Causal inference with missing values

Effect of tranexamic acid on mortality for head trauma patient

Julie Josse, (INRIA XPOP - X) - Imke Mayer

22 January, 2019

Statistic seminar Nice

Research activities

- Dimensionality reduction methods to visualize complex data (PCA based) : multi-sources, textual, arrays, questionnaire
- Low rank estimation, selection of regularization parameters
- Missing values - matrix completion
- Causal inference
- Fields of application : bio-sciences (agronomy, sensory analysis), health data (hospital data)
- R community : book R for Stat, R foundation, taskforce, packages :
  FactoMineR explore continuous, categorical, multiple contingency tables (correspondence analysis), combine clustering and PC, ..
  MissMDA for single and multiple imputation, PCA with missing
  denoiseR to denoise data with low-rank estimation
  R-miss-tastic missing values plateform

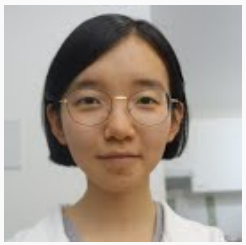
Overview

1. Introduction
2. Causal inference
    - Inverse-propensity weighting
    - Double robust methods
3. Handling missing values
    - Single imputation with PCA
    - Supervised learning with missing values
        - Logistic regression with missing values
4. Results
5. Conclusion

Introduction

Collaborators

Imke Mayer, Wei Jiang, Genevieve Robin, polytechnique students,
Jean-Pierre Nadal,
Traumabase (APHP) : Tobias Gauss, Sophie Hamada, Jean-denis Moyer
Capgemini



Traumabase

15000 patients/ 250 variables/ 11 hospitals, from 2011 (4000 new patients/ year)

	Center	Accident	Age	Sex	Weight	Height	BMI	BP	SBP
1	Beaujon	Fall	54	m	85	NR	NR	180	110
2	Lille	Other	33	m	80	1.8	24.69	130	62
3	Pitie Salpetriere	Gun	26	m	NR	NR	NR	131	62
4	Beaujon	AVP moto	63	m	80	1.8	24.69	145	89
6	Pitie Salpetriere	AVP bicycle	33	m	75	NR	NR	104	86
7	Pitie Salpetriere	AVP pedestrian	30	w	NR	NR	NR	107	66
9	HEGP	White weapon	16	m	98	1.92	26.58	118	54
10	Toulon	White weapon	20	m	NR	NR	NR	124	73

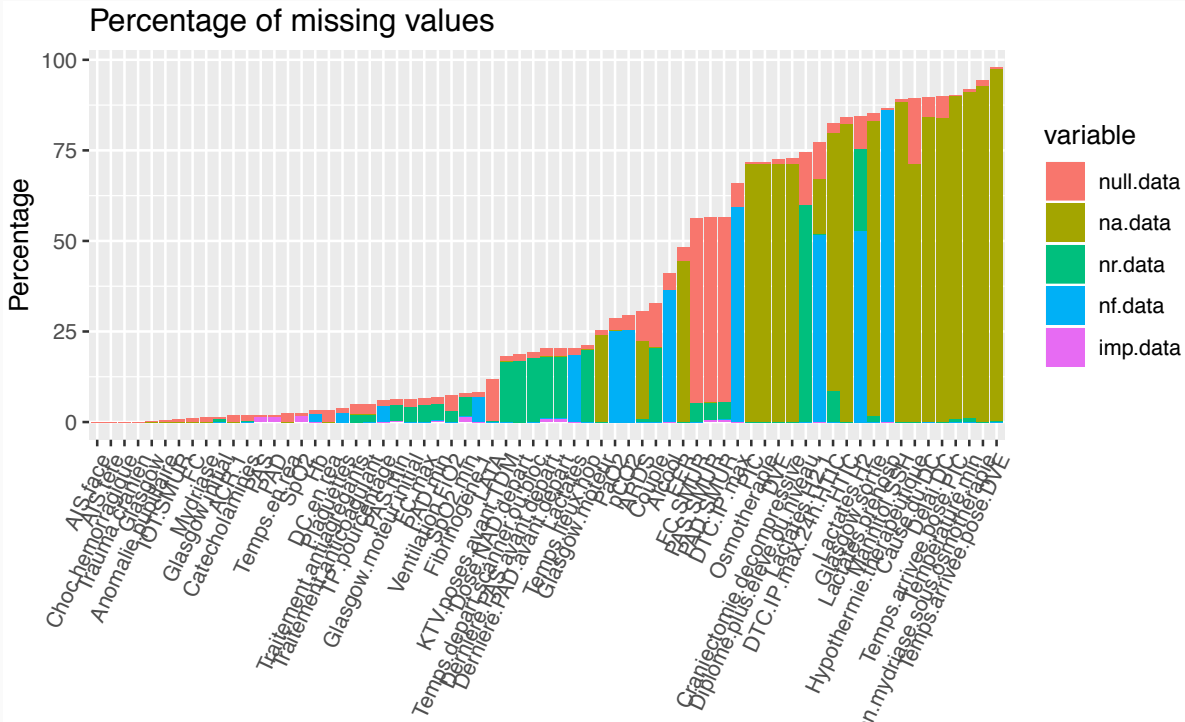
.....

	SpO2	Temperature	Lactates	Hb	Glasgow	Transfusion .....
1	97	35.6	<NA>	12.7	12	yes
2	100	36.5	4.8	11.1	15	no
3	100	36	3.9	11.4	3	no
4	100	36.7	1.66	13	15	yes
6	100	36	NM	14.4	15	no
7	100	36.6	NM	14.3	15	yes
9	100	37.5	13	15.9	15	yes
10	100	36.9	NM	13.7	15	no

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

Causal inference for traumatic brain injury with missing values

- 3050 patients with a brain injury (a lesion visible on the CT scan)
- Treatment : tranexamic acid (binary)
- Outcome : in-ICU death (binary), causes : brain death, withdrawal of care, head injury and multiple organ failure.
- 45 **quantitative** & **categorical** covariates selected by experts (Delphi process). Pre-hospital (blood pressure, patients reactivity, type of accident, anamnesis, etc.) and hospital data



Outline

$\Rightarrow$ Causal inference

Causal inference methodology : estimate causal relationships between an intervention (acid administration) and an outcome (mortality), when the study is potentially confounded by selection bias due to the absence of randomization.

$\Rightarrow$ How to handle missing values ?
$\Rightarrow$ Causal inference with missing values, analysis of the data

Causal inference

Potential outcome framework (Rubin, 1974)

Causal effect

Binary treatment $w \in \{0, 1\}$ on *i-th* individual with potential outcomes $Y_i(1)$ and $Y_i(0)$. Individual causal effect of the treatment :

$$\Delta_i = Y_i(1) - Y_i(0)$$

Potential outcome framework (Rubin, 1974)

Causal effect

Binary treatment $w \in \{0, 1\}$ on *i-th* individual with potential outcomes $Y_i(1)$ and $Y_i(0)$. Individual causal effect of the treatment :

$$\Delta_i = Y_i(1) - Y_i(0)$$

- Problem : Δ_i never observed (only observe one outcome/indiv). Causal inference as a missing value pb ?
- **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 treated.

$\Rightarrow$ First solution : estimate τ with randomized controlled trials (RCT).

Average treatment effect estimation in RTCs

Assumptions :

Observe n iid samples (Y_i, W_i) each satisfying :

- $Y_i = Y_i(W_i)$ (SUTVA)
- $W_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\}$ (random treatment assignment)

Difference-in-means estimator

$$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{W_1=1} Y_i - \frac{1}{n_0} \sum_{W_1=0} Y_i$$

Properties of $\hat{\tau}_{DM}$

$\hat{\tau}_{DM}$ is unbiased and $\sqrt{n}$-consistent. $\sqrt{n}\,(\hat{\tau}_{DM} - \tau) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V_{DM})$, where $V_{DM} = \frac{Var(Y_i(0))}{\mathbb{P}(W_i=0)} + \frac{Var(Y_i(1))}{\mathbb{P}(W_i=1)}$.

Average treatment effect estimation in RTCs

$$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{W_1=1} Y_i - \frac{1}{n_0} \sum_{W_1=0} Y_i$$

Furthermore assume a linear model for the two potential outcomes :

Linear assumptions n iid samples (X_i, Y_i, W_i)

- $Y_i(w) = c_{(w)} + X_i \beta_{(w)} + \varepsilon_i(w)$, $w \in \{0, 1\}$,
  $Y_i(w) = \mu_{(w)}(X_i) + \varepsilon_i(w)$
- $\mathbb{E}[\varepsilon_i(w)|X_i] = 0$ and $Var(\varepsilon_i(w)|X_i) = \sigma^2$.

OLS estimator

$\hat{\tau}_{OLS} = \hat{c}_{(1)} - \hat{c}_{(0)} + \bar{X}(\hat{\beta}_{(1)} - \hat{\beta}_{(0)}) =$
$\frac{1}{n} \sum_i \left((\hat{c}_{(1)} + X_i \hat{\beta}_{(1)}) - (\hat{c}_{(0)} - X_i \hat{\beta}_{(0)}) \right) = \frac{1}{n} \sum_i \left(\hat{\mu}_{(1)} X_i - \hat{\mu}_{(0)}(X_i) \right)$

Properties of $\hat{\tau}_{OLS}$

$\sqrt{n}\,(\hat{\tau}_{OLS} - \tau) \xrightarrow[n \to \infty]{d} \mathcal{N}(0, V_{OLS})$. And $V_{DM} = V_{OLS} + \|\beta_{(0)} + \beta_{(1)})\|_A^2$.

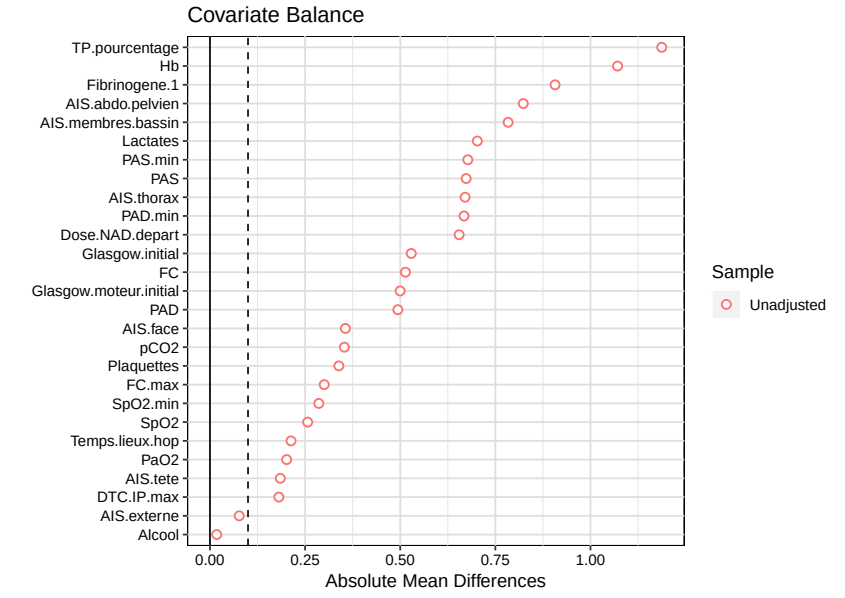
Observational data. Non random assignment : confusion

Mortality rate 16% - treated 28 - not treated 13 : treatment kills ?

	Died		P(Outcome | Treatment)	
Treated	0	1	0	1
FALSE	2225	340	0.867	0.133
TRUE	436	168	0.722	0.278

Strong indication for confounding factors that need to be controlled for.

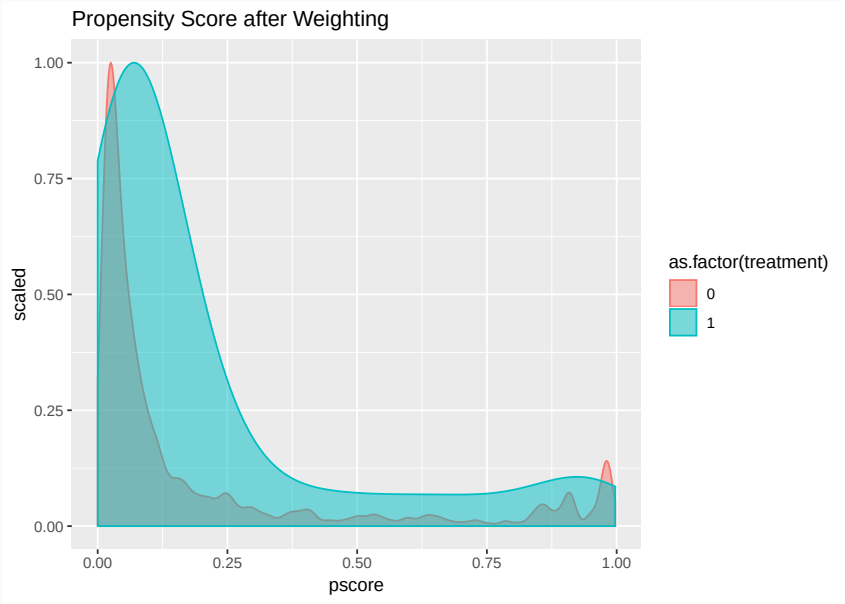
Standardized mean differences between treated and control.



Treated patients are more severe with higher risk of death (graphical model)

Solutions to estimate ATE with observational data

- **Matching** : pair each treated (resp. untreated) patient with one or more similar untreated (resp. treated) patient (R package `Match`)
- **Inverse-propensity weighting** : to adjust for biases in the treatment assignment



- **Double robust methods** for model misspecifications : covariate balancing propensity score, augmented IPW. (Robins *et al.*, 1994)
- **Regression adjustment**, **regression-adjusted matching**, etc.

Unconfoundedness and the propensity score

Assumptions

- n iid samples (X_i, Y_i, W_i),
- Treatment assignment is random conditionally on X_i :
  $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, X_i \quad \equiv$ **unconfoundedness** assumption.

Measure enough covariates to capture any dependence between W_i and the PO

Propensity score

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

Key property

e is a balancing score, i.e. under unconfoundedness, it satisfies

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

As a consequence, it suffices to control for $e(X)$ (rather than X), to remove biases associated with non-random treatment assignment.

Unconfoundedness and the propensity score

Propensity score

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

Key property

Under unconfoundedness, $e(x)$ satisfies $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid e(X_i)$.

Proof

To prove this balancing property, we note that the distribution of W is fully specified by its mean. Therefore we need to prove that :

$\mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, X_i] = \mathbb{E}[W_i|X_i] \Rightarrow \mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, e(X_i)] = \mathbb{E}[W_i|e(X_i)]$

Unconfoundedness and the propensity score

Propensity score

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

Key property

Under unconfoundedness, $e(x)$ satisfies $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid e(X_i)$.

Proof

To prove this balancing property, we note that the distribution of W is fully specified by its mean. Therefore we need to prove that :
$\mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, X_i] = \mathbb{E}[W_i|X_i] \Rightarrow \mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, e(X_i)] = \mathbb{E}[W_i|e(X_i)]$

a) By the law of total expectation we have :

$$\mathbb{E}[W_i|e(X_i)] = \mathbb{E}[\mathbb{E}[W_i|X_i, e(X_i)]|e(X_i)] = \mathbb{E}[\mathbb{E}[W_i|X_i]|e(X_i)] = e(X_i)$$

Unconfoundedness and the propensity score

Propensity score

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

Key property

Under unconfoundedness, $e(x)$ satisfies $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid e(X_i)$.

Proof

To prove this balancing property, we note that the distribution of W is fully specified by its mean. Therefore we need to prove that :

$\mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, X_i] = \mathbb{E}[W_i|X_i] \Rightarrow \mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, e(X_i)] = \mathbb{E}[W_i|e(X_i)]$

a) By the law of total expectation we have :

$$\mathbb{E}[W_i|e(X_i)] = \mathbb{E}[\mathbb{E}[W_i|X_i, e(X_i)]|e(X_i)] = \mathbb{E}[\mathbb{E}[W_i|X_i]|e(X_i)] = e(X_i)$$

b) And again using the law of total expectation we have the following :

$$\begin{aligned}
\mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, e(X_i)] &= \mathbb{E}[\mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, X_i, e(X_i)]|\{Y_i(0), Y_i(1)\}, e(X_i)] \\
&= \mathbb{E}[\mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, X_i]|\{Y_i(0), Y_i(1)\}, e(X_i)] \\
&= \mathbb{E}[\mathbb{E}[W_i|X_i]|\{Y_i(0), Y_i(1)\}, e(X_i)] \ (\textit{unconfoundedness}) \\
&= \mathbb{E}[e(X_i)|\{Y_i(0), Y_i(1)\}, e(X_i)] = e(X_i)
\end{aligned}$$

■

Inverse-propensity weighting estimation of ATE

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

$\Rightarrow$ Balance the difference between the two groups

The quality of this estimator depends on the estimation quality of $\hat{e}(x)$/on the postulated propensity score model. Indeed we have :

$$\begin{aligned}
\mathbb{E}\left[\frac{WY}{e(X)}\right] &= \mathbb{E}\left[\frac{WY(1)}{e(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{WY(1)}{e(X)}|\, Y(1), X\right]\right] \\
&= \mathbb{E}\left[\frac{Y(1)}{e(X)}\mathbb{E}[W|Y(1), X]\right] = \mathbb{E}\left[\frac{Y(1)}{e(X)}\mathbb{E}[W|X]\right] \\
&= \mathbb{E}\left[\frac{Y(1)}{e(X)}e(X)\right] = \mathbb{E}[Y(1)].
\end{aligned}$$

This holds if $e(X) = \mathbb{P}(W = 1|X)$, therefore if $\hat{e}(X)$ is not the true propensity score then $\hat{\tau}_{IPW}$ is not necessarily a (consistent) estimate of τ. Variance of the oracle estimate is bad !

Covariate balancing propensity score (CBPS)

Assume a linear-logistic model :

1. $e(x) = \mathbb{P}(W_i = 1 \,|\, X_i = x) = \frac{1}{1+e^{-x^T\alpha}}$
2. $\mu_{(w)}(x) = x^T\beta_{(w)}$ (for $w \in \{0,1\}$).
3. $Y_i(w) = \mu_{(W_i)}(X_i) + \varepsilon_i$.

Decompose ATE $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\gamma}_{(1)}(X_i)W_iY_i - \hat{\gamma}_{(0)}(X_i)(1-W_i)Y_i\right)$:

$$\hat{\tau} = \bar{X}(\beta_{(1)} - \beta_{(0)}) + [\text{term for } \varepsilon] + \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_{(1)}(X_i)W_iX_i - \bar{X}\right)\beta_{(1)} - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_{(0)}(X_i)(1-W_i)X_i - \bar{X}\right)$$

$$= \bar{X}(\beta_{(1)} - \beta_{(0)}) + \frac{W_i(Y_i - \mu_{(1)}(X_i))}{e(X_i)} - \frac{(1-W_i)(Y_i - \mu_{(0)}(X_i))}{1-e(X_i)}$$

What happens when models are mis-specified ? Double robustness

For specific $\hat{\gamma}_{(1)}$ and $\hat{\gamma}_{(0)}$ (functions of α), $\hat{\tau}$ is the CPBS and it is doubly robust, i.e. it is consistent in either one of the following cases :

1. Outcome model is linear but propensity score $e(x)$ is not logistic.
2. Propensity score $e(x)$ is logistic but outcome model is not linear.

Another doubly robust ATE estimator

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

Doubly robust estimator

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

is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent. Furthermore $\hat{\tau}_{DR^*}$ has good asymptotic variance.

Another doubly robust ATE estimator

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

Doubly robust estimator

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

is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent. Furthermore $\hat{\tau}_{DR^*}$ has good asymptotic variance.

Remark 1 : 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.

Remark 2 : In case of overparametrization or non-parametric estimation $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$ should be learned/estimated by **cross-splitting** to avoid overfitting. Package `grf`. (Wager, Tibshirani)

Semiparametric efficiency for ATE estimation

Efficient score estimator

Given unconfoundedness ($\{Y_i(1), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, X_i$) but no further parametric assumptions on $\mu_{(w)}(x)$ and $e(x)$, the previously attained asymptotic variance,

$$V^* := Var(\tau(X)) + \mathbb{E}\left[\frac{\sigma^2(X)}{e(X)(1-e(X))}\right],$$

is optimal and any estimator τ^* that attains it is asymptotically equivalent to $\hat{\tau}_{DR^*}$.

V^* is the semiparametric efficient variance for ATE estimation.

Semiparametric *: we are interested in a parametric estimand, τ, which we estimate using nonparametric estimates ($\hat{\tau}_{DR}$ depends on nonparametric estimates $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$).*

Handling missing values

Solutions to handle missing values

Litterature : Schaefer (2002) ; Little & Rubin (2002) ; Gelman & Meng (2004) ; Kim & Shao (2013) ; Carpenter & Kenward (2013) ; van Buuren (2015)

$\Rightarrow$ Modify the estimation process to deal with missing values. Maximum likelihood : **EM algorithm** to obtain point estimates + Supplemented EM (Meng & Rubin, 1991) ; Louis for their variability
Difficult to establish ?
Not many implementations, even for simple models
One specific algorithm for each statistical method...

$\Rightarrow$ **Imputation** (multiple) to get a completed data set on which you can perform any statistical method (Rubin, 1976)
Famous imputation based on SVD (PCA) - quantitative

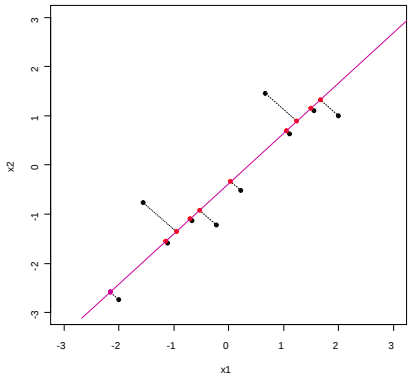
PCA reconstruction

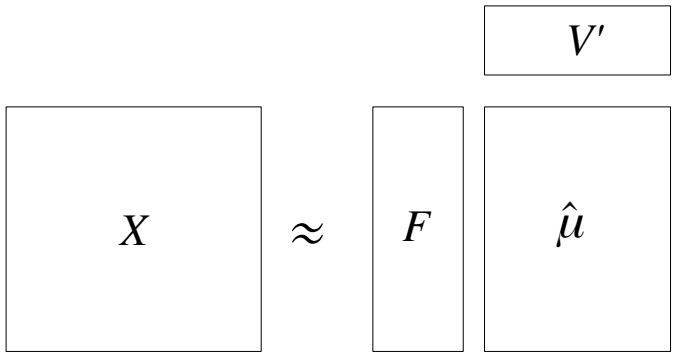
X

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

$\hat{\mu}$

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





$\Rightarrow$ Minimizes distance between observations and their projection

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

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

SVD X : $\hat{\mu}^{\text{PCA}} = U_{n\times k} D_{k\times k} V'_{p\times k}$ $\quad F = UD$ PC - scores

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

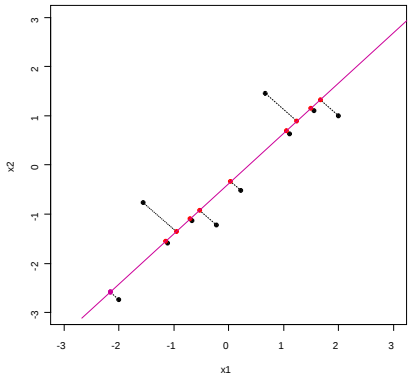
PCA reconstruction

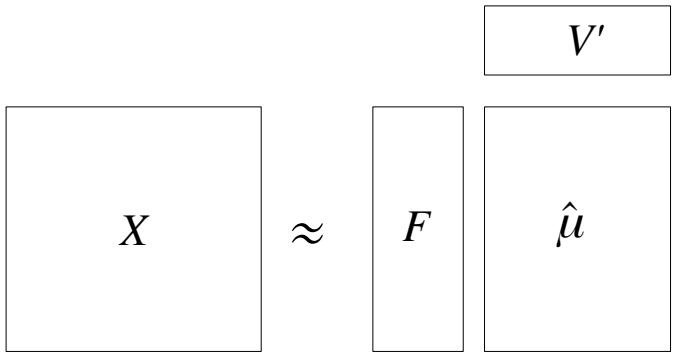
X

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

$\hat{\mu}$

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





$\Rightarrow$ Minimizes distance between observations and their projection

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

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

SVD X : $\hat{\mu}^{\text{PCA}} = U_{n\times k} D_{k\times k} V'_{p\times k}$ $\quad F = UD$ PC - scores

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

Missing values in PCA

$\Rightarrow$ PCA : least squares

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

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

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

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

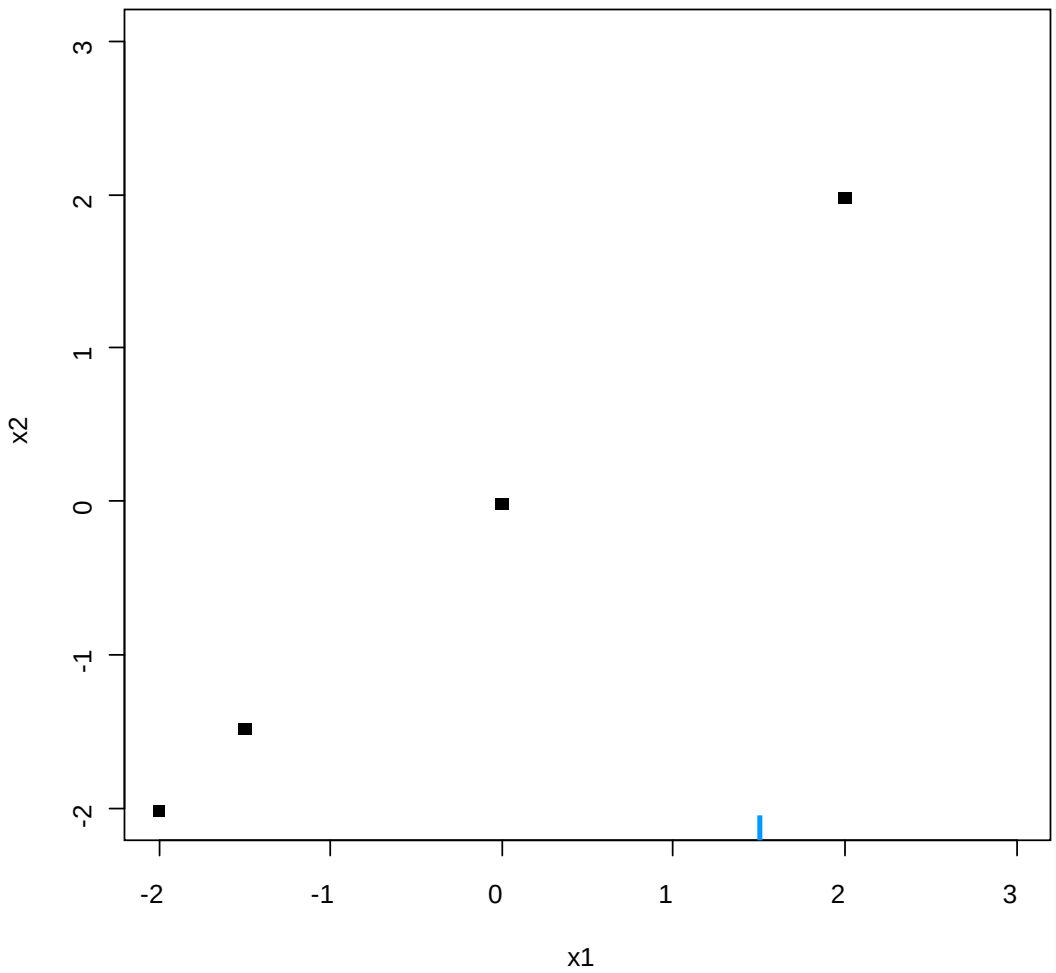
Many algorithms :

Gabriel & Zamir, 1979 : weighted alternating least squares (without explicit imputation)

Kiers, 1997 : iterative PCA (with imputation)

Iterative PCA

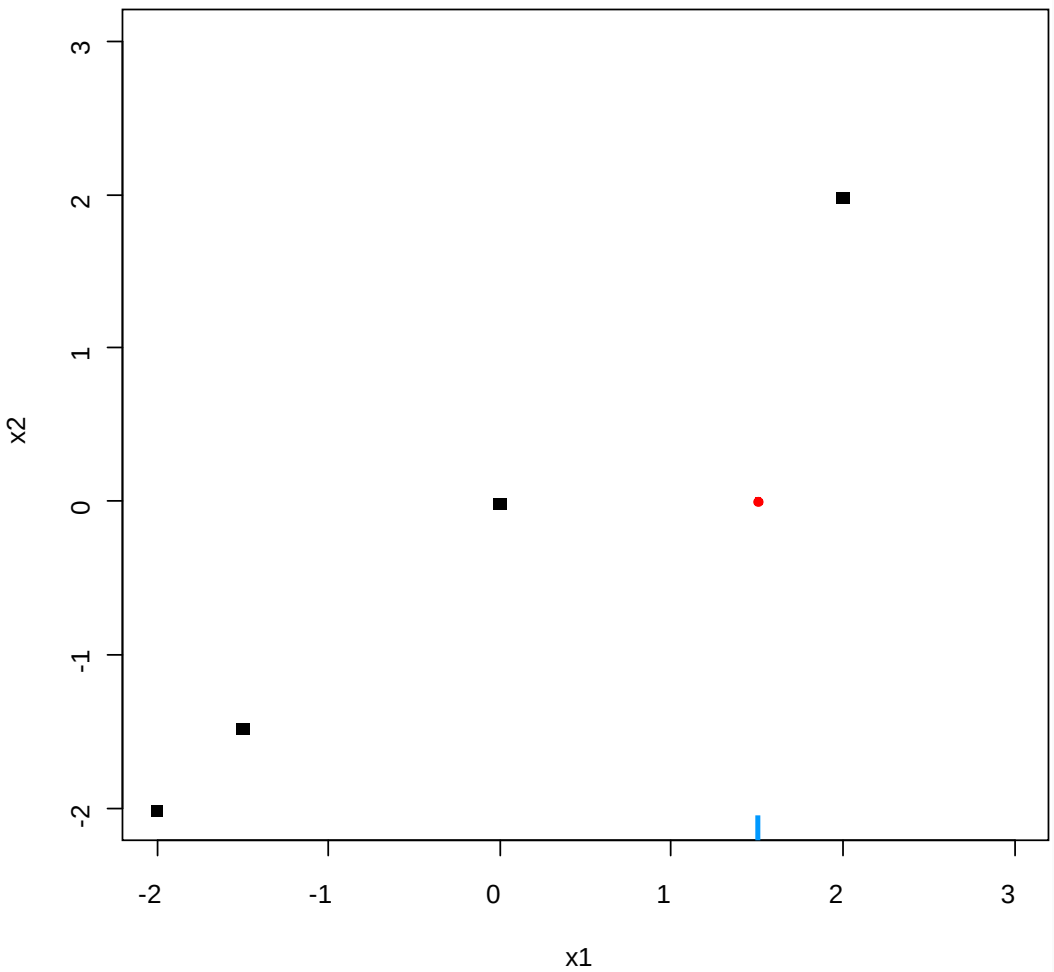
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98



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



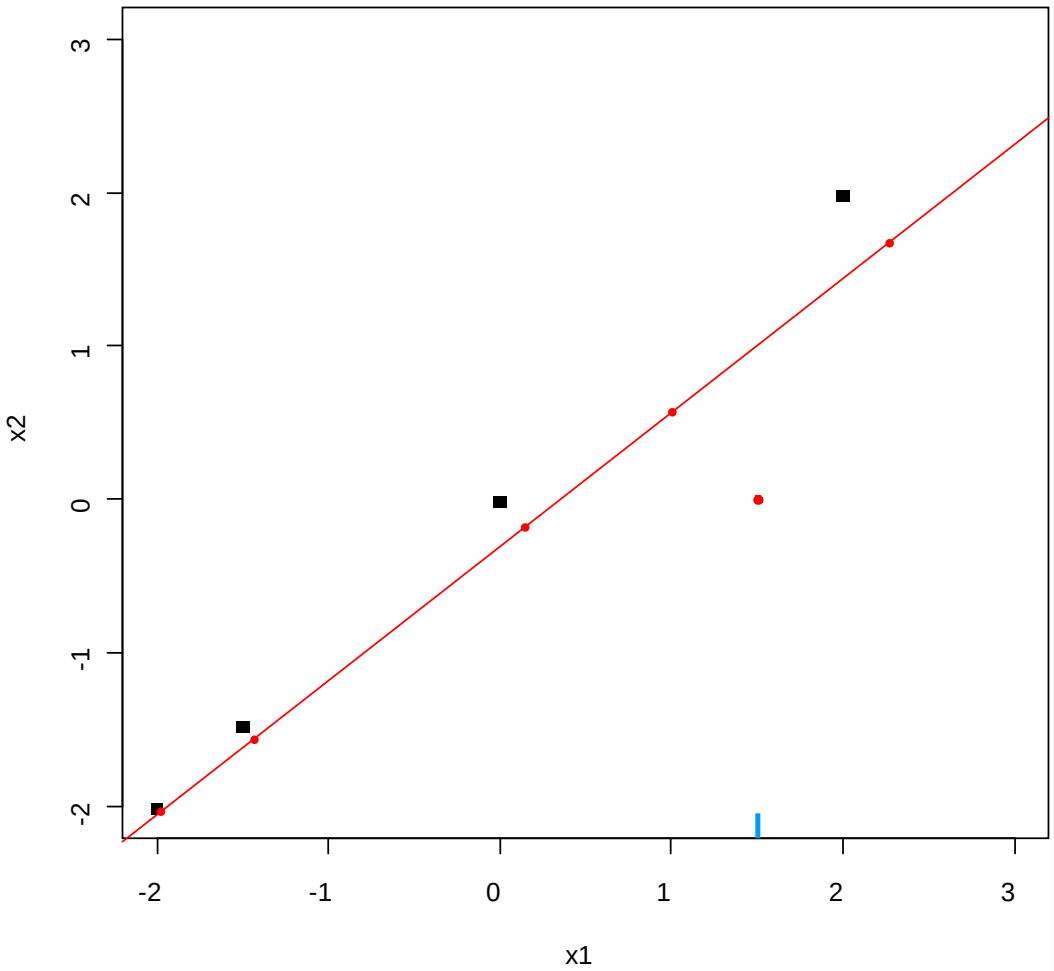
Initialization $\ell = 0 : X^0$ (mean imputation)

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



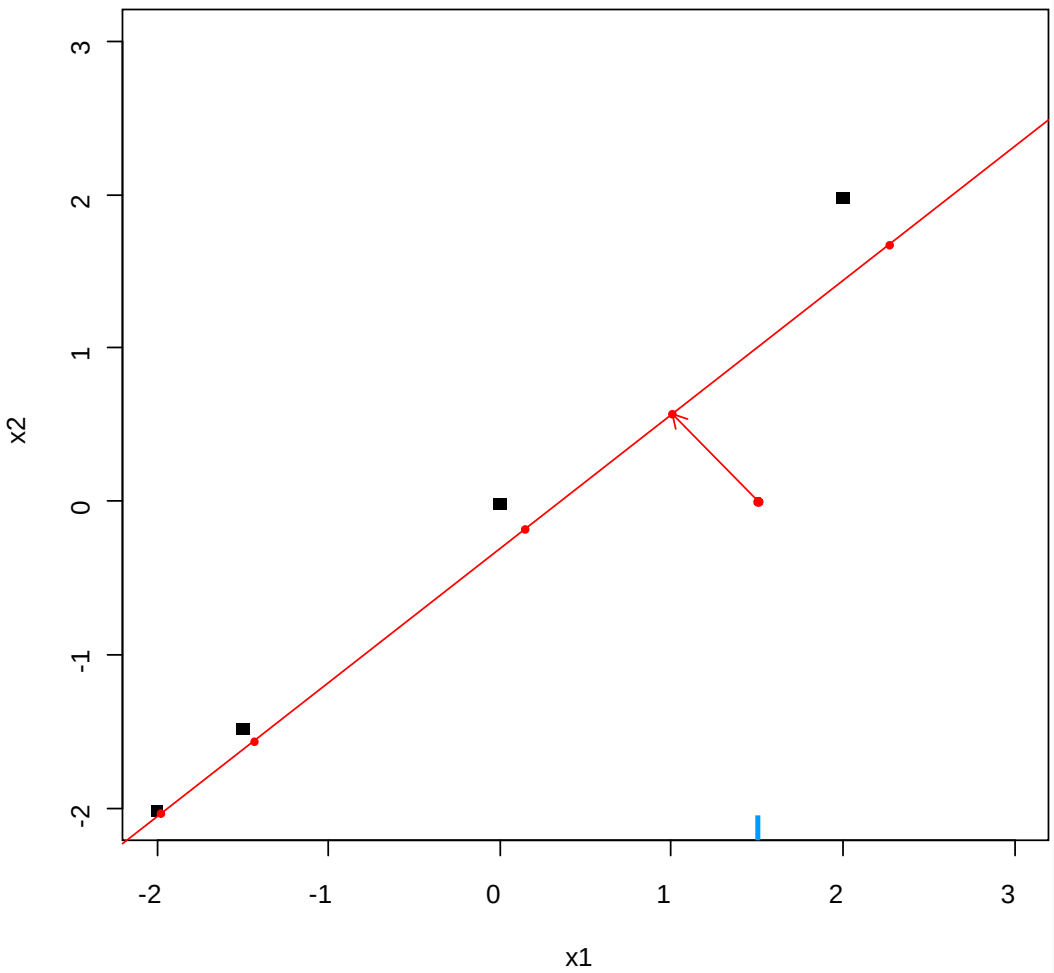
PCA on the completed data set $\rightarrow (U^\ell, \Lambda^\ell, D^\ell)$;

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
```

```
   ^     ^
  x1    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 $\hat{\mu}^{\ell} = U^{\ell} D^{\ell} V^{\ell\prime}$

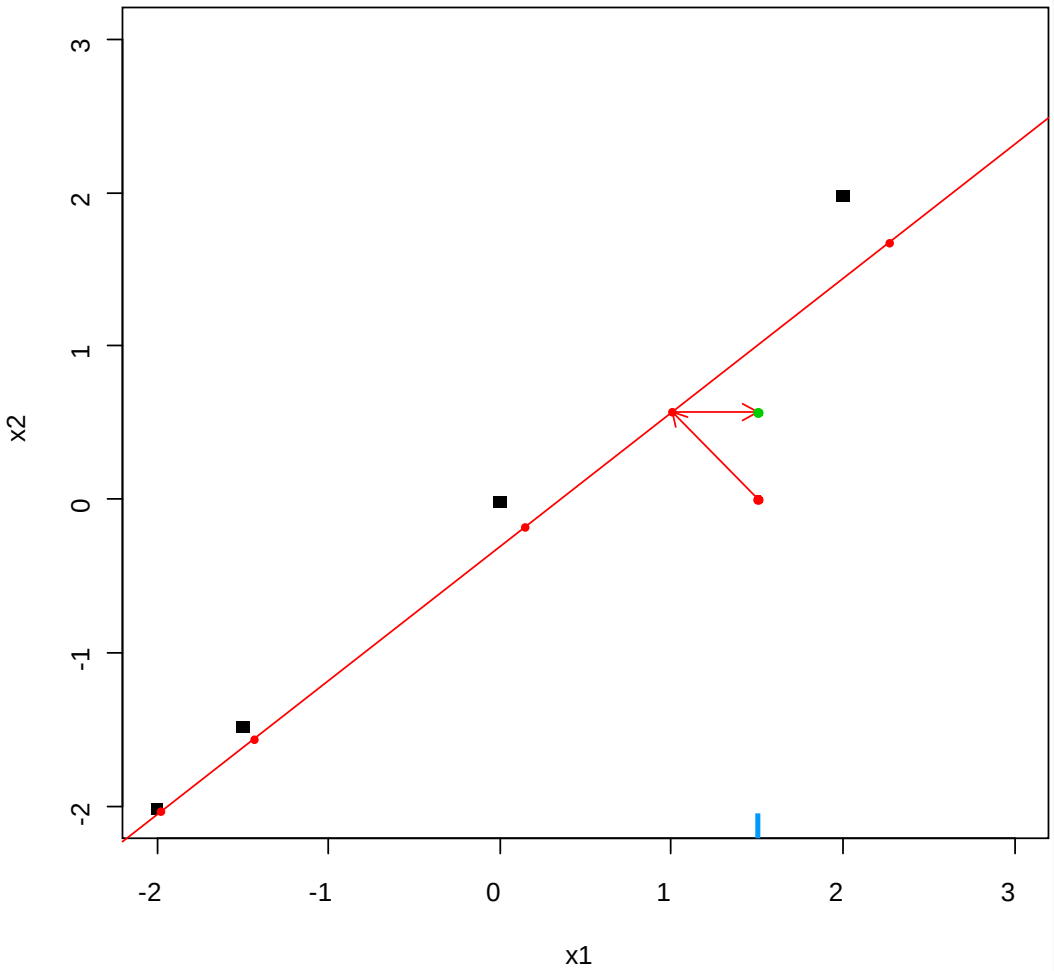
Iterative PCA

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



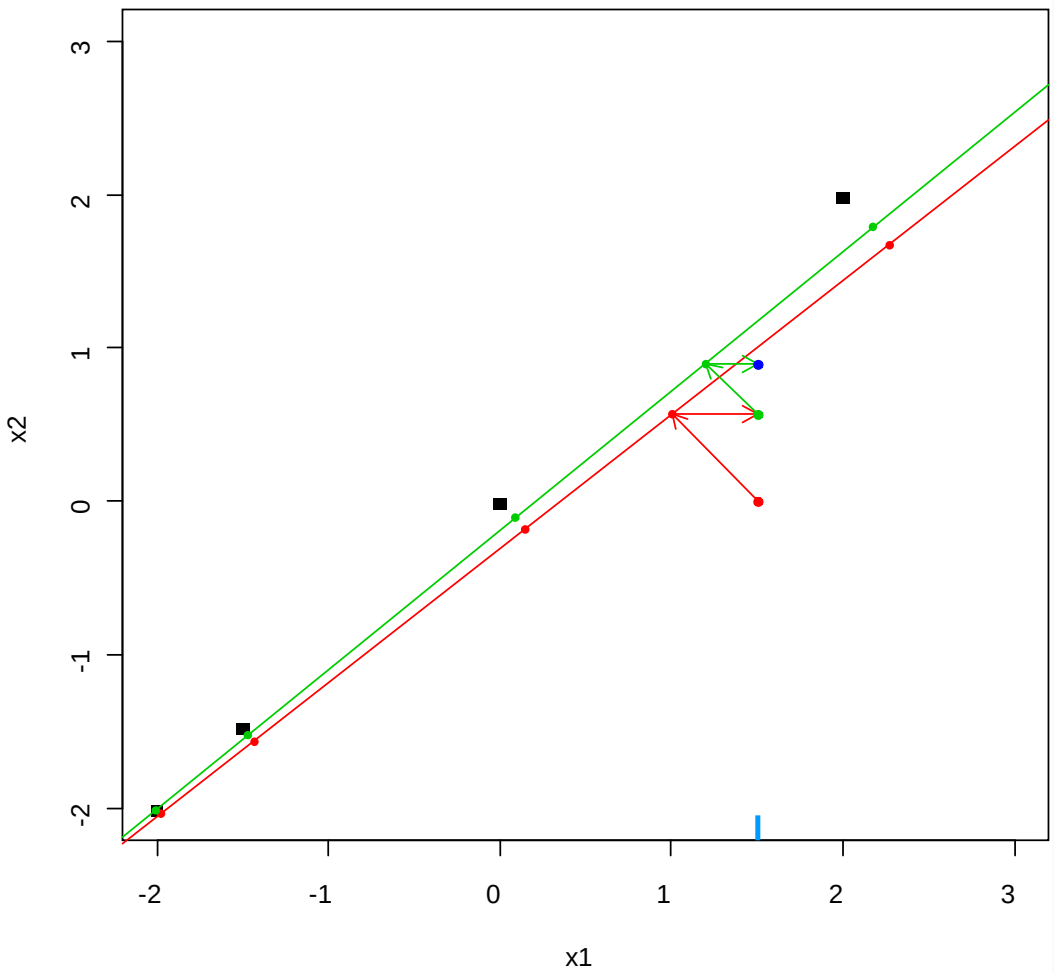
The new imputed dataset is $\hat{X}^{\ell} = W \odot X + (\mathbf{1} - W) \odot \hat{\mu}^{\ell}$

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
```



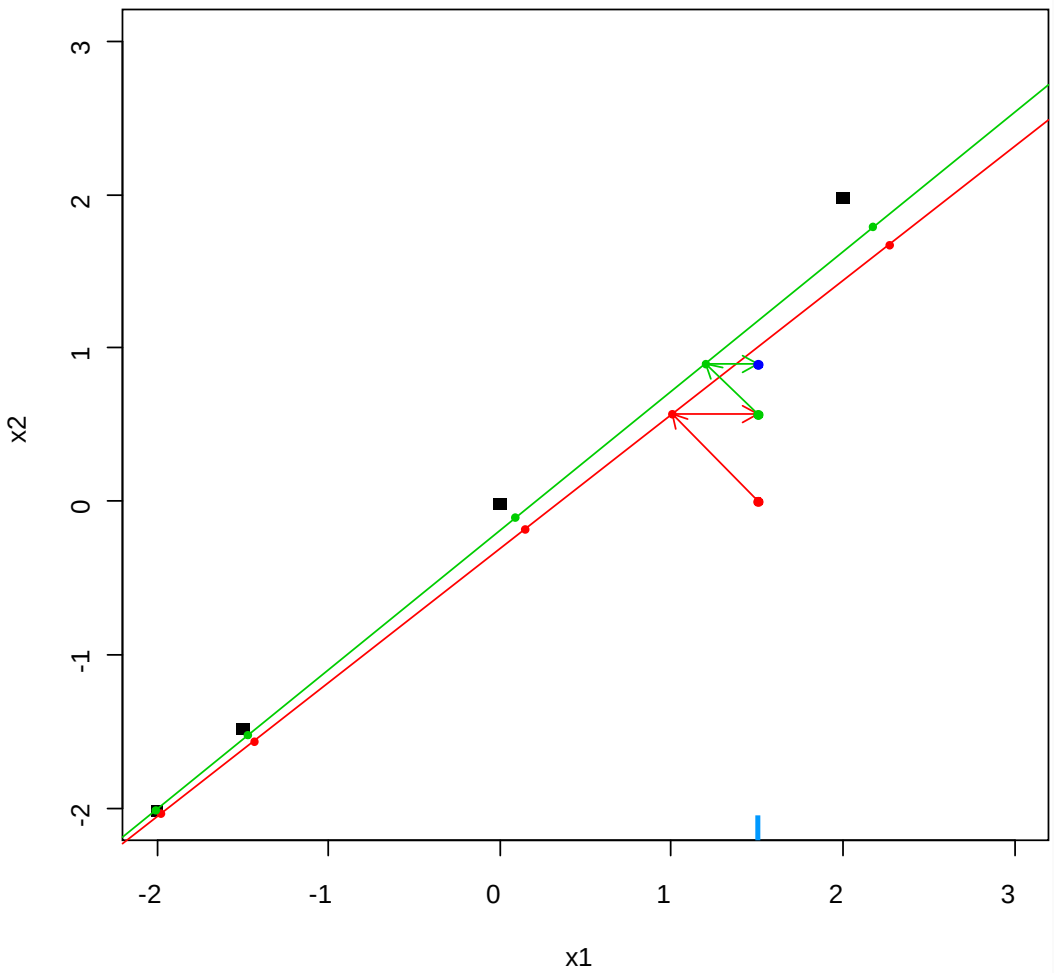
Iterative PCA

```
  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
```

```
   ^     ^
  x1    x2
-2.00 -2.01
-1.47 -1.52
 0.09 -0.11
 1.20  0.90
 2.18  1.78
```

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



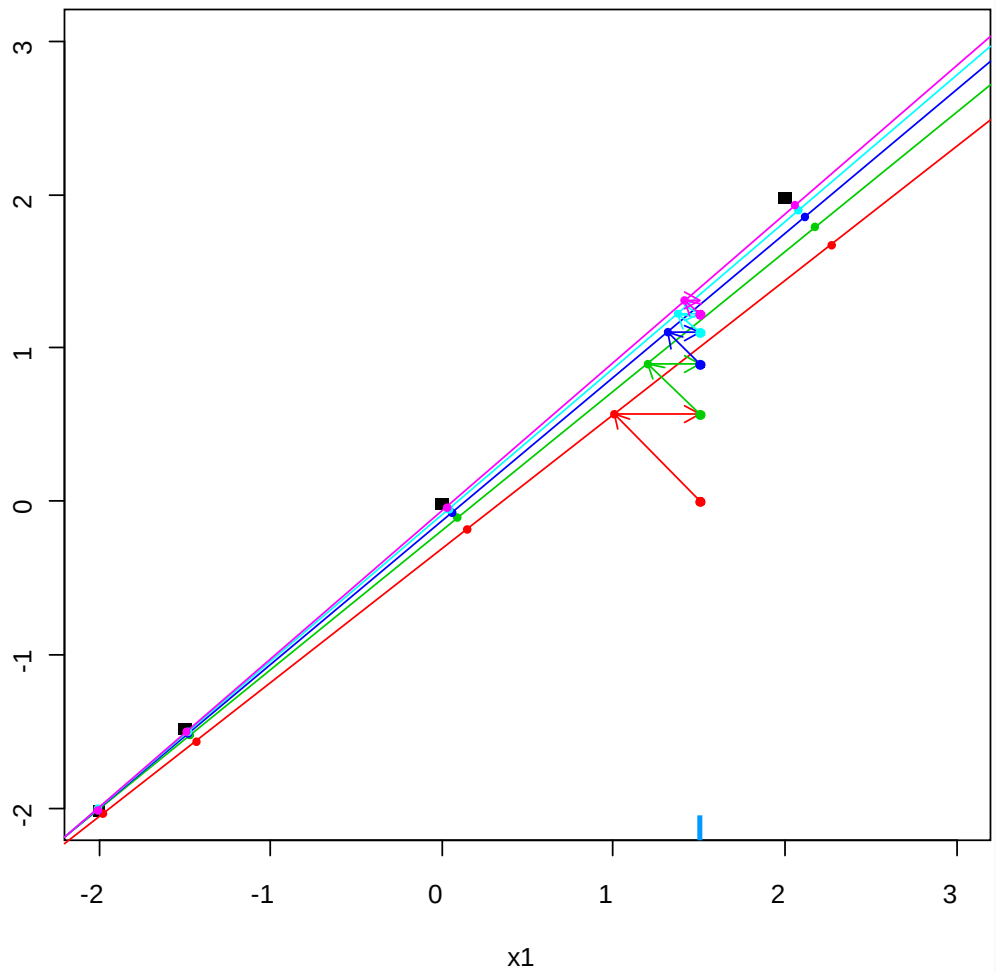
Iterative PCA

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

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

$\hat{x}1$	$\hat{x}2$
-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

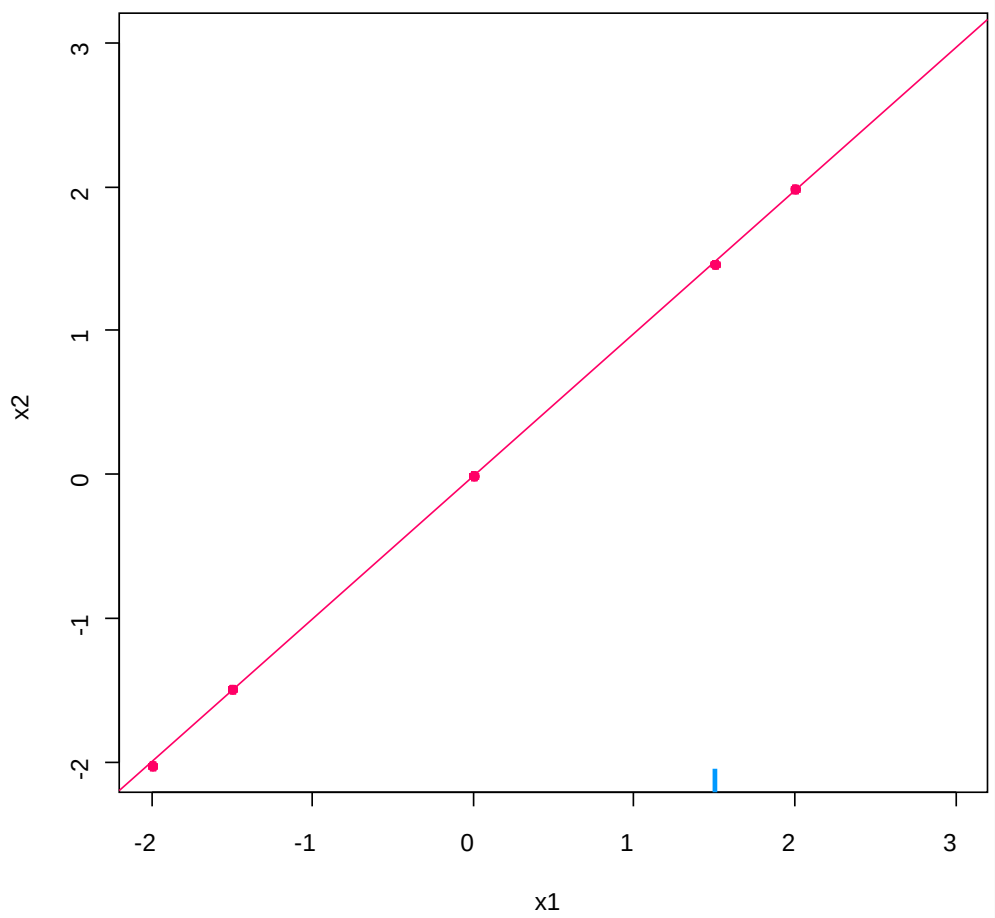


Steps are repeated until convergence

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 $\rightarrow (U^\ell, D^\ell, V^\ell)$

Missing values imputed with the fitted matrix $\hat{\mu}^\ell = U^\ell D^\ell V^{\ell\prime}$

Iterative PCA

1. initialization $\ell = 0$: X^0 (mean imputation)
2. step ℓ :
   (a) PCA on the completed data $\rightarrow (U^\ell, D^\ell, V^\ell)$; k dim kept
   (b) $\hat{\mu}^{\text{PCA}} = \sum_{q=1}^{k} d_q u_q v_q^{'}$ $\qquad X^\ell = W \odot X + (1 - W) \odot \hat{\mu}^\ell$
3. steps of estimation and imputation are repeated

$\Rightarrow$ Overfitting : nb param $(U_{n\times k}, V_{k\times p})$/obs values : k large - NA ; noisy

Regularized versions. Imputation is replaced by

$$(\hat{\mu})_\lambda = \sum_{q=1}^{p} (d_q - \lambda)_+ u_q v_q^{'} \arg\min_\mu \left\{ \|W \odot (X - \mu)\|_2^2 + \lambda \|\mu\|_* \right\}$$

Different regularization : Hastie et.al. (2015) (softimpute), Verbank, J. & Husson (2013) ; Gavish & Donoho (2014), J. & Wager (2015), J. & Sardy (2014), etc.

$\Rightarrow$ Iterative SVD algo good to impute data (matrix completion, Netflix)

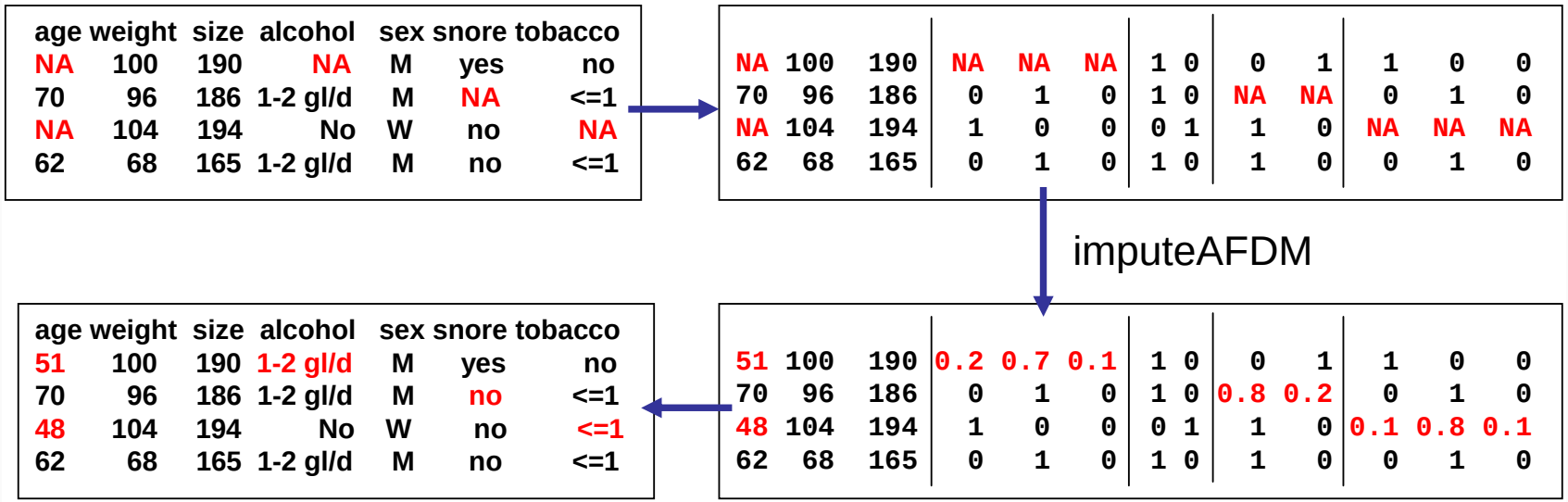
$\Rightarrow$ Model makes sense : data = rank k signal+ noise

$X = \mu + \varepsilon$ $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ with μ of low rank

(Udell & Townsend, 2017)

Iterative SVD

⇒ Imputation with FAMD for mixed data :



⇒ Multilevel imputation : hospital effect with patient nested in hospital.

(J., Husson, Robin & Balasu., 2018, Imputation of mixed data with multilevel SVD. *JCGS*)

package `MissMDA`.

Iterative random forests imputation

Imputation with fully conditional specification (FCS). Impute with a joint model defined implicitely through the conditional distributions (`mice`). Here, imputation model for each variable is a forest.

1. Initial imputation : mean imputation - random category
2. for t in $1 : T$ loop through iterations t
3. for j in $1 : p$ loop through variables j

   Define currently complete data set except $X^t_{-j} = (X^t_1, X^t_{j-1}, X^{t-1}_{j+1}, X^{t-1}_p)$, then X^t_j is obtained by
   - fitting a RF X^{obs}_j on the other variables X^t_{-j}
   - predicting X^{miss}_j using the trained RF on X^t_{-j}

R package `missForest` (Stekhoven & Buhlmann, 2011)

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

$\Rightarrow$ 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

$\Rightarrow$ Random forests

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

$\Rightarrow$ PCA

$\Rightarrow$ RF good for non linear relationship / PCA linear relation

$\Rightarrow$ RF computationally costly $\Rightarrow$ Imputation inherits from the method

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

$x = (x_{ij})$ a $n \times p$ matrix of quantitative covariates

$y = (y_i)$ an n-vector of binary responses $\{0, 1\}$

Logistic regression model

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

Covariables

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

Log-likelihood for complete-data with $\theta = (\mu, \Sigma, \beta)$

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

Decomposition : $x = (x_{\text{obs}}, x_{\text{mis}})$

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

Stochastic Approximation EM

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

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

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

$$\Rightarrow \textit{Unfeasible computation of expectation}$$

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

$$\Rightarrow \textit{Require a huge number of samples}$$

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

Comparison with competitors (mice). Unbiased, good coverage.
Time : $n = 1000$, MCEM : 700s - SAEM : 13s - mice 1s

An integrated procedure

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

How to estimate *observed likelihood* ?

$$\mathrm{p}(y_i, x_{i,\mathrm{obs}}; \theta) = \int \mathrm{p}(y_i, x_{i,\mathrm{obs}} | x_{i,\mathrm{mis}}; \theta) \mathrm{p}(x_{i,\mathrm{mis}}; \theta) dx_{i,\mathrm{mis}}$$

Empirical mean using sample from the proposal distribution in SAEM.

$\Rightarrow$ Prediction on a test set (with missing entries !).

$$\begin{aligned}
\hat{y} &= \arg\max_y \mathrm{p}(y | x_{\mathrm{obs}}) \\
&= \arg\max_y \int \mathrm{p}(y|x) \mathrm{p}(x_{\mathrm{mis}} | x_{\mathrm{obs}}) dx_{\mathrm{mis}} \\
&= \arg\max_y \mathbb{E}_{\mathrm{p}_{x_{\mathrm{mis}}|x_{\mathrm{obs}}}} \mathrm{p}(y|x) \\
&= \arg\max_y \sum_{m=1}^{M} \mathrm{p}\left(y | x_{\mathrm{obs}}, x_{\mathrm{mis}}^{(m)}\right).
\end{aligned}$$

$\Rightarrow$ R package `misaem`

Random forests with missing values

Erwan Scornet, Nicolas Prost, Gael Varoquaux,

Stefan Wager

Missing values and causal inference

Confounders estimated by matrix factorization on the observed covariates (Kallus, Mao and Udell (2018)) .

Main assumptions

- Linear regression model $Y_i = U_i^T \alpha + \tau T_i + \varepsilon_i$.
- Covariates X are noisy and incomplete proxies of true confounders U.
- X is a noisy realization of a low rank matrix $X = UV' + \varepsilon$.
- MCAR.

Results :

Under appropriate assumptions on α, $\|U\|$, the relationship between U and T and the estimation of $col(U)$ by $col(\hat{U})$ and assuming unconfoundedness, then the resulting ATE estimator is consistent.

In practice : perform MF on X, keep U, perform the linear regression and estimate ATE with $\hat{\tau}$

Results

Many choices, issues in practice....

- Coding issues : recode certain not really missing values, for ex Glasgow score ($\in \{3, \ldots, 15\}$) is missing for deceased patients. Recode by a category or a constant (lower bound min(GCS)=3).

Many choices, issues in practice....

- Coding issues : recode certain not really missing values, for ex Glasgow score ($\in \{3, \ldots, 15\}$) is missing for deceased patients. Recode by a category or a constant (lower bound min(GCS)=3).
- Impute with iterative FAMD (out-of-range imputation), Random forests (computational costly), mice (invertibility pbs with many categories) ?
- Which observations ? All individuals (TBI and no-TBI patients)
- Which variables ? All available variables or the pre-selected ones
- Impute with treatment, covariates and outcome ? (Impute with Y ?)

Many choices, issues in practice....

- Coding issues : recode certain not really missing values, for ex Glasgow score ($\in \{3, \ldots, 15\}$) is missing for deceased patients. Recode by a category or a constant (lower bound min(GCS)=3).
- Impute with iterative FAMD (out-of-range imputation), Random forests (computational costly), mice (invertibility pbs with many categories) ?
- Which observations ? All individuals (TBI and no-TBI patients)
- Which variables ? All available variables or the pre-selected ones
- Impute with treatment, covariates and outcome ? (Impute with Y ?)

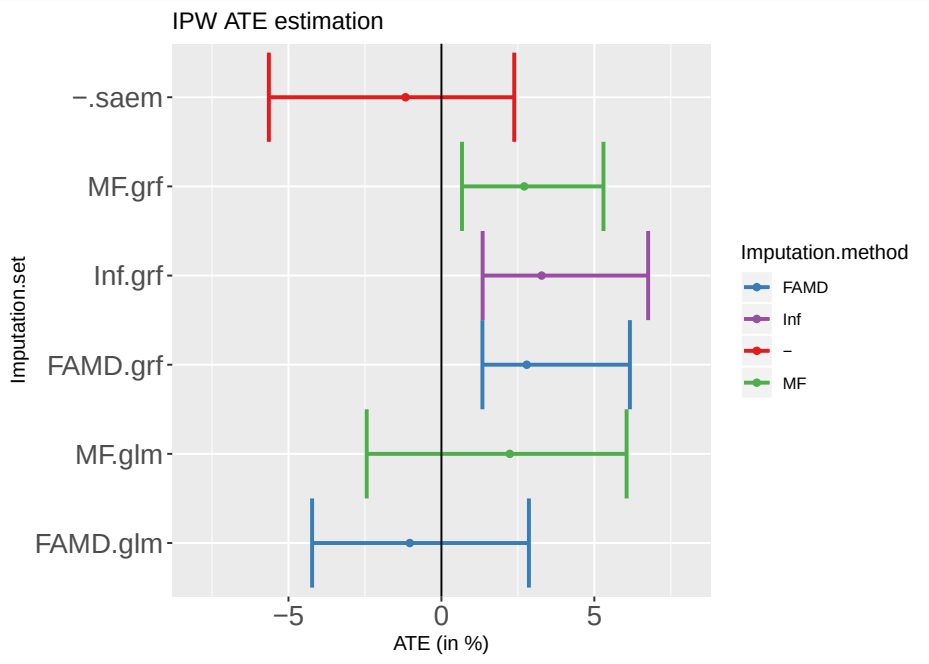
Imputation (FAMD, RF)+ IPW : $\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right)$
Model treatment on covariates $e(x) = \mathbb{P}(W_i = 1 \,|\, X_i = x)$ weights : **GLM, GRF, GBM**. Trimming (0.1% & 99.9% quantiles to threshold the weights).

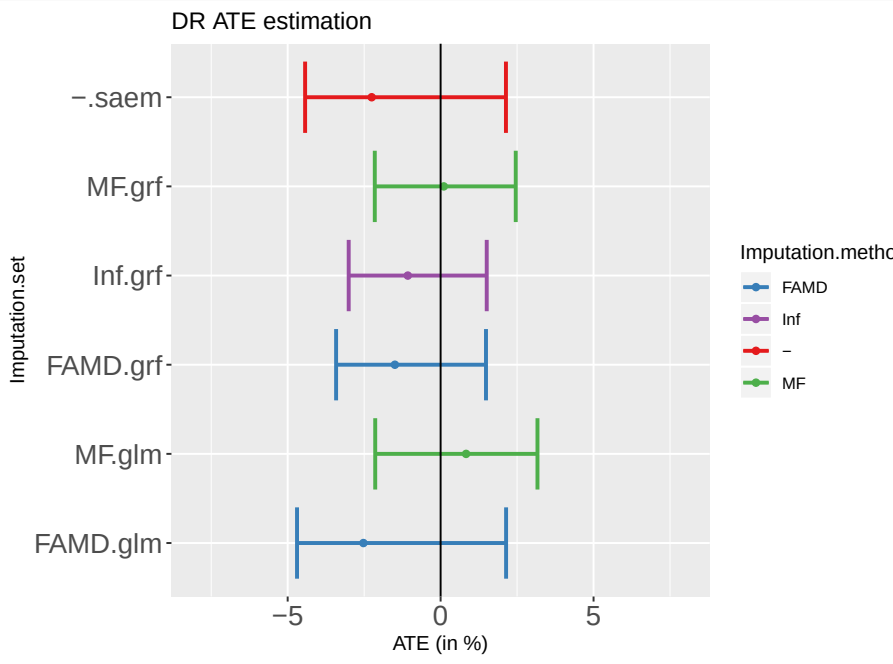
SAEM (quanti) + IPW (weights glm, trimming, grf)

Imputation (FAMD, RF) + double robusts : models outcome on covariates and treatment on covariates (**GLM, RF, GBM**)

Results

ATE estimations (bootstrap CI) for the effect of Tranexamic acid on in-ICU mortality for TBI patients. Imputations/SAEM on all patients (TBI + no-TBI).





(*y*-axis : `imputation . ps estimation`), (*x*-axis : ATE estimation with bootstrap CI)

We compute the mortality rate in the treated group and the mortality rate in the control group (after covariate balancing). The value obtained corresponds to the **difference in percentage points between mortality rates in treatment and control**.

Conclusion

On-going work, perspectives

Methodology/Theory

- Different missing values mechanism. Sportisse, A., Boyer, C. and Josse, J. Low-rank estimation with missing non at random data.
- Logistic regression for mixed variables.
- Identify subgroups of patients who could benefit from treatment ? Optimal Prescription Trees (Bertsimas et al., 2018).
- Heterogeneous treatment effects (Athey and Imbens, 2015) and optimal policy learning (Imai and Ratkovic, 2013).
- Multiple imputation.
- Towards more complex treatment strategies : Do certain treatment strategies, i.e. bundles of treatments (administration of noradrenaline and SSH and tranexamic acid, etc.), have an effect on 24h mortality, on 14d mortality, etc. ?
- Consistency of ATE estimators with missing values.

On-going work, perspectives

Traumabase - Traumatic brain injury

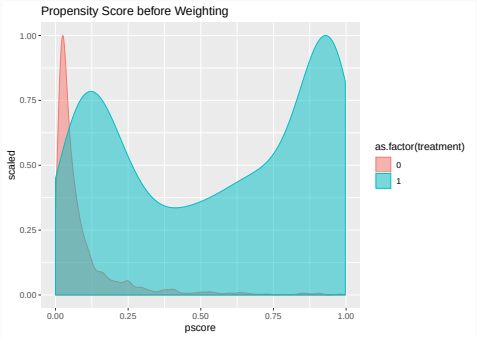
- Bias of mortality (dead before receiving ?)
- Unconfoundedness ?
- Choice of pre- and post-treatment covariates. Depending on future application. Ideally real-time treatment decision $\rightarrow$ learning optimal treatment policies.

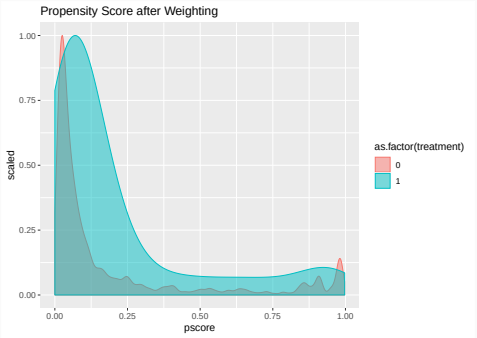
Do you have any questions or comments ?

Results

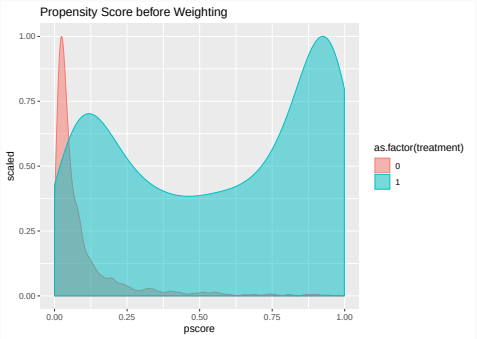
Imputations/SAEM **on all patients**. PS estimated with **logistic regr.**

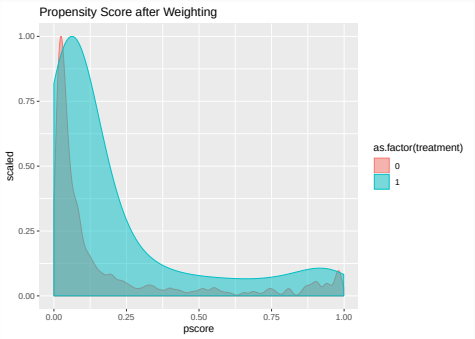
- (SAEM)



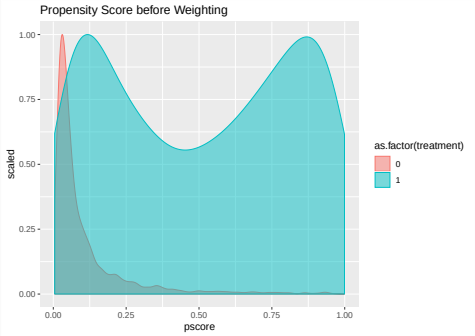


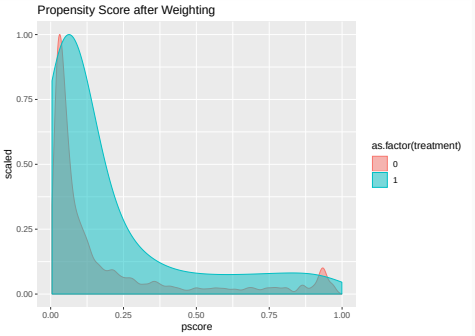
FAMD





MF (Udell)

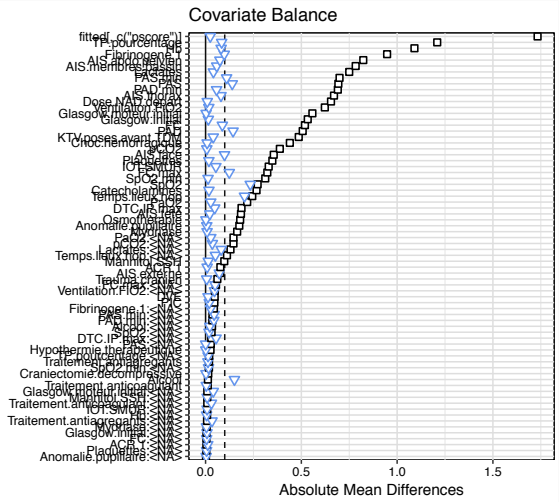


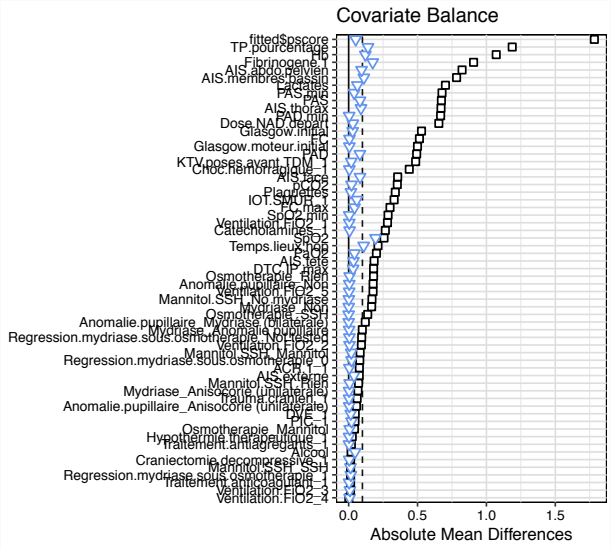


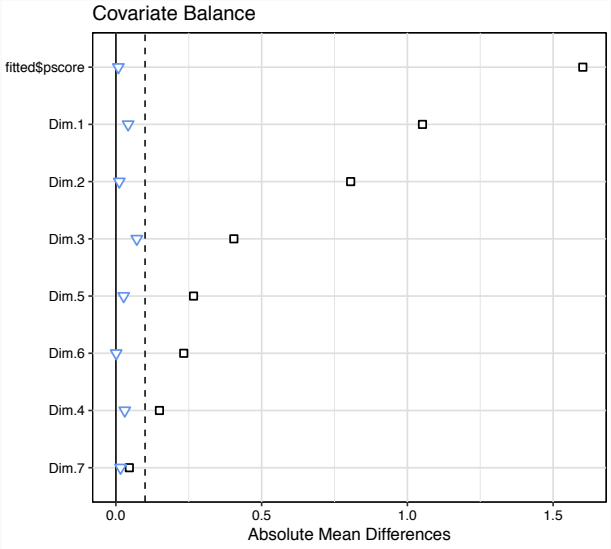
Results

Imputations/SAEM on all patients. PS estimated with **logistic regr.**

- (SAEM) FAMD MF (Udell)



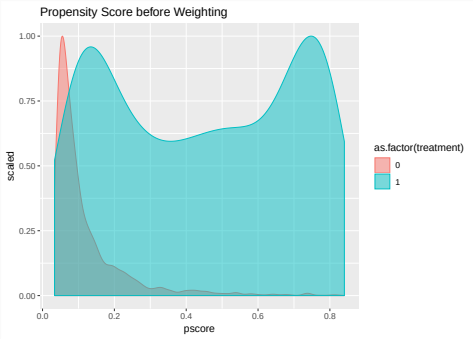


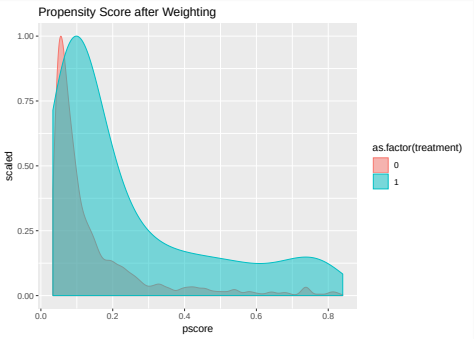


Results

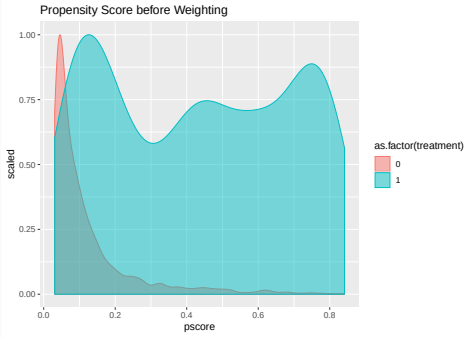
Imputations on all patients. PS estimated with **grf**.

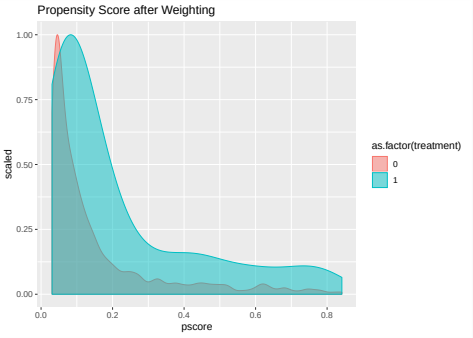
FAMD





MF (Udell)

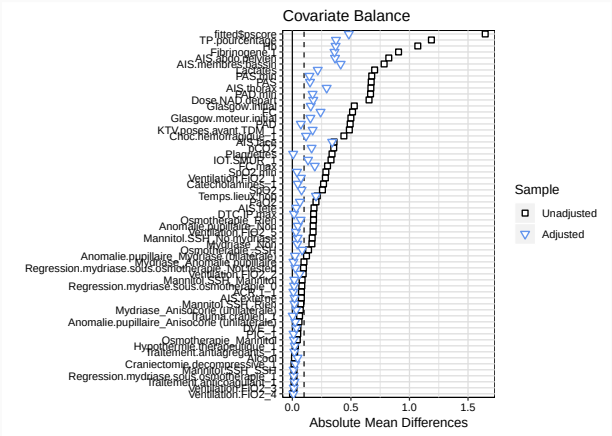




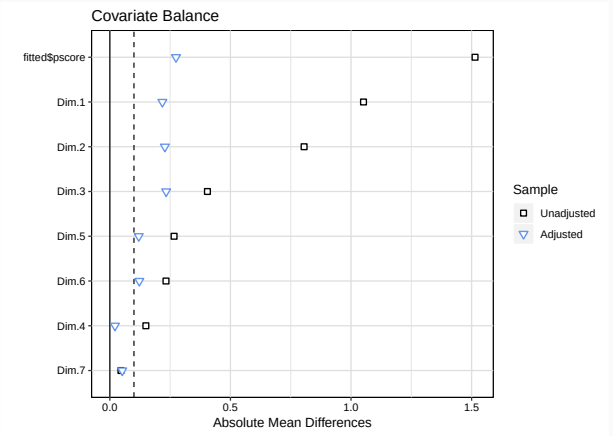
Results

Imputations **on all patients**. PS estimated with **grf**.

FAMD



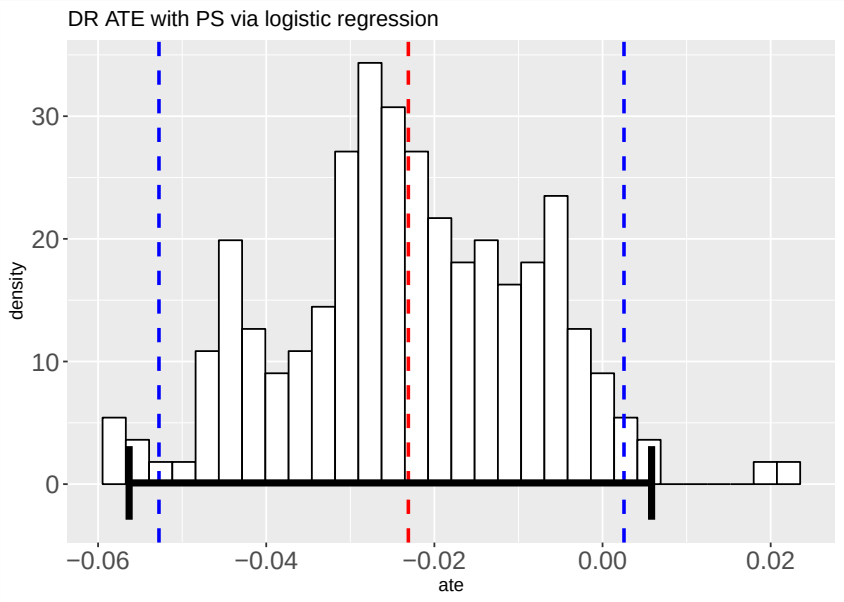
MF (Udell)



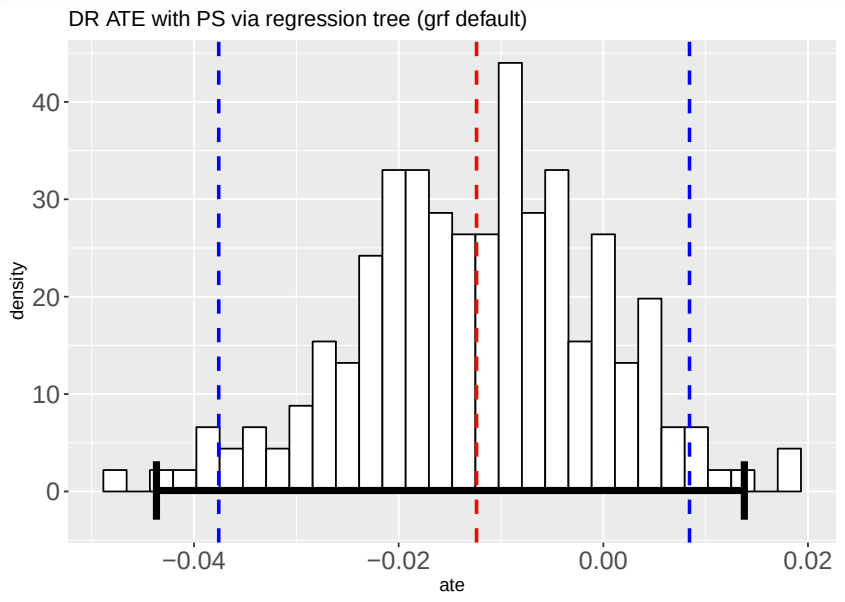
Results

FAMD imputation on all patients. Bootstrap CI for DR ATE estimations. ($N = 200$).

Logistic regression



(Generalized) random forest



blue dotted line : Bootstrap quantiles (2.5% and 97.5%)

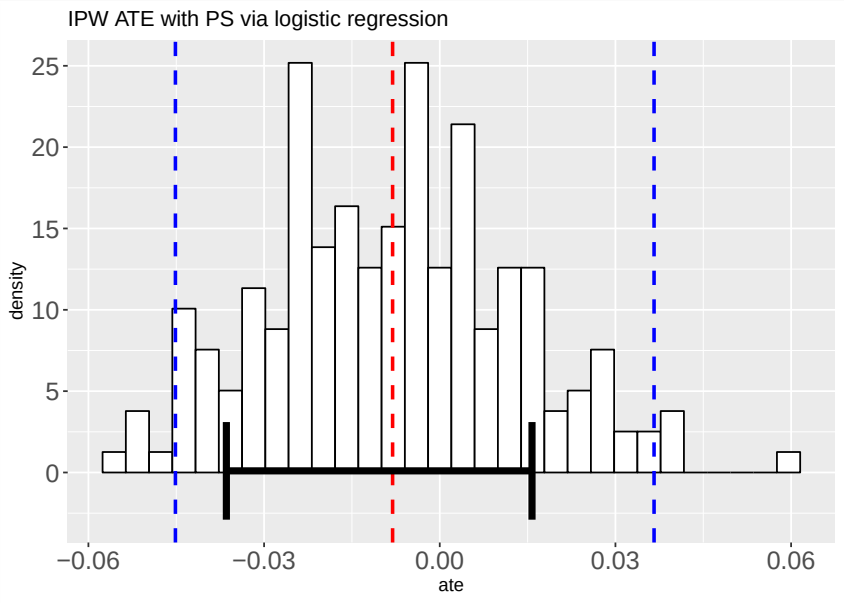
red dotted line : Bootstrap mean

black segment : ATE estimation with $\pm 1.96SE$.

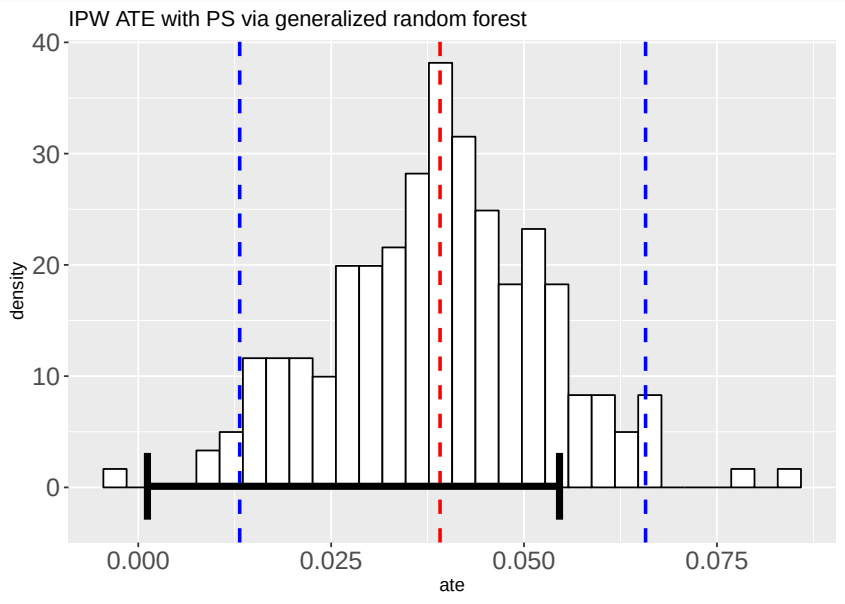
Results

FAMD imputation on all patients. Bootstrap CI for **IPW** ATE estimations. ($N = 200$).

Logistic regression



Generalized random forest



blue dotted line : Bootstrap quantiles (2.5% and 97.5%)

red dotted line : Bootstrap mean

black segment : ATE estimation with $\pm 1.96SE$.

References i


S. Athey, J. Tibshirani, and S. Wager.
Generalized random forests.
Ann. Statist., 47(2) :1148–1178, 2019.


J. Carpenter and M. Kenward.
Multiple Imputation and its Application.
Wiley, Chichester, West Sussex, UK, 2013.


S. R. Hamada, J. Josse, S. Wager, T. Gauss, et al.
Effect of fibrinogen administration on early mortality in traumatic haemorrhagic shock : a propensity score analysis.
Submitted, 2018.

References ii


W. Jiang, J. Josse, M. Lavielle, et al.
Stochastic approximation em for logistic regression with missing values.
arXiv preprint arXiv :1805.04602, 2018.


J. Josse, J. Pagès, and F. Husson.
Multiple imputation in principal component analysis.
Advances in Data Analysis and Classification, 5(3) :231–246, 2011.


N. Kallus, X. Mao, and M. Udell.
Causal inference with noisy and missing covariats via matrix factorization.
arXiv preprint, 2018.


J. K. Kim and J. Shao.
Statistical Methods for Handling Incomplete Data.
Chapman and Hall/CRC, Boca Raton, FL, USA, 2013.

References iii


R. J. A. Little and D. B. Rubin.
Statistical Analysis with Missing Data.
Wiley, 2002.


D. K. Menon and A. Ercole.
Critical care management of traumatic brain injury.
In *Handbook of clinical neurology*, volume 140, pages 239–274. Elsevier, 2017.


J. M. Robins, A. Rotnitzky, and L. P. Zhao.
Estimation of regression coefficients when some regressors are not always observed.
Journal of the American Statistical Association, 89(427) :846–866, 1994.

References iv


J. L. Schafer and J. W. Graham.
Missing data : our view of the state of the art.
Psychological Methods, 7(2) :147–177, 2002.


A. Sportisse, C. Boyer, and J. Josse.
Imputation and low-rank estimation with missing non at random data.
arXiv preprint arXiv :1812.11409, 2018.


S. van Buuren.
Flexible Imputation of Missing Data.
Chapman and Hall/CRC, Boca Raton, FL, 2018.


S. Wager.
Lecture notes in causal inference and treatment effect estimation (oit 661), 2018.