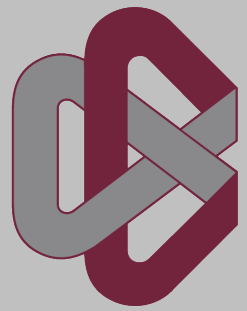
Regresión con datos perdidos usando Bosques Aleatorios

Irving Gómez Méndez

13 mayo 2019



CIMAT

Problema

- Variable de interés $Y \in \mathbb{R}$
- Vector de predictores $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^p$

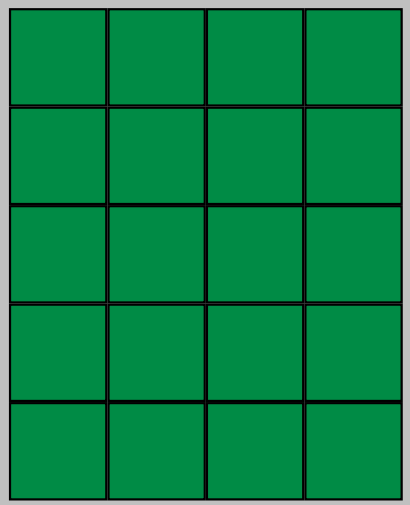
$$\mathbf{X} = \left(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)}\right)$$

- Muestra de entrenamiento $\mathcal{D}_n = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$

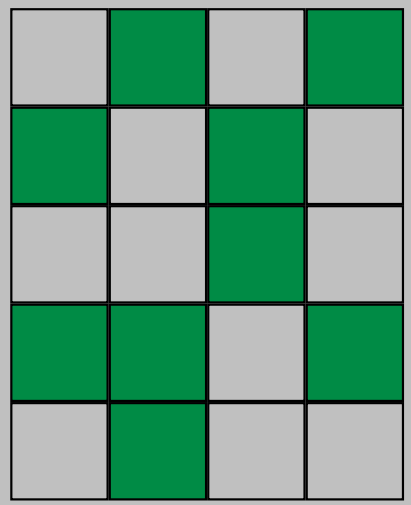
Deseamos estimar $m(\cdot) \equiv \mathbb{E}[Y|\mathbf{X} = \cdot]$ usando $\mathcal{D}_n$

Idealmente la matriz de predictores debería estar completa, pero en la práctica podría tener valores perdidos ("huecos").

$\mathbf{X}^{(1)}\ \mathbf{X}^{(2)}\ \mathbf{X}^{(3)}\ \mathbf{X}^{(4)}$



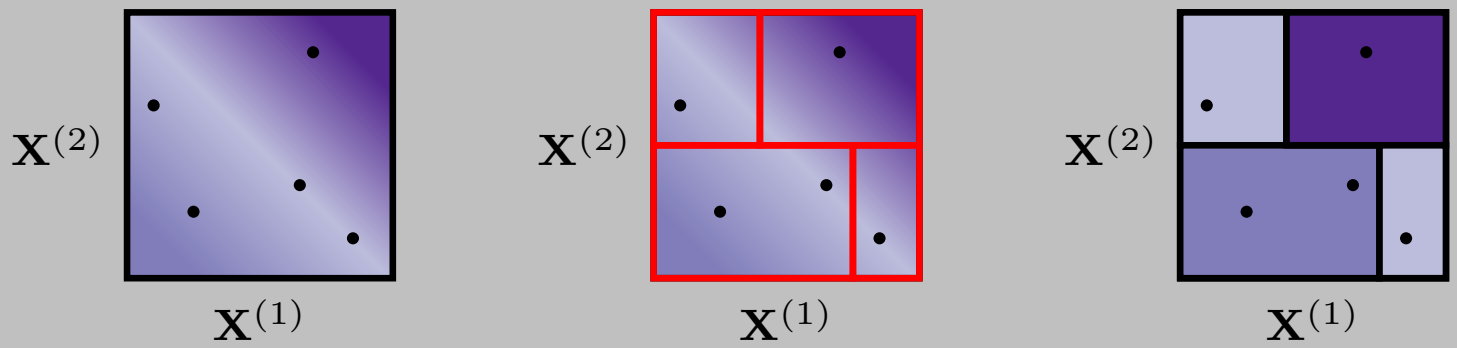
$\mathbf{X}^{(1)}\ \mathbf{X}^{(2)}\ \mathbf{X}^{(3)}\ \mathbf{X}^{(4)}$



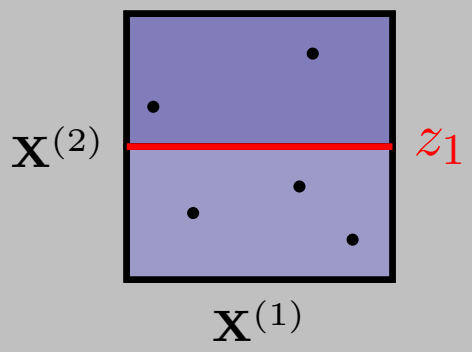
Deseamos estimar $m(\cdot)$ cuando hay valores perdidos.

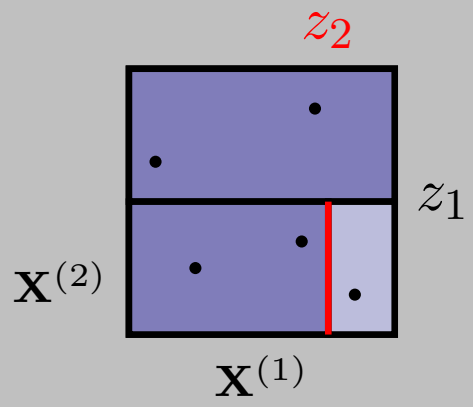
Árboles de regresión

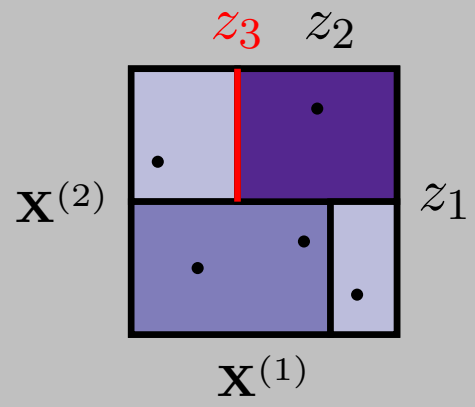
Dividimos el espacio $\mathcal{X}$ en regiones disjuntas y tomamos el promedio de los puntos que están en cada región.

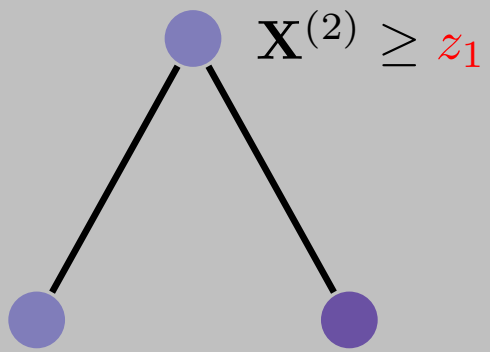


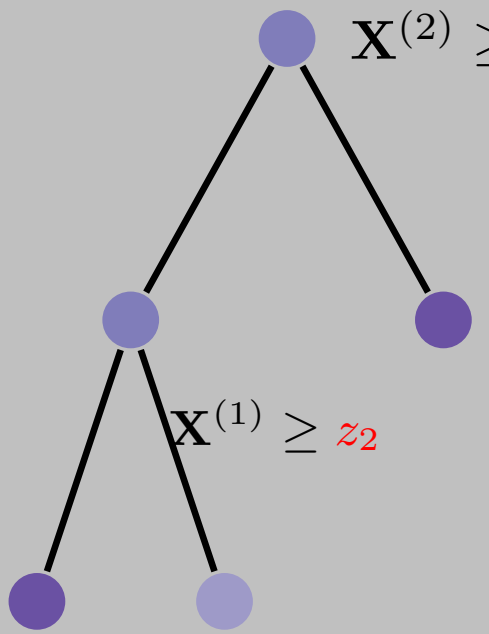
Las celdas se construyen de manera recursiva, lo que da la estructura de árbol al estimador.

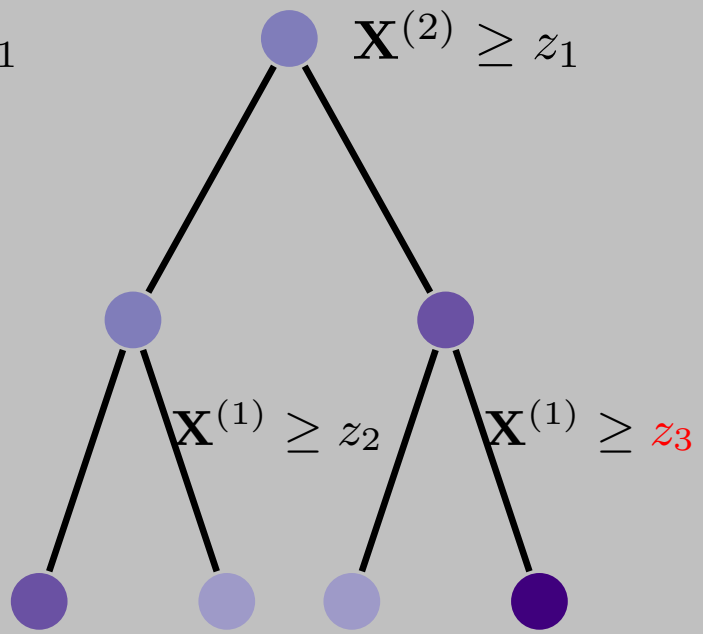

$\mathbf{X}^{(2)}$
z_1
$\mathbf{X}^{(1)}$


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


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

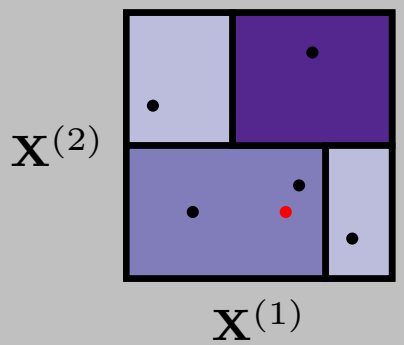

$\mathbf{X}^{(2)} \geq z_1$


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


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

- $A_n(\mathbf{x}, \Theta, \mathcal{D}_n)$ celda que contiene a $\mathbf{x}$.
- $N_n(\mathbf{x}, \Theta, \mathcal{D}_n)$ número de observaciones en la celda $A_n(\mathbf{x}, \Theta, \mathcal{D}_n)$.
- Θ variable que caracteriza al árbol.

$$m_n(\mathbf{x}; \Theta, \mathcal{D}_n) = \sum_{i=1}^{n} \frac{Y_i \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}; \Theta, \mathcal{D}_n)}}{N_n(\mathbf{x}, \Theta, \mathcal{D}_n)}$$



$m_n(\ \cdot\ ; \Theta, \mathcal{D}_n) =$

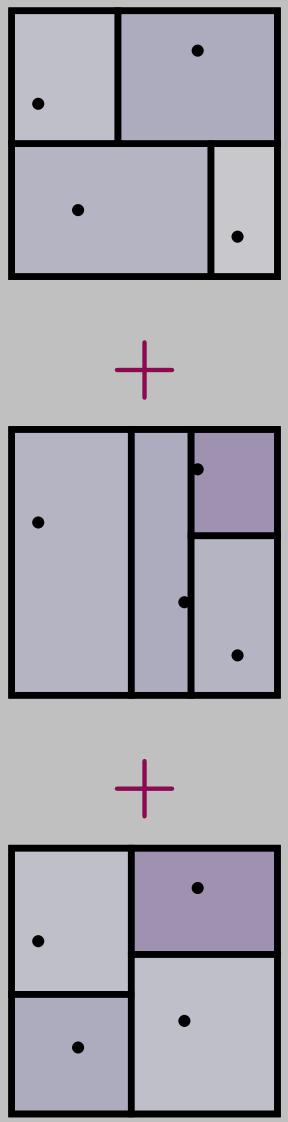
Criterio CART

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

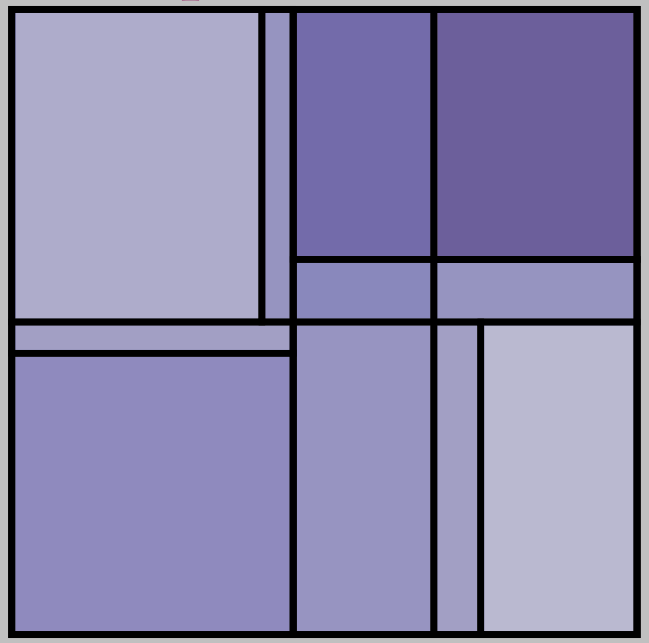
Construimos las celdas maximizando el criterio CART sobre todos los posibles cortes en la celda $\mathcal{C}_A$.

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

Bosques de regresión



Bosque Aleatorio



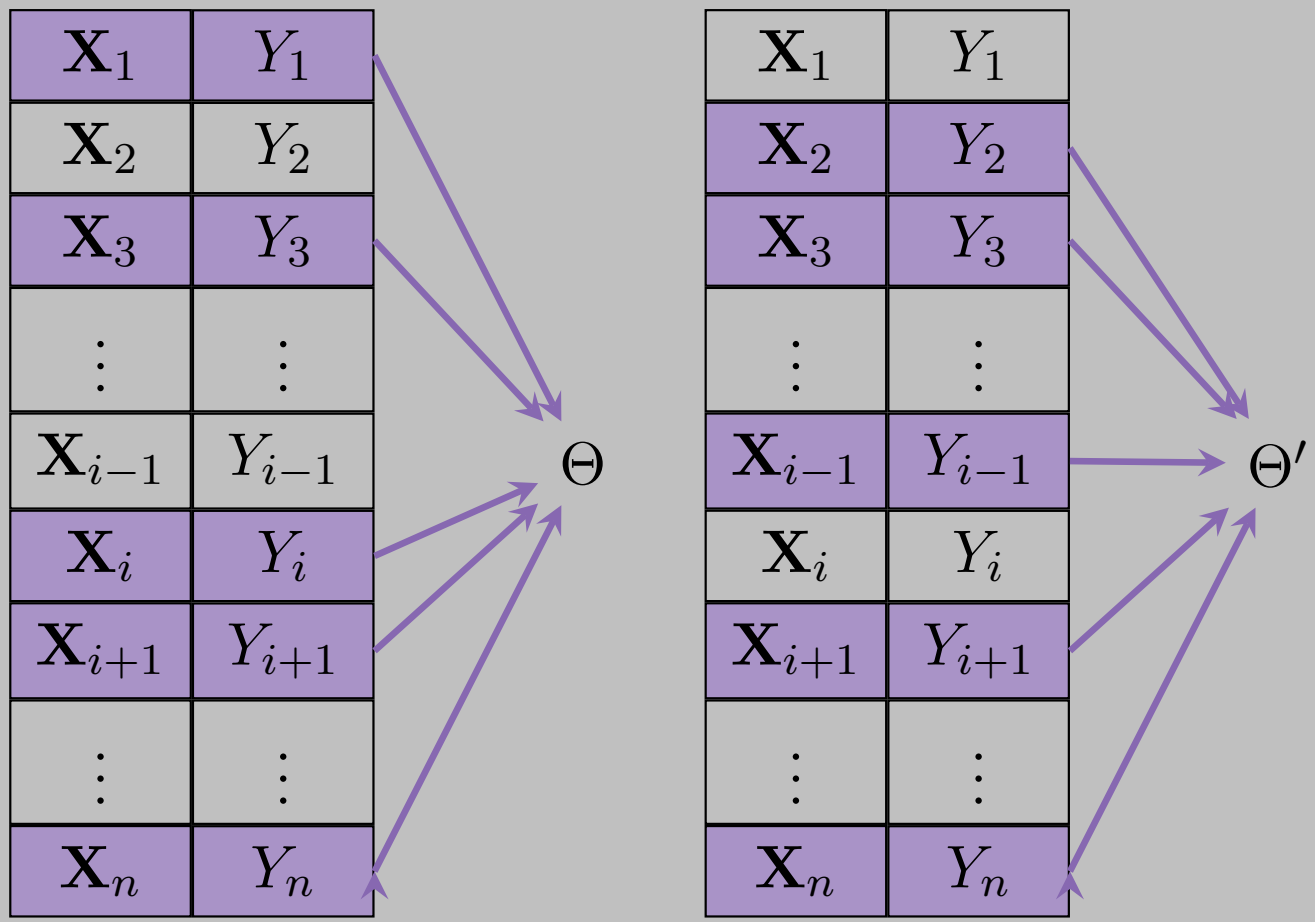
El bosque es un ensamble de varios árboles, i.e.

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

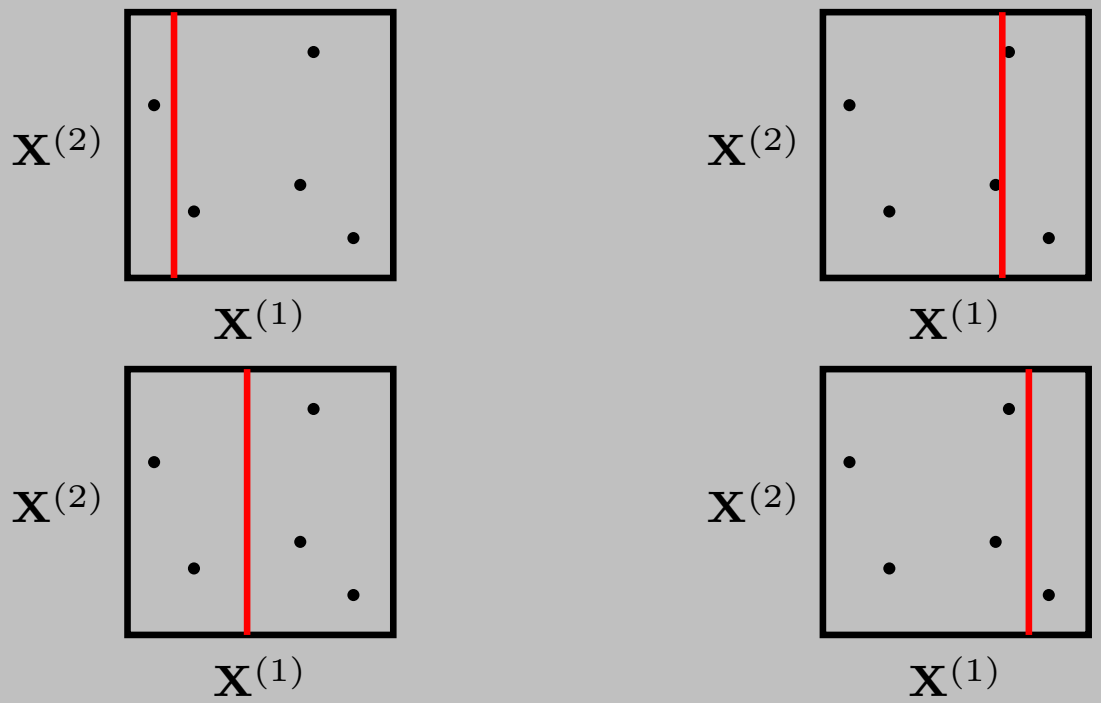
Pero agregamos dos fuentes de variabilidad al construir los árboles, que son

1. Las observaciones usadas por cada árbol.
2. Direcciones candidatas en el criterio CART.

Previo a la construcción de cadad árbol seleccionamos al azar con o sin reemplazo a_n observaciones para cada árbol.



Previo a maximizar el criterio CART para dividir una celda, seleccionamos al azar `mtry` posibles direcciones.



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

Los datos son usados dos veces:

1. Construir las celdas.
2. Estimar la función de regresión en cada celda.

Este doble uso dificulta el análisis teórico del algoritmo original.

La mayoría de los resultados teóricos estudian modelos más simples.

(H1) *Se tiene un modelo aditivo de regresión*

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

$\mathbf{X} = \left(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)}\right)$ *se distribuye uniformemente en* $[0,1]^p$, ε *es ruido gaussiano, independiente, centrado con varianza finita* $\sigma^2 > 0$ *y cada componente* m_j *es continua.*

Teorema 1[1]

Suponga que (H1) se satisface. Entonces, si $a_n \to \infty$, $t_n \to \infty$ y $t_n (\log a_n)^9 / a_n \to 0$, el bosque aleatorio es consistente, i.e.

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

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

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

Error de Estimación

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

Error de Aproximación

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

Error de Estimación

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

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

Usamos desigualdades de concentración para procesos empíricos.

Si $g : \mathbb{R}^{d+1} \to [0, B]$, por la Desigualdad de Hoeffding

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

Usando la Desigualdad de Boole

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

Por Lema de Borel-Cantelli, si

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

entonces

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

Eror de Aproximación

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

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

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

Lema 1

Suponga que (H1) se satisface. Entonces, $\forall\, \mathbf{x} \in [0,1]^p$,

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

Lema 2

Suponga que (H1) se satisface. Fijando $\mathbf{x} \in [0,1]^p$, $k \in \mathbb{N}^\star$ y sea $\xi > 0$, $L_{n,k}(\mathbf{x}, \cdot)$ es estocaśticamente equicontinua en $\bar{\mathcal{A}}_k^\xi(\mathbf{x})$; es decir, $\forall\, \alpha, \rho > 0$, $\exists\, \delta > 0$ t.q.

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

Lema 3

Suponga que (H1) se satisface. Fijando $\xi, \rho > 0$ y $k \in \mathbb{N}^\star$, $\exists\, N \in \mathbb{N}^\star$ t.q. $\forall\, n \geq N$

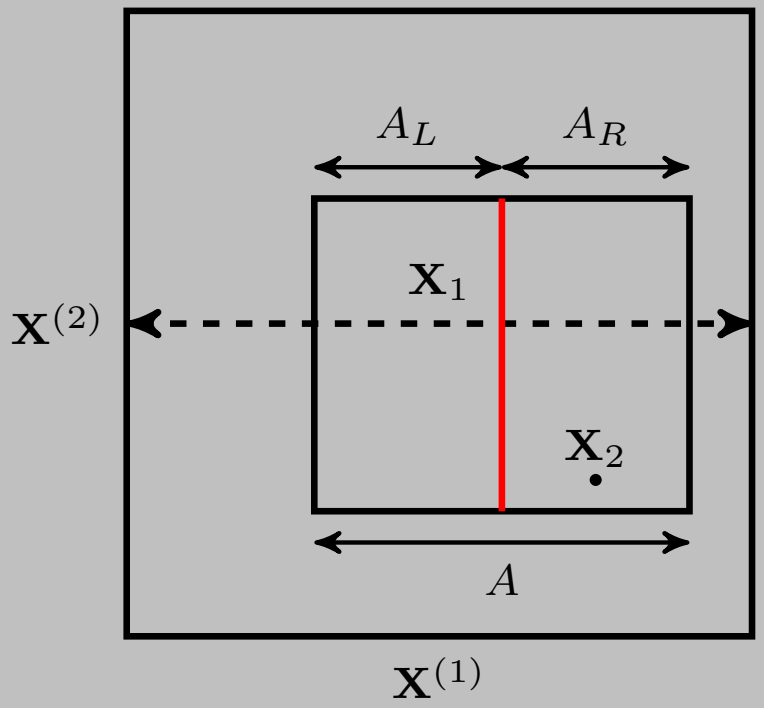
$$\mathbb{P}[d_\infty(\hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta), \mathcal{A}_k^\star(\mathbf{X}, \Theta)) \leq \xi] \geq 1 - \rho$$

En el Lema 1 se estudian los bosques aleatorios teóricos.
El Lema 3 prueba (via el Lema 2) que los cortes teóricos y empíricos son cercanos.
Finalmente,

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

como consecuencia de los Lemas 1 y 3.

Bosques Aleatorios con Datos Perdidos



	$\mathbf{X}^{(1)}$	$\mathbf{X}^{(2)}$
$\mathbf{X}_1$		0.5
$\mathbf{X}_2$	0.75	0.25
A	$[0.3, 0.9]$	$[0.2, 0.7]$
A_L	$[0.3, 0.6]$	$[0.2, 0.7]$
A_R	$[0.6, 0.9]$	$[0.2, 0.7]$

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

Imputación

Los métodos que hacen imputación mediante bosques aleatorios suelen ser iterativos.
Comenzando con una iteración burda que se va mejorando en los siguientes etapas.

Definimos

$$M_i^{(h)} = \begin{cases} 1 & \text{if } \mathbf{X}_i^{(h)} \text{ está perdida} \\ 0 & \text{en otro caso} \end{cases}, \quad 1 \leq h \leq p$$

Sean

$$\mathbf{i}_{obs}^{(h)} = \{i \in \{1, \ldots, n\};\ M_i^{(h)} = 0\}$$
$$\mathbf{i}_{miss}^{(h)} = \{i \in \{1, \ldots, n\};\ M_i^{(h)} = 1\}$$

$$\hat{\mathbf{X}}_{j,t_1}^{(h)} = \text{mediana}\left(\mathbf{X}_{\mathbf{i}_{obs}^{(h)}}^{(h)}\right), \qquad \begin{array}{l} h = 1, \ldots, p \\ j \in \mathbf{i}_{miss}^{(h)} \end{array}$$

Definimos

$$\mathbf{X}_{i,t_1}^{(h)} = \begin{cases} \mathbf{X}_i^{(h)} & \text{si } i \in \mathbf{i}_{obs}^{(h)} \\ \hat{\mathbf{X}}_{i,t_1}^{(h)} & \text{si } i \in \mathbf{i}_{miss}^{(h)} \end{cases}, \quad h = 1, \ldots, p$$

$$\mathbf{X}_{i,t_1} = \left(\mathbf{X}_{i,t_1}^{(1)}, \ldots, \mathbf{X}_{i,t_1}^{(p)}\right)$$

Y

$$\mathcal{D}_{n,t_1} = ((\mathbf{X}_{1,t_1}, Y_1), \ldots, (\mathbf{X}_{n,t_1}, Y_n))$$

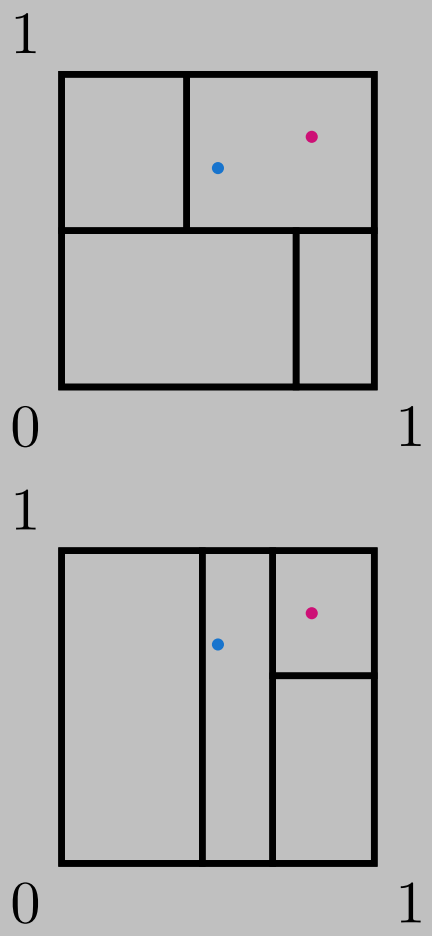
1. Construimos el bosque con $\mathcal{D}_{n,t_1}$, i.e.

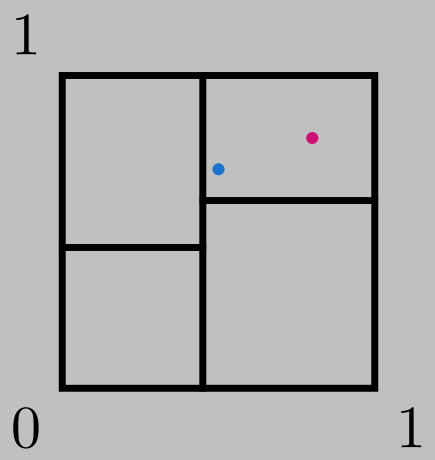
$$m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_{n,t_1}) = \frac{1}{M} \sum_{k=1}^{M} m_n(\mathbf{x}; \Theta_k, \mathcal{D}_{n,t_1})$$

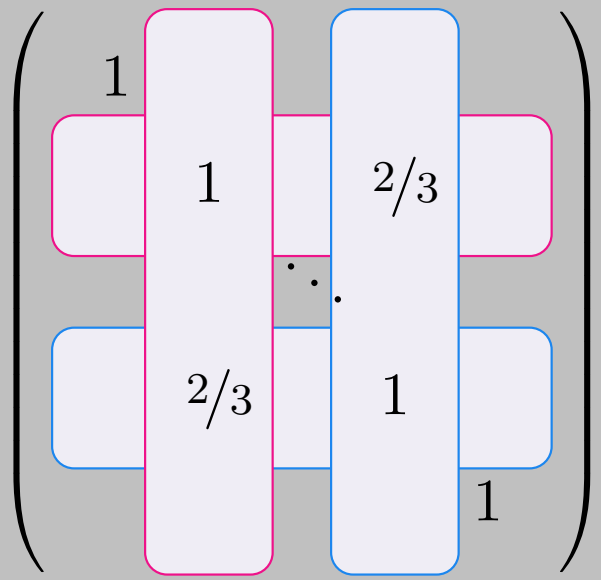
2. Calculamos la matriz de proximidad del bosque
3. Usamos la matriz de proximidad para mejorar la imputación[23] y tener $\mathcal{D}_{n,t_2}$
4. Iteramos

[2] Tsunenori Ishioka. "Imputation of missing values for unsupervised data using the proximity in random forests". In: *International Conference on Mobile, Hybrid, and On-line Learning. Nice.* 2013, pp. 30–36.

[3] Leo Breiman. *Setting up, using, and understanding random forests V4.0*. 2003. URL: https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf.







La matriz de proximidad tiene la proporción de árboles en que los puntos están “conectados”.

Sea $q_{t_\ell}(i,j)$ la entrada i,j de la matriz de proximidad del bosque $m_{M,n}(\mathbf{x};\Theta_1,\ldots,\Theta_M,\mathcal{D}_{n,t_\ell})$, $\ell \geq 1$

- **Breiman**

$$\hat{\mathbf{X}}^{(h)}_{j,t_\ell+1} = \frac{\sum_{i\in \mathbf{i}_{obs}} q_{t_\ell}(i,j)\mathbf{X}^{(h)}_i}{\sum_{i\in \mathbf{i}_{obs}} q_{t_\ell}(i,j)}, \quad j \in \mathbf{i}^{(h)}_{miss}$$

- **Ishioka**

$$\hat{\mathbf{X}}^{(h)}_{j,t_\ell+1} = \frac{\sum_{\substack{i\in \text{neigh}_k \\ i\neq j}} q_{t_\ell}(i,j)\mathbf{X}^{(h)}_{i,t_\ell}}{\sum_{\substack{i\in \text{neigh}_k \\ i\neq j}} q_{t_\ell}(i,j)}, \quad j \in \mathbf{i}^{(h)}_{miss}$$

Nuestro Método

Sea

$$A_1 = \mathcal{X}$$

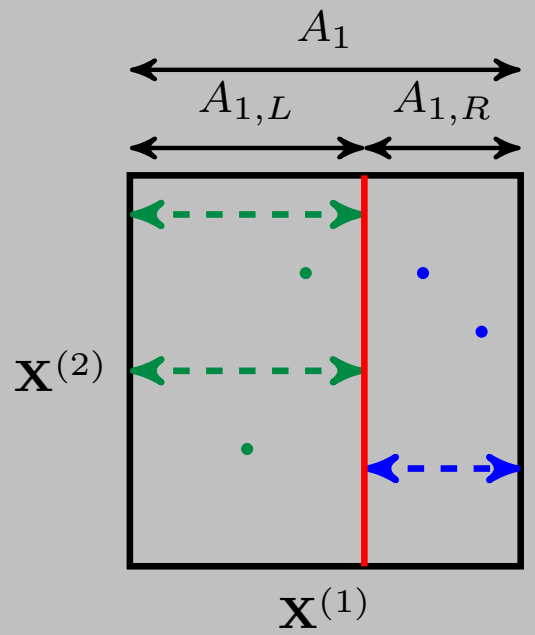
Y

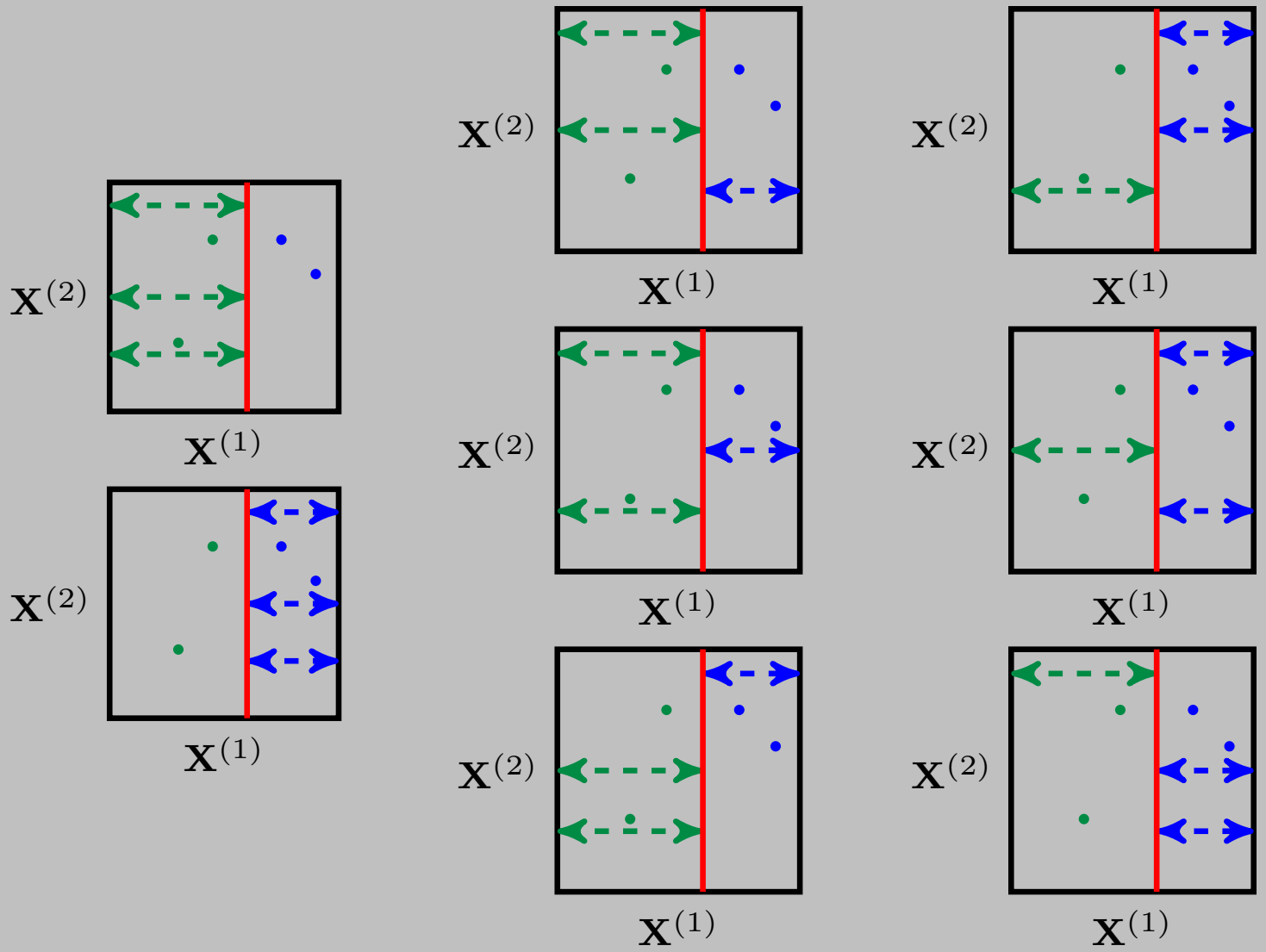
$$\mathbf{X}_{i,1}^{(h)} = \begin{cases} \mathbf{X}_i^{(h)} & \text{si } M_i^{(h)} = 0 \\ \mathcal{X}^{(h)} & \text{si } M_i^{(h)} = 1 \end{cases}, \quad 1 \leq h \leq p$$

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

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

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

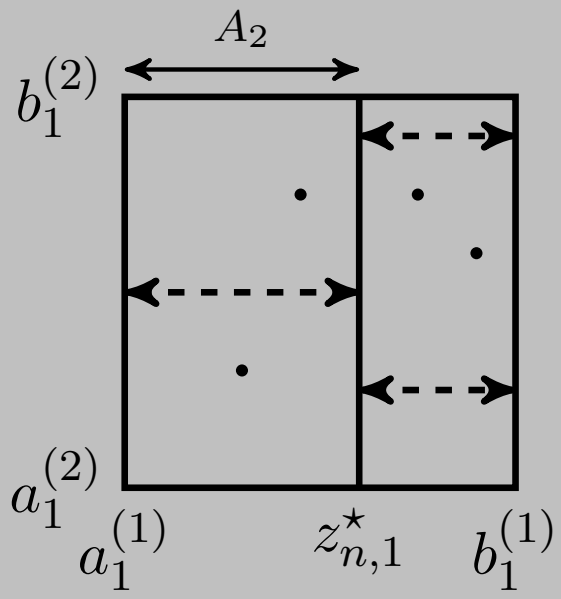



$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$
$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$
$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$
$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$
$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$
$\mathbf{X}^{(2)}$
$\mathbf{X}^{(1)}$

1. Calculamos el criterio CART con todos los posibles valores de las variables perdidas.
2. Seleccionamos la combinación que maximiza el criterio CART, entre $\mathcal{M}_{try}$ y $\mathcal{C}_{A_1}$

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

3. Moverse a A_2 y tomar $\mathbf{X}_{i,2}$



$$\mathbf{X}_{i,2}^{(h_{n,1}^{\star})} = \begin{cases} \left[a_1^{(h_{n,1}^{\star})}, z_{n,1}^{\star}\right] & \text{si } M_i^{(h_{n,1}^{\star})} = 1 \text{ y } \mathbf{X}_{i,1}^{\star(h_{n,1}^{\star})} < z_{n,1}^{\star} \\ \left[z_{n,1}^{\star}, b_1^{(h_{n,1}^{\star})}\right] & \text{si } M_i^{(h_{n,1}^{\star})} = 1 \text{ y } \mathbf{X}_{i,1}^{\star(h_{n,1}^{\star})} \geq z_{n,1}^{\star} \\ \mathbf{X}_i^{(h_{n,1}^{\star})} & \text{si } M_i^{(h_{n,1}^{\star})} = 0 \end{cases}$$

Mecanismos de Pérdida

- **Pérdida Completamente al Azar (MCAR)** Una variable está perdida completamente al azar si la probabilidad de pérdida es la misma para todas las unidades.
- **Pérdida al Azar (MAR)** Una variable está perdida al azar si la probabilidad de pérdida sólo depende de información observada.
- **Pérdida No al Azar (NMAR)** Si la pérdida depende de información que ha sido recolectada, deja de ser una pérdida al azar.

Simulaciones

- Tomamos la función de regresión "friedman1":

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

Bases de entrenamiento

- Creamos 10 bases de datos de entrenamiento.
- Simulamos 200 observaciones de $\mathbf{X} \sim \mathcal{U}[0,1]^5$.

Base de prueba

- Creamos 1 base de datos de prueba.
- Simulamos 2000 observaciones de $\mathbf{X} \sim \mathcal{U}[0,1]^5$.
- Todos los valores están observados.

- Los valores perdidos son creados siguiendo 7 mecanismos de pérdida diferentes:
  - 1 completamente al azar (MCAR).
  - 5 al azar (MAR1, MAR2, MAR3, MAR4, Depy).
  - 1 no al azar (LOG).

Variable Determinante	Variable Perdida	% Datos Perdidos
$\mathbf{X}^{(2)}$, Y	$\mathbf{X}^{(1)}$	20%
$\mathbf{X}^{(5)}$, Y	$\mathbf{X}^{(3)}$	10%
	$\mathbf{X}^{(4)}$	20%

- **MAR1** La probabilidad de `NA` es

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

- **MAR2** Creamos dos grupos en var. det. El primer grupo son mayores a mediana, el segundo menores.
  La probabilidad de `NA` para cada grupo es

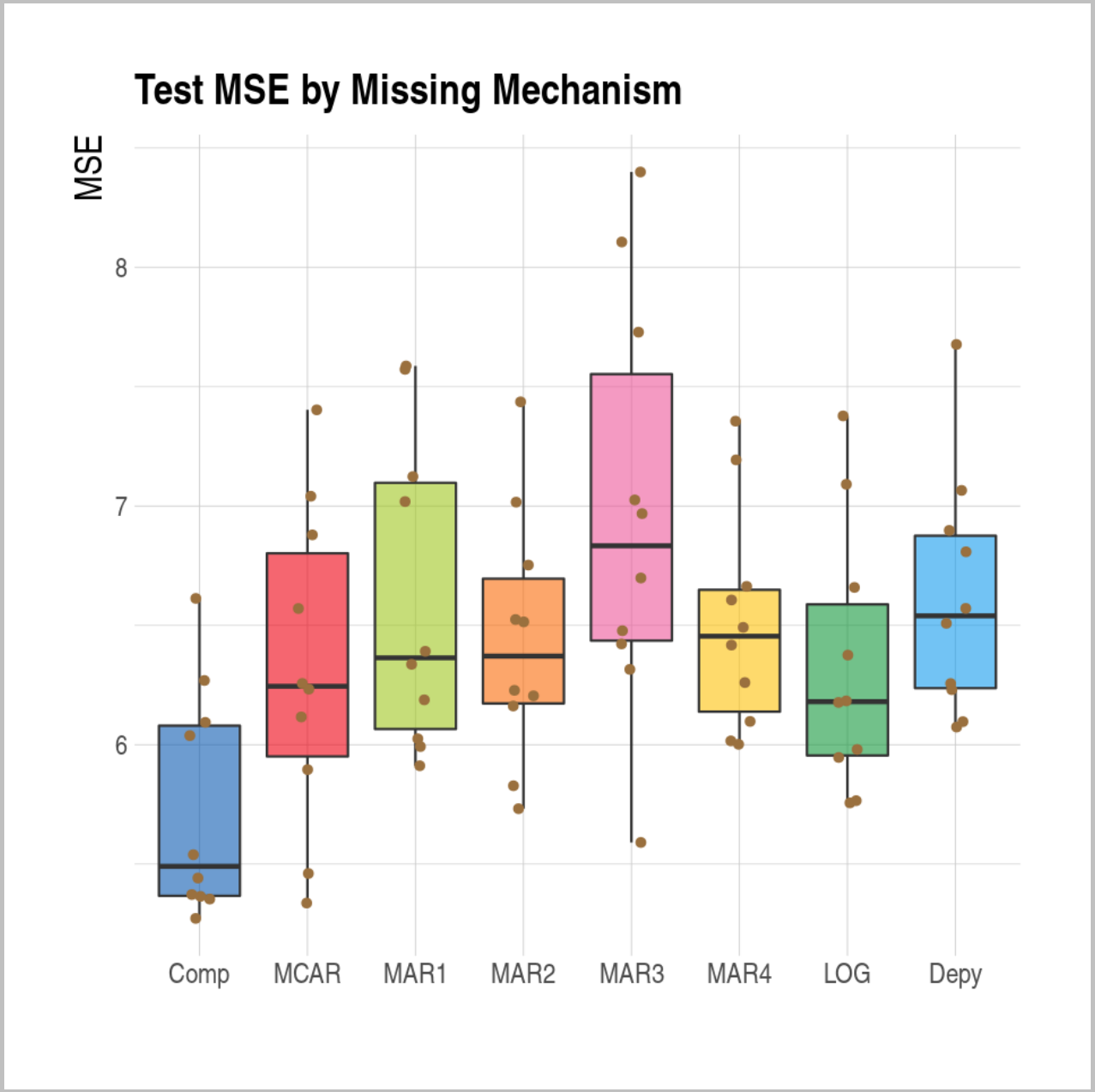
$$0.9/\#(\text{obs. en 1er grupo}) \quad 0.1/\#(\text{obs. en 2do grupo})$$

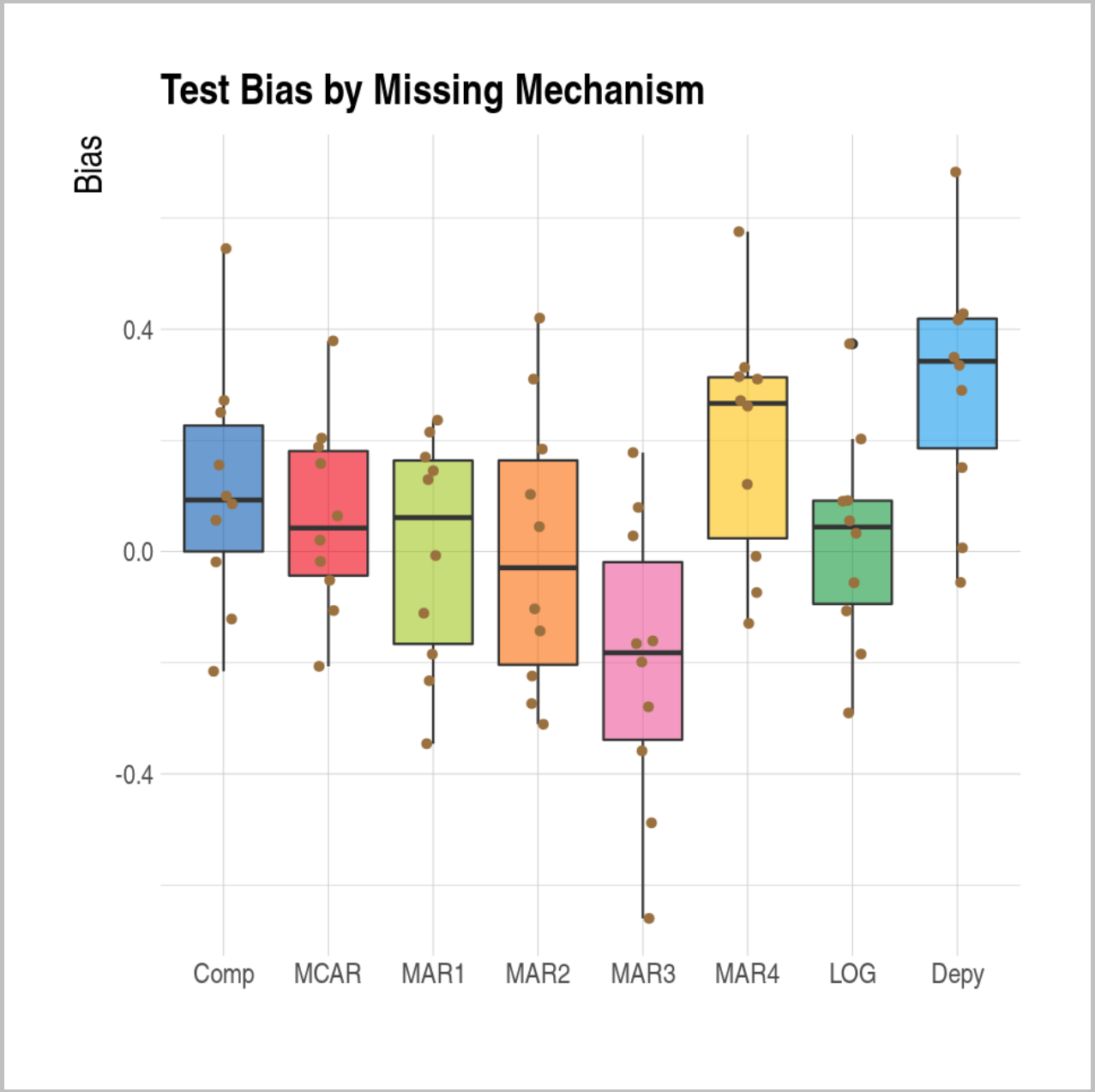
- **MAR3** Valores más grandes en var. det. son `NA` en var. perdida
- **MAR4** Valores más grandes y más pequeños en var. det. son `NA` en var. perdida
- **Depy** Probabilidad de `NA` es 0.1 si $Y \geq 13$, si no, es 0.4
- **LOG**

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

Bosques aleatorios

- Por cada base de datos y cada mecanismo de pérdida (incluyendo los datos completos) creamos 1 bosque aleatorio.
- Cada bosque es construido con los parámetros:
  - $M = 50$ árboles.
  - `mtry` $= 1$ variable seleccionada al azar para hacer el corte.
  - $a_n = 127$ observaciones seleccionadas al azar y sin reemplazo para cada árbol.
  - `nodesize` $= 5$, máximo número de observaciones en nodos finales.


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

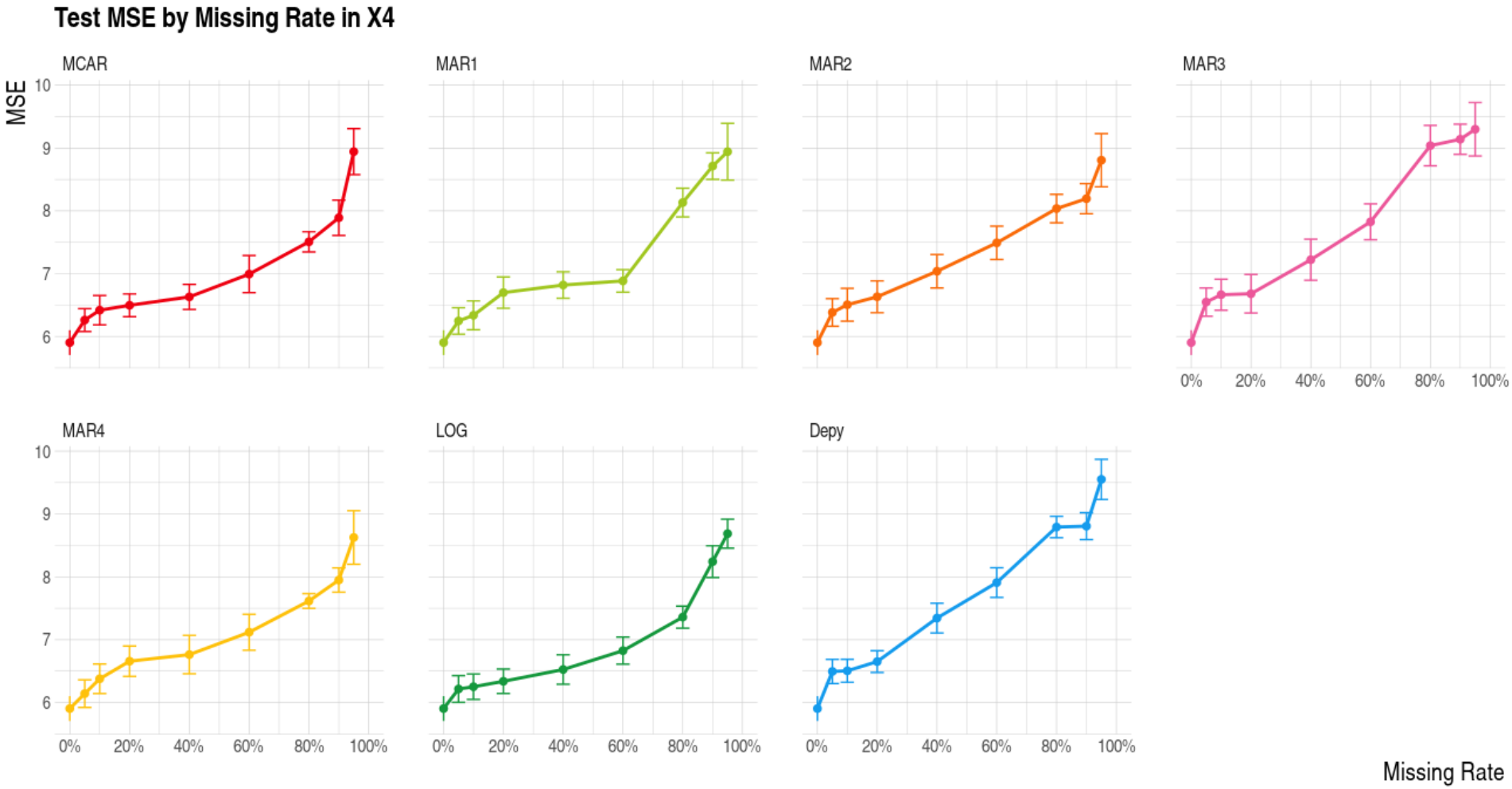

Test Bias by Missing Mechanism
Bias
0.4
0.0
-0.4
Comp
MCAR
MAR1
MAR2
MAR3
MAR4
LOG
Depy

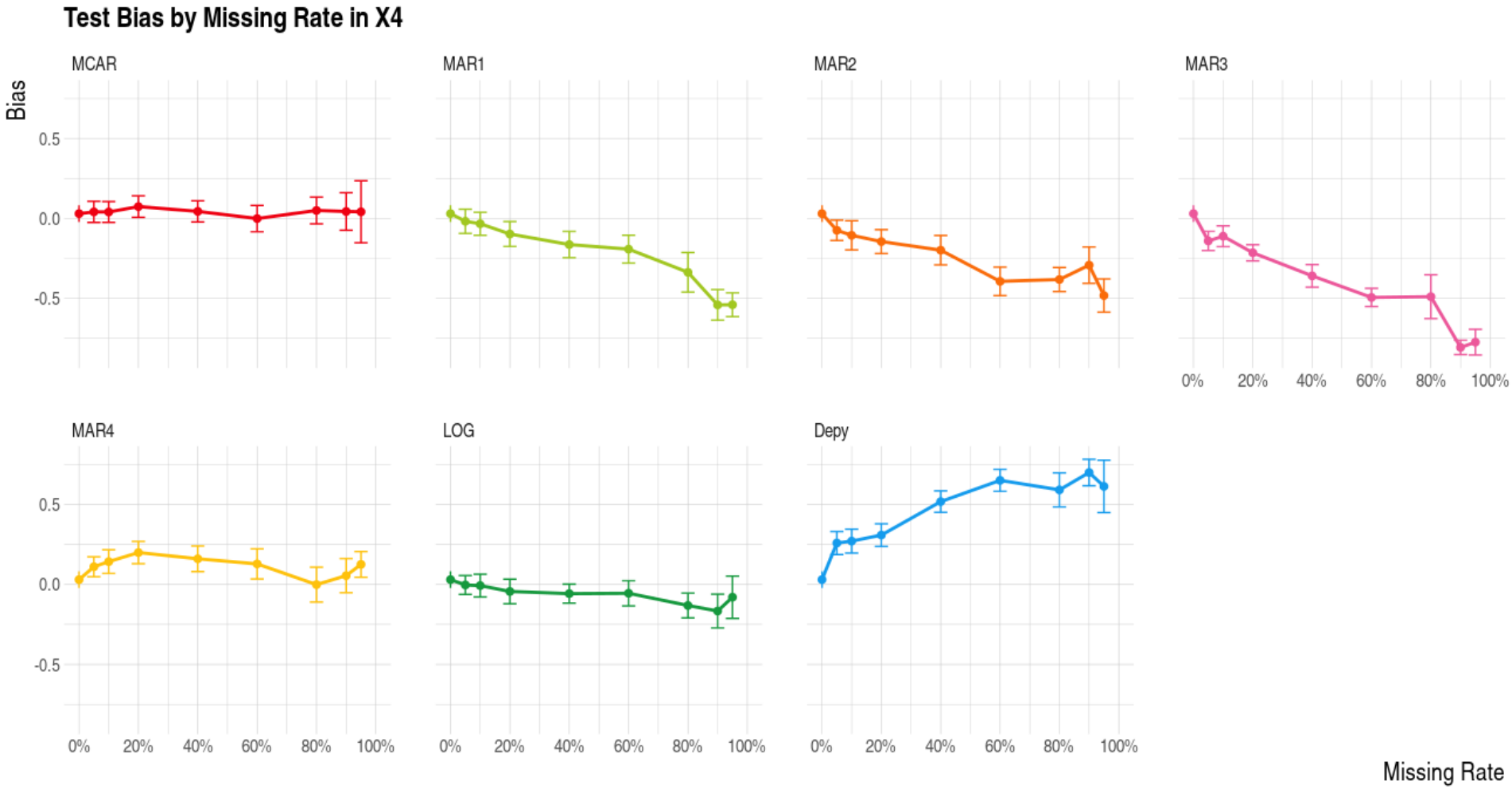
Variamos el porcentaje de pérdida de cada variable por separando, dejando el mismo porcentae para las otras dos vaariables.

$\mathbf{X}^{(1)}$	$\mathbf{X}^{(3)}$	$\mathbf{X}^{(4)}$
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
MAR1
MAR2
MAR3
MAR4
LOG
Depy
MSE
10
9
8
7
6
0%
20%
40%
60%
80%
100%
Missing Rate


Test Bias by Missing Rate in X4
MCAR
MAR1
MAR2
MAR3
MAR4
LOG
Depy
Bias
0.5
0.0
-0.5
0%
20%
40%
60%
80%
100%
Missing Rate

Conclusiones

Nuestro método:

- No "explota" aun con un alto porcentaje de pérdida. ▲
- No requiere la matriz de proximidad y construye un sólo bosque. ▲
- Tiene un MSE similar a métodos que hace imputación. ▲
- No requiere tratar los datos de manera previa. ▲
- No hace imputación. ◆