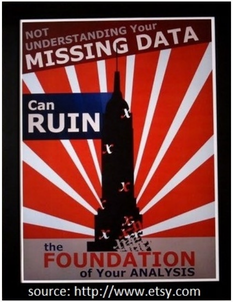
On the consistency of supervised learning with missing values

Julie Josse, Nicolas Prost, Erwan Scornet, Gael Varoquaux
CMAP, INRIA Parietal

4 April 2019

https://hal.archives-ouvertes.fr/hal-02024202

Google Brain, Paris



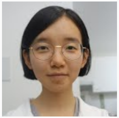
Overview

1. Introduction
2. Handling missing values (inferential framework)
3. Supervised learning with missing values
4. Decision trees
5. Discussion

Introduction

Collaborators

- PhD students : G. Robin, W. Jiang, I. Mayer, **N. Prost**, (X)
- Colleagues : J-P Nadal (EHESS), **E. Scornet (X)**, **G. Varoquaux (INRIA)**, S. Wager (Stanford), B. Naras (Stanford)
- Traumabase (hospital) : T. Gauss, S. Hamada, J-D Moyer
- Capgemini







Traumabase

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

Center	Accident	Age	Sex	Weight	Height	BMI	BP	SBP
Beaujon	Fall	54	m	85	NR	NR	180	110
Pitie Salpetriere	Gun	26	m	NR	NR	NR	131	62
Beaujon	AVP moto	63	m	80	1.8	24.69	145	89
Pitie Salpetriere	AVP pedestrian	30	w	NR	NR	NR	107	66
HEGP	White weapon	16	m	98	1.92	26.58	118	54

....

SpO2	Temperature	Lactates	Hb	Glasgow	Transfusion	....
97	35.6	<NA>	12.7	12	yes	
100	36	3.9	11.4	3	no	
100	36	NM	14.4	15	no	
100	36.6	NM	14.3	15	yes	

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

Traumabase

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

Center	Accident	Age	Sex	Weight	Height	BMI	BP	SBP
Beaujon	Fall	54	m	85	NR	NR	180	110
Pitie Salpetriere	Gun	26	m	NR	NR	NR	131	62
Beaujon	AVP moto	63	m	80	1.8	24.69	145	89
Pitie Salpetriere	AVP pedestrian	30	w	NR	NR	NR	107	66
HEGP	White weapon	16	m	98	1.92	26.58	118	54

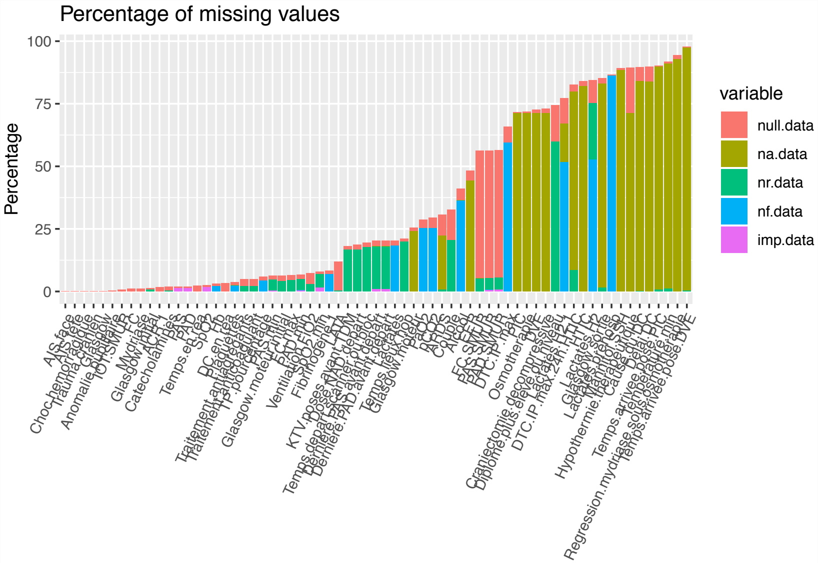
....

SpO2	Temperature	Lactates	Hb	Glasgow	Transfusion	....
97	35.6	<NA>	12.7	12	yes	
100	36	3.9	11.4	3	no	
100	36	NM	14.4	15	no	
100	36.6	NM	14.3	15	yes	

⇒ **Predict** whether to start a blood transfusion, the risk of hemorrhagic shock, etc...

⇒ (Logistic) regressions with missing categorical/continuous values

Missing values



Handling missing values (inferential framework)

Solutions to handle missing values

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

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

Aim : **estimate parameters** and their variance from an incomplete data.
Inferential framework

Solutions to handle missing values

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

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

Difficult to establish ? Not many software even for simple models
One specific algorithm for each statistical method...

Aim : **estimate parameters** and their variance from an incomplete data.
Inferential framework

Solutions to handle missing values

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

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

Difficult to establish ? Not many software even for simple models
One specific algorithm for each statistical method...

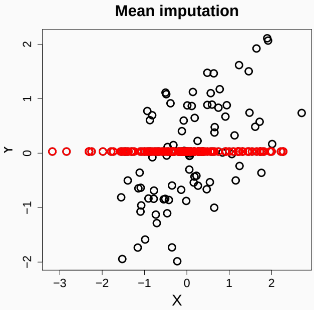
Imputation (multiple) to get a complete data set

on which you can perform any statistical method (Rubin, 1976)
Ex logistic regression : impute and apply logistic model to get $\hat{\beta}$, $\hat{V}(\hat{\beta})$

Aim : **estimate parameters** and their variance from an incomplete data.
Inferential framework

Dealing with missing values

$\Rightarrow$ Imputation to get a complete data set

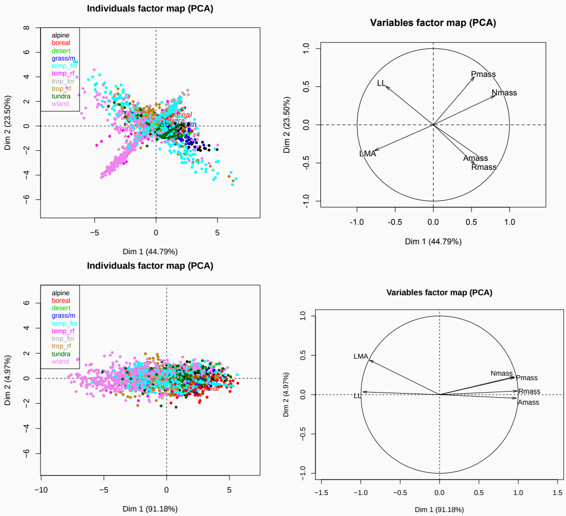


$\mu_y = 0$
$\sigma_y = 1$
$\rho = 0.6$

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

Mean imputation deforms joint and marginal distributions

Dealing with missing values

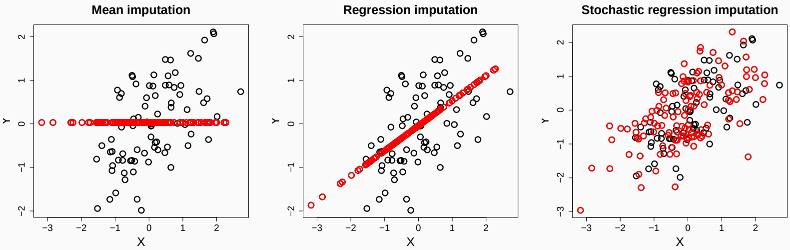


$\Rightarrow$ Mean imputation is bad for estimation

Wright IJ, et al. (2004). The worldwide leaf economics spectrum. *Nature*, 69 000 species - LMA (leaf mass per area), LL (leaf lifespan), Amass (photosynthetic assimilation), Nmass (leaf nitrogen), Pmass (leaf phosphorus), Rmass (dark respiration rate)

Imputation methods

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



	Mean	Regression	Stochastic regression
$\mu_y = 0$	0.01	0.01	0.01
$\sigma_y = 1$	0.5	0.72	0.99
$\rho = 0.6$	0.30	0.78	0.59

Imputation methods

Assuming a joint model

- Gaussian distribution : $x_{i.} \sim \mathcal{N}(\mu, \Sigma)$ (package Amelia)
- low rank : $X_{n \times d} = \mu_{n \times d} + \varepsilon \; \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ with μ of low rank k
  (package softimpute, Hastie; missMDA, Josse)
- latent class - nonparametric Bayesian (package dpmpm, Reiter)
- deep learning using variational autoencoders (MIWAE, Mattei, 2018)

Using conditional models (joint implicitly defined)

- with logistic, multinomial, poisson regressions (mice, Van Buuren)
- iterative impute each variable by random forests (missForest, Buhlmann)

Imputation for categorical, mixed, multilevel/blocks data, etc.

$\Rightarrow$ R-miss-tastic missing values plateform

Aim is not to impute but estimate parameters & variance (multiple imputation)

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

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

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

Covariables

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

Log-likelihood 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 : $x = (x_{\mathrm{obs}}, x_{\mathrm{mis}})$

Under MAR, possibility to ignore the missing value mechanism
Observed likelihood $\arg\max \mathcal{LL}(\theta; x_{\mathrm{obs}}, y) = \int \mathcal{LL}(\theta; x, y) dx_{\mathrm{mis}}$

Stochastic Approximation EM

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

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

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

$\Rightarrow$ *Unfeasible computation of expectation*

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

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

Supervised learning with missing values

Supervised learning

- A feature matrix $\mathbf{X}$ and a response vector Y
- Find a prediction function that minimizes the expected risk.
  Bayes rule : $f^\star \in \underset{f\colon \mathcal{X}\to\mathcal{Y}}{\arg\min}\ \mathbb{E}\left[\ell(f(\mathbf{X}), Y)\right]$ $f^\star(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$
- Empirical risk minimization :

$$\hat{f}_{\mathcal{D}_{n,\mathrm{train}}} \in \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\min} \left(\frac{1}{n}\sum_{i=1}^{n} \ell\left(f(\mathbf{X}_i), Y_i\right) \right)$$

  A new data $\mathcal{D}_{n,\mathrm{test}}$ to estimate the generalization error rate
- Bayes consistent : $\mathbb{E}[\ell(\hat{f}_n(\mathbf{X}), Y)] \xrightarrow[n\to\infty]{} \mathbb{E}[\ell(f^\star(\mathbf{X}), Y)]$

Supervised learning

- A feature matrix $\mathbf{X}$ and a response vector Y
- Find a prediction function that minimizes the expected risk.

  Bayes rule : $f^\star \in \underset{f\colon \mathcal{X}\to\mathcal{Y}}{\arg\min}\, \mathbb{E}\left[\ell(f(\mathbf{X}), Y)\right]$ $f^\star(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$
- Empirical risk minimization :

$$\hat{f}_{\mathcal{D}_{n,\mathrm{train}}} \in \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\min} \left(\frac{1}{n}\sum_{i=1}^{n} \ell\left(f(\mathbf{X}_i), Y_i\right)\right)$$

  A new data $\mathcal{D}_{n,\mathrm{test}}$ to estimate the generalization error rate
- Bayes consistent : $\mathbb{E}[\ell(\hat{f}_n(\mathbf{X}), Y)] \xrightarrow[n\to\infty]{} \mathbb{E}[\ell(f^\star(\mathbf{X}), Y)]$

Differences with classical litterature

- response variable Y - Aim : Prediction
- two data sets (out of sample) with missing values : train & test sets

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

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

EM and out-of sample prediction

$\mathbb{P}\left(y_i = 1 | x_i; \beta\right) = \frac{\exp(\sum_{j=1}^{d} \beta_j x_{ij})}{1+\exp(\sum_{j=1}^{d} \beta_j x_{ij})}$ After EM $\hat{\theta}_n = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_d, \hat{\mu}, \hat{\Sigma})$

New obs : $x_{n+1} = (x_{(n+1)1}, NA, NA, x_{(n+1)4}, \ldots, x_{(n+1)d})$

Predict Y on **a test set with missing entries** $x_{\text{test}} = (x_{obs}, x_{miss})$

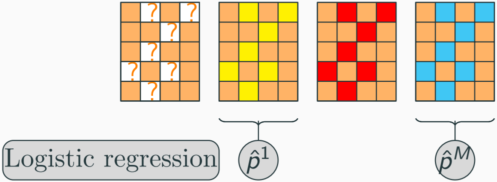
EM and out-of sample prediction

$\mathbb{P}(y_i = 1 | x_i; \beta) = \frac{\exp(\sum_{j=1}^{d} \beta_j x_{ij})}{1+\exp(\sum_{j=1}^{d} \beta_j x_{ij})}$ After EM $\hat{\theta}_n = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_d, \hat{\mu}, \hat{\Sigma})$

New obs : $x_{n+1} = (x_{(n+1)1}, NA, NA, x_{(n+1)4}, \ldots, x_{(n+1)d})$

Predict Y on **a test set with missing entries** $x_{\text{test}} = (x_{obs}, x_{miss})$

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



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

Prediction on test incomplete data with a full data model

- Let a Bayes-consistent predictor f for complete data : $f(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$
- Note the data : $\widetilde{\mathbf{X}} = \mathbf{X} \odot (\mathbf{1} - \mathbf{M}) + \mathtt{NA} \odot \mathbf{M}$ (takes value in $\mathbb{R} \cup \{\mathtt{NA}\}$)
- Perform multiple imputation :

$$f^{\star}_{mult\ \ imput}(\widetilde{\mathbf{x}}) = \mathbb{E}_{\mathbf{X}_m|\mathbf{X}_o=\mathbf{x}_o}[f(\mathbf{X}_m, \mathbf{x}_o)]$$

same as out-of sample EM but assuming know f

Theorem

Consider the regression model $Y = f(\mathbf{X}) + \varepsilon$, where

- we assume **MAR** $\forall \mathcal{S} \subset \{1, \ldots, d\}$, $(M_j)_{j\in\mathcal{S}} \perp\!\!\!\perp (X_j)_{j\in\mathcal{S}} \mid (X_k)_{k\in\mathcal{S}^c}$
- $\varepsilon \perp\!\!\!\perp (M_1, X_1, \ldots, M_d, X_d)$ is a centred noise

Then multiple imputation is consistent :

$$f^{\star}_{mult\ \ imput}(\widetilde{\mathbf{x}}) = \mathbb{E}[Y|\widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}]$$

Proof

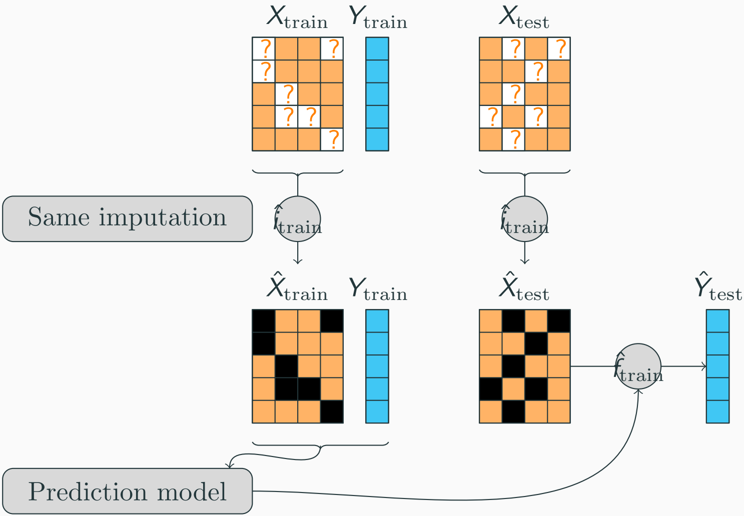
Let $\widetilde{\mathbf{x}} \in (\mathbb{R} \cup \mathtt{NA})^d$. Without loss of generality, assume only $\tilde{x}_1, \dots, \tilde{x}_j$ are NA.

$$\begin{aligned}
f^\star_{mult\ \ imput}(\widetilde{\mathbf{x}}) &= \mathbb{E}_{\mathbf{X}_m|X_o=\mathbf{x}_o}[f(\mathbf{X}_m, X_o = \mathbf{x}_o)] \\
&= \mathbb{E}[f(\mathbf{X}_m, X_o = \mathbf{x}_o)|X_o = \mathbf{x}_o] \\
&= \mathbb{E}[Y|X_o = \mathbf{x}_o] \\
&= \mathbb{E}[Y|\widetilde{X}_{j+1} = \tilde{x}_{j+1}, \dots, \widetilde{X}_d = \tilde{x}_d]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[Y|\widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}] &= \mathbb{E}[Y|\tilde{X}_1 = NA, \dots, \tilde{X}_j = NA, \tilde{X}_{j+1} = \tilde{x}_{j+1}, \dots, \tilde{X}_d = \tilde{x}_d] \\
&= \mathbb{E}[Y|M_1 = 1, \dots, M_j = 1, \tilde{X}_{j+1} = \tilde{x}_{j+1}, \dots, \tilde{X}_d = \tilde{x}_d] \\
&= \mathbb{E}[Y|\tilde{X}_{j+1} = \tilde{x}_{j+1}, \dots, \tilde{X}_d = \tilde{x}_d]
\end{aligned}$$

Imputation prior to learning

Impute the train, learn a model with $\hat{X}_{\text{train}}, Y_{\text{train}}$. Impute the test with the same imputation and predict with $\hat{X}_{\text{test}}$ and $\hat{f}_{train}$



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 imput 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 not the training set

Imputation with the same model : mean imputation is consistent

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**

Framework - assumptions

- $Y = f(\mathbf{X}) + \varepsilon$
- $\mathbf{X} = (X_1, \ldots, X_d)$ has a continuous density $g > 0$ on $[0,1]^d$
- $\|f\|_\infty < \infty$
- Missing data 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 $(\mathbf{X}, M_1)$

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

Imputation with the same model : mean imputation is consistent

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**

Imputed entry $\mathbf{x}' = (x'_1, x_2, \ldots, x_d)$: $x'_1 = x_1 \mathbb{1}_{M_1=0} + \mathbb{E}[X_1]\mathbb{1}_{M_1=1}$

Theorem

$$\begin{aligned} f^\star_{impute}(x') =& \mathbb{E}[Y|X_2 = x_2, \ldots, X_d = x_d, M_1 = 1] \\ & \quad \mathbb{1}_{x'_1=\mathbb{E}[X_1]} \mathbb{1}_{\mathbb{P}[M_1=1|X_2=x_2,\ldots,X_d=x_d]>0} \\ & + \mathbb{E}[Y|\mathbf{X} = \mathbf{x}'] \mathbb{1}_{x'_1=\mathbb{E}[X_1]} \mathbb{1}_{\mathbb{P}[M_1=1|X_2=x_2,\ldots,X_d=x_d]=0} \\ & + \mathbb{E}[Y|X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d, M_1 = 0] \mathbb{1}_{x'_1 \neq \mathbb{E}[X_1]}. \end{aligned}$$

Prediction with mean is equal to the Bayes function almost everywhere

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

Imputation with the same model : mean imputation is consistent

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**

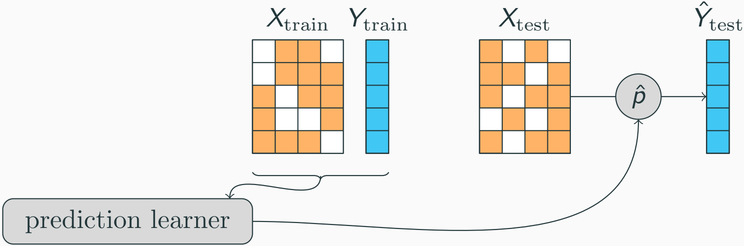
Rationale

The learning algorithm learns the imputed value (here the mean) and use that information to detect that the entry was initially missing. If the imputed value changes from train to test set the learning algorithm may fail, since imputed data distribution differs between train and test sets.

$\Rightarrow$ Other values than the mean are possible. Mean not a bad choice for prediction despite its drawbacks for estimation.

Trees - Simulations

End-to-end learning with missing values

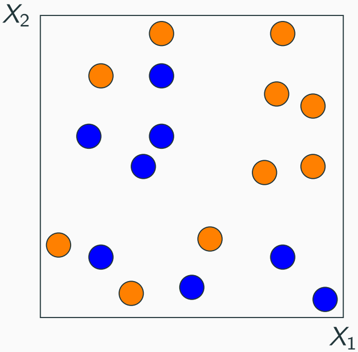


Trees natural for empirical risk minimization with NA : handle half discrete data

CART (Breiman, 1984)

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

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

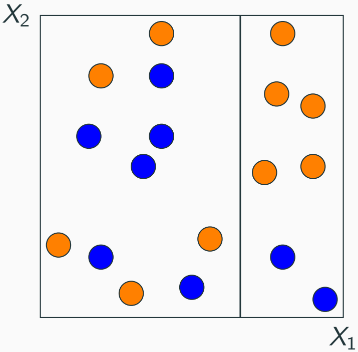


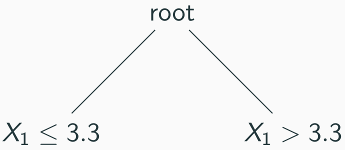
root

CART (Breiman, 1984)

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

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

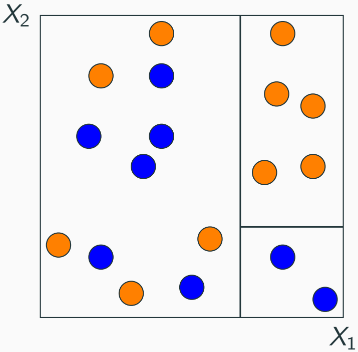


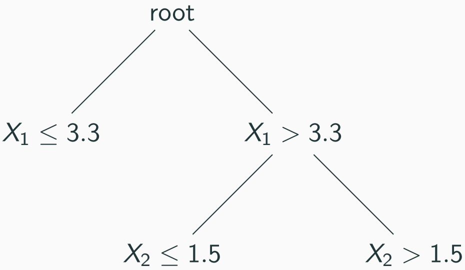


CART (Breiman, 1984)

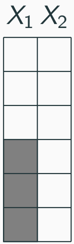
Built recursively by splitting the current cell into two children : find the feature $j^\star$, the threshold $z^\star$ which minimises the (quadratic) loss

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





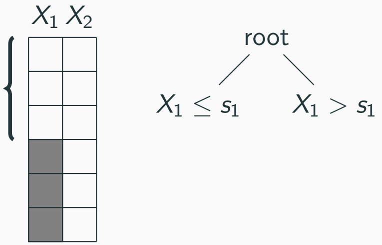
CART with missing values : split on available cases



root

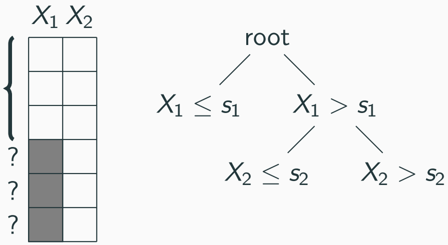
$$\mathbb{E}\Big[\big(Y - \mathbb{E}[Y|X_j \leq z, M_j = 0]\big)^2 \cdot \mathbb{1}_{X_j \leq z, M_j = 0} + \big(Y - \mathbb{E}[Y|X_j > z, M_j = 0]\big)^2 \cdot \mathbb{1}_{X_j > z, M_j = 0}\Big].$$

CART with missing values : split on available cases



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

CART with missing values : split on available cases



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

Propagate observation with missing values ?

Probabilistic splits : $\mathcal{B}ernouilli(\frac{\#L}{\#L+\#R})$ (C4.5 algorithm)

Block : send all to a side by minimizing the error (`xgboost`, `lightgbm`)

Surrogate split : search for a split on another variable that induces a partition close to the original one (rpart)

Rk : Implicit impute by an interval (missing values assigned to the left or right)
Variable selection bias (not a problem to predict) : conditional trees (Hothorn)

Missing incorporated in attribute, Twala et al 2008

Selection of the variable, threshold and propagation of missing values

$$f^\star \in \underset{f \in \mathcal{P}_{c,miss}}{\arg\min}\ \mathbb{E}\Big[\big(Y - f(\widetilde{\mathbf{X}})\big)^2\Big],$$

where $\mathcal{P}_{c,miss} = \mathcal{P}_{c,miss,L} \cup \mathcal{P}_{c,miss,R} \cup \mathcal{P}_{c,miss,sep}$ with

- $\mathcal{P}_{c,miss,L} \ \to\ \{\{\widetilde{X}_j \le z \vee \widetilde{X}_j = \mathtt{NA}\}, \{\widetilde{X}_j > z\}\}$
- $\mathcal{P}_{c,miss,R} \ \to\ \{\{\widetilde{X}_j \le z\}, \{\widetilde{X}_j > z \vee \widetilde{X}_j = \mathtt{NA}\}\}$
- $\mathcal{P}_{c,miss,sep} \ \to\ \{\{\widetilde{X}_j \ne \mathtt{NA}\}, \{\widetilde{X}_j = \mathtt{NA}\}\}$.

$\Rightarrow$ Missing values treated like a category (well to handle $\mathbb{R} \cup \mathtt{NA}$)

$\Rightarrow$ Target $\mathbb{E}\left[Y \middle| \widetilde{\mathbf{X}}\right] = \sum_{\mathbf{m} \in \{0,1\}^d} \mathbb{E}\left[Y | o(\mathbf{X}, \mathbf{m}), \mathbf{M} = \mathbf{m}\right]\ \mathbb{1}_{\mathbf{M}=\mathbf{m}}$

$\Rightarrow$ Good for informative pattern ($\mathbf{M}$ explains Y)

$\Rightarrow$ Implementation : duplicate the incomplete columns, and replace the missing entries once by $+\infty$ and once by $-\infty$.

Simulations : 20% missing values

Quadratic : $Y = X_1^2 + \varepsilon$, $x_{i.} \in \mathcal{N}(\mu, \Sigma_{4\times 4})$, $\rho = 0.5$, $n = 1000$

$$\widetilde{d}_n = \begin{bmatrix} 2 & 3 & \texttt{NA} & 0 & 15 \\ 1 & \texttt{NA} & 3 & 5 & 13 \\ 9 & 4 & 2 & \texttt{NA} & 18 \\ 7 & 6 & \texttt{NA} & \texttt{NA} & 10 \end{bmatrix}$$

$$\widetilde{d}_n + \text{mask} = \begin{bmatrix} 2 & 3 & \texttt{NA} & 0 & 0 & 0 & 1 & 0 & 15 \\ 1 & \texttt{NA} & 3 & 5 & 0 & 1 & 0 & 0 & 13 \\ 9 & 4 & 2 & \texttt{NA} & 0 & 0 & 0 & 1 & 18 \\ 7 & 6 & \texttt{NA} & \texttt{NA} & 0 & 0 & 1 & 1 & 10 \end{bmatrix}$$

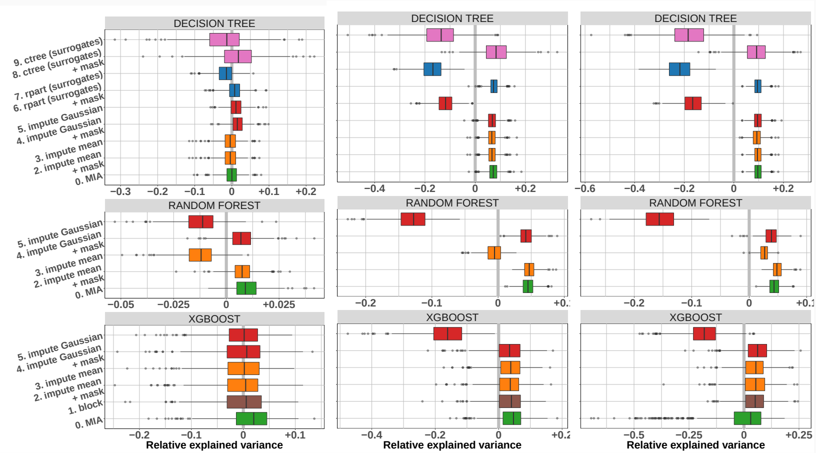
Imputation (mean, gaussian) + prediction with trees
Imputation (mean, gaussian) + mask+ prediction with trees
Trees MIA

Simulations : 20% missing values

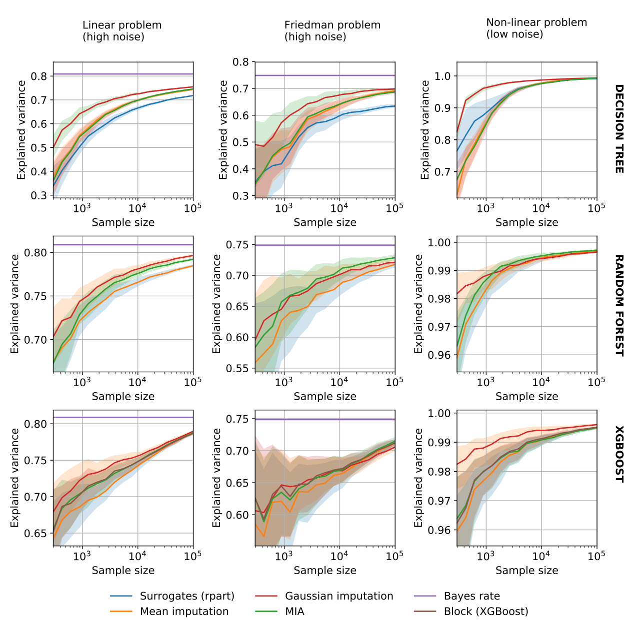
Quadratic : $Y = X_1^2 + \varepsilon,\ x_{i.} \in \mathcal{N}(\mu, \Sigma_{4\times4}),\ \rho = 0.5,\ n = 1000$

MCAR (MAR) | MNAR | Predictive

$M_{i,1} \sim \mathcal{B}(p)$ | $M_{i,1} = \mathbb{1}_{X_{i,1} > [X_1]_{(1-p)n}}$ | $Y = X_1^2 + 3M_1 + \varepsilon$



Consistency : 40% missing values MCAR



Discussion

Discussion

Take-home

- Consistent learner for the fully observed data $\rightarrow$ **multiple imputation on the test set**
- Incomplete train and test $\rightarrow$ **same imputation model**
- **Single mean imputation is consistent, provided a powerful learner**
- tree-based models $\rightarrow$ **Missing Incorporated in Attribute** optimizes not only the split but also the handling of the missing values
- Empirically, good imputation methods reduce the number of samples required to reach good prediction
- Informative missing data **Adding the mask** helps imputation - MIA

To be done

- Nonasymptotic results
- Prove the usefulness of methods in MNAR
- Uncertainty associated with the prediction
- Distributional shift : no missing values in the test set ?

Context

Major trauma : any injury that endangers the life or the functional integrity of a person. Road traffic accidents, interpersonal violence, self-harm, falls, etc $\rightarrow$ hemorrhage and traumatic brain injury.

Major source of mortality and handicap in France and worldwide (3rd cause of death, 1st cause for 16-45 - 2-3th cause of disability)

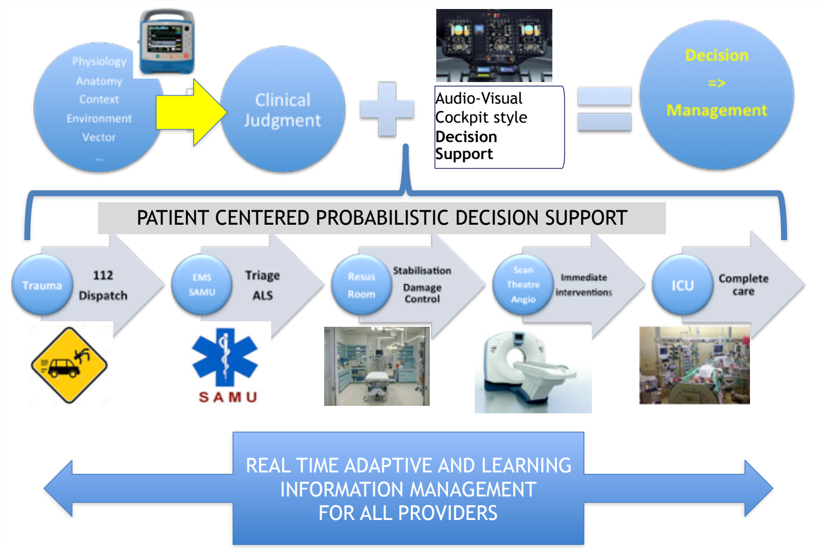
$\Rightarrow$ A public health challenge

Patient prognosis can be improved : **standardized and reproducible procedures** but **personalized** for the patient and the trauma system.

Trauma decision making : rapid and **complex decisions** under **time pressure** in a dynamic and multi-player environment (fragmentation : loss or distortion of information) with high levels of uncertainty and **stress**. Issues : patient management exceeds time frames, diagnostic errors, decisions not reproducible, etc

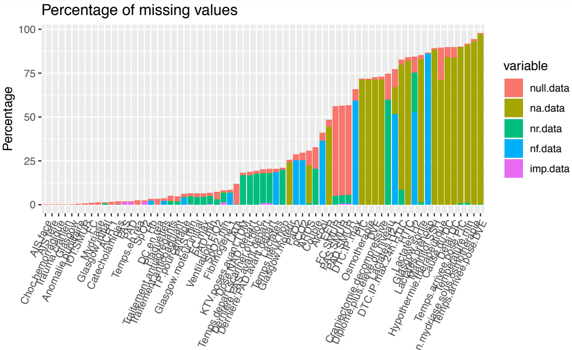
$\Rightarrow$ Can Machine Learning, AI help ?

Decision support tool for the management of severe trauma : Traumamatrix



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



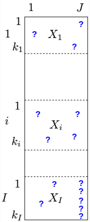
Missing values

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" Gertrude Mary Cox.

⇒ Still an issue with "big data"

Data integration : data from different sources



Multilevel data : sporadically - systematic (one variable missing in one hospital)

Imputation assuming a joint modeling with gaussian distribution

based on Gaussian assumption : $x_{i.} \sim \mathcal{N}(\mu, \Sigma)$

• Bivariate with missing on $x_{.1}$ (stochastic reg) : estimate β and σ - impute from the predictive $x_{i1} \sim \mathcal{N}\left(x_{i2}\hat{\beta}, \hat{\sigma}^2\right)$
• Extension to multivariate case : estimate μ and Σ from an incomplete data with EM - impute by drawing from $\mathcal{N}\left(\hat{\mu}, \hat{\Sigma}\right)$ equivalence conditional expectation and regression (complement Schur)

packages `Amelia`, `mice` (conditional)

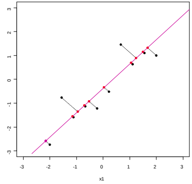
PCA reconstruction

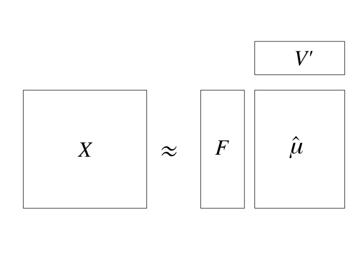
X

-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 $k < p$ $\|A\|_2^2 = \mathrm{tr}(AA^\top)$:

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

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

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

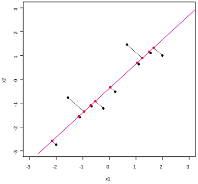
PCA reconstruction

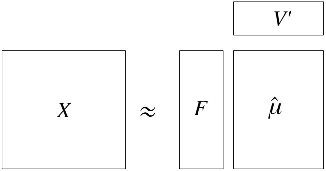
X

-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 $k < p$ $\|A\|_2^2 = \mathrm{tr}(AA^\top)$:

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

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

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

Missing values in PCA

$\Rightarrow$ PCA : least squares

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

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

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

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

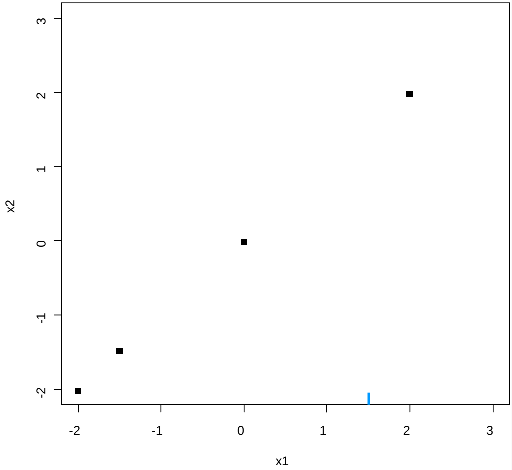
Many algorithms :

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

Kiers, 1997 : iterative PCA (with imputation)

Iterative PCA

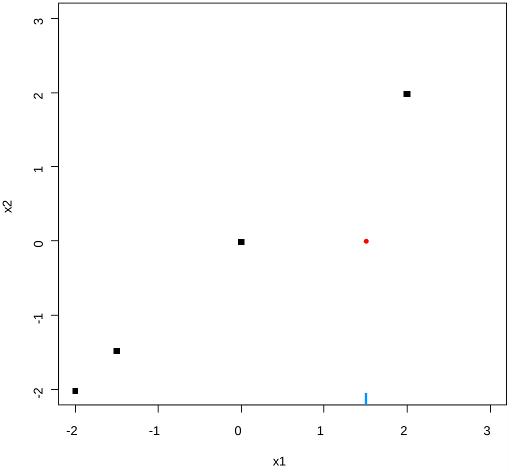
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	NA
2.0	1.98



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



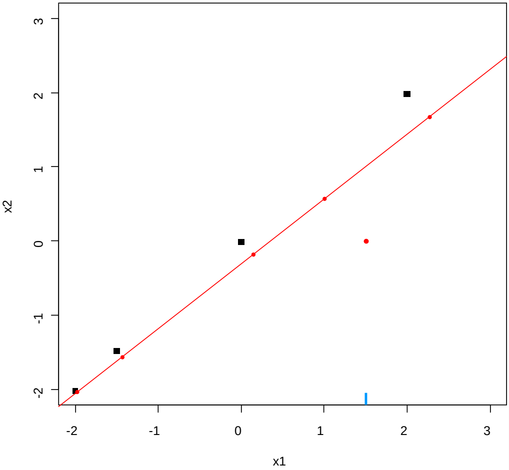
Initialization $\ell = 0 : X^0$ (mean imputation)

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



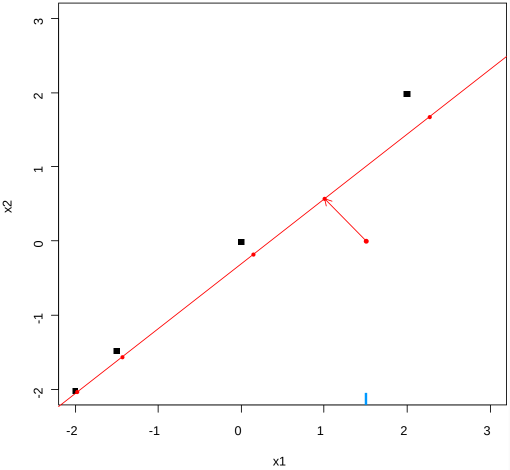
PCA on the completed data set $\rightarrow (U^\ell, \Lambda^\ell, D^\ell)$;

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

$\widehat{x1}$	$\widehat{x2}$
-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 $\hat{\mu}^{\ell} = U^{\ell} D^{\ell} V^{\ell\prime}$

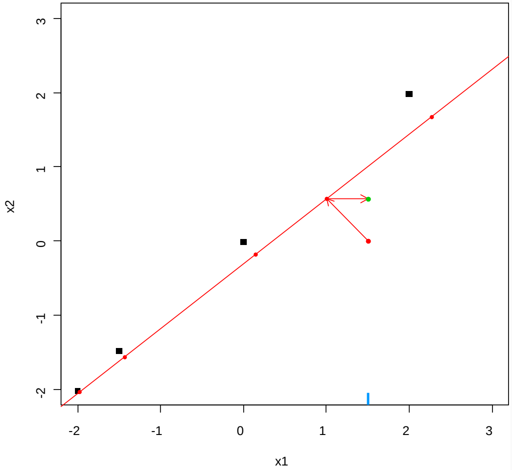
Iterative PCA

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



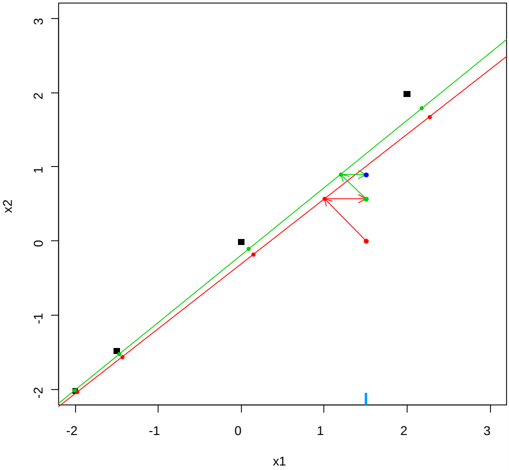
The new imputed dataset is $\hat{X}^{\ell} = W \odot X + (\mathbf{1} - W) \odot \hat{\mu}^{\ell}$

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



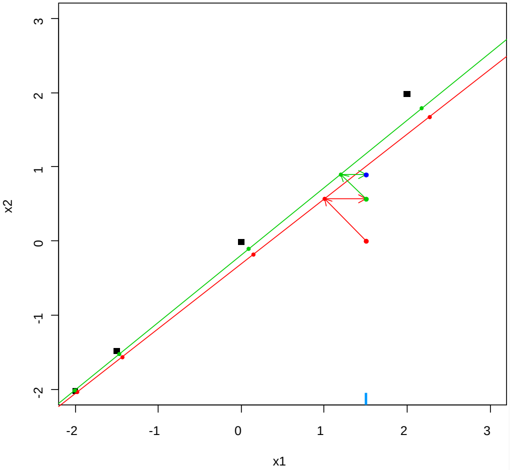
Iterative PCA

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

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

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



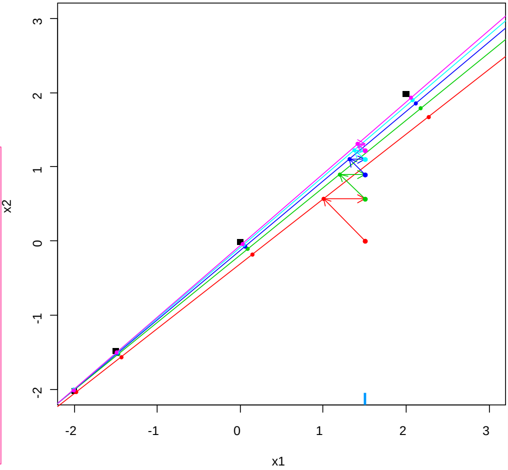
Iterative PCA

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

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

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

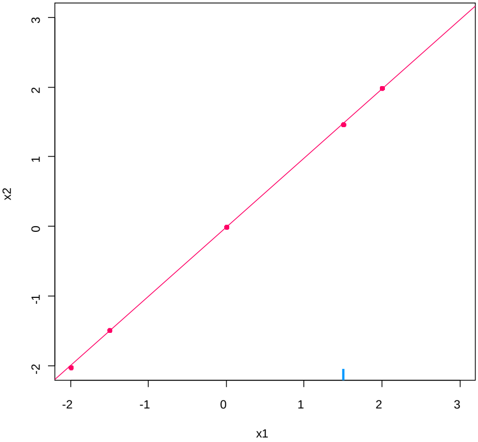
x1	x2
-2.0	-2.01
-1.5	-1.48
0.0	-0.01
1.5	**0.57**
2.0	1.98



Steps are repeated until convergence

Iterative PCA

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



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

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

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) $\hat{\mu}^{\text{PCA}} = \sum_{q=1}^{k} d_q u_q v_q^{'}$ $\quad X^\ell = W \odot X + (1 - W) \odot \hat{\mu}^\ell$
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)

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

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

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

$\Rightarrow$ Missing $\Rightarrow$ Random forests (mice) $\Rightarrow$ PCA

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

$\Rightarrow$ Aim is not to impute as well as possible but estimate parameters and their variance (multiple imputation).

Iterative SVD

⇒ Imputation with FAMD for mixed data :

age	weight	size	alcohol	sex	snore	tobacco
NA	100	190	NA	M	yes	no
70	96	186	1-2 gl/d	M	NA	<=1
NA	104	194	No	W	no	NA
62	68	165	1-2 gl/d	M	no	<=1

NA	100	190	NA	NA	NA	1	0	0	1	1	0	0
70	96	186	0	1	0	1	0	NA	NA	0	1	0
NA	104	194	1	0	0	0	1	1	0	NA	NA	NA
62	68	165	0	1	0	1	0	1	0	0	1	0

imputeAFDM

51	100	190	0.2	0.7	0.1	1	0	0	1	1	0	0
70	96	186	0	1	0	1	0	0.8	0.2	0	1	0
48	104	194	1	0	0	0	1	1	0	0.1	0.8	0.1
62	68	165	0	1	0	1	0	1	0	0	1	0

age	weight	size	alcohol	sex	snore	tobacco
51	100	190	1-2 gl/d	M	yes	no
70	96	186	1-2 gl/d	M	no	<=1
48	104	194	No	W	no	<=1
62	68	165	1-2 gl/d	M	no	<=1

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

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

package `MissMDA`.

Imputation methods : conditional model

Imputation with fully conditional specification (FCS). Impute with a joint model defined implicitely through the conditional distributions (`mice`).

$\Rightarrow$ 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
   $X_{-j}^t = (X_1^t, X_{j-1}^t, X_{j+1}^{t-1}, X_p^{t-1})$, then X_j^t is obtained by
   - fitting a RF X_j^{obs} on the other variables X_{-j}^t
   - predicting X_j^{miss} using the trained RF on X_{-j}^t

`package missForest` (Stekhoven & Buhlmann, 2011)

Mechanism

$\mathbf{M} = (M_1, \ldots, M_d)$: indicator of missing values in $\mathbf{X} = (X_1, \ldots, X_d)$.

Missing value mechanisms (Rubin, 1976)

MCAR	$\forall \phi, \forall \mathbf{m}, \mathbf{x}, g_\phi(\mathbf{m} \vert \mathbf{x}) = g_\phi(\mathbf{m})$
MAR	$\forall \phi, \forall i, \forall \mathbf{x}', o(\mathbf{x}', \mathbf{m}_i) = o(\mathbf{x}_i, \mathbf{m}_i) \Rightarrow g_\phi(\mathbf{m}_i \vert \mathbf{x}') = g_\phi(\mathbf{m}_i \vert \mathbf{x}_i)$
	$(e.g.\ g_\phi((0,0,1,0) \mid (3,2,4,8)) = g_\phi((0,0,1,0) \mid (3,2,7,8)))$
MNAR	Not MAR

$\rightarrow$ useful for likelihoods

Missing value mechanisms – variable level

MCAR	$\mathbf{M} \perp\!\!\!\perp \mathbf{X}$
MAR (bis)	$\forall \mathcal{S} \subset \{1, \ldots, d\}$, $(M_j)_{j \in \mathcal{S}} \perp\!\!\!\perp (X_j)_{j \in \mathcal{S}} \mid (X_k)_{k \in \mathcal{S}^c}$
MNAR	Not MAR

$\rightarrow$ useful for our results

Parametric estimation

Let $\mathbf{X} \sim f_{\theta^\star}$.

Observed log-likelihood $$\ell_{\mathrm{obs}}(\theta) = \sum_{i=1}^{n} \log \int f_\theta(\mathbf{x}) \, \mathrm{d}\delta_{o(\cdot,\mathbf{m}_i)=o(\mathbf{x}_i,\mathbf{m}_i)}(\mathbf{x}).$$

(inspired by Seaman 2013)

Example

$$X_1, X_2 \sim f_\theta(x_1) g_\theta(x_2|x_1)$$

$$M_{1,2}, \ldots, M_{r,2} = 1$$

$$\ell_{\mathrm{obs}}(\theta) = \sum_{i=1}^{r} \log f_\theta(x_1) + \sum_{i=r+1}^{n} \log f_\theta(x_1) g_\theta(x_2|x_1).$$

Parametric estimation

Let $\mathbf{X} \sim f_{\theta^\star}$.

Observed log-likelihood $$\ell_{\mathrm{obs}}(\theta) = \sum_{i=1}^{n} \log \int f_\theta(\mathbf{x}) \, \mathrm{d}\delta_{o(\cdot,\mathbf{m}_i)=o(\mathbf{x}_i,\mathbf{m}_i)}(\mathbf{x}).$$

Full log-likelihood $$\ell_{\mathrm{full}}(\theta, \phi) = \sum_{i=1}^{n} \log \int f_\theta(\mathbf{x}) g_\phi(\mathbf{m}_i|\mathbf{x}) \, \mathrm{d}\delta_{o(\cdot,\mathbf{m}_i)=o(\mathbf{x}_i,\mathbf{m}_i)}(\mathbf{x}).$$

Theorem (Theorem 7.1 in Rubin 1976)
θ can be infered from ℓ_{obs}, assuming MAR.

Ignorable mechanism

Full log-likelihood :

$$\ell_{\text{full}}(\theta) = \sum_{i=1}^{n} \log \int f_\theta(\mathbf{x}) g_\phi(\mathbf{m}_i|\mathbf{x}) \, \mathrm{d}\delta_{o(\cdot,\mathbf{m}_i)=o(\mathbf{x}_i,\mathbf{m}_i)}(\mathbf{x}).$$

Observed log-likelihood :

$$\ell_{\text{obs}}(\theta) = \sum_{i=1}^{n} \log \int f_\theta(\mathbf{x}) \, \mathrm{d}\delta_{o(\cdot,\mathbf{m}_i)=o(\mathbf{x}_i,\mathbf{m}_i)}(\mathbf{x}).$$

Assuming MAR,

$$\begin{aligned}
\ell_{\text{full}}(\theta, \phi) &= \sum_{i=1}^{n} \log \int f_\theta(\mathbf{x}) g_\phi(\mathbf{m}_i|\mathbf{x}_i) \, \mathrm{d}\delta_{o(\cdot,\mathbf{m}_i)=o(\mathbf{x}_i,\mathbf{m}_i)}(\mathbf{x}) \\
&= \ell_{\text{obs}}(\theta) + \sum_{i=1}^{n} \log g_\phi(\mathbf{m}_i|\mathbf{x}_i).
\end{aligned}$$

Parametric estimation

Let $\mathbf{X} \sim f_{\theta^\star}$.

Observed log-likelihood $\qquad \ell_{\text{obs}}(\theta) = \sum_{i=1}^{n} \log \int f_\theta(\mathbf{x}) \, \mathrm{d}\delta_{o(\cdot,\mathbf{m}_i)=o(\mathbf{x}_i,\mathbf{m}_i)}(\mathbf{x}).$

EM algorithm (Dempster, 1977)

Starting from an initial parameter $\theta^{(0)}$, the algorithm alternates the two following steps,

(E-step) $\qquad Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n} \int (\log f_\theta(\mathbf{x})) f_{\theta^{(t)}}(\mathbf{x}) \, \mathrm{d}\delta_{o(\cdot,\mathbf{m}_i)=o(\mathbf{x}_i,\mathbf{m}_i)}(\mathbf{x}).$

(M-step) $\qquad \theta^{(t+1)} \in \underset{\theta \in \Theta}{\operatorname{argmax}} \; Q(\theta|\theta^{(t)}).$

The likelihood is guaranteed to increase.

Missing values

$\widetilde{\mathbf{X}} = \mathbf{X} \odot (\mathbf{1} - \mathbf{M}) + \mathtt{NA} \odot \mathbf{M}$ takes value in $\mathbb{R} \cup \{\mathtt{NA}\}$

The (unobserved) complete sample $\mathcal{D}_n = (\mathbf{X}_i, \mathbf{M}_i, Y_i)_{1 \leq i \leq n} \sim (\mathbf{X}, \mathbf{M}, Y)$

$$d_n = \begin{bmatrix} 2 & 3 & 1 & 0 & 0 & 0 & 1 & 0 & 15 \\ 1 & 0 & 3 & 5 & 0 & 1 & 0 & 0 & 13 \\ 9 & 4 & 2 & 5 & 0 & 0 & 0 & 1 & 18 \\ 7 & 6 & 3 & 2 & 0 & 0 & 1 & 1 & 10 \end{bmatrix},$$

The observed training set $\widetilde{\mathcal{D}}_{n,\mathrm{train}} = (\widetilde{\mathbf{X}}_i, Y_i)_{1 \leq i \leq n}$

$$\widetilde{d}_n = \begin{bmatrix} 2 & 3 & \mathtt{NA} & 0 & 15 \\ 1 & \mathtt{NA} & 3 & 5 & 13 \\ 9 & 4 & 2 & \mathtt{NA} & 18 \\ 7 & 6 & \mathtt{NA} & \mathtt{NA} & 10 \end{bmatrix}.$$

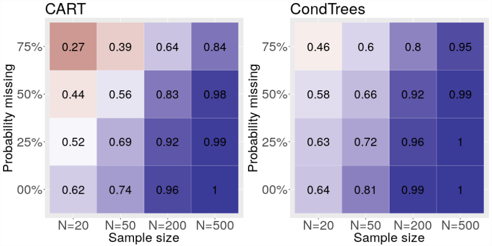
Split on available observations

⇒ Biais in variable selection : tendency to underselect variables with missing values (favor variables where many splits are available)

⇒ Conditional tree (Hothorn, 2006) Ctree selects variables with a test

$$\begin{cases} X_1 \perp\!\!\!\perp X_2 & \sim \mathcal{N}(0,1) \\ Y & = 0.25X_1 + \varepsilon \text{ with } \varepsilon \sim \mathcal{N}(0,1) \end{cases}$$

Frequency of selection of X_1 when there are missing values on X_1 :



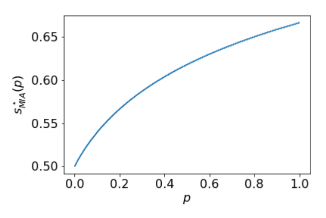
CART selects the non-informative variable X_2 more frequently

Split comparison

$$\begin{cases} Y & = & X_1 \\ X_1 & \sim & U([0,1]) \end{cases}, \quad \begin{cases} \mathbb{P}[M_1 = 0] & = & 1-p \\ \mathbb{P}[M_1 = 1] & = & p \end{cases},$$

The best split CART $s^\star = 1/2$

The split chosen by the MIA

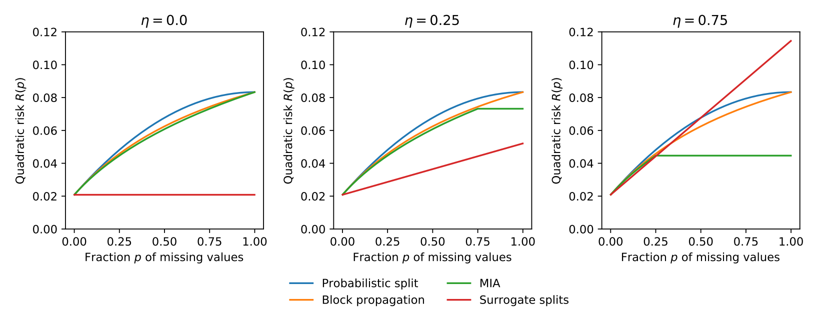


Risk comparison

Consider the regression model

$$\begin{cases} Y &= X_1 \\ X_1 &\sim U([0,1]) \\ X_2 &= X_1 \mathbb{1}_{W=1} \end{cases}, \quad \begin{cases} \mathbb{P}[W=0] &= \eta \\ \mathbb{P}[W=1] &= 1-\eta \end{cases}, \quad \begin{cases} \mathbb{P}[M_1=0] &= 1-p \\ \mathbb{P}[M_1=1] &= p \end{cases},$$

where $(M_1, W) \perp\!\!\!\perp (X_1, Y)$.



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