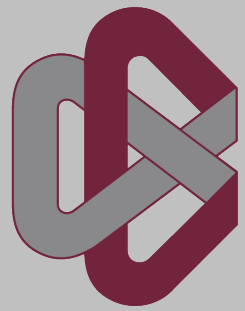
Some ideas for random forests algorithms with missing values

Irving Gómez Méndez

January 7th 2020



CIMAT

Regression Trees

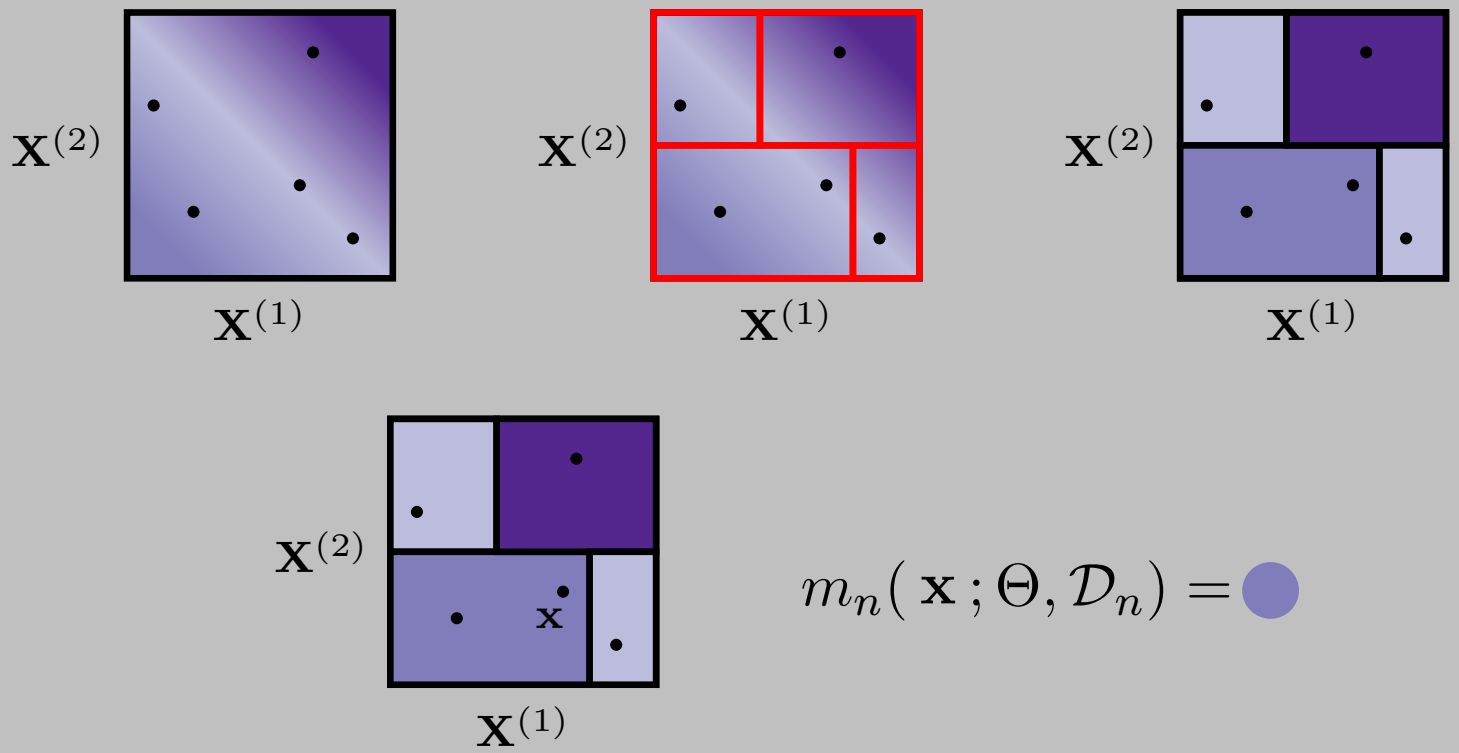
- Variable of interest $Y \in \mathbb{R}$
- Vector of predictors $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^p$

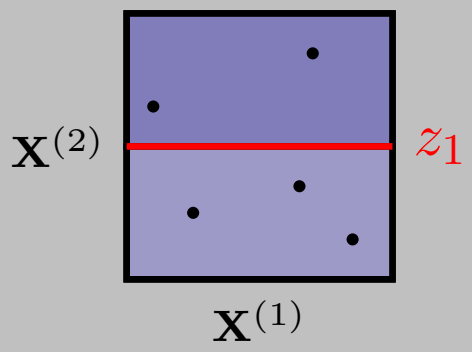
$$\mathbf{X} = \left(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)}\right)$$

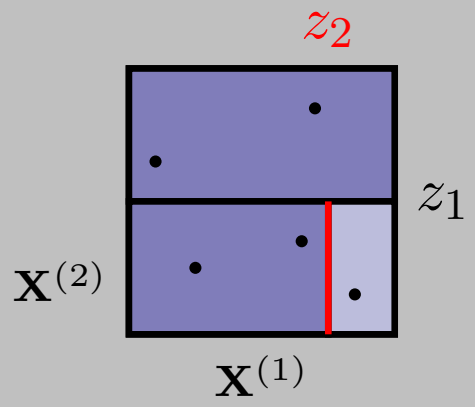
- training data set $\mathcal{D}_n = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$

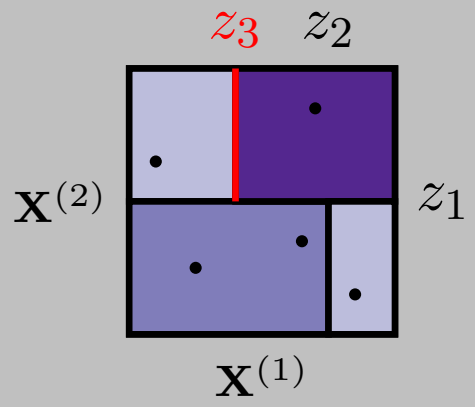
We want to estimate $m(\cdot) \equiv \mathbb{E}[Y|\mathbf{X} = \cdot]$ using $\mathcal{D}_n$

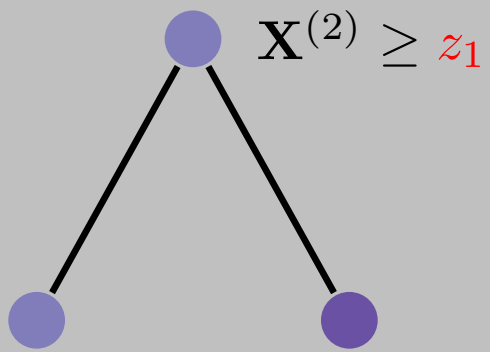
We can estimate $m(\cdot) \equiv \mathbb{E}[Y|\mathbf{X} = \cdot]$ dividing (somehow) $\mathcal{X}$ in disjoint regions and predicting with a constant in each region.

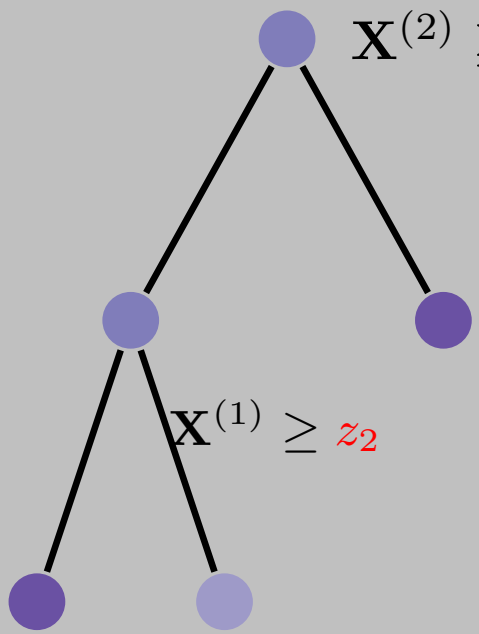


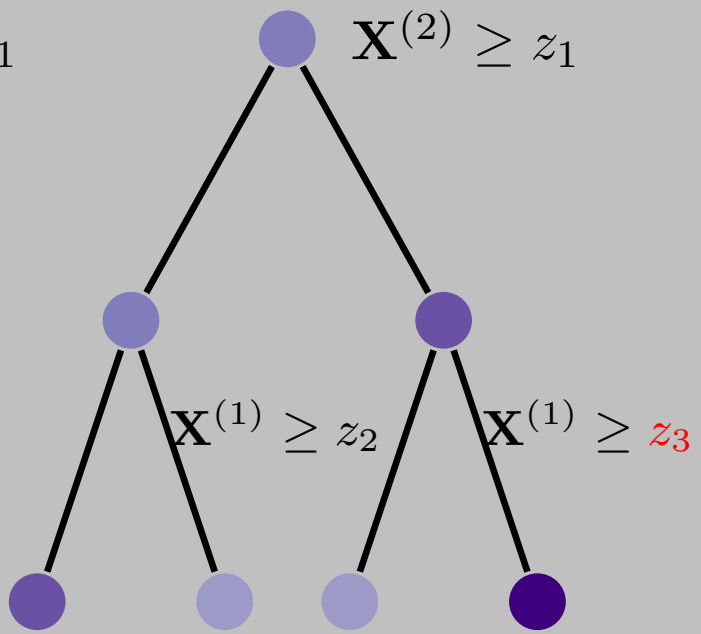

$\mathbf{X}^{(2)}$
z_1
$\mathbf{X}^{(1)}$


z_2
z_1
$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$


z_3
z_2
z_1
$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$


$\mathbf{X}^{(2)} \geq z_1$


$\mathbf{X}^{(2)} \geq z_1$
$\mathbf{X}^{(1)} \geq z_2$


$\mathbf{X}^{(2)} \geq z_1$
$\mathbf{X}^{(1)} \geq z_2$
$\mathbf{X}^{(1)} \geq z_3$

CART-split Criterion

$$
\begin{aligned}
L_{n,A}(h,z) =& \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_A \right)^2 \mathbb{1}_{\mathbf{X}_i \in A} \\
& - \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(h)} < z} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(h)} \geq z} \right)^2 \mathbb{1}_{\mathbf{X}_i \in A} \\
& \vdots \\
=& \frac{N_n(A_L) N_n(A_R)}{N_n(A) N_n(A)} \left(\bar{Y}_{A_L} - \bar{Y}_{A_R} \right)^2
\end{aligned}
$$

We construct the cells maximizing the CART criterion over all possible cuts in the cell $\mathcal{C}_A$.

$$
(h_n^\star, z_n^\star) \in \underset{(h,z) \in \mathcal{C}_A}{\arg\max}\, L_n(h,z)
$$

Decision Trees

They are easy to be interpreted (not exactly true).
Shallow trees are easy to be interpreted. ◆

They are unstable. ▼

(Shallow) trees are inaccurate. ▼

They can be used in regression or supervised classification problems. ▲

They handle categorical or numerical variables, or a mix of both. ▲

They can handle complex relations between variable. ▲

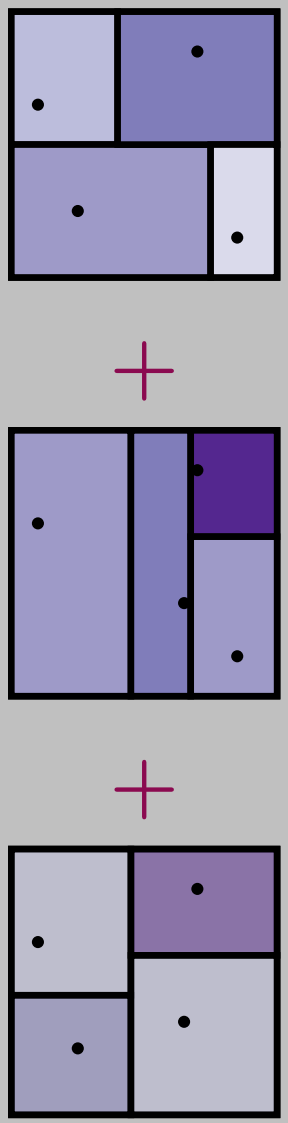
Regression Random Forests

A random forest is an ensemble of many trees, i.e.

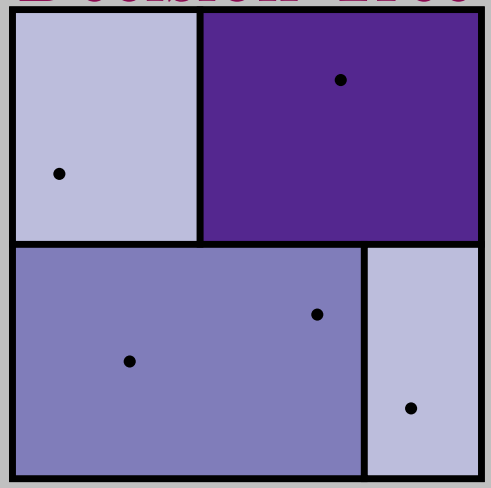
$$m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{k=1}^{M} m_n(\mathbf{x}; \Theta_k, \mathcal{D}_n)$$

We add two sources of randomness in each tree.

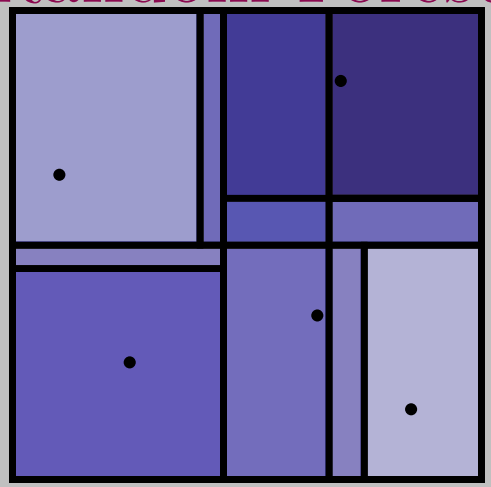
1. We select randomly a_n (with or without) replacement observations prior to the construction of each tree.
2. We select randomly `mtry` candidate directions to perform the cut.



Decision Tree



Random Forest



Random Forests

They are difficult to be interpreted. ▼

They are stable. ▲

They are accurate. ▲

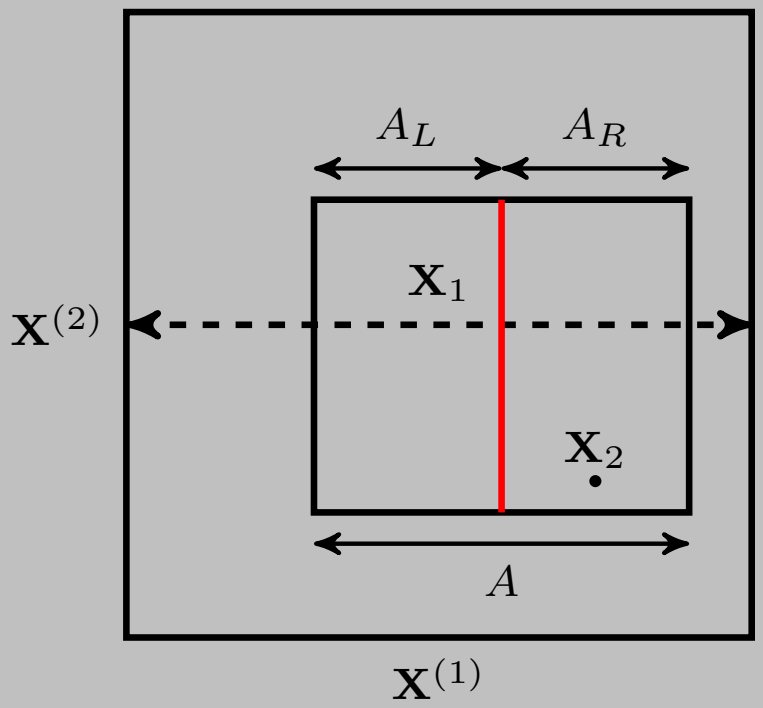
They can be used in regression or supervised classification problems. ▲

They handle categorical or numerical variables, or a mix of both. ▲

They can handle complex relations between variable. ▲

They are easily parallelized. ▲

Decision Trees with Missing Values



$$L_{n,A}(h,z) = \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_A\right)^2 \mathbb{1}_{\mathbf{X}_i \in A}$$
$$- \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(h)} < z} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(h)} \geq z}\right)^2 \mathbb{1}_{\mathbf{X}_i \in A}$$

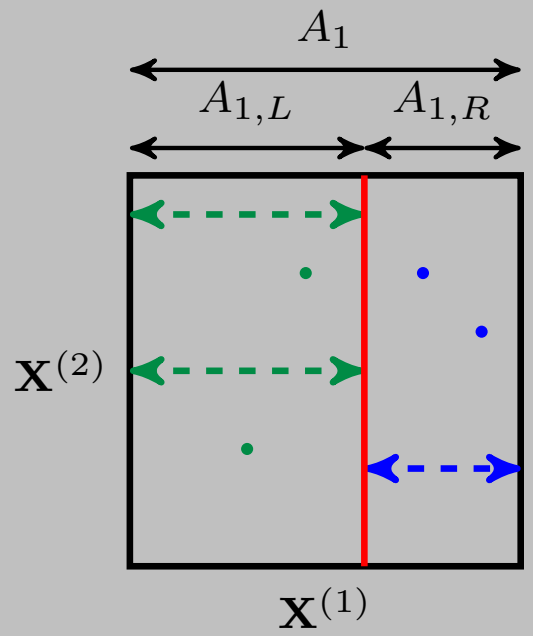
Let be

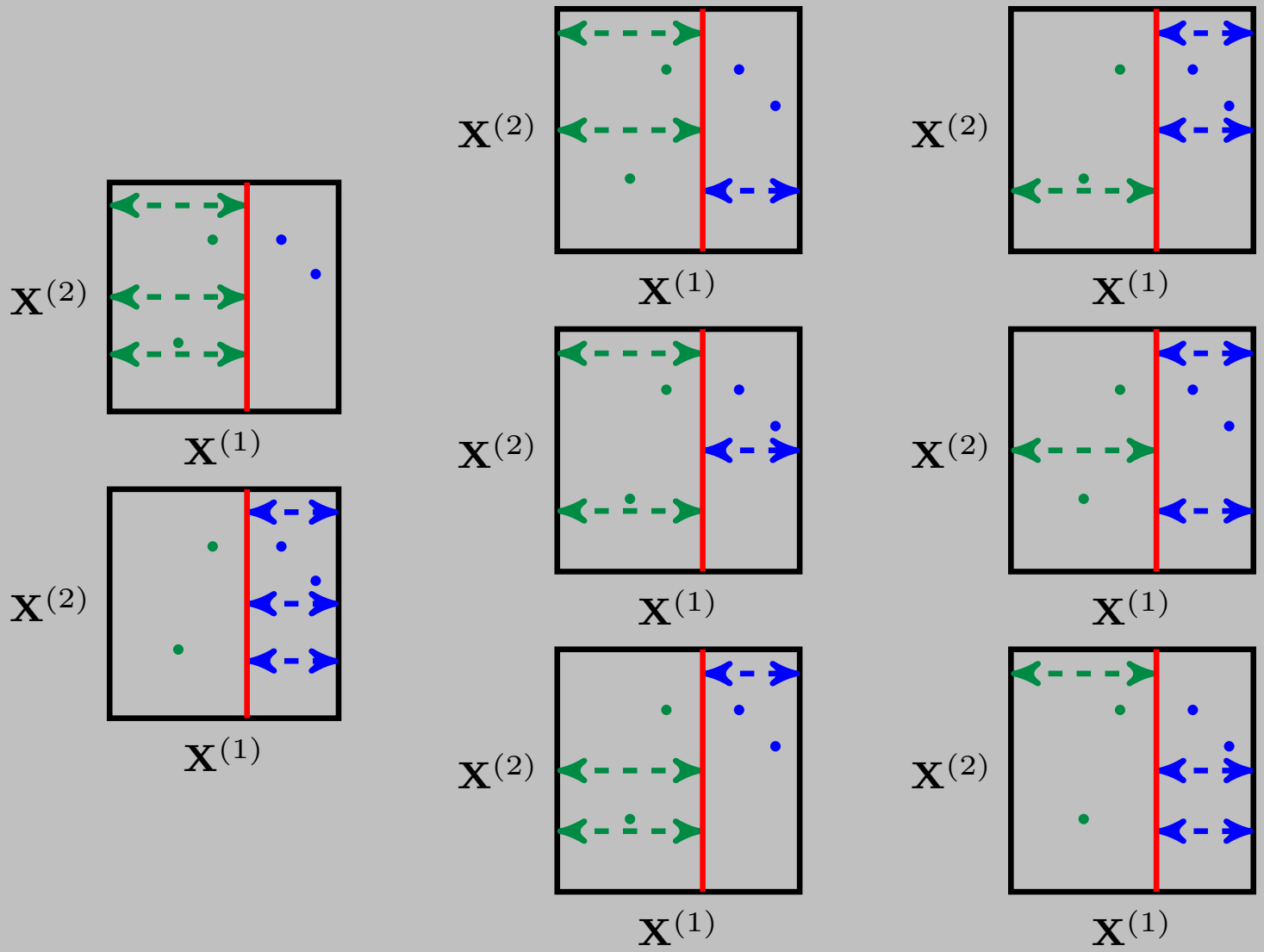
$$A_1 = \mathcal{X}$$

and

$$\mathbf{X}_{i,1}^{(h)} = \begin{cases} \mathbf{X}_i^{(h)} & \text{si } M_i^{(h)} = 0 \\ \mathcal{X}^{(h)} & \text{si } M_i^{(h)} = 1 \end{cases}, \quad 1 \leq h \leq p$$

$$L_{n,A_1}\left(h,z,\mathbf{X}^{(h)}_{miss}\right) = \frac{1}{N_n(A_1)}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_{A_1}\right)^2 \mathbb{1}_{\mathbf{X}_{i,1}\in A_1}$$
$$- \frac{1}{N_n(A_1)}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_{A_{1,L}}\right)^2 \mathbb{1}_{\mathbf{X}_{i,1}\in A_1,\, \mathbf{X}_i^{(h)}<z}$$
$$- \frac{1}{N_n(A_1)}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_{A_{1,R}}\right)^2 \mathbb{1}_{\mathbf{X}_{i,1}A_1,\, \mathbf{X}_i^{(h)}\geq z}$$




$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$
$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$
$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$
$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$
$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$
$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$

1. For each candidate cut we calculate the CART criterion with all possibilities for the missing values.
2. We select the combination that maximizes the CART criterion between $\mathcal{M}_{try}$, $\mathcal{C}_{A_1}$ and $\mathbf{X}_{miss}^{(h_{n,1})}$

$$\left(h_{n,1}^{\star}, z_{n,1}^{\star}, \mathbf{X}_{miss}^{\star\left(h_{n,1}^{\star}\right)}\right) \in \underset{\substack{h \in \mathcal{M}_{try} \\ (h,z) \in \mathcal{C}_{A_1} \\ \mathbf{x}_{miss}^{(h)} \in \left[a_1^{(h)}, b_1^{(h)}\right]}}{\arg\max} L_{n,A_1}\left(h, z, \mathbf{X}_{miss}^{(h)}\right)$$

3. We move to A_2 taking $\mathbf{X}_{i,2}$.

Mechanisms of Missingness[1]

- **Missing Completely at Random (MCAR)** A variable is missing completely at random if the probability of missingness is the same for all units.
- **Missing at Random (MAR)** A variable is missing at random if the probability of missingness just depends on observed information.
- **Not Missing at Random (NMAR)** If the probability of missingness depends on unobserved information, it is called not missing at random.

[1] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. 2nd ed. John Wiley & Sons, 2002.

Simulation[2]

[2] Anna Rieger, Torsten Hothorn, and Carolin Strobl. "Random forests with missing values in the covariates". In: (2010).

- We take the regression function "friedman1":

$$m(\mathbf{x}) = 10 \sin\left(\pi \mathbf{x}^{(1)} \mathbf{x}^{(2)}\right) + 20\left(\mathbf{x}^{(3)} - 0.5\right)^2 + 10\mathbf{x}^{(4)} + 5\mathbf{x}^{(5)}$$

Training data sets

- We created 10 training data sets.
- We simulated 200 observations from $\mathbf{X} \sim \mathcal{U}[0,1]^5$.

Test data set

- We create 1 data set.
- We simulated 2000 observations from $\mathbf{X} \sim \mathcal{U}[0,1]^5$.
- All values are observed.

- For the missing values we create 7 different mechanisms of missingness:
  - 1 Missing Completely At Random (MCAR).
  - 5 Missing At Random (MAR1, MAR2, MAR3, MAR4, Depy).
  - 1 Not Missing At Random (LOG).

Determinant Variable	Missing Variable	% Missing Data
$\mathbf{X}^{(2)}$, Y	$\mathbf{X}^{(1)}$	20%
$\mathbf{X}^{(5)}$, Y	$\mathbf{X}^{(3)}$	10%
	$\mathbf{X}^{(4)}$	20%

- **MAR1** The probability of `NA` is

$$\frac{2 \times \text{rango(var.det.)}}{n(n+1)}$$

- **MAR2** We create two groups in the det. var. An observation belongs to the first group if it is bigger to the the median, otherwise it belongs to the second group. The probability of `NA` for each group is

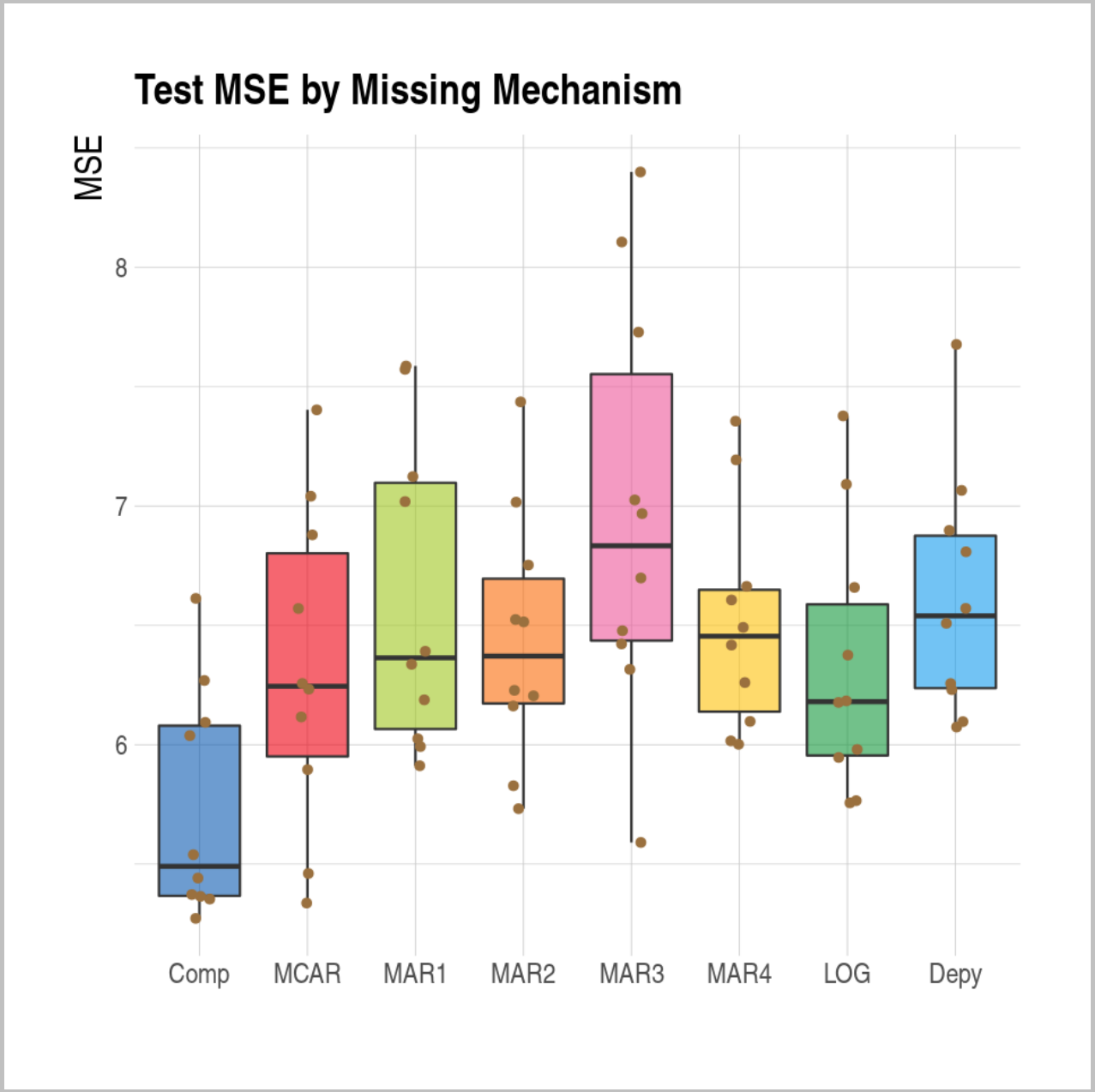
$$0.9/\#(\text{obs. in 1st group}) \quad 0.1/\#(\text{obs. in 2nd group})$$

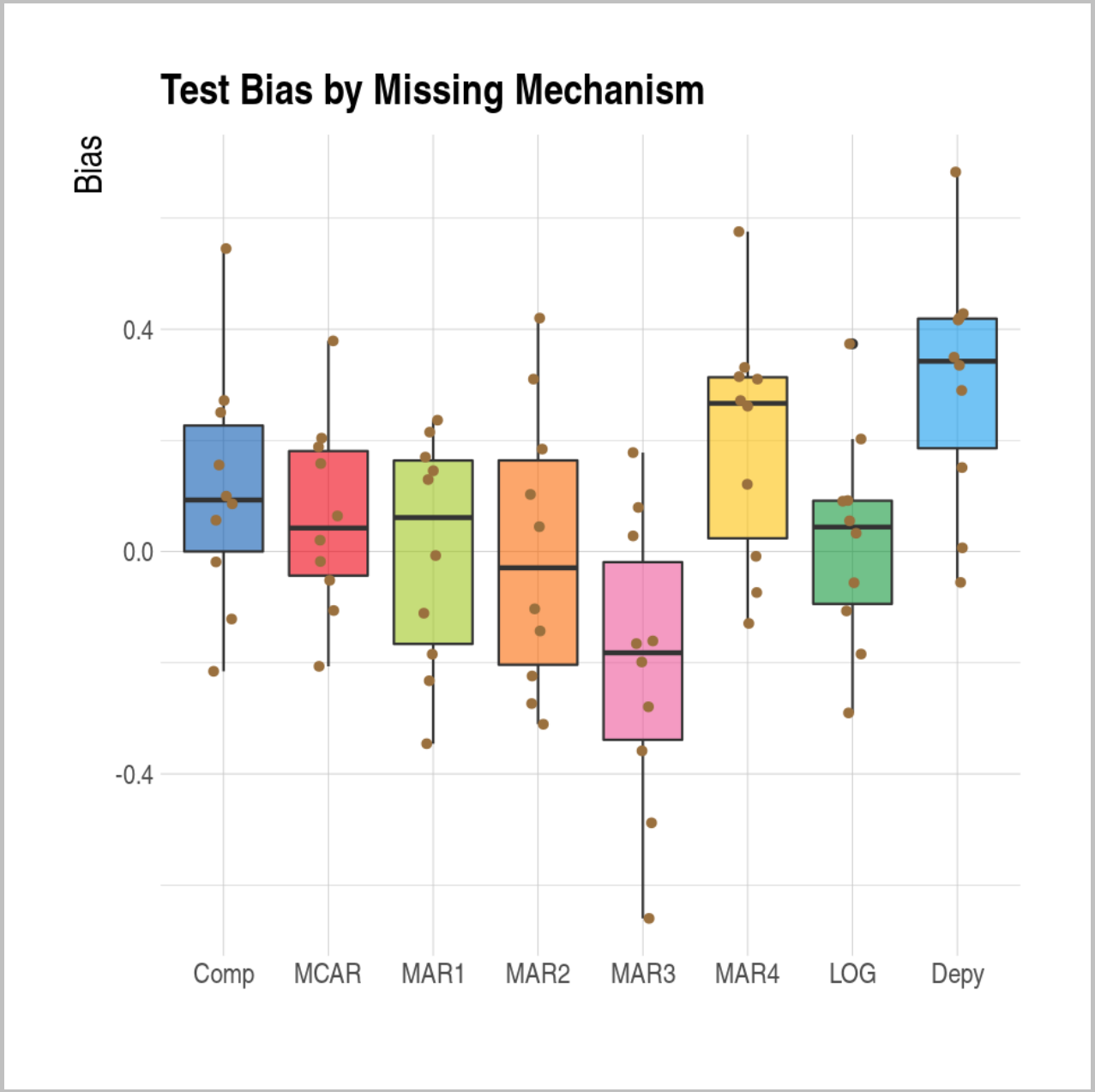
- **MAR3** The biggest values in det. var. are `NA` in missing var.
- **MAR4** The biggest and smallest values in det. var. are `NA` in missing var.
- **Depy** Probability of `NA` is 0.1 if $Y \geq 13$, otherwise is 0.4
- **LOG**

$$\text{logit}(\mathbb{P}[\mathbf{M}^{(h)} = 1]) = -0.5 + \sum_{\substack{k=1 \\ k \neq h}}^{5} \mathbf{X}^{(h)}$$

Random Forests

- For each data set and each mechanism of missingness (including complete data) we create 1 random forest.
- Each forest is built with the parameters:
  - $M = 50$ trees.
  - `mtry` $= 1$ variable selected at random to perform the cut.
  - $a_n = 127$ observations selected at random and without replacement for each tree.
  - `nodesize` $= 5$, maximum number of observations o¿inf the final nodes.


Test MSE by Missing Mechanism
MSE
8
7
6
Comp
MCAR
MAR1
MAR2
MAR3
MAR4
LOG
Depy

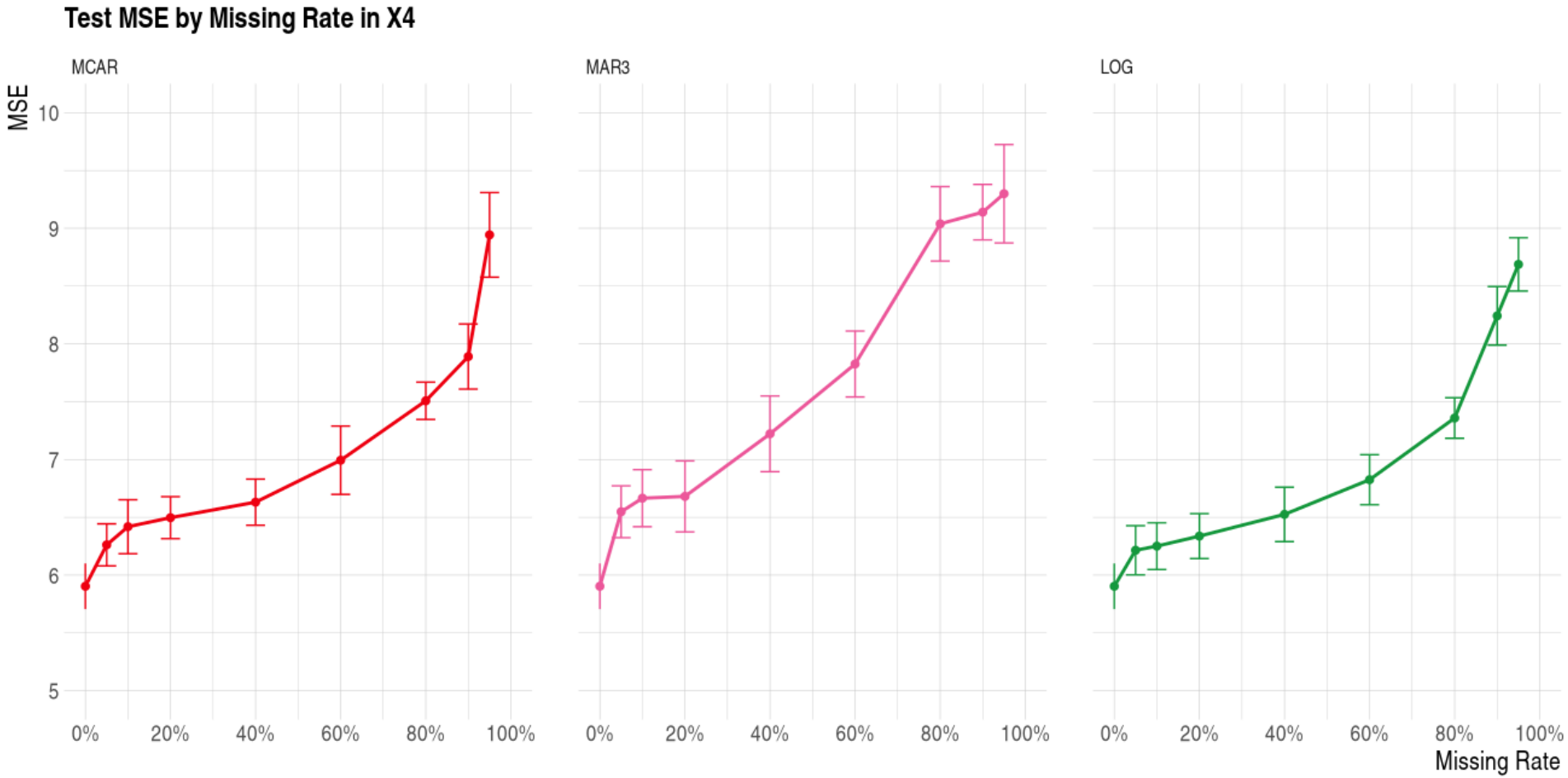

Test Bias by Missing Mechanism
Bias
0.4
0.0
-0.4
Comp
MCAR
MAR1
MAR2
MAR3
MAR4
LOG
Depy

We also change the percentage of missingness of each variable, without changing the percentage of the other variables.

$\mathbf{X}^{(1)}$	$\mathbf{X}^{(3)}$	$\mathbf{X}^{(4)}$
20%	10%	5%, 10%, 20%, 40%, 60%, 80%, 90%, 95%

$\mathbf{X}^{(4)}$	$\mathbf{X}^{(3)}$	$\mathbf{X}^{(1)}$
20%	10%	5%, 10%, 20%, 40%, 60%, 80%, 90%, 95%

$\mathbf{X}^{(1)}$	$\mathbf{X}^{(4)}$	$\mathbf{X}^{(3)}$
20%	20%	5%, 10%, 20%, 40%, 60%, 80%, 90%, 95%


Test MSE by Missing Rate in X4
MCAR
MAR3
LOG
MSE
10
9
8
7
6
5
0%
20%
40%
60%
80%
100%
Missing Rate

Test Bias by Missing Rate in X4
MCAR
MAR3
LOG
Bias
0.5
0.0
-0.5
-1.0
0%
20%
40%
60%
80%
100%
Missing Rate

Consistency

$$\lim_{n\to\infty} \mathbb{E}_{\mathbf{X},\mathcal{D}_n}[m_n(\mathbf{X}) - m(\mathbf{X})]^2 = 0$$

$$\mathbb{E}_{Y,\mathbf{X}|\mathcal{D}_n}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 = \text{Estimation Error} + \text{Approximation Error}$$

Estimation Error

$$\mathbb{E}_{Y,\mathbf{X}|\mathcal{D}_n}\left[m_n(\mathbf{X}) - Y\right]^2 - \inf_{f\in\mathcal{F}_n} \mathbb{E}_{Y,\mathbf{X}}\left[f(\mathbf{X}) - Y\right]^2$$

Approximation Error

$$\inf_{f\in\mathcal{F}_n} \mathbb{E}_{Y,\mathbf{X}}\left[f(\mathbf{X}) - Y\right]^2 - \mathbb{E}_{Y,\mathbf{X}}\left[m(\mathbf{X}) - Y\right]$$

Estimation Error

$$\mathbb{E}_{Y,\mathbf{X}|\mathcal{D}_n}\left[m_n(\mathbf{X}) - Y\right]^2 - \inf_{f\in\mathcal{F}_n} \mathbb{E}_{Y,\mathbf{X}}\left[f(\mathbf{X}) - Y\right]^2$$

$$\leq 2 \sup_{f\in\mathcal{F}_n} \left| \frac{1}{n}\sum_{i=1}^{n} \left[f(\mathbf{X}_i) - Y_i\right]^2 - \mathbb{E}_{Y,\mathbf{X}}\left[f(\mathbf{X} - Y)\right]^2 \right|$$

We use concentration inequalities for empirical processes.

If $g : \mathbb{R}^{d+1} \to [0, B]$, by Hoeffding's Inequality

$$\mathbb{P}\left\{ \left| \frac{1}{n}\sum g(Z_i) - \mathbb{E}g(Z) \right| > \varepsilon \right\} \leq 2e^{-\frac{2n\varepsilon^2}{B^2}}$$

Using Boole's Inequality

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum g(Z_i)-\mathbb{E}g(Z)\right|>\varepsilon\right\}\leq 2|\mathcal{G}_n|e^{-\frac{2n\varepsilon^2}{B^2}}$$

By Lemma of Borel-Cantelli, if

$$\sum_{n=1}^{\infty}|\mathcal{G}_n|e^{-\frac{2n\varepsilon^2}{B^2}}<\infty$$

then

$$\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum g(Z_i)-\mathbb{E}g(Z)\right|\xrightarrow[n\to\infty]{a.s.}0$$

Approximation Error

$$\inf_{f \in \mathcal{F}_n} \mathbb{E}_{Y,\mathbf{X}} \left[f(\mathbf{X}) - Y \right]^2 - \mathbb{E}_{Y,\mathbf{X}} \left[m(\mathbf{X}) - Y \right]^2$$
$$= \inf_{f \in \mathcal{F}_n} \mathbb{E}_{\mathbf{X}} \left[f(\mathbf{X}) - m(\mathbf{X}) \right]^2$$

$$\leq \mathbb{E} \left[\Delta \left(m, A_n(\mathbf{X}, \Theta) \right) \right]^2$$
$$\leq \xi^2 + 4 \|m\|_\infty^2 \mathbb{P} \left[\Delta \left(m, A_n(\mathbf{X}, \Theta) \right) > \xi \right]$$

$$\Delta(m, A) = \sup_{x, x' \in A} |m(x) - m(x')|$$

(H1) *he response Y follows*

$$Y = \sum_{j=1}^{p} m_j \left(\mathbf{X}^{(j)} \right) + \varepsilon$$

where $\mathbf{X}$ is uniformly distributed over $[0,1]^p$, there is a MCAR mechanism for all input variables, ε is an independent centered Gaussian noise with finite variance $\sigma^2 > 0$ and each component m_j is continuous.[3]

[3] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. "Consistency of random forests". In: *The Annals of Statistics* 43.4 (2015), pp. 1716–1741.

Theorem 1

Assume that (H1) is satisfied. Then, if $a_n \to \infty$, $t_n \to \infty$ and $t_n \left(\log a_n\right)^9 / a_n \to 0$, the random forest is consistent, i.e.

$$\lim_{n \to \infty} \mathbb{E}_{\mathbf{X}, \mathcal{D}_n} \left[m_n(\mathbf{X}) - m(\mathbf{X}) \right]^2 = 0$$

Lemma 1

Assume that (H1) is satisfied. Then, $\forall\, \mathbf{x} \in [0,1]^p$,

$$\Delta\left(m, A_k^\star(\mathbf{x}, \Theta)\right) \to 0 \quad \text{casi seguro, con } k \to \infty$$

Lemma 2

Assume that (H1) is satisfied. Fix $\mathbf{x} \in [0,1]^p$, $k \in \mathbb{N}^\star$ and let be $\xi > 0$, $L_{n,k}(\mathbf{x}, \cdot)$ is stochastically equicontinuous in $\bar{\mathcal{A}}_k^\xi(\mathbf{x})$; that is, $\forall\, \alpha, \rho > 0$, $\exists\, \delta > 0$ t.q.

$$\lim_{n\to\infty} \mathbb{P}\Big[\sup_{\substack{\|\mathbf{d}_k - \mathbf{d}_k'\|_\infty \leq \delta \\ \mathbf{d}_k, \mathbf{d}_k' \in \bar{\mathcal{A}}_k^\xi(\mathbf{x})}} |L_{n,k}(\mathbf{x}, \mathbf{d}_k) - L_{n,k}(\mathbf{x}, \mathbf{d}_k')| > \alpha \Big] \leq \rho$$

Lemma 3

Assume that (H1) is satisfied. Fix $\xi, \rho > 0$ and $k \in \mathbb{N}^\star$, $\exists\, N \in \mathbb{N}^\star$ s.t. $\forall\, n \geq N$

$$\mathbb{P}[d_\infty(\hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta), \mathcal{A}_k^\star(\mathbf{X}, \Theta)) \leq \xi] \geq 1 - \rho$$

Lemma 1 studies the theoretical random forests.
Lemma 3 Proves (via Lemma 2) that theoretical cuts and empirical cuts are close.
Finally,

$$\mathbb{P}\left[\Delta\left(m, A_n(\mathbf{X}, \Theta)\right) > \xi\right] \xrightarrow{n\to\infty} 0$$

as consequence of Lemmas 1 y 3.

Conclusions

The proposed algorithm:

- Does not "explode" even with a high percentage of missingness. ▲
- Does not required the proximity matrix and just build one forest. ▲
- The assignation of the missing values can be done in parallel. ▲
- Has a similar MSE to other imputation methods. ▲
- Does not require a previous processing of the data. ▲
- Uses extensive computer power. ◆

Thank you!

irving.gomez@cimat.mx
irvinggomez.netlify.com