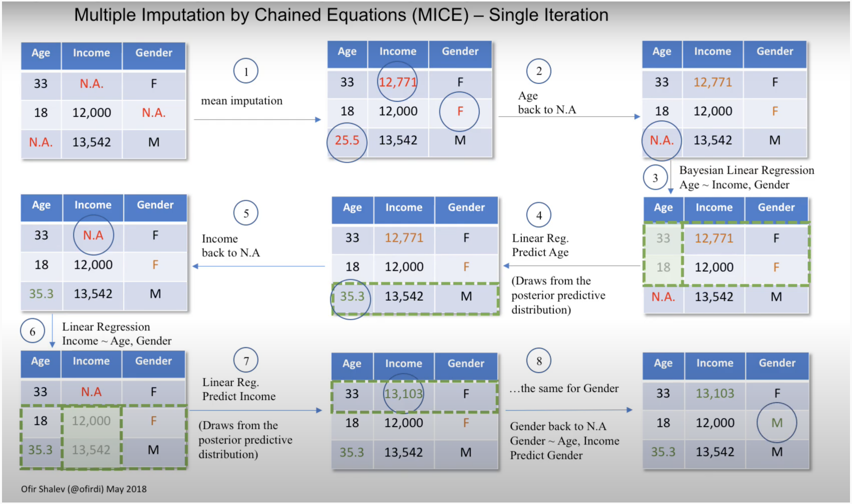
Amelia Earhart

James Honaker

Gary King

Matt Blackwell

Fully conditional modeling [13]



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

Fully conditional modeling: one model/variable

1. Initial imputation: mean imputation
2. For a variable j
   2.1 $(\hat{\beta}^{-j}, \hat{\sigma}^{-j})$ drawn from a Bootstrap: $(\hat{\beta}^{-j}, \hat{\sigma}^{-j})^1, \ldots, (\hat{\beta}^{-j}, \hat{\sigma}^{-j})^M$
   2.2 Imputation of the missing values in variable j with a model of X_j on the other X_{-j}: stochastic regression imputation from $\mathcal{N}\left((x_{i,-j})'\hat{\beta}^{-j}, \hat{\sigma}^{-j}\right)$
3. Cycling through variables

$\Rightarrow$ Iteratively refine the imputation.

$\Rightarrow$ With continuous variables & regression/variable: gibbs $\mathcal{N}(\mu, \Sigma)$ [14] [15]

Implemented in `mice` (website) and Python*

"There is no clear-cut method for determining whether the MICE algorithm has converged"

Stef van Buuren

* IterativeImputer by default does single imputation with iterative ridge regression

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

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

Single Iterative Random Forests Imputation [16]

1. Initial imputation: mean imputation - random category
   Sort the variables according to the amount of missing values
2. Fit a RF X_j^{obs} on variables X_{-j}^{obs} and then predict X_j^{miss}
3. Cycling through variables
4. Repeat step 2.2 and 3 until convergence

- number of trees: 100
- number of variables randomly selected at each node $\sqrt{d}$
- number of iterations: 4-5

Implemented in the R package `missForest`

[16] Stekhoven, Buhlmann. 2012. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*

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... tedious? **Computational costly**.

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

Outline

Ignorable missing values mechanism

- The full joint data distribution of (Z, M) with density $p(z, m|\theta, \phi)$ [17]

- The (full) observed distribution [18] :

$$\begin{aligned} p(z_{\text{obs}}, m; \theta, \phi) &= \int p(z, m; \theta, \phi) dz_{\text{mis}} \\ &= \int p(z; \theta) p(m|z; \phi) dz_{\text{mis}} \end{aligned}$$

- With M(C)AR data:

$$\begin{aligned} p(z_{\text{obs}}, m; \theta, \phi) &= \int p(z; \theta) p(m|z_{\text{obs}}; \phi) dz_{\text{mis}}, \\ &= p(m|z_{\text{obs}}; \phi) \int p(z; \theta) dz_{\text{miss}}, \\ &= p(m|z_{\text{obs}}; \phi) p(z_{\text{obs}}; \theta). \end{aligned}$$

$\Rightarrow$ Likelihood inference can be based on $p(z_{\text{obs}}; \theta)$

[17] We assume separability of θ and ϕ
[18] $z_{\text{obs}}(m)$ is denoted z_{obs}

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.

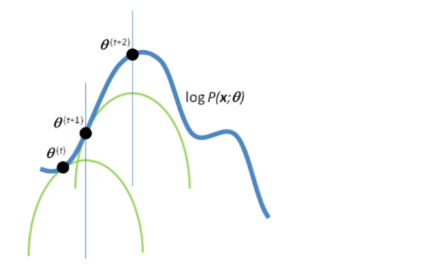
E step (conditional expectation):

$$Q(\theta, \theta^{\ell}) = \int log(p(z;\theta))p(z_{miss}|z_{obs};\theta^{\ell})dz_{miss}$$

M step (maximization):

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

Result: when $\theta^{\ell+1}$ max $Q(\theta, \theta^{\ell})$ then $L_{obs}(\theta^{\ell+1}) \geq L_{obs}(\theta^{\ell})$.



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

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

$\Rightarrow$ Variances:

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

Logistic regression with missing covariates: Parameter estimation, model selection and prediction (Jiang, J., et al, CSDA, 2018)

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

Logistic regression model: $\mathbb{P}(y_i = 1 | x_i; \beta) = \frac{\exp(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij})}$

Covariables: $x_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}_d(\mu, \Sigma)$

Log-likelihood with $\theta = (\mu, \Sigma, \beta)$:

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

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

Logistic regression with missing covariates: Parameter estimation, model selection and prediction (Jiang, J., et al, CSDA, 2018)

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

Logistic regression model: $\mathbb{P}(y_i = 1 | x_i; \beta) = \frac{\exp(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij})}$

Covariables: $x_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}_d(\mu, \Sigma)$

Log-likelihood with $\theta = (\mu, \Sigma, \beta)$:

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

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

Stochastic Approximation EM - package misaem

argmax $\mathcal{LL}(\theta; x_{\rm obs}, y) = \int \mathcal{LL}(\theta; x, y) dx_{\rm mis}$

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

$$\begin{aligned} Q(\theta, \theta^\ell) &= \mathbb{E}[\mathcal{LL}(\theta; x, y) | x_{\rm obs}, y; \theta^\ell] \\ &= \int \mathcal{LL}(\theta; x, y) \mathrm{p}(x_{\rm mis} | x_{\rm obs}, y; \theta^\ell) dx_{\rm mis} \end{aligned}$$

- **M-step:** $\theta^{\ell+1} = \text{argmax}_\theta Q(\theta, \theta^\ell)$

$\Rightarrow$ *Unfeasible computation of expectation*

MCEM (Wei & Tanner, 1990): Generate samples of missing data from $\mathrm{p}(x_{\rm mis} | x_{\rm obs}, y; \theta^\ell)$ and replace the expectation by an empirical mean

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

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

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

Stochastic Approximation EM

Starting from an initial guess θ_0, the kth iteration consists of three steps:

- **Simulation:** For $i = 1, 2, \cdots, n$, draw one sample $x_{i,\mathrm{mis}}^{(k)}$ from

$$\mathrm{p}(x_{i,\mathrm{mis}} | x_{i,\mathrm{obs}}, y_i; \theta_{k-1}).$$

- **Stochastic approximation:** Update the function Q

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left(\mathcal{LL}(\theta; x_{\mathrm{obs}}, x_{\mathrm{mis}}^{(k)}, y) - Q_{k-1}(\theta) \right),$$

  where (γ_k) is a decreasing sequence of positive numbers.

- **Maximization:** $\theta_k = \mathrm{argmax}_\theta\, Q_k(\theta)$.

Stochastic Approximation EM

Starting from an initial guess θ_0, the kth iteration consists of three steps:

- **Simulation:** For $i = 1, 2, \cdots, n$, draw one sample $x_{i,\mathrm{mis}}^{(k)}$ from

$$\mathrm{p}(x_{i,\mathrm{mis}} | x_{i,\mathrm{obs}}, y_i; \theta_{k-1}).$$

- **Stochastic approximation:** Update the function Q

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left(\mathcal{LL}(\theta; x_{\mathrm{obs}}, x_{\mathrm{mis}}^{(k)}, y) - Q_{k-1}(\theta) \right),$$

  where (γ_k) is a decreasing sequence of positive numbers.

- **Maximization:** $\theta_k = \mathrm{argmax}_\theta\, Q_k(\theta)$.

Convergence: Allassonniere et al. 2010)
The choice of the sequence (γ_k) is important for ensuring the almost sure convergence of SAEM to a MLE.

Metropolis-Hastings algorithm

Target distribution

$$\begin{aligned} f_i(x_{i,\mathrm{mis}}) &= \mathtt{p}(x_{i,\mathrm{mis}}|x_{i,\mathrm{obs}}, y_i; \theta) \\ &\propto \mathtt{p}(y_i|x_i; \beta)\, \color{red}{\mathtt{p}(x_{i,\mathrm{mis}}|x_{i,\mathrm{obs}}; \mu, \Sigma)}. \end{aligned}$$

Metropolis-Hastings algorithm

Target distribution

$$\begin{aligned} f_i(x_{i,\mathrm{mis}}) &= \mathrm{p}(x_{i,\mathrm{mis}}|x_{i,\mathrm{obs}}, y_i; \theta) \\ &\propto \mathrm{p}(y_i|x_i; \beta)\, \mathrm{p}(x_{i,\mathrm{mis}}|x_{i,\mathrm{obs}}; \mu, \Sigma). \end{aligned}$$

Proposal distribution $g_i(x_{i,\mathrm{mis}}) = \mathrm{p}(x_{i,\mathrm{mis}}|x_{i,\mathrm{obs}}; \mu, \Sigma)$

$$x_{i,\mathrm{mis}}|x_{i,\mathrm{obs}} \sim \mathcal{N}_p(\mu_i, \Sigma_i)$$

$$\mu_i = \mu_{i,\mathrm{mis}} + \Sigma_{i,\mathrm{mis},\mathrm{obs}}\Sigma^{-1}_{i,\mathrm{obs},\mathrm{obs}}(x_{i,\mathrm{obs}} - \mu_{i,\mathrm{obs}}),$$

$$\Sigma_i = \Sigma_{i,\mathrm{mis},\mathrm{mis}} - \Sigma_{i,\mathrm{mis},\mathrm{obs}}\Sigma^{-1}_{i,\mathrm{obs},\mathrm{obs}}\Sigma_{i,\mathrm{obs},\mathrm{mis}},$$

Metropolis-Hastings algorithm

Target distribution

$$\begin{aligned} f_i(x_{i,\mathrm{mis}}) &= \mathrm{p}(x_{i,\mathrm{mis}} | x_{i,\mathrm{obs}}, y_i; \theta) \\ &\propto \mathrm{p}(y_i | x_i; \beta)\, \mathrm{p}(x_{i,\mathrm{mis}} | x_{i,\mathrm{obs}}; \mu, \Sigma). \end{aligned}$$

Proposal distribution $g_i(x_{i,\mathrm{mis}}) = \mathrm{p}(x_{i,\mathrm{mis}} | x_{i,\mathrm{obs}}; \mu, \Sigma)$

$$x_{i,\mathrm{mis}} | x_{i,\mathrm{obs}} \sim \mathcal{N}_p(\mu_i, \Sigma_i)$$

$$\begin{aligned} \mu_i &= \mu_{i,\mathrm{mis}} + \Sigma_{i,\mathrm{mis},\mathrm{obs}} \Sigma_{i,\mathrm{obs},\mathrm{obs}}^{-1} (x_{i,\mathrm{obs}} - \mu_{i,\mathrm{obs}}), \\ \Sigma_i &= \Sigma_{i,\mathrm{mis},\mathrm{mis}} - \Sigma_{i,\mathrm{mis},\mathrm{obs}} \Sigma_{i,\mathrm{obs},\mathrm{obs}}^{-1} \Sigma_{i,\mathrm{obs},\mathrm{mis}}, \end{aligned}$$

Metropolis

- $z_{im}^{(k)} \sim g_i(x_{i,mis}),\ u \sim \mathcal{U}[0,1]$
- $r = \frac{f_i(z_{im}^{(k)}) / g_i(z_{im}^{(k)})}{f_i(z_{i,m-1}^{(k)}) / g_i(z_{i,m-1}^{(k)})}$
- If $u < r$, accept $z_{im}^{(k)}$

Only need a few steps of Markov chains in each iteration of SAEM!

Variance estimation

Observed Fisher information matrix (FIM) *wrt* β

$$\mathcal{I}(\theta) = -\frac{\partial^2 \mathcal{LL}(\theta; x_{\mathrm{obs}}, y)}{\partial\theta\partial\theta^T}.$$

Variance estimation

Observed Fisher information matrix (FIM) *wrt* β

$$\mathcal{I}(\theta) = -\frac{\partial^2 \mathcal{LL}(\theta; x_{\text{obs}}, y)}{\partial\theta\partial\theta^T}.$$

Louis formula

$$\begin{aligned}
\mathcal{I}(\theta) = & -\mathbb{E}\left(\frac{\partial^2 \mathcal{LL}(\theta; x, y)}{\partial\theta\partial\theta^T}\Big| x_{\text{obs}}, y; \theta\right) \\
& -\mathbb{E}\left(\frac{\partial \mathcal{LL}(\theta; x, y)}{\partial\theta}\frac{\partial \mathcal{LL}(\theta; x, y)^T}{\partial\theta}\Big| x_{\text{obs}}, y; \theta\right) \\
& +\mathbb{E}\left(\frac{\partial \mathcal{LL}(\theta; x, y)}{\partial\theta}| x_{\text{obs}}, y; \theta\right)\mathbb{E}\left(\frac{\partial \mathcal{LL}(\theta; x, y)}{\partial\theta}| x_{\text{obs}}, y; \theta\right)^T.
\end{aligned}$$

Given the MH samples of unobserved data $(x_{i,\text{mis}}^{(m)}, 1 \le i \le n, 1 \le m \le M)$, and the SAEM estimate $\hat{\theta}$

$\Rightarrow$ Estimate FIM by empirical means.

Model selection: criterion BIC

With $\tilde{p}_\theta$ the number of estimated parameters in a given model $\mathcal{M}$, model selection criterion (*penalized likelihood*) :

$$\mathrm{BIC}(\mathcal{M}) = -2\mathcal{L}\mathcal{L}(\hat{\theta}_\mathcal{M}; x_{\mathrm{obs}}, y) + \log(n) d(\mathcal{M}),$$

How to estimate *observed likelihood* ?

Model selection: criterion BIC

With $\tilde{p}_\theta$ the number of estimated parameters in a given model $\mathcal{M}$, model selection criterion (*penalized likelihood*) :

$$\mathrm{BIC}(\mathcal{M}) = -2\mathcal{LL}(\hat{\theta}_{\mathcal{M}}; x_{\mathrm{obs}}, y) + \log(n) d(\mathcal{M}),$$

How to estimate *observed likelihood* ?

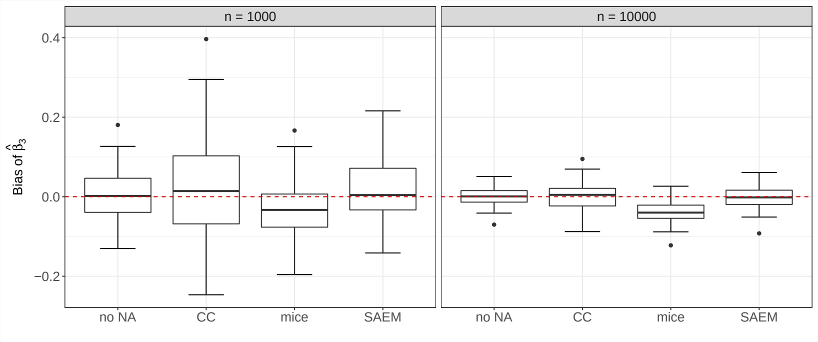
$$\begin{aligned}
\mathtt{p}(y_i, x_{i,\mathrm{obs}}; \theta) &= \int \mathtt{p}(y_i, x_{i,\mathrm{obs}} | x_{i,\mathrm{mis}}; \theta) \mathtt{p}(x_{i,\mathrm{mis}}; \theta) dx_{i,\mathrm{mis}} \\
&= \int \mathtt{p}(y_i, x_{i,\mathrm{obs}} | x_{i,\mathrm{mis}}; \theta) \frac{\mathtt{p}(x_{i,\mathrm{mis}}; \theta)}{g_i(x_{i,\mathrm{mis}})} g_i(x_{i,\mathrm{mis}}) dx_{i,\mathrm{mis}} \\
&= \mathbb{E}_{g_i} \left(\mathtt{p}(y_i, x_{i,\mathrm{obs}} | x_{i,\mathrm{mis}}; \theta) \frac{\mathtt{p}(x_{i,\mathrm{mis}}; \theta)}{g_i(x_{i,\mathrm{mis}})} \right).
\end{aligned}$$

Sample from g_i (the proposal distribution in SAEM)

$\Rightarrow$ Empirical mean.

Comparison with competitors: estimates

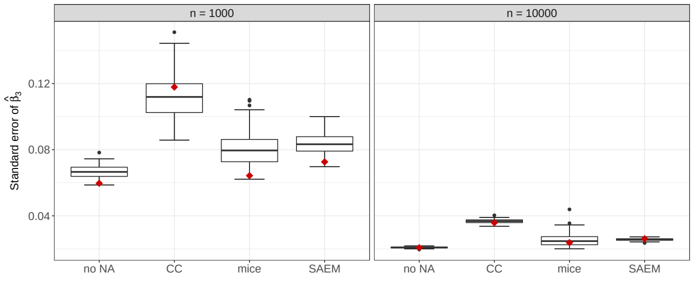
x: $d = 5$, $n = 1000$ / $n = 10\,000 \Rightarrow y \in \{0, 1\}$
percentage of missingness = 10%.
Repeat 1000 times for each setting.



Comparison with competitors: coverage

Table 1: Coverage (%) for $n = 10\,000$, calculated over 1000 simulations.

parameter	no NA	CC	mice	SAEM
β_0	95.2	94.4	95.2	94.9
β_1	96.0	94.7	93.9	95.1
β_2	95.5	94.6	94.0	94.3
β_3	94.9	94.3	*86.5*	94.7
β_4	94.6	94.2	96.2	95.4
β_5	95.9	94.4	*89.6*	94.7



Comparison with competitors: execution time

Table 2: Comparison of execution time between no NA, MCEM, mice, and SAEM with $n = 1000$ calculated over 1000 simulations.

Execution time (seconds)	no NA	MCEM	mice	SAEM
min	2.87×10^{-3}	492	0.64	9.96
mean	4.65×10^{-3}	773	0.70	13.50
max	43.50×10^{-3}	1077	0.76	16.79

Application on TraumaBase

- 6384 patients, 14 variables, percentage of NA from 0 to *60%*
- Prediction of hemoragic shock
- Selection of 8 variables, interpretation of coefficients (age, low glasgow score positive effect)

```
> library(misaem)
> reg <- miss.glm(y~., data = don)
> regBIC <- miss.glm.model.select(don$y, subset(don,-c("y")))
> pr.saem <- predict(reg, newdata = dontest)
```

Take home message inference/imputation

- Few implementation of EM strategies

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

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

Take home message inference/imputation

⇒ Challenges with multiple imputation

- Multiple imputation in high dimension ?
- Aggregating lasso regressions
- Aggregating different models
- Theory with other asymptotic small n, large p ?

⇒ Other contributions:

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

Muzelec, Cuturi, Boyer, J. 2020. Missing Data Imputation using Optimal Transport. *ICML*.

Outline

Outline

PCA (complete)

Find the subspace that best represents the data

Figure 2: Camel or dromedary?

⇒ Best approximation with projection
⇒ Best representation of the variability
⇒ Do not distort the distances between observations

PCA (complete)

Find the subspace that best represents the data

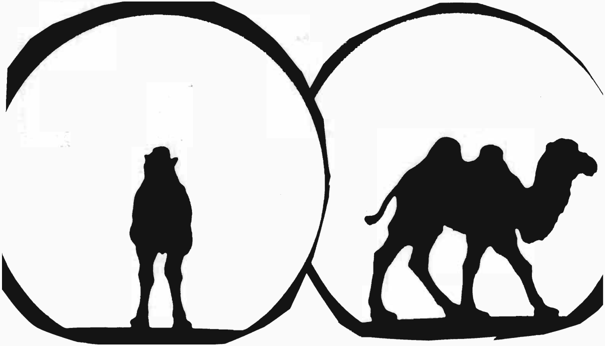


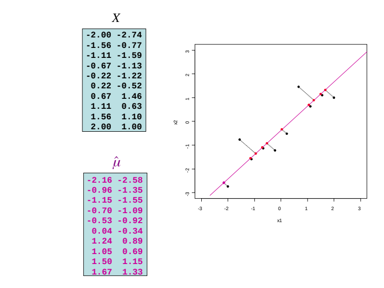
Figure 2: Camel or dromedary? source J.P. Fénelon

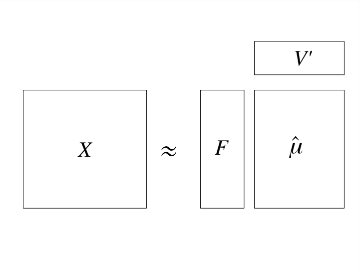
⇒ Best approximation with projection

⇒ Best representation of the variability

⇒ Do not distort the distances between observations

PCA reconstruction



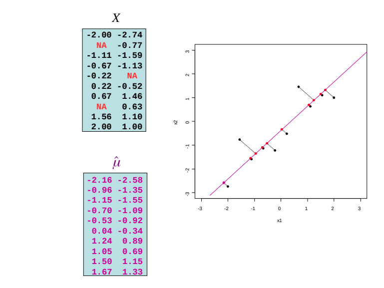


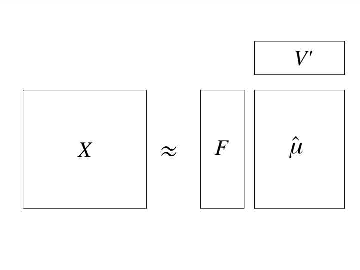
$\Rightarrow$ Minimizes distance between observations and their projection

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

$$\mathrm{argmin}_{\mu} \left\{ \|X - \mu\|_2^2 : \mathrm{rank}\,(\mu) \leq S \right\}$$

PCA reconstruction





$\Rightarrow$ Minimizes distance between observations and their projection

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

$$\mathrm{argmin}_\mu \left\{ \|X - \mu\|_2^2 : \mathrm{rank}\,(\mu) \le S \right\}$$

SVD X: $\hat{\mu}^{\mathrm{PCA}} = U_{n\times S}\Lambda^{\frac{1}{2}}_{S\times S}V'_{p\times S}$ $\quad F = U\Lambda^{\frac{1}{2}}$ PC - scores

$= F_{n\times S}V'_{p\times S}$ $\quad V$ principal axes - loadings

Missing values in PCA

$\Rightarrow$ PCA: least squares

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

$\Rightarrow$ PCA with missing values: weighted least squares

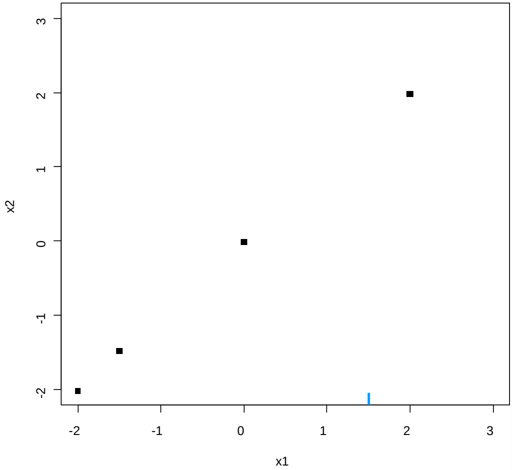
$$\text{argmin}_\mu \left\{ \|W_{n\times p} \odot (X - \mu)\|_2^2 : \text{rank}\,(\mu) \leq S \right\}$$

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

Many algorithms: weighted alternating least squares (Gabriel & Zamir, 1979); iterative PCA (Kiers, 1997)

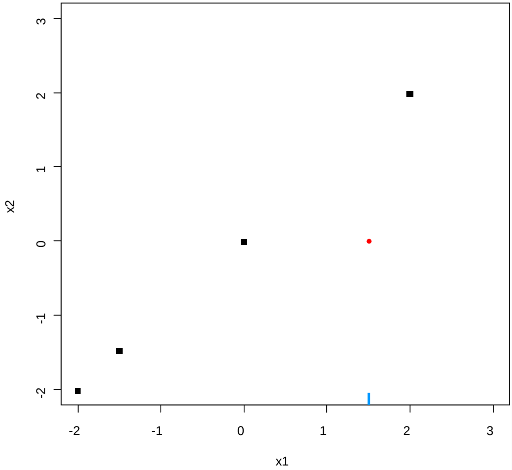
Iterative PCA

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
```



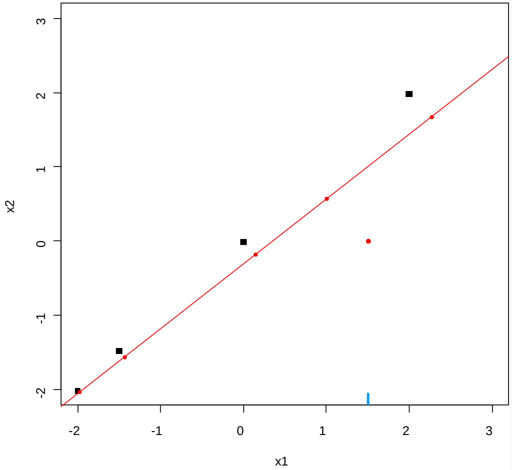
Initialization $\ell = 0$: X^0 (mean imputation)

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



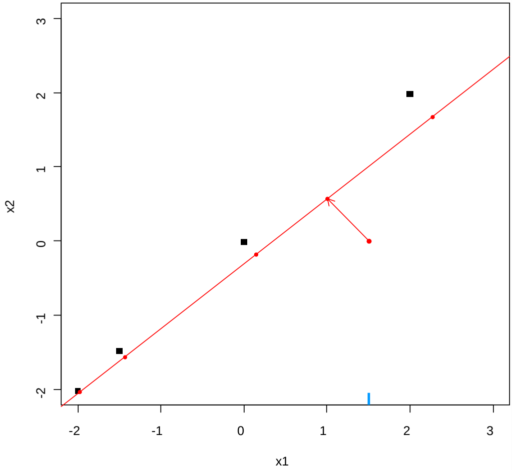
PCA on the completed data set $\rightarrow (U^\ell, \Lambda^\ell, V^\ell)$;

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 $\hat{\mu}^{\ell} = U^{\ell} \Lambda^{1/2^{\ell}} V^{\ell\prime}$

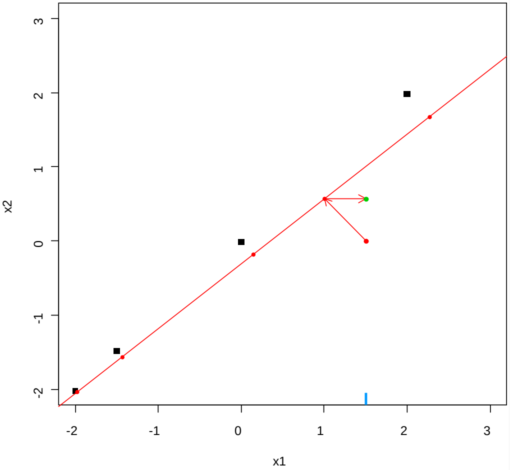
Iterative PCA

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



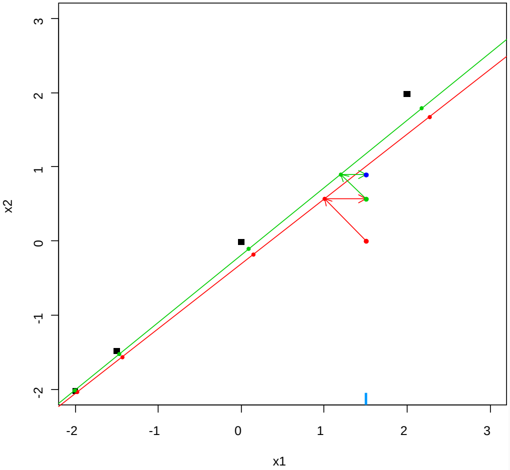
The new imputed dataset is $\hat{X}^\ell = W \odot X + (\mathbf{1} - W) \odot \hat{\mu}^\ell$

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
```



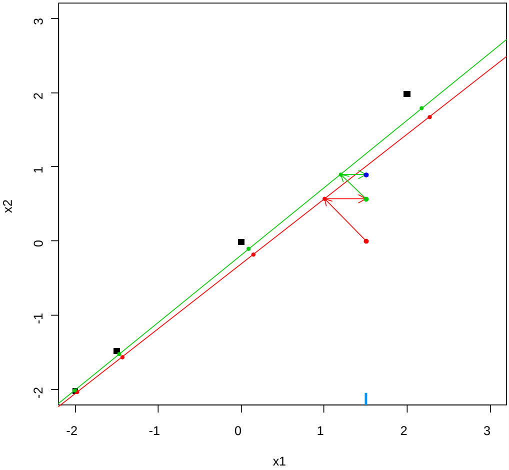
Iterative PCA

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

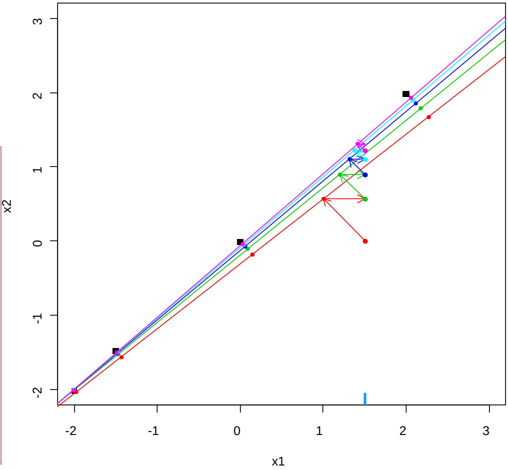


x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98

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

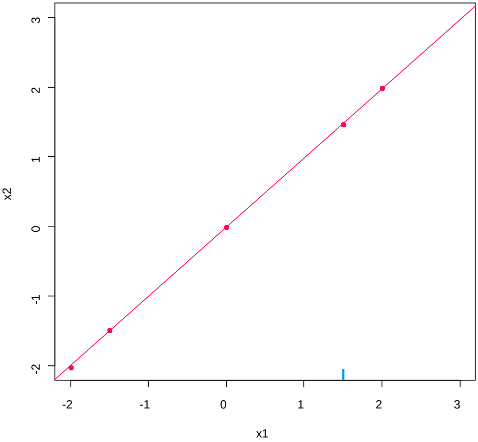
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



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 dimensions 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

[19] 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 dimensions 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

$\Rightarrow$ $\hat{\mu}$ 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)

[19] 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 dimensions 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

$\Rightarrow$ $\hat{\mu}$ 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? Generalized cross-validation [19]

[19] 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 [20]

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

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

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

Implemented in `softImpute`

[20] 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 [21],[22], [23] :

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

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

σ^2 large $\rightarrow$ mean imputation

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

Implemented in `missMDA` (Youtube link)

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

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

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

Properties

⇒ 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 [24]: Data = structure of rank S + noise

⇒ Different noise regime [25], [26]

- 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` [27]

Imputed data should be analysed with caution by other methods

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

Random Forests versus PCA

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

Random forests versus PCA

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

Missing

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

`missForest`

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

`imputePCA`

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

Multiple imputation with Bootstrap PCA [28]

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

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

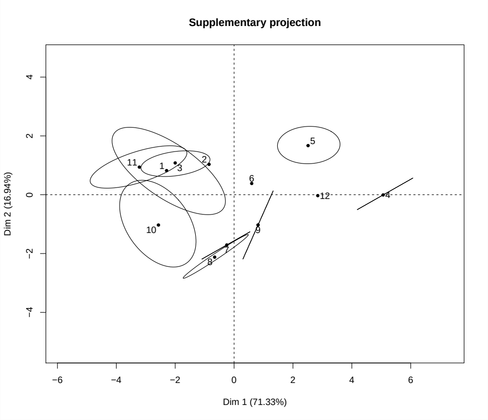
[28] J. Pages. Husson. 2011. Multiple imputation in principal component analysis. *ADAC.*

Visualization of the imputed values

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

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

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




```
library(missMDA)
MIPCA(traumadata)
```

Percentage of NA?

Joint, conditional and PCA

$\Rightarrow$ 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

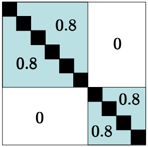
`Amelia` & `mice` have difficulties with large correlations or $n < p$
`missMDA` does not but requires a tuning parameter: number of dim.

`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

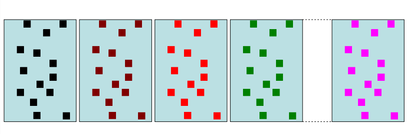
Simulations

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, JM, CM



$\Rightarrow$ 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

Outline

Incomplete ozone

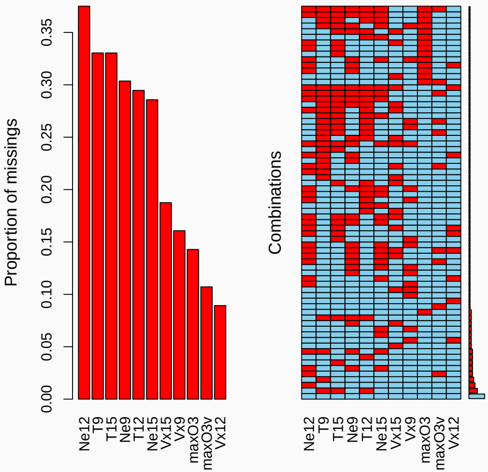
	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
0601	87	15.6	18.5	18.4	4	4	8	NA	-1.7101	-0.6946	84
0602	82	NA	18.4	17.7	5	5	7	NA	NA	NA	87
0603	92	NA	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82
0604	114	16.2	NA	NA	1	1	0	NA	NA	NA	92
0605	94	17.4	20.5	NA	8	8	7	-0.5	NA	-4.3301	114
0606	80	17.7	NA	18.3	NA	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

Complete ozone

```
         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
```

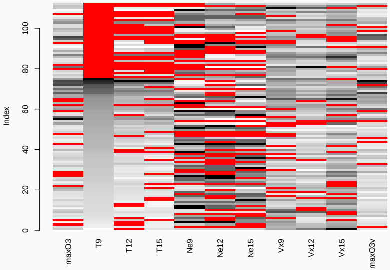
```
> library(missMDA)
> res.comp <- imputePCA(ozo[, 1:11])
> res.comp$comp
```

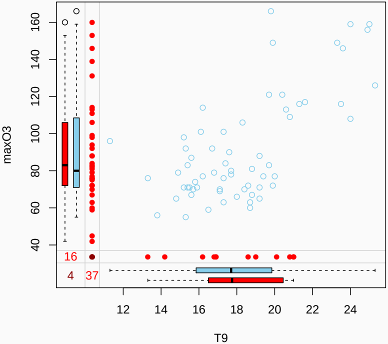
Pattern visualization




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

Visualization





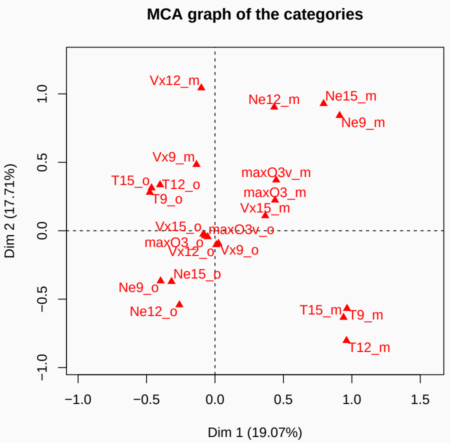

```
> library(VIM)
> matrixplot(don, sortby = 2)
> marginplot(don[ ,c("T9", "maxO3")])
```

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

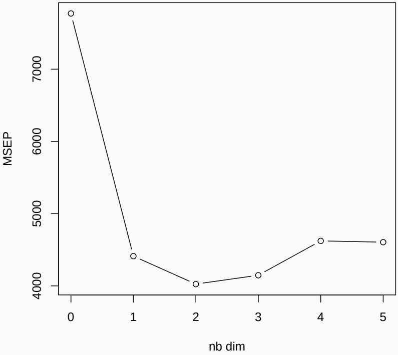



```
> library(FactoMineR)
> resMCA <- MCA(mis.ind)
> plot(resMCA, invis = "ind", title = "MCA graph of the categories")
```

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")
```

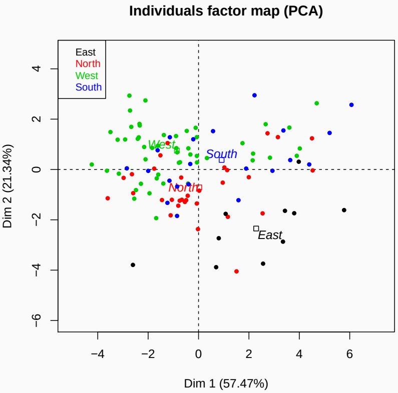


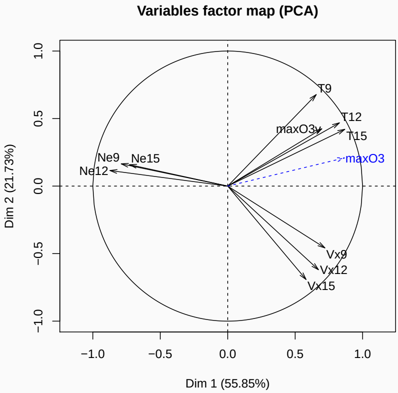
Imputation with PCA in practice

⇒ 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)
```

Imputation 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
```

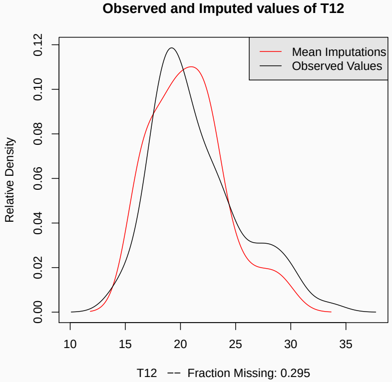
Multiple imputation in practice

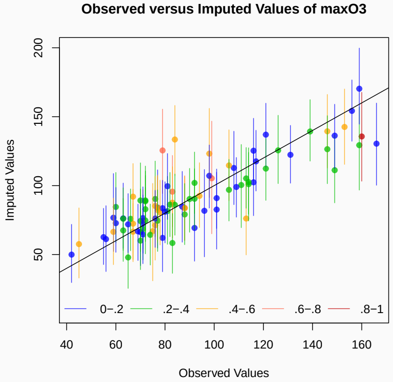
⇒ Step 1: Generate M imputed data sets

```
> 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



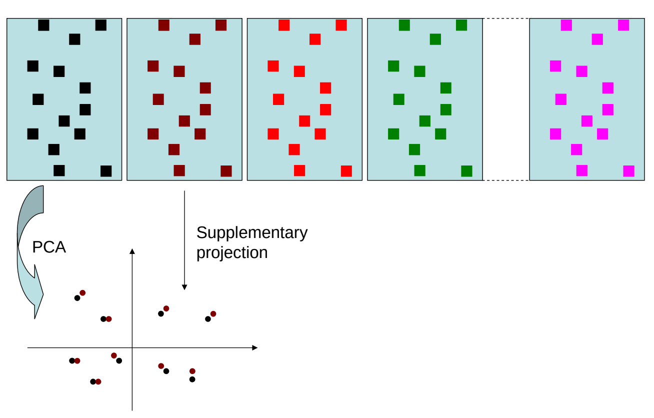



```
>  library(Amelia)
> res.amelia <- amelia(don, m = 100)
> compare.density(res.amelia, var = "T12")
> overimpute(res.amelia, var = "maxO3")

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

Multiple imputation in practice

⇒ Step 2: visualization

⇒ Individuals position (and variables) with other predictions



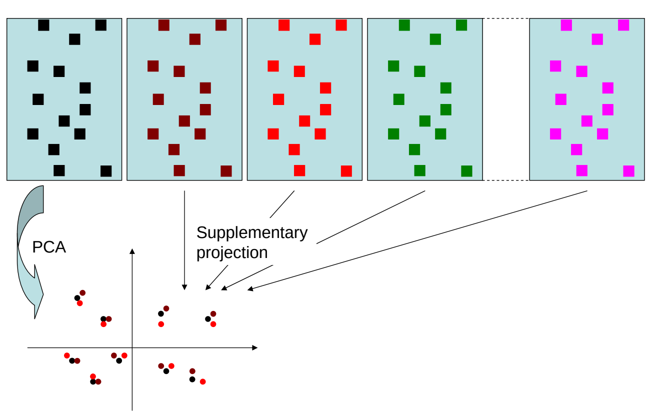
Regularized iterative PCA

⇒ reference configuration

Multiple imputation in practice

⇒ Step 2: visualization

⇒ Individuals position (and variables) with other predictions



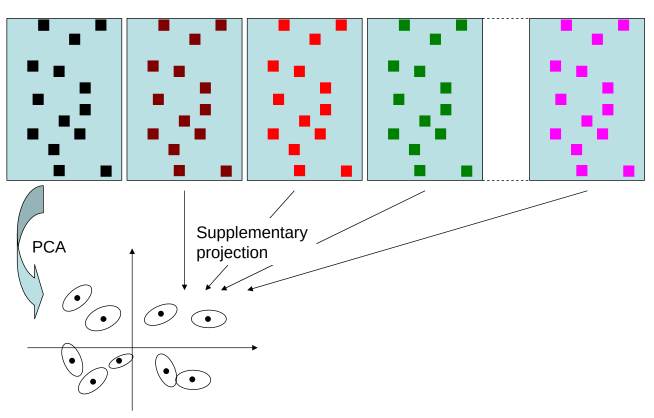
Regularized iterative PCA

⇒ reference configuration

Multiple imputation in practice

⇒ Step 2: visualization

⇒ Individuals position (and variables) with other predictions

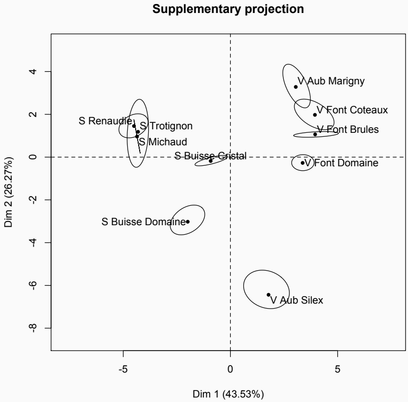


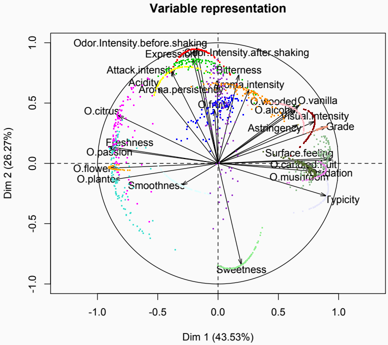
Regularized iterative PCA
⇒ reference configuration

Multiple imputation in practice

⇒ Step 2: visualization

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





Multiple imputation in practice

$\Rightarrow$ 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
```

Outline

Low rank estimation with MNAR data

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

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

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

- How to estimate μ ?
- How to impute the unknown entries of X ?

Data distribution

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

MNAR missing-data mechanism via a Logistic Model

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

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

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

Method 1: EM algo with MNAR (self-mask logistic)[29]

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

$$\hat{\mu} \in \mathrm{argmin}_{\mu} \|(X-\mu) \odot M\|_2^2 + \lambda \|\mu\|_\star,$$

Algo: iterative soft-thresholding SVD (ISTA), **accelerated version: FISTA**

MNAR (non ignorable) $L(\mu, \phi; x_{\mathrm{obs}}, m) = \int p(x;\mu) p(m|x;\phi) dx_{\mathrm{mis}}$.

- **E-step:**
$$Q(\mu, \phi | \hat{\mu}^{(/ell)}, \hat{\phi}^{(\ell)}) = -\mathbb{E}_{X_{\mathrm{mis}}} \left[\ell(\mu, \phi; x, \mu) | X_{\mathrm{obs}}, M; \mu = \hat{\mu}^{(\ell)}, \phi = \hat{\phi}^{(t)} \right]$$
- **M-step:**
$$\hat{\mu}^{(\ell+1)}, \hat{\phi}^{(\ell+1)} \in \mathrm{argmin}_{\mu,\phi} Q(\mu, \phi | \hat{\mu}^{(t)}, \hat{\phi}^{(\ell)}) + \lambda \|\mu\|_\star$$

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

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

[29] Sportisse, Boyer, J. 2018. Low-rank estimation with missing non at random data. *Statistics & Computing.*

Method 2: implicitly modelling the mechanism

Add the mask !

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

Solve the classical MAR optimization problem

$$\hat{\Theta} \in \mathrm{argmin}_{\Theta} \frac{1}{2} \| \, [(1-M) \odot X | M] \, - [M|1] \odot \Theta \|_2^2 + \lambda \|\Theta\|_\star,$$

- `softImpute`, FISTA.
- taking into account the mask binary type, with a Penalized Iteratively Reweighted Least Squares algorithm [30].

Computationally efficient but no theoretical guaranties.

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

Outline

Categorical data

Questionnaire data from health institute `http://www.inpes.sante.fr`

```
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
```

Principal components method to explore categorical data: Multiple Correpondence Analysis[31]

[31] 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: $p_c = n_c/n$.

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

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 [32]

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
```

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

Regularized iterative MCA [32]

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
```

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

Regularized iterative MCA [32]

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
```

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

Regularized iterative MCA [32]

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
```

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

Regularized iterative MCA [32]

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
```

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

Regularized iterative MCA [32]

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
```

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

Regularized iterative MCA [32]

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
```

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

Multiple imputation with MCA[33]

❶ 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

❷ Categories drawn from multinomial disribution 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()
```

[33] Audigier, Husson, J. MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis. 2017. *Statistic & 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)

$\Rightarrow$ 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

Categorical imputation in practice

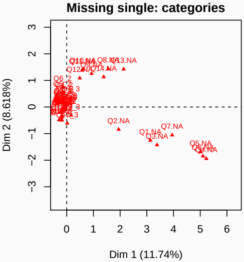
- 1232 respondents, 14 questions, 35 categories, 9% of missing values concerning 42% of respondents

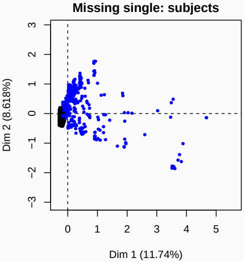
In `missMDA` (Youtube)

```
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 in practice

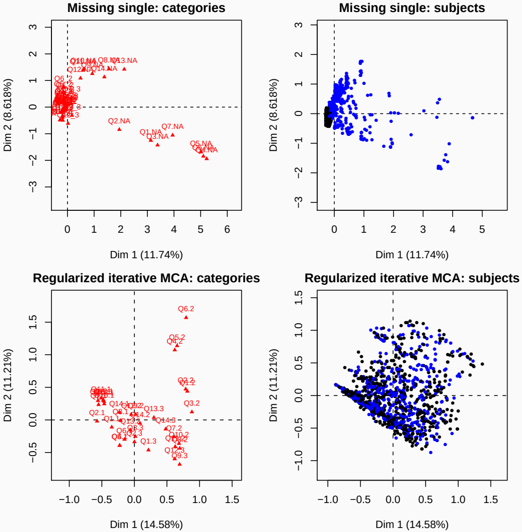
- 1232 respondents, 14 questions, 35 categories, 9% of missing values concerning 42% of respondents





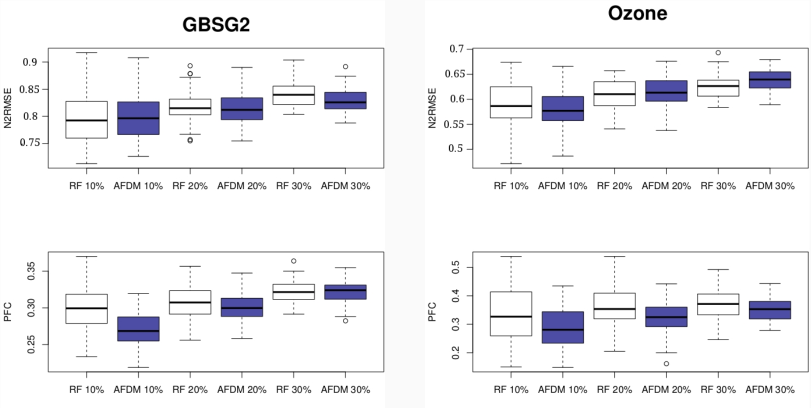
Categorical imputation in practice

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



Comparison with respect to the imputation

- Mixed data: imputation with Factorial Analysis for Mixed Data[34] FAMD.[35]
- Comparison with Random Forest imputation with RMSE for continuous data & proportion of falsely classified entries for categorical data.



[34] F. Husson, et. al. 2017. Exploratory Multivariate Analysis by Example Using R. Chapman & Hall.

Comparison with respect to the imputation

Imputations with PC methods are appropriate

- for strong linear relationships
- for categorical variables
- especially for rare categories (weights of MCA)

$\Rightarrow$ Tuning: number of components S (Cross-Validation)

Imputations with RF are appropriate

- for strong non-linear relationships between continuous variables
- when there are interactions

$\Rightarrow$ No tuning parameters?

Rq: categorical data improve the imputation on continuous data and continuous data improve the imputation on categorical data

Comparison with respect to the imputation

Imputations with PC methods are appropriate

- for strong linear relationships
- for categorical variables
- especially for rare categories (weights of MCA)

$\Rightarrow$ Tuning: number of components S (Cross-Validation)

Imputations with RF are appropriate

- for strong non-linear relationships between continuous variables (cutting continuous variables into categories)
- when there are interactions (creating interactions)

$\Rightarrow$ No tuning parameters?

Rq: categorical data improve the imputation on continuous data and continuous data improve the imputation on categorical data

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
```

Missing values in multi-source, multi-scale data

		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	

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

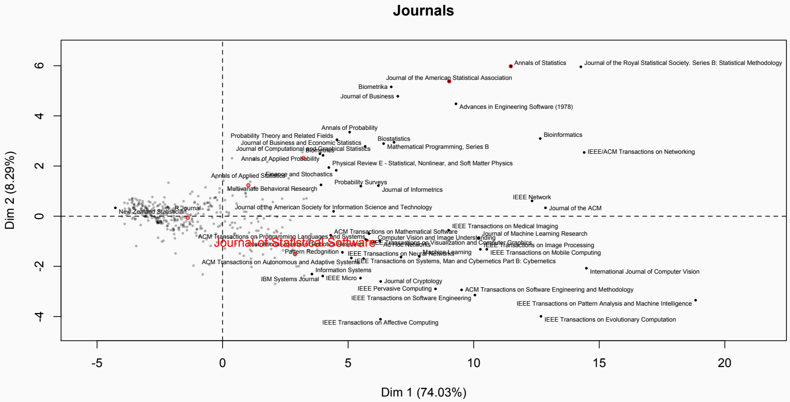
Data from journalmetrics.com

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.

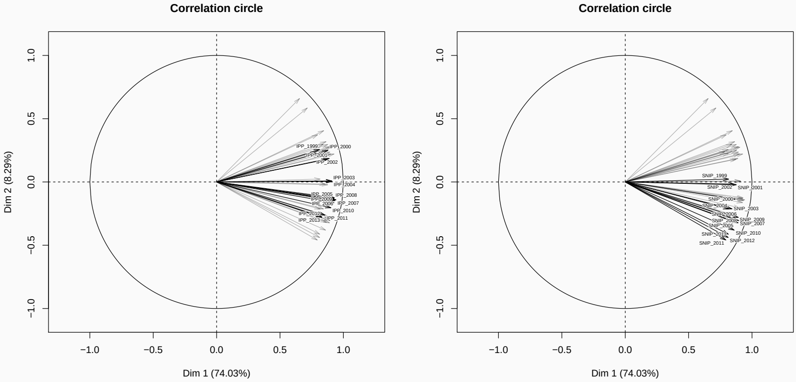
Many missing values per block of years.

Multiple Factor Analysis (MFA) with missing values [36]



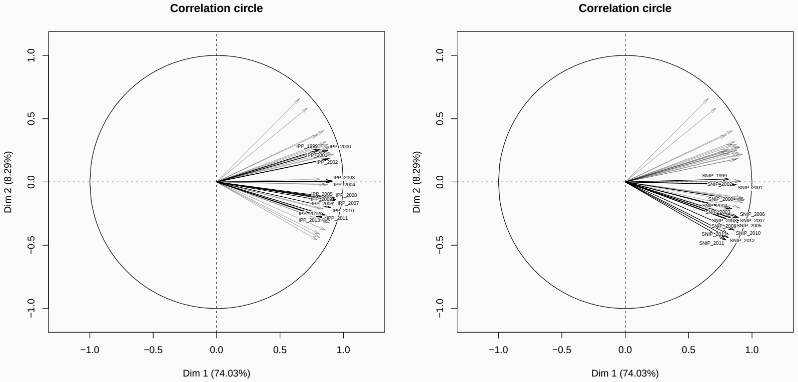
[36] Husson, J. 2013. Handling missing values in Multiple Factor Analysis. *FQP.*

Multiple Factor Analysis (MFA) with missing values [36]



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

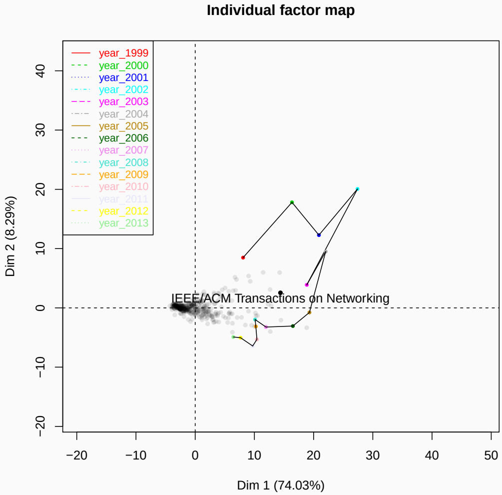
Multiple Factor Analysis (MFA) with missing values [36]



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

Multiple Factor Analysis (MFA) with missing values [36]

ACM Transactions on Networking trajectory



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

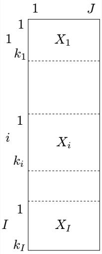
MFA imputation in practice

```
> 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)
```

Multilevel component analysis for group of observations

Ex: inhabitants nested within countries $X \in \mathbb{R}^{K \times J}$

- similarities between countries? level 1
- similarities between inhabitants within each country? level 2
- relationship between variables at each level



$$\begin{array}{rcl} x_{ijk_i} = x_{.j.} + (x_{ij.} - x_{.j.}) & + & (x_{ijk_i} - x_{ij.}) \\ \text{Between} & + & \text{Within} \end{array}$$

Analysis of variance: split the sum of squares for each variable j

$$\sum_{i=1}^{I} \sum_{k=1}^{k_i} (x_{ijk_i})^2 = \sum_{i=1}^{I} k_i (x_{.j.})^2 + \sum_{i=1}^{I} k_i (x_{ij.} - x_{.j.})^2 + \sum_{i=1}^{I} \sum_{k=1}^{k_i} (x_{ijk_i} - x_{ij.})^2$$

Multilevel PCA (MLPCA)

$\Rightarrow$ Model for the between and within part $i = 1, \dots, I$ groups, J var

$$X_{i_{(k_i \times J)}} = 1_{k_i} m' + 1_{k_i} F_i^{b'} V^{b'} + F_i^w V^{w'} + E_i$$

- F_i^b $(Q_b \times 1)$ between component scores of group i
- V^b $(J \times Q_b)$ between loading matrix
- F_i^w $(k_i \times Q_w)$ within component scores of group i
- V_w $(J \times Q_w)$ within loading matrix. Constant across groups

Fitted by minimizing the least squares [37]

$$\text{argmin}_{(m, F_i^b, V^b, F_i^w, V^w)} \sum_{i=1}^{I} \left\| X_i - 1_{k_i} m' - 1_{k_i} F_i^{b'} V^{b'} - F_i^w V^{w'} \right\|^2,$$

$\sum_{i=1}^{I} k_i F_i^b = 0_{Q_b}$ and $1'_{k_i} F_i^w = 0_{Q_w}$, $\forall i$ for identifiability.

[37] Timmerman. 2006. Multilevel component analysis. *Br J Math Stat Psychol.*

MLPCA - quantitative data

$i = 1, ..., I$ groups, J var, k_i nb obs in group i

$$\text{argmin}_{(m, F_i^b, V^b, F_i^w, V^w)} \sum_{i=1}^{I} \left\| X_i - 1_{k_i} m' - 1_{k_i} F_i^{b'} V^{b'} - F_i^w V^{w'} \right\|^2,$$

$\sum_{i=1}^{I} k_i F_i^b = 0_{Q_b}$ and $1'_{k_i} F_i^w = 0_{Q_w}$, $\forall i$ for identifiability.

$(\hat{F}^b, \hat{V}^b)$: Weigthed PCA on the between part: SVD on $D_w X_m$; X_m $(I \times J)$ the means of the variables per group, D_w $(I \times I)$ $D_{w\,ii} = \sqrt{k_i}$

$(\hat{F}^w, \hat{V}^w)$ PCA on the within part: SVD on the centered data per group X^w $(K \times J)$, $K = \sum_i k_i$

$\Rightarrow$ With missing values: Weighted Least Squares

$\Rightarrow$ Iterative imputation algorithm (imputation - estimation)

Iterative MLPCA

2. iteration ℓ: *estimation of the between structure*
   - SVD $D_w X_m^\ell = PDQ'$; Q_b eigenvectors are kept:
     $\hat{F}_i^b = [D_w^{-1} P_{Q_b}]_i$, $\hat{F}^b$ concatenation by row of $[\mathbf{1}_{k_i} \hat{F}_i^b]$
     $\hat{V}^b = Q_{Q_b} D_{Q_b}$, $(J \times Q_b)$
   - the between hat matrix is computed: $(\hat{X}^b)^\ell = \hat{F}^b \hat{V}^{b'}$
3. iteration ℓ: *imputation of the missing values with the fitted values*
   - $\hat{X}^\ell = \mathbf{1}_K \hat{m}^{(\ell-1)'} + (\hat{X}^b)^\ell + (\hat{X}^w)^{(\ell-1)}$. The newly imputed dataset is
     $X^\ell = W \odot X + (\mathbf{1}_K \times \mathbf{1}_J' - W) \odot \hat{X}^\ell$
   - $\hat{m}^\ell$ is computed on X^ℓ
4. iteration ℓ: *estimation of the within structure*
   - SVD $(X^w)^\ell = PDQ'$; Q_w eigenvectors are kept:
     $F^w = P_{Qw}$ $(K \times Q_w)$
     $V^w = Q_{Qw} D_{Qw}$ $(J \times Q_w)$
   - the within hat matrix is computed $(\hat{X}^w)^\ell = \hat{F}^w \hat{V}^{w'}$
5. iteration ℓ: *imputation of the missing values with the fitted values*
   - $X^{\ell+1} = W \odot X + (\mathbf{1}_K \times \mathbf{1}_J' - W) \odot \left(\mathbf{1}_K \hat{m}^{(\ell)'} + (\hat{X}^b)^\ell + (\hat{X}^w)^\ell\right)$
   - $\hat{m}^{\ell+1}$ is computed on $X^{\ell+1}$

Simulations design

The simulated data:

- $X_{i_{(k_i \times J)}} = 1_{k_i} m' + 1_{k_i} F_i^{b'} V^{b'} + F_i^{w} V^{w'} + E_i$, with $E_{ijk_i} \sim \mathcal{N}(0, \sigma)$
- 5 groups, 10 variables, $Q_b = 2$, $Q_w = 2$

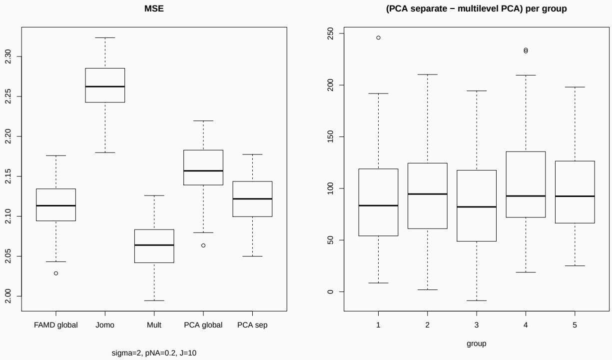
Many scenarios are considered:

- number of individuals per group: 10-20, 70-100
- level of noise: low, strong
- percentage of missing values: 10%, 25%, 40%
- missing values mechanism: MCAR, MAR

$\Rightarrow$ Prediction error: $\frac{1}{KJ} \sum (x_{ijk_i} - \hat{x_{ijk_i}})^2$

Comparison with competitors in terms of imputation

- Conditional model with random effect regression [38], implemented in micemd
- Random forests imputation
- Global PCA - Separate PCA on each table
- Global mixed PCA (FAMD) with hospital as a variable



[38] Audigier, White, Jolani, Debray, Quartagno, Carpenter, van Buuren, S. & Resche-Rigon. 2018. Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science.*

Comparison with competitors in terms of imputation

- PCA mixed give similar results than Random Forest
- mice (random effect model): difficulties with large dimensions
- Separate PCA: pb with many missing values
- Multilevel PCA is equivalent to global PCA when no group effect
- Other methods do not handle categorical variables

$\Rightarrow$ Multilevel PCA Computationaly fast in comparison to mice or RF.
Implemented R package missMDA

• Numbers of components Q_b and Q_w ?

• Inference after imputation. Underestimation of the variance with single imputation

Comparison with competitors in terms of imputation

- PCA mixed give similar results than Random Forest
- mice (random effect model): difficulties with large dimensions
- Separate PCA: pb with many missing values
- Multilevel PCA is equivalent to global PCA when no group effect
- Other methods do not handle categorical variables

$\Rightarrow$ Multilevel PCA Computationaly fast in comparison to mice or RF. Implemented R package missMDA

• Numbers of components Q_b and Q_w ?
cross-validation?

• Inference after imputation. Underestimation of the variance with single imputation

Comparison with competitors in terms of imputation

- PCA mixed give similar results than Random Forest
- mice (random effect model): difficulties with large dimensions
- Separate PCA: pb with many missing values
- Multilevel PCA is equivalent to global PCA when no group effect
- Other methods do not handle categorical variables

$\Rightarrow$ Multilevel PCA Computationaly fast in comparison to mice or RF.
Implemented R package missMDA

• Numbers of components Q_b and Q_w ?
cross-validation?

• Inference after imputation. Underestimation of the variance with single imputation
Multiple imputation: bootstrap + drawn from the predictive distribution
$\mathcal{N}\left(\mathbf{1}_K \hat{m}' + \hat{F}^b \hat{B}^{b'} + \hat{F}^w \hat{B}^{w'}, \hat{\sigma}^2\right)$

Aggregation of medical data

Combining data from different institutional databases promises many advantages in personalizing medical care (large n, more chance for finding patients like me)

Aggregation of medical data

Combining data from different institutional databases promises many advantages in personalizing medical care (large *n*, more chance for finding patients like me)

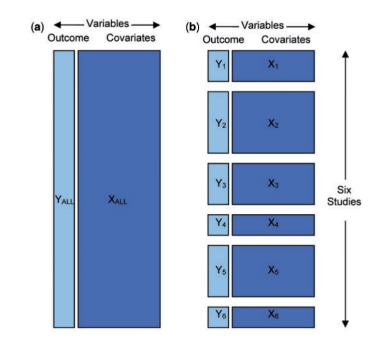
$\Rightarrow$ The problem: high barriers to aggregation of medical data

- lack of standardization of ontologies
- privacy concerns
- proprietary attitude towards data, reluctance to cede control
- complexity/size of aggregated data, updates problems

Solution: distributed computation

⇒ Data aggregation is not always necessary

⇒ NIH splits the storage of aggregated data across several centers



⇒ Data can stay at site

⇒ Computations can be distributed (share burden)

⇒ Hospitals only share intermediate results instead of the raw data

Topology: master-workers (Wolfson, et. al (2010))

⇒ Ex: Each site share the sum of age $\tilde{X}_i$ and the number of patients n_i.
The master computes $\bar{X} = \sum n_i \tilde{X}_i / \sum n_i$

Solution: distributed computation

⇒ Many models fitting can be implemented:

- Maximizing a likelihood. Intermediate computations break up into sums of quantities computed on local data at sites. Log-likelihood, score function and Fisher information can partition into sums. (OK for logistic regression)
- Singular Value Decomposition (ex power method involve inner product and sum). Iterative algorithms available for SVD using quantities computed on local data at sites.
- And more.

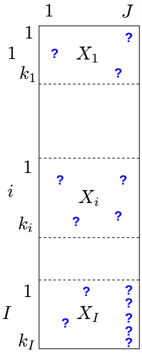
Implemented in the R package discomp[39]

[39] Narasimhan et. al. 2017. *Software for Distributed Computation on Medical Databases: A Demonstration Project.*

Privacy preserving rank k SVD

Data: each worker has private data $\boldsymbol{X}_i \in \mathcal{R}^{n_i \times p}$
Result: $V \in \mathcal{R}^{p \times k}$, and $d_1 \geq \ldots d_k \geq 0$
$V \leftarrow 0,\ d \leftarrow 0$ **foreach** *worker site* j **do**
    $U^{[j]} = 0$;
    transmit n_j to master;
end
for $i \leftarrow 1$ **to** k **do**
    foreach *worker site* j **do** $u^{[j]} \leftarrow (1, 1, \ldots, 1)$ of length n_j;
    $\|u\| \leftarrow \sqrt{\sum_j n_j}$;
    transmit $\|u\|, V$, and D to workers;
    repeat
        foreach *worker site* j **do**
            $u^{[j]} \leftarrow u^{[j]} / \|u\|$;
            calculate $v^{[j]} \leftarrow (\boldsymbol{X}^{[j]} - U^{[j]} D V^\top)^\top u^{[j]}$;
            transmit $v^{[j]}$ to master;
        end
        $v \leftarrow \sum_j v^{[j]}$;
        $v \leftarrow v / \|v\|$;
        transmit v to workers;
        foreach *worker site* j **do**
            calculate $u^{[j]} \leftarrow \boldsymbol{X}^{[j]} v$;
            transmit $\|u^{[j]}\|$ to master;
        end
        $\|u\| \leftarrow \sum_j \|u^{[j]}\|$;
        transmit $\|u\|$ to workers;
        $d_i \leftarrow \|u\|$;
    until *convergence*;
    $V \leftarrow \texttt{cbind}(V, v)$;
    foreach *worker site* j **do** $U^{[j]} \leftarrow \texttt{cbind}(U^{[j]}, u^{[j]})$;
end

Iterative multilevel distributed imputation (distributed iterative MLPCA) [40]



⇒ Impute the data of one hospital using the data of the others

⇒ Incentive to encourage the hospitals to participate in the project

[40] Robin, Husson, Narasimhan, J. (2018). Imputation of mixed data with multilevel singular value decomposition *JCGS*

Low rank matrix completion for heterogeneous (count data)

- Robin, J., Moulines & Sardy. Low-rank model with covariates for count data with missing values. 2019. *Journal of Multivariate Analysis* (slides)
- Robin, Klopp, J., Moulines & Tibshirani. Main effects and interactions in mixed and incomplete data frames. 2019. *JASA*.
- Sportisse, Boyer, J. Estimation and imputation in Probabilistic Principal Component Analysis with Missing Not At Random data. 2020. *NeurIPS*.

Works of Madeleine Udell:

- Missing Value Imputation for Mixed Data Through Gaussian Copula. 2020. *ACM SIGKDD conference*.
- Matrix Completion with Quantified Uncertainty through Low Rank Gaussian Copula. 2020. *NeurIPS*.

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
- Handling missing values in PCA, MCA, FAMD, MFA, 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

Package FactoShiny (Shiny interface), FactoInvestigate (for automatic reporting)

Outline

Outline

Collaborators on supervised learning with missing values

- M. Le Morvan, Researcher, INRIA, Paris.
- E. Scornet, Asso. Pr., Ecole Polytechnique, Paris. Topic: random forests.
- G. Varoquaux, Researcher, INRIA, Paris. Topic: machine learning/ Scikitlearn



⇒ **Random Forests with missing values**

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

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

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

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

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

Supervised learning framework

- A feature matrix X and a response vector Y
- Find a prediction function that minimizes the expected risk

  Bayes rule: $f^\star \in \underset{f:\,\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}} \mathbb{E}\left[\ell(f(X), Y)\right]; \quad f^\star(X) = \mathbb{E}[Y|X]$

- Empirical risk: $\hat{f}_{\mathcal{D}_{n,\mathrm{train}}} \in \underset{f:\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}} \left(\frac{1}{n}\sum_{i=1}^{n} \ell\left(f(X_i), Y_i\right)\right)$

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

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

Supervised learning framework

- A feature matrix X and a response vector Y
- Find a prediction function that minimizes the expected risk

  Bayes rule: $f^\star \in \underset{f:\mathcal{X}\rightarrow\mathcal{Y}}{\operatorname{argmin}} \mathbb{E}\left[\ell(f(X), Y)\right]$; $\quad f^\star(X) = \mathbb{E}[Y|X]$

- Empirical risk: $\hat{f}_{\mathcal{D}_{n,\text{train}}} \in \underset{f:\mathcal{X}\rightarrow\mathcal{Y}}{\operatorname{argmin}} \left(\frac{1}{n}\sum_{i=1}^{n} \ell\left(f(X_i), Y_i\right)\right)$

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

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

Differences with classical litterature

Aim: target an outcome Y (not estimate parameters and their variance)

Specificities: train & test sets with missing values

$\Rightarrow$ Is it possible to use previous approaches (EM - impute), consistent?

$\Rightarrow$ Do we need to design new ones?

Imputation prior to learning

Imputation with the same model

Easy to implement for univariate imputation: The means $(\hat{\mu}_1, ..., \hat{\mu}_d)$ of each colum of the train. Also OK for Gaussian imputation.

Issue: Many methods are "black-boxes" and take as an input the incomplete data and output the completed data (`mice`, `missForest`)

Separate imputation

Impute train and test separately (with a different model)

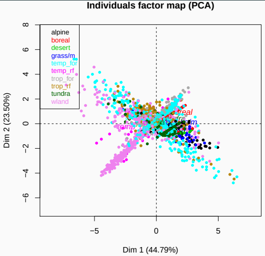
Issue: Depends on the size of the test set? one observation?

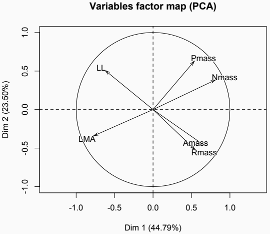
Group imputation/ semi-supervised

Impute train and test simultaneously but the predictive model is learned only on the training imputed data set

Issue: Sometimes no training set at test time

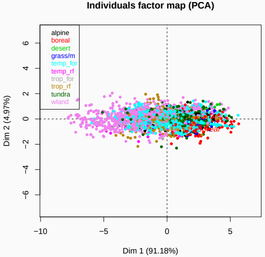
Mean imputation is bad for estimation

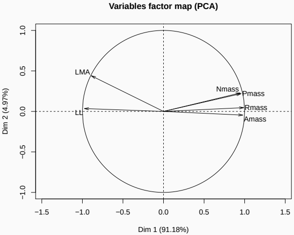




PCA with mean imputation

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





EM-PCA

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

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

Ecological data: [41] $n = 69000$ species - 6 traits. Estimated correlation between Pmass & Rmass ≈ 0 (mean imputation) or ≈ 1 (EM PCA)

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

Constant (mean) imputation is consistent for prediction

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

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

Find a prediction function that minimizes the risk.

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

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

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

Constant (mean) imputation is consistent for prediction

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

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

Find a prediction function that minimizes the risk.

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

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

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

Constant (mean) imputation is consistent

Framework - assumptions

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

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

Constant (mean) imputation is consistent

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

Theorem. (J. et al. 2019)

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

Prediction with mean is equal to the Bayes function almost everywhere

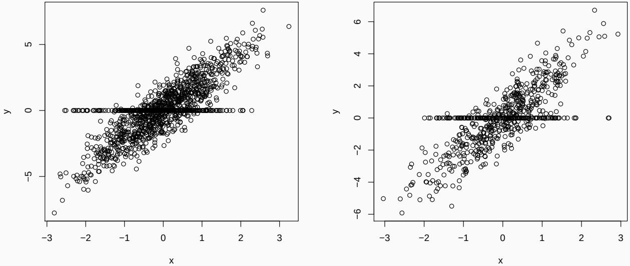
$$f^\star_{impute}(X') = f^\star(\tilde{X}) = \mathbb{E}[Y|\tilde{X} = \tilde{x}]$$

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

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

Consistency of supervised learning with NA: Rationale

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



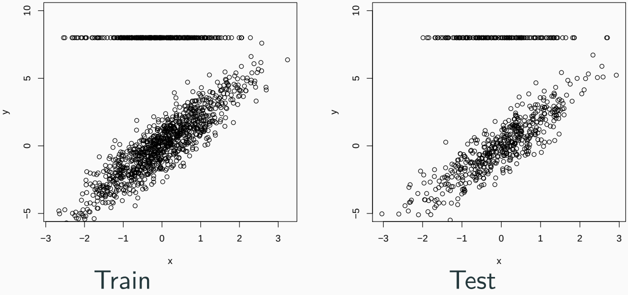
Train Test

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

Empirically good results for MNAR

Consistency of supervised learning with NA: Rationale

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



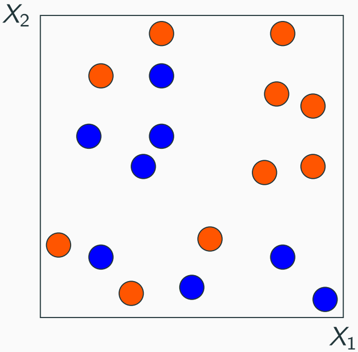
Mean imputation not bad for prediction; it is consistent; despite its drawbacks for estimation - Useful in practice!

Empirically good results for MNAR

CART (Breiman, 1984)

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

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

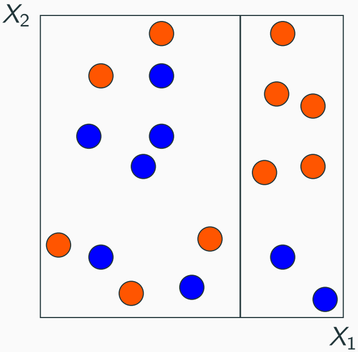


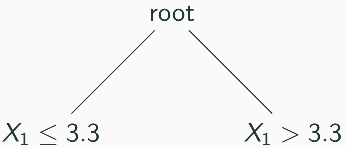
root

CART (Breiman, 1984)

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

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

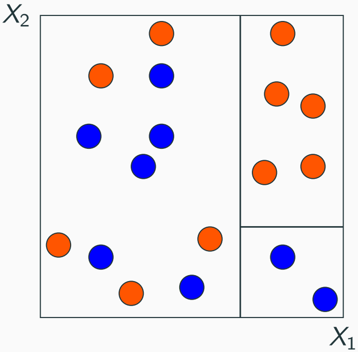


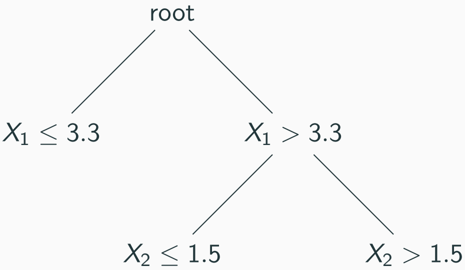


CART (Breiman, 1984)

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

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





CART with missing values

root

	X_1	X_2	Y
1			
2	NA		
3	NA		
4			

CART with missing values

root

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

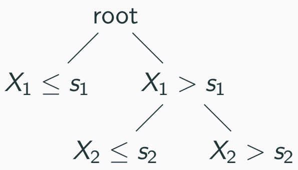
	X_1	X_2	Y
1			
~~2~~	~~NA~~		
~~3~~	~~NA~~		
4			

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

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

CART with missing values

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



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

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

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

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

Missing incorporated in attribute (Twala *et al.* 2008)

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

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

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

Targets one model per pattern:

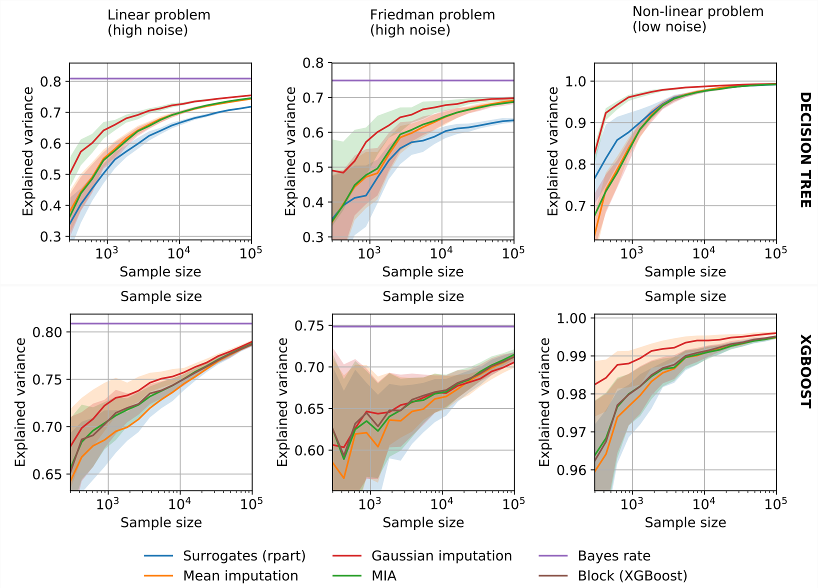
$$\mathbb{E}\left[Y \middle| \tilde{X}\right] = \sum_{m \in \{0,1\}^d} \mathbb{E}\left[Y | X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$$

- Implementation [42]: `grf package`, `scikit-learn`, `partykit`

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

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

Consistency: 40% missing values MCAR



Outline

Linear model with missing values

Linear model:

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

Existing solutions

- ML with EM algo. (available implementation struggles for large d)
- Multiple imputation (few aggregation strategies for predictive models)

$\Rightarrow$ Mainly to estimate parameters in Missing At Random setting

Aim: Predict Y (out of sample) with any missing value mechanism

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

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

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

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

Linear model with missing values not necessarely linear

Example

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

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

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

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

Explicit Bayes predictor with missing values

Linear model:

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

Bayes predictor for the linear model:

$$\begin{aligned} f^\star(\tilde{X}) &= \mathbb{E}[Y|\tilde{X}] = \mathbb{E}[\beta_0 + \beta^\top X \mid \ M, X_{obs(M)}] \\ &= \beta_0 + \beta_{obs(M)}^\top X_{obs(M)} + \beta_{mis(M)}^\top \ \mathbb{E}[X_{mis(M)} \mid M, X_{obs(M)}] \end{aligned}$$

Assumptions on covariates and missing values

1. Gaussian pattern mixture model, PMM: $X \mid (M = m) \sim \mathcal{N}(\mu_m, \Sigma_m)$
2. 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

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

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

Expanded Bayes predictor

Under GPMM, bayes predictor is linear per pattern $\Leftrightarrow$ linear model in W

$$f^*(\tilde{X}) = \langle W, \delta \rangle$$

W an expansion (2^d blocks) & parameters $\delta \in \mathbb{R}^d$ function of β, μ_m, Σ_m

$$\tilde{X} = \left(\begin{array}{c|cc} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ \hline 1 & x_{3,1} & \text{NA} \\ 1 & x_{4,1} & \text{NA} \\ \hline 1 & \text{NA} & x_{5,2} \\ 1 & \text{NA} & x_{6,2} \\ \hline 1 & \text{NA} & \text{NA} \\ 1 & \text{NA} & \text{NA} \end{array}\right) \quad W = \left(\begin{array}{ccc|cc|cc|c} 1 & x_{1,1} & x_{1,2} & 0 & 0 & 0 & 0 & 0 \\ 1 & x_{2,1} & x_{2,2} & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & x_{3,1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_{4,1} & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & x_{5,2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & x_{6,2} & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

$W = (\mathbb{1}_{M=(0,0)}, X_1\mathbb{1}_{M=(0,0)}, X_2\mathbb{1}_{M=(0,0)}, \mathbb{1}_{M=(0,1)}, X_1\mathbb{1}_{M=(0,1)}, \mathbb{1}_{M=(1,0)}, X_2\mathbb{1}_{M=(1,0)}, \mathbb{1}_{M=(1,1)})$.

Expanded Bayes predictor

Under GPMM, bayes predictor is linear per pattern $\Leftrightarrow$ linear model in W

$$f^*(\tilde{X}) = \langle W, \delta \rangle$$

W an expansion (2^d blocks) & parameters $\delta \in \mathbb{R}^d$ function of β, μ_m, Σ_m

$$\tilde{X} = \left(\begin{array}{c|cc} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ \hline 1 & x_{3,1} & \mathrm{NA} \\ 1 & x_{4,1} & \mathrm{NA} \\ \hline 1 & \mathrm{NA} & x_{5,2} \\ 1 & \mathrm{NA} & x_{6,2} \\ \hline 1 & \mathrm{NA} & \mathrm{NA} \\ 1 & \mathrm{NA} & \mathrm{NA} \end{array}\right) \quad W = \left(\begin{array}{ccc|cc|cc|c} 1 & x_{1,1} & x_{1,2} & 0 & 0 & 0 & 0 & 0 \\ 1 & x_{2,1} & x_{2,2} & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & x_{3,1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_{4,1} & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & x_{5,2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & x_{6,2} & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

$W = (\mathbb{1}_{M=(0,0)}, X_1\mathbb{1}_{M=(0,0)}, X_2\mathbb{1}_{M=(0,0)}, \mathbb{1}_{M=(0,1)}, X_1\mathbb{1}_{M=(0,1)}, \mathbb{1}_{M=(1,0)}, X_2\mathbb{1}_{M=(1,0)}, \mathbb{1}_{M=(1,1)})$.

Problem: Dim of W is $p = \sum_{k=0}^{d} \binom{d}{k} \times (k+1) = 2^{d-1} \times (d+2)$.

Need to approximate it: Linear + MLP approximation + Neumiss

Linear Approximation

The Bayes predictor can be expressed as a polynome of X and M, which can be truncated to a lower-dimensional approximation.

$$f^\star_{\mathrm{approx}}(\tilde{X}) = \beta^\star_{0,0} + \sum_{j=1}^{d} \beta^\star_{j,0} M_j + \sum_{j=1}^{d} \beta^\star_j X_j(1 - M_j).$$

$$\left(\begin{array}{c|cc|cc}
1 & X_1 \odot (1 - M_1) & X_2 \odot (1 - M_2) & \color{red}{M_1} & \color{red}{M_2} \\
\hline
1 & x_{1,1} & x_{1,2} & 0 & 0 \\
1 & x_{2,1} & x_{2,2} & 0 & 0 \\
\hline
1 & x_{3,1} & \color{red}{0} & 0 & 1 \\
1 & x_{4,1} & \color{red}{0} & 0 & 1 \\
\hline
1 & \color{red}{0} & x_{5,2} & 1 & 0 \\
1 & \color{red}{0} & x_{6,2} & 1 & 0 \\
\hline
1 & \color{red}{0} & \color{red}{0} & 1 & 1 \\
1 & \color{red}{0} & \color{red}{0} & 1 & 1
\end{array}\right)$$

Imputing X by 0 and concatenate M

Linear Approximation

Impute X by 0 and concatenate $M \Leftrightarrow$ optimize an imputation constant.

$$\text{Given } \begin{array}{c} \begin{matrix} X_1 & X_2 \end{matrix} \\ \begin{pmatrix} 1.1 & 3.2 \\ \texttt{NA} & 0.1 \\ 4.6 & \texttt{NA} \\ 4.0 & 0.9 \\ \texttt{NA} & 2.2 \end{pmatrix} \end{array}, \quad \begin{array}{c} \begin{matrix} X_1 & X_2 & M_1 & M_2 \end{matrix} \\ \begin{pmatrix} 1.1 & 3.2 & 0 & 0 \\ 0 & 0.1 & 1 & 0 \\ 4.6 & 0 & 0 & 1 \\ 4.0 & 0.9 & 0 & 0 \\ 0 & 2.2 & 1 & 0 \end{pmatrix} \end{array} \Leftrightarrow \begin{array}{c} \begin{matrix} X_1 & X_2 \end{matrix} \\ \begin{pmatrix} 1.1 & 3.2 \\ C_1 & 0.1 \\ 4.6 & C_2 \\ 4.0 & 0.9 \\ C_1 & 2.2 \end{pmatrix} \end{array}$$

Indeed,

$$\beta_j \left\{ X_j(1 - M_j) + c_j M_j \right\} = \beta_j X_j (1 - M_j) + \left\{ \beta_j c_j \right\} M_j.$$

Expanded model VS Linear approximation

$$
\begin{array}{c}
\texttt{expanded} \\
\left(\begin{array}{ccc|cc|cc|c}
1 & x_{1,1} & x_{1,2} & 0 & 0 & 0 & 0 & 0 \\
1 & x_{2,1} & x_{2,2} & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 1 & x_{3,1} & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & x_{4,1} & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 1 & x_{5,2} & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & x_{6,2} & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right)
\end{array}
\;\textbf{vs}\;
\begin{array}{c}
\texttt{linear approximation} \\
\left(\begin{array}{c|cc|cc}
1 & x_{1,1} & x_{1,2} & 0 & 0 \\
1 & x_{2,1} & x_{2,2} & 0 & 0 \\
\hline
1 & x_{3,1} & 0 & 0 & 1 \\
1 & x_{4,1} & 0 & 0 & 1 \\
\hline
1 & 0 & x_{5,2} & 1 & 0 \\
1 & 0 & x_{6,2} & 1 & 0 \\
\hline
1 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 1 & 1
\end{array}\right)
\end{array}
$$

Two estimations strategies:

• Linear reg. to estimate the expanded bayes predictor: rich model, powerful in low dimension. Costly, large variance in high dimension

• Linear approximation: lower approximation capacity smaller variance since it contains fewer parameters

Finite sample bounds - Excess of risk

• Expanded: $\mathcal{O}\left(\frac{2^d}{n}\right)$

• Linear approximation: $\mathcal{O}\left(d^2 + \frac{d}{n}\right)$

Comparing the upper bounds: Risk of expanded is lower than risk of approximation when $n >> \frac{2^d}{d}$

Bayes consistency of the MLP

Theorem. Bayes consistency of a MLP. Le Morvan et al. (2020)

Under linear model + Gaussian pattern mixture model, a MLP:

- with one hidden layer containing 2^d hidden units
- ReLU activation functions
- fed with $[X \odot (1 - M), M]$ ($\tilde{X}$ imputed by 0 concatenated with mask)

can achieve the Bayes rate.

Rationale: The MLP produces a prediction function piecewise affine. Since the Bayes predictor is linear per pattern, MLP good candidate.

We show that there exists a configuration of the parameters of the MLP so that the resulting predictor is the Bayes predictor.

Number of parameters: $(d+1)2^{d+1} + 1$.

$\Rightarrow$ Provides a natural way to reduce the model capacity by reducing the number of hidden units. (Trading off estimation and approximation error)

Neumiss Networks to approximate the covariance matrix

The Bayes predictor is linear per pattern (Gaussian+ M(C)AR)

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

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

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

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

Neumiss Networks to approximate the covariance matrix

Order-ℓ approx of the Bayes predictor in MAR

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

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

Proposition (Risk of the Order-ℓ approx)

Let ν be the smallest eigenvalue of Σ. Assume linear model with Gaussian covariates, M(C)AR, and that the spectral radius of Σ is < 1. Then, for all $\ell \geq 1$,

$$\mathbb{E}\left[\left(f^\star_\ell(X_{obs}, M) - f^\star(X_{obs}, M)\right)^2\right] \leq \frac{(1-\nu)^{2\ell} \|\beta^\star\|_2^2}{\nu} \mathbb{E}\left[\left\| Id - S^{(0)}_{obs(M)} \Sigma_{obs(M)} \right\|_2^2\right]$$

The error of the order-ℓ approximation decays exponentially fast with ℓ.

Neumiss Networks to approximate the covariance matrix

Order-ℓ approx of the Bayes predictor in MAR

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

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**

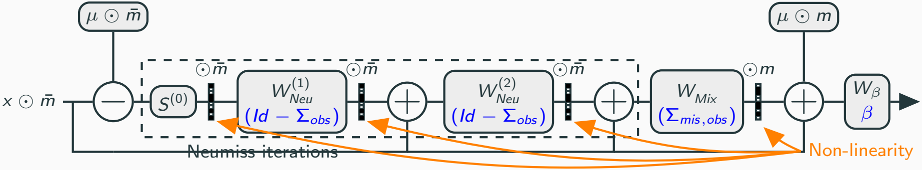


Figure 1: Depth of 3, $\bar{m} = 1 - m$. Each weight matrix $W^{(k)}$ corresponds to a simple transformation of the covariance matrix indicated in blue.

Neumiss Networks to approximate the covariance matrix

Order-ℓ approx of the Bayes predictor in MAR

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

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**

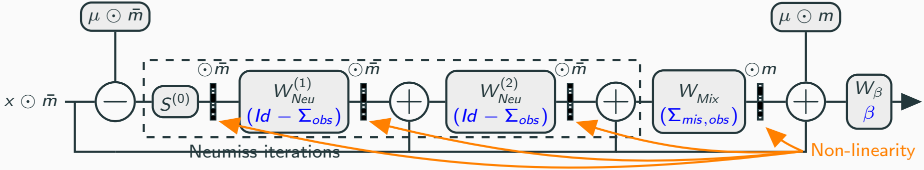
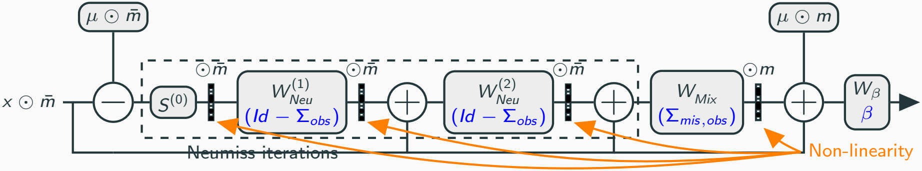


Figure 1: Depth of 3, $\bar{m} = 1 - m$. Each weight matrix $W^{(k)}$ corresponds to a simple transformation of the covariance matrix indicated in blue.

Networks with missing values: $\odot M$ nonlinearity

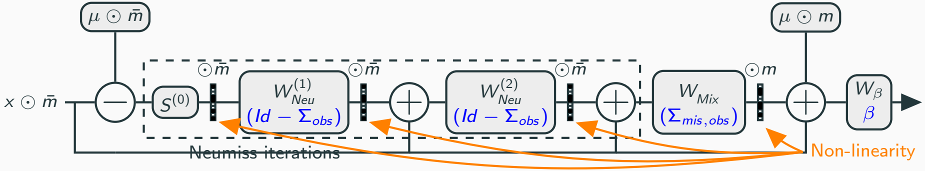


- Implementing a network with the matrix **weights** $W^{(k)} = (I - \Sigma_{obs(m)})$ **masked differently for each sample** can be challenging
- Masked weights is **equivalent to masking input & output vector.**

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 $\odot M$

Networks with missing values: $\odot M$ nonlinearity



- 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 $\odot M$

Proposition (equivalence MLP - depth-0 Neumiss network)

A MLP with ReLU activations, one hidden layer of d hidden units, and which operates on the $[X \odot (1 - M), M]$, the input X imputed by 0 concatenated with the mask M, is equivalent to the 0-depth NN

Experiments for linear regression with missing values

- $Y = X\beta^* + \varepsilon$, ε chosen such as SNR = 10.
- $X \sim \mathcal{N}(\mu, \Sigma)$
- $\Sigma = UU^\top + \text{diag}(\epsilon')$, $U \in \mathbb{R}^{d \times \frac{d}{2}}$, $U_ij \sim \mathcal{N}(0, 1)$ $\epsilon' \sim \mathcal{U}(10^{-2}, 10^{-1})$
- 50% of MCAR, MAR, Probit self-masking.

- **Max Likelihood**: to estimate the parameters of the joint Gaussian distribution $(X_1, ..., X_d, Y)$ with EM. Predict by conditional expectation of Y given X_{obs}.

- **ICE + LR**: conditional imputation with an iterative imputer followed by linear regression.

- **MLP**: take as input the data imputed by 0 concatenated with the mask $[X \odot (1 - M), M]$ with ReLU nonlinearity,
  - **MLP-Wide**: one hidden layer with width increased (between d & 2^d)
  - **MLP-Deep**: 1 to 10 hidden layers of d hidden units

- **Neumiss**: The Neumiss architecture with the $\odot M$, choosing the depth on a validation set.

Results

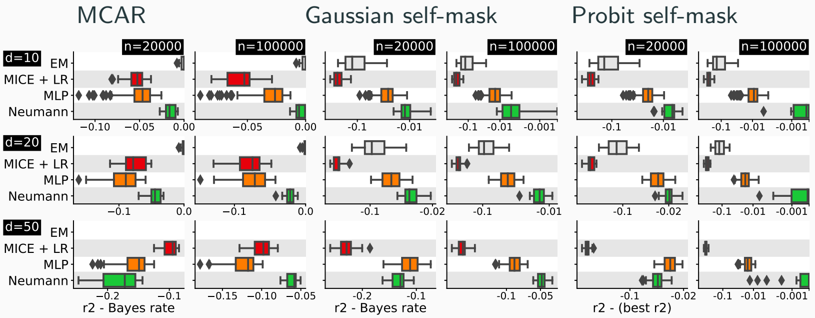


Figure 2: Predictive performances in various scenarios — varying missing-value mechanisms, number of samples n, and number of features d.

$\Rightarrow$ Best performances for MNAR scenario (50% of NA on all variables)

- More effective to increase the capacity of the Neumiss network (depth) than to increase the capacity (width) of MLP Wide.

Discussion - challenges

Take-home message. Supervised learning with missing values.

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

Powerful learner with missing values

- Incomplete train and test $\rightarrow$ same imputation model
- Single constant imputation is consistent with a powerful learner
- Tree-based models : Missing Incorporated in Attribute
- To be done: nonasymptotic results, uncertainty, distributional shift: No NA in the test? Proofs in MNAR

Linear regression with missing values

- The Bayes predictor is explicit under Gaussian assumptions/ MAR and gaussian self mask but high-dimensional.
- Approx include MLP which can be consistent and Neumiss Network
- New architecture for network with missing data: $\odot M$ nonlinearity.

Outline

Collaborators

- Imke Mayer (Postdoc Charité Universittsmedizin Berlin)
- Stefan Wager, Erik Sverdrup (Stanford)
- Tobias Gauss, Jean-Denis Moyer (Assistance Publique Hopitaux de Paris, Traumabase)

Mayer, et al. Doubly robust treatment effect estimation with missing attributes. *Annals of Applied Statistics*, 14(3), 2020

Traumabase

- 30000 patients
- 250 continuous and categorical variables: **heterogeneous**
- 24 hospitals
- 4000 new patients/ year

Center	Accident	Age	Sex	Weight	Lactactes	BP	Acid Tran.	Y
Beaujon	fall	54	m	85	NM	180	treated	0
Pitie	gun	26	m	NR	NA	131	control	1
Beaujon	moto	63	m	80	3.9	145	treated	1
Pitie	moto	30	w	NR	Imp	107	control	0
HEGP	knife	16	m	98	2.5	118	treated	1
$\vdots$								$\ddots$

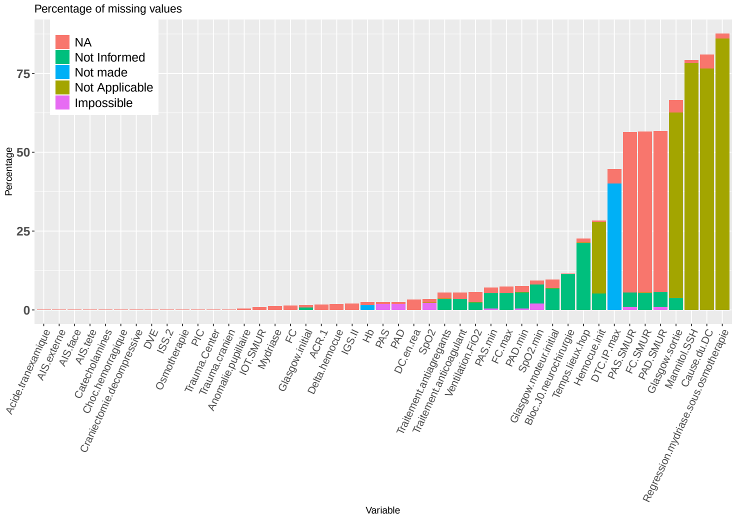
Traumabase

- 30000 patients
- 250 continuous and categorical variables: **heterogeneous**
- 24 hospitals
- 4000 new patients/ year

Center	Accident	Age	Sex	Weight	Lactactes	BP	Acid Tran.	Y
Beaujon	fall	54	m	85	NM	180	treated	0
Pitie	gun	26	m	NR	NA	131	control	1
Beaujon	moto	63	m	80	3.9	145	treated	1
Pitie	moto	30	w	NR	Imp	107	control	0
HEGP	knife	16	m	98	2.5	118	treated	1
$\vdots$								$\ddots$

$\Rightarrow$ **Estimate causal effect**: Administration of the **treatment** "tranexamic acid" (within 3 hours after the accident) on the **outcome** mortality for traumatic brain injury patients.

Missing values



Different types of missing values

Multilevel data/ data integration: Systematic missing variable in one hospital

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

Causal effect for a binary treatment

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

Missing 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	**?**

Cov.			Treat.	Out.
X_1	X_2	X_3	W	Y
1.1	20	F	1	200
-6	45	F	0	10
0	15	M	1	150
	...		...	...
-2	52	M	0	100

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

Causal effect for a binary treatment

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

Missing 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	**?**

Cov.			Treat.	Out.
X_1	X_2	X_3	W	Y
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): $\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$

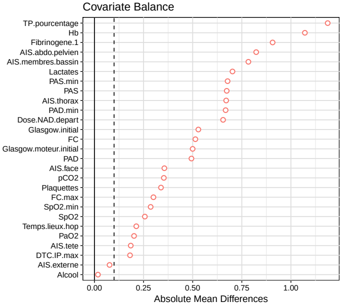
The ATE is the difference of the average outcome had everyone gotten treated and the average outcome had nobody gotten treatment

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 control group does not look like treatment group, difference in response may be **confounded** by differences between the groups.

Assumption for ATE identifiability in observational data

Unconfoundedness - selection on observables

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$$

Treatment assignment W_i is random conditionally on covariates X_i
Measure enough covariates to capture dependence between W_i and outcomes

Overlap

Propensity score: probability of treatment given observed covariates.

$$e(x) \triangleq \mathbb{P}(W_i = 1 \mid X_i = x) \quad \forall\, x \in \mathcal{X}.$$

We assume overlap, i.e. $\eta < e(x) < 1 - \eta, \quad \forall\, x \in \mathcal{X}$ and some $\eta > 0$

ATE not identifiable without assumptions: it is not a sample size problem.

Assumption for ATE identifiability in observational data

Unconfoundedness - selection on observables

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$$

Treatment assignment W_i is random conditionally on covariates X_i
Measure enough covariates to capture dependence between W_i and outcomes

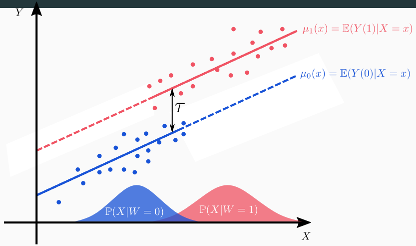
Overlap

Left: Non smoker and never treated   Right: Smokers and all treated

If proba to be treated when smoker $e(x) = 1$, how to estimate the outcome for smokers when not treated $Y(0)$? How to extrapolate if total confusion?

Regression adjustment: g-estimator

$\mu_{(w)}(x) \triangleq \mathbb{E}[Y(w)|X = x]$



OLS model $w \in \{0, 1\}$

$Y_i(w) = c_{(w)} + X_i\beta_{(w)} + \varepsilon_i(w)$

Identifiability (using $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$)

$$\begin{aligned}
\tau &= \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0)|X_i] = \mathbb{E}[\mu_{(1)}(X_i) - \mu_{(0)}(X_i)] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|W_i = 1, X_i = x] - \mathbb{E}[Y_i(0)|W_i = 0, |X_i = x]](\text{uncounfoud}) \\
&= \mathbb{E}[\mathbb{E}[Y_i|W_i = 1, X_i] - \mathbb{E}[Y_i|W_i = 0, X_i]](\text{consistency})
\end{aligned}$$

$\mathbb{E}[Y_i|W_i = 1, X_i]$ can be estimated from data but $\mathbb{E}[Y_i(1)|X_i]$ not.

$$\hat{\tau}_{OLS} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)) = \frac{1}{n}\sum_{i=1}^{n}(\hat{c}_{(1)} + X_i\hat{\beta}_{(1)}) - (\hat{c}_{(0)} + X_i\hat{\beta}_{(0)})$$

$\Rightarrow$ Consistent if $\hat{\mu}_{(w)}$ consistent

Inverse-propensity weighting estimator

Average treatment effect (ATE): $\tau \triangleq \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$

Propensity score (proba treated|covariates): $e(x) \triangleq \mathbb{P}(W_i = 1 \,|\, X_i = x)$

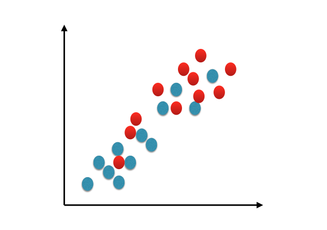
IPW estimator (Horvitz-Thomson, survey)

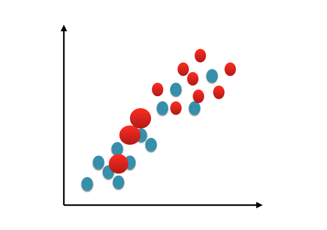
$$\hat{\tau}_{IPW} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)} \right)$$

$\Rightarrow$ Balance the differences between the two groups

$\Rightarrow$ Consistent estimator of τ when $\hat{e}(\cdot)$ consistent (logistic regression).

$\Rightarrow$ High variance (divide by probability)





Doubly robust estimator

Define $\mu_{(w)}(x) \triangleq \mathbb{E}[Y_i(w) \mid X_i = x]$ and $e(x) \triangleq \mathbb{P}(W_i = 1 \mid X_i = x)$.

Augmented IPW - Double Robust (DR)

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

is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent.

- $\hat{\tau}_{IPW} \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right)$: Treatment assignment $\sim$ covariates
- $\hat{\tau}_{OLS} \triangleq \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$: Outcome $\sim$ covariates

$\Rightarrow$ Both sensitive to misspecification. DR: combine ols + ipw of residuals

Doubly robust estimator

Define $\mu_{(w)}(x) \triangleq \mathbb{E}[Y_i(w) \,|\, X_i = x]$ and $e(x) \triangleq \mathbb{P}(W_i = 1 \,|\, X_i = x)$.

Augmented IPW - Double Robust (DR)

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

is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent.

- $\hat{\tau}_{IPW} \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right)$: Treatment assignment $\sim$ covariates
- $\hat{\tau}_{OLS} \triangleq \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$: 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 $e(x) \triangleq \mathbb{P}(W_i = 1 \mid X_i = x)$
Model Outcome on Covariates $\mu_{(w)}(x) \triangleq \mathbb{E}[Y_i(w) \mid X_i = x]$

Augmented IPW - Double Robust (DR)

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

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.

Causal inference with missing attributes?

Missing (informative) values in the covariates

Straightforward – but often biased – solution is complete-case

Covariates			Treatment	Outcome(s)	
X_1^*	X_2^*	X_3^*	W	Y(0)	Y(1)
NA	20	F	1	?	Survived
-6	45	NA	0	Dead	?
0	NA	M	1	?	Survived
NA	32	F	1	?	Dead
1	63	M	1	Dead	?
-2	NA	M	0	Survived	?

→ Often not a good idea! What are the alternatives?

Three families of methods - different assumptions

- Classical unconfoundedness + classical missing values mechanisms
- Unconfoundedness with missingness + (no) missing values mechanisms Mayer, J., Wager, Sverdrup, Moyer, Gauss. *AOAS* 2020.
- Latent unconfoundedness + classical missing values mechanisms Mayer, J., Raimundo, Vert. 2020.

Under 1: Multiple Imputation

Consistency of IPW with missing values (Seaman, White 2014)

Assume **Missing At Random (MAR)** mechanism. Multiple imputation (MICE using (X^*, W, Y)) with IPW on each imputed data is consistent when Gaussian covariates and logistic/linear treatment/oucome model

X_1^*	X_2^*	X_3^*	...	W	Y
NA	20	10	...	1	survived
-6	45	NA	...	1	survived
0	NA	30	...	0	died
NA	32	35	...	0	survived
-2	NA	12	...	0	died
1	63	40	...	1	survived

1) Generate M plausible values for each missing value

X_1	X_2	X_3	...	W	Y
3	20	10	...	1	s
-6	45	6	...	1	s
0	4	30	...	0	d
-4	32	35	...	0	s
-2	15	12	...	0	d
1	63	40	...	1	s

X_1	X_2	X_3	...	W	Y
-7	20	10	...	1	s
-6	45	9	...	1	s
0	12	30	...	0	d
13	32	35	...	0	s
-2	10	12	...	0	d
1	63	40	...	1	s

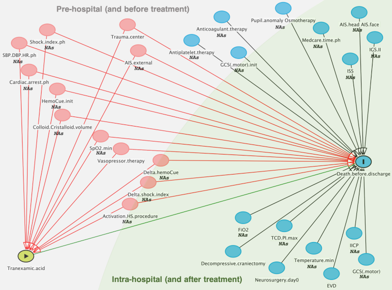
X_1	X_2	X_3	...	W	Y
7	20	10	...	1	s
-6	45	12	...	1	s
0	-5	30	...	0	d
2	32	35	...	0	s
-2	20	12	...	0	d
1	63	40	...	1	s

2) Estimate ATE on each imputed data set: $\hat{\tau}_m$, $\widehat{Var}\left(\hat{\tau}_m\right)$

3) Combine the results (Rubin's rules): $\hat{\tau} = \frac{1}{M}\sum_{m=1}^{M} \hat{\tau}_m$

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

2. Unconfoundedness with missing + (no) missing hypothesis



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: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \,|\, X$ not testable from the data.
$\Rightarrow$ Doctors give us the DAG (covariates relevant for either treatment decision and for predicting the outcome)

Unconfoundedness with missing values: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \,|\, X^*$
$X^* \triangleq (1 - M) \odot X + M \odot NA$; with $M_{ij} = 1$ if X_{ij} is missing, 0 otherwise.

$\Rightarrow$ Doctors decide to treat a patient based on what they observe/record. We have access to the same information as the doctors.

Under 2: Double Robust with missing values

AIPW with missing values

$$\hat{\tau}^* \triangleq \frac{1}{n}\sum_i \left(\widehat{\mu^*_{(1)}}(X_i) - \widehat{\mu^*_{(0)}}(X_i) + W_i \frac{Y_i - \widehat{\mu^*_{(1)}}(X_i)}{\widehat{e^*}(X_i)} - (1 - W_i)\frac{Y_i - \widehat{\mu^*_{(0)}}(X_i)}{1 - \widehat{e^*}(X_i)}\right)$$

Generalized propensity score **(Rosenbaum, Rubin JASA 1984)**

$$e^*(x^*) \triangleq \mathbb{P}(W = 1 \,|\, X^* = x^*)$$

One model per pattern: $\sum_{m \in \{0,1\}^d} \mathbb{E}\left[W | X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$

$\Rightarrow$ Supervised learning with missing values. [1] [2] [3]

- Mean imputation is consistent with a universally consistent learner.
- Missing Incorporate in Attributes (MIA) for trees methods.

[1] Consistency of supervised learning with missing values J., Prost, Scornet, Varoquaux. *2020*

[2] Neumiss networks: differential programming for supervised learning with missing values. Le Morvan, J. et al. *Neurips2020*

[3] What's a good imputation to predict with missing values? Le Morvan, J. et al. *Neurips 2021*

Under 2: Double Robust with missing values

AIPW with missing values

$$\hat{\tau}^* \triangleq \frac{1}{n}\sum_i \left(\widehat{\mu^*_{(1)}}(X_i) - \widehat{\mu^*_{(0)}}(X_i) + W_i \frac{Y_i - \widehat{\mu^*_{(1)}}(X_i)}{\widehat{e^*}(X_i)} - (1 - W_i)\frac{Y_i - \widehat{\mu^*_{(0)}}(X_i)}{1 - \widehat{e^*}(X_i)} \right)$$

Generalized propensity score **(Rosenbaum, Rubin JASA 1984)**

$$e^*(x^*) \triangleq \mathbb{P}(W = 1 \mid X^* = x^*)$$

One model per pattern: $\sum_{m \in \{0,1\}^d} \mathbb{E}\left[W | X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$

$\Rightarrow$ Supervised learning with missing values.

- Mean imputation is consistent with a universally consistent learner.
- Missing Incorporate in Attributes (MIA) for trees methods.

Implemented in `grf` package: combine two non-parametrics models, forests (conditional outcome and treatment assignment) adapted to **any** missing values with MIA.

$\hat{\tau}_{AIPW^*}$ is $\sqrt{n}$-consistent, asymptotically normal given the product of RMSE of the nuisance estimates decay as $o(n^{-1/2})$ Mayer, J. et al. AOAS 2020

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* [4]	✓	✗	✓	✗	✓	✗	✗	✓	✗
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

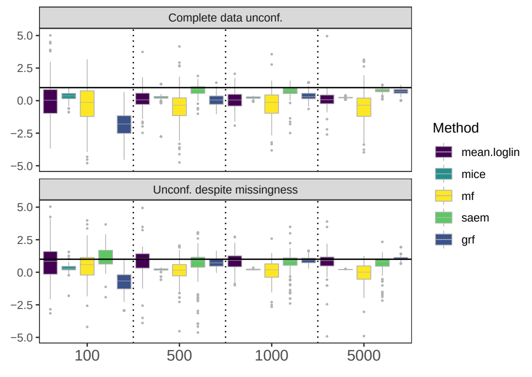
(✗): no theoretical guarantees but heuristics.

[4] 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 1: 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

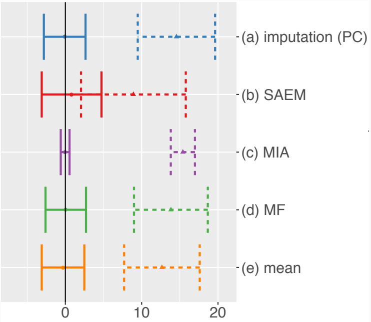
- AIPW estimators outperform their IPW counterparts.
- For $\hat{\tau}_{mia}$, the *unconfoundedness despite missingness* is indeed necessary.
- $\hat{\tau}_{mia}$ unbiased for all missingness mechanisms, especially for MNAR.
- Multiple imputation (mice) only requires standard unconfoundedness, but needs MAR

Results for Trauma Brain Injuries (TBI)

40 covariates, 18 confounders. 8,248 patients.

Overlap: cannot be tested but high level of uncertainty at diagnosing severe (internal bleeding) makes it likely
Many MNAR missing values

ATE estimations ($\times 100$): effect of tranexamic acid on in-ICU mortality



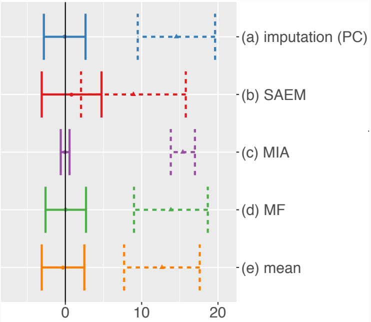
(y-axis: estimation approach, solid: **Double Robust AIPW**, dotted: **IPW**), (x-axis: ATE estimation with bootstrap CI)
The obtained value corresponds to the **difference in percentage points between mortality rates in treatment and control**.

Results for Trauma Brain Injuries (TBI)

40 covariates, 18 confounders. 8,248 patients.

Overlap: cannot be tested but high level of uncertainty at diagnosing severe (internal bleeding) makes it likely
Many MNAR missing values

ATE estimations ($\times 100$): effect of tranexamic acid on in-ICU mortality



(y-axis: estimation approach, solid: **Double Robust AIPW**, dotted: **IPW**), (x-axis: ATE estimation with bootstrap CI)

Comparison with CRASH-3 study same conclusion of "no average treatment effect".

Conclusion and perspectives

Take-away messages

- Missing attributes alter causal analyses.
- Additional assumptions on appropriate unconfoundedness.
- New proposals to handle missing values in causal inference.
- Prefer AIPW to IPW estimators, in theory and in practice.
- Heterogeneous treatment effects with missing values (causal forest)

implemented in the grf R package

Ongoing work

- Causal survival analysis, Policy learning (with missing values)
- Combine RCT and observational data to generalize the ATE to a (broader) target population [5] [6]

	Set	S	X_1	X_2	X_3	W	Y
1	$\mathcal{R}$	1	1.1	20	NA	1	24.1
...	$\mathcal{R}$	1		...		...	...
$n-1$	$\mathcal{R}$	1	-6	45	8.3	0	26.3
n	$\mathcal{R}$	1	0	15	6.2	1	23.5
$n+1$	$\mathcal{O}$	?	-2	NA	7.1	NA	NA
$n+2$	$\mathcal{O}$	?	-1	NA	2.4	NA	NA
...	$\mathcal{O}$	?		...		NA	NA
$n+m$	$\mathcal{O}$	?	-2	NA	3.4	NA	NA

Data with observed treatment W and outcome Y only in the RCT.

CRASH3

- Multi-centric RCT over 29 counties
- No effect of TXA with difference in means (-0.3 with [95% CI -0.8 0.2])

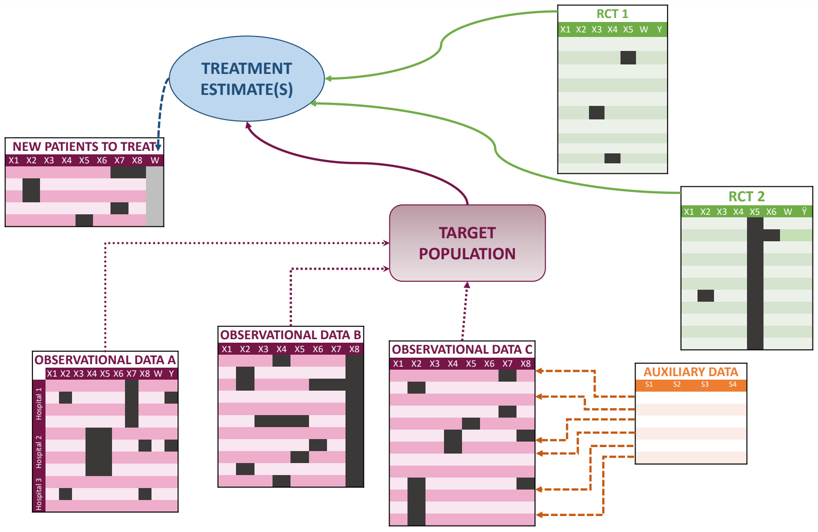
Traumabase

- Representative sample
- 8200 patients with TBI

ATE = -0.035, 95% CI [-0.38 0.28] when generalizing with g-estimator.
Treatment effect modifiers "time to treatment" is missing in Traumabase

[5] Colnet, J. et al. (2021). Causal inference methods for combining RCT and observational studies: a review. *In revision in Statistical Science* - Causal effect on a target population: a sensitivity analysis to handle missing covariates. *Submitted*

[6] Mayer, J. et al. Transporting treatment effects with missing attributes (2021) *Submitted*


TREATMENT ESTIMATE(S)
RCT 1
X1 X2 X3 X4 X5 W Y
RCT 2
X1 X2 X3 X4 X5 X6 W Ÿ
NEW PATIENTS TO TREAT
X1 X2 X3 X4 X5 X6 X7 X8 W
TARGET POPULATION
OBSERVATIONAL DATA A
X1 X2 X3 X4 X5 X6 X7 X8 W Y
Hospital 1
Hospital 2
Hospital 3
OBSERVATIONAL DATA B
X1 X2 X3 X4 X5 X6 X7 X8
OBSERVATIONAL DATA C
X1 X2 X3 X4 X5 X6 X7 X8
AUXILIARY DATA
S1 S2 S3 S4

Missing value website

More information and details on missing values: **R-miss-tastic** platform. Mayer, J. et al., 2019

→ Theoretical and practical tutorials, popular datasets, bibliography, workflows (in R and in python), active contributors/researchers in the community, etc.

`rmisstastic.netlify.com`

Interested in contribute to our platform? Feel free to contact us!

MERCI

An active area of research! Join this exciting field!

Challenges:

- SGD with missing values for linear regression and MCAR [43]. Difficult to extend to logistic or MAR.
  - Naively impute the missing values, get $\tilde{X}$,
  - Adapt algorithm to account for the error & apply this debiased version to the complete dataset $\tilde{X}$.

  Naive imputation + debiasing also used for Lasso [44]

Current works

- Times series with missing values for classification
- Model-based Clustering with Missing Not At Random Data
- **MNAR missing values** - CV with MNAR data? Contribution of causality for missing data

  Mohan, Pearl. 2021. Graphical Models for Processing Missing Data. *JASA.*

  Sportisse, Boyer, J. Estimation and imputation in Probabilistic Principal Component Analysis with Missing Not At Random data. Neurips2020.

[43] Sportisse, Boyer, Dieuleveut, J. Debiasing Stochastic Gradient Descent to handle missing values. *Neurips2020*

Ressources

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

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

Project funded by the R consortium (Infrastructure Steering Committee)[45]

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

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

$\Rightarrow$ Federate the community

$\Rightarrow$ Contribute!

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

Ressources

Examples:

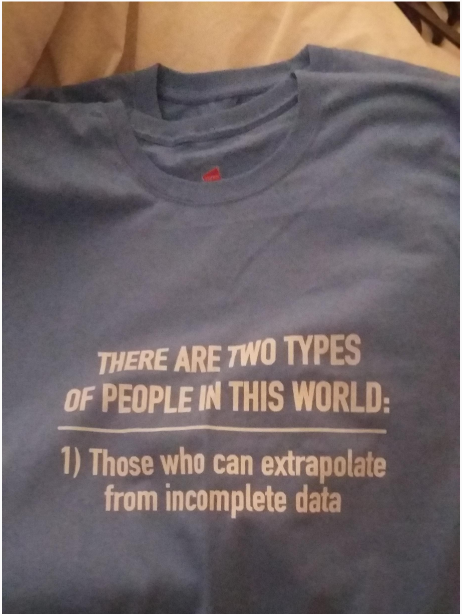
- Lecture [46] - General tutorial : Statistical Methods for Analysis with Missing Data (Mauricio Sadinle)
- Lecture - Multiple Imputation: `mice` by Nicole Erler [47]
- Longitudinal data, Time Series Imputation (Steffen Moritz - very active contributor of r-miss-tastic), Principal Component Methods[48]

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

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

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

Thank you


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