MissingBigData

Missing data in the big data era

Gaël Varoquaux[†], Nicolas Prost[†⋆]
Julie Josse[⋆], Erwan Scornet[⋆]
⋆ CMAP, École Polytechnique † Inria

MissingBigData

Missing data in the big data era

Gaël Varoquaux[†], Nicolas Prost[†⋆]
Julie Josse[⋆], Erwan Scornet[⋆]

⋆ CMAP, École Polytechnique † Inria

1. **Context**
2. **Random forests with missing values**

1 Context

“big data”

More samples Higher dimensional Observational

Missing data is central to “big data”

1 Context: big data in health and social sciences

- More and more missing data due to:
  - high dimensionality (one feature may be missing)
  - difficulty of fine control on the acquisition process

- Causal conclusions from analysis challenging:
  - observational data (as opposed to experiments)
  - missing data induces selection biases

New data sources challenge missing-data methodology:

- **high-dimensional**
- **observational**
- **uncontrolled confounds**

1 Motivating data in health

Traumabase: 15 000 patients/ 250 var/ 15 hospitals

Center	Age	Sex	Weight	Height	BMI	T°	Lactates	Glasgow
Beaujon	54	m	85	NR	NR	35.6	NA	12
Lille	33	m	80	1.8	24.69	36.5	4.8	15
Pitie	26	m	NR	NR	NR	36	3.9	3
Beaujon	63	m	80	1.8	24.69	36.7	1.66	15
Pitie	30	w	NR	NR	NR	36.6	NM	15

- missing: `Not Recorded`, `Made`, `Applicable`, etc.
- predict the Glasgow score, start of a transfusion
- study the effect of a treatment on survival

1 Motivating data in health

Traumabase: 15 000 patients/ 250 var/ 15 hospitals

Center	Age	Sex	Weight	Height	BMI	T°	Lactates	Glasgow
Beaujon	54	m	85	NR	NR	35.6	NA	12
Lille	33	m	80	1.8	24.69	36.5	4.8	15
Pitie	26	m	NR	NR	NR	36	3.9	3
Beaujon	63	m	80	1.8	24.69	36.7	1.66	15
Pitie	30	w	NR	NR	NR	36.6	NM	15

- missing: `Not Recorded`, `Made`, `Applicable`, etc.
- predict the Glasgow score, start of a transfusion
- study the effect of a treatment on survival

UK Biobank: prospective epidemiology

- 1 Million patients of a normal aging population
- 10% have medical imaging data
- observational data to study risk factors

1 State of the art to handle missing values

Single imputation: complete the data

$\Rightarrow$ Need to reflect the uncertainty in the analyses

Multiple imputation: generate different imputed data and apply the analysis on each imputed data

$\Rightarrow$ Impute by approximating the joint distribution

1 State of the art to handle missing values

Single imputation: complete the data

⇒ Need to reflect the uncertainty in the analyses

Multiple imputation: generate different imputed data and apply the analysis on each imputed data

⇒ Impute by approximating the joint distribution

Solutions:	SVD (+bootstrap) [Josse... 2016]	Nonparametric Bayes
Benefits:	low-rank [Udell, 2017]	flexible
Drawbacks:	struggle with complex relationships	do not scale

Age	Height	T°	Glasgow score
26	1.84	36.0	3
16	1.92	37.5	4
54	1.6	35.6	10
33	1.69	36.0	5
63	1.8	36.7	12
33	1.73	36.5	15

1 Missing value mechanisms

- missing at random everywhere MCAR

Easily unbiased

Age	Height	T°	Glasgow score
26	NA	36.0	3
NA	1.92	37.5	4
54	1.6	35.6	10
33	1.69	NA	5
NA	1.8	36.7	12
33	1.73	NA	15

1 Missing value mechanisms

- missing at random everywhere — MCAR
- missing at random on certain variables — MCAR

(Missingness on X_1) $\perp\!\!\!\perp X_1 | X_{i \neq 1}$

$\Rightarrow$ max likelihood imputation unbiased

Age	Height	T°	Glasgow score
26	1.84	36.0	3
16	1.92	**NA**	4
54	1.6	35.6	10
33	1.69	**NA**	5
63	1.8	36.7	12
33	1.73	**NA**	15

1 Missing value mechanisms

- missing at random everywhere MCAR
- missing at random on certain variables MCAR

  (Missingness on X_1) $\perp\!\!\!\perp X_1 | X_{i \neq 1}$

  $\Rightarrow$ max likelihood imputation unbiased
- missingness not independent of data MNAR

  non-ignorable pattern

Age	Height	T°	Glasgow score
26	1.84	**NA**	$\leftarrow$ 3
16	1.92	**NA**	$\leftarrow$ 4
54	1.6	35.6	10
33	1.69	**NA**	$\leftarrow$ 5
63	1.8	36.7	12
33	1.73	36.5	15

1 Missing not at random and causal interpretation

- Missingness depends on the underlying value (eg income)
  - **problem**: selection bias
  - **solution**: model for the missing values mechanism
  - **state of the art**: only 1 variable with missing values

- Graphical models for missing values [Pearl 2018]
  - Explicit distribution (X, R_X)
  - Ex: Y years of work experience, I income

  $Y \rightarrow I \rightarrow R_I$ but $P(Y|I)$ may be recovered

$\Rightarrow$ Powerful models

to capture interactions between variables

1 Objectives of the MissingBigData project

- **Broad models**: avoid underfitting but also scalable
- Modeling the **dependency structure in missingness** across covariates (not at random)
- Control possible **biases** (non ignorable missingness)

Enable statistical analysis
⇒ Combining predictive models with causal inference

1 Ongoing: causal conclusions with missing values

- Causal conclusions:

  Y outcome, X covariates, W treatment 0 or 1

  Average Treatment Effect $\tau = E[Y_i(1) - Y_i(0)]$

  - experimental design: $\bar{Y}_1 - \bar{Y}_0$
  - observational data: adjust for the covariate

  Unconfoundness: $(Y_i \perp\!\!\!\perp W_i | X_i)$

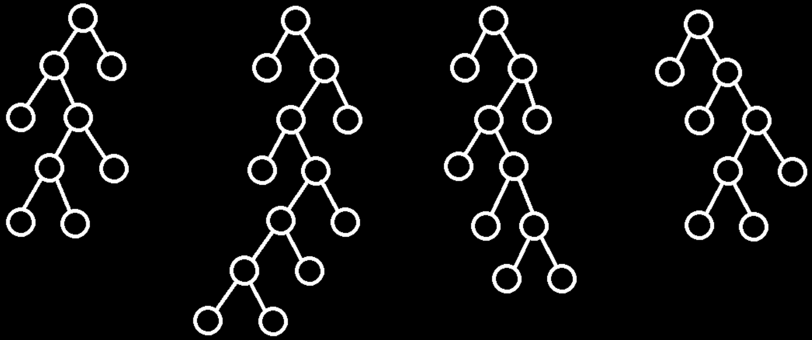
- Inverse probability weighting — "Doubly robusts"

  Estimates weights: $e(x) = P(W_i = 1 | X = x)$

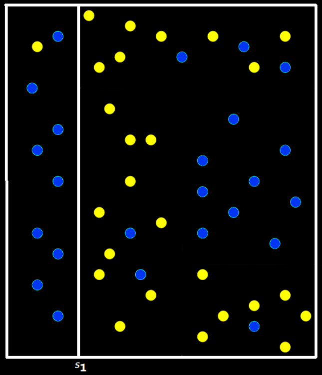
  Average Treatment Effect $\hat{\tau} = \frac{1}{n} \sum_i \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i) Y_i}{1-\hat{e}(X_i)} \right)$

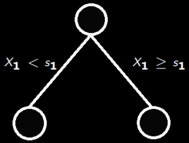
$\Rightarrow$ Random Forests with missing values

2 Random forests with missing values



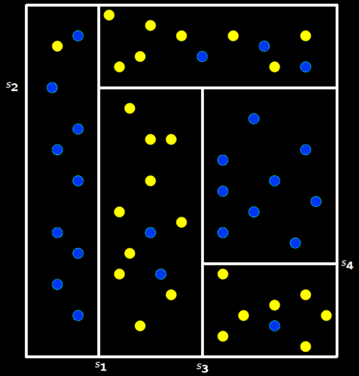
- A split point s_1 is selected at each iteration.

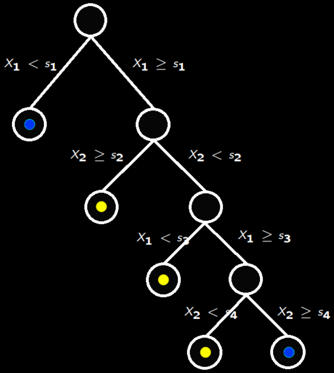




2 Random forest: constructing the trees

- A split point s_1 is selected at each iteration.
- The average of Y in each leaf is the prediction.





$\Rightarrow$ How to split?

2 How to split? Two classic strategies

"Classic" CART

- **Exhaustive search**
- **Impurity of a node:**

$$\mathcal{I} = \sum (Y_i - \overline{Y})^2$$

Conditional trees [Hothorn... 2006]

- Variable choice:

$$T(X_j) = \sum X_i^j Y_i$$

- Threshold choice: impurity

Splitting criterion:

$$\mathcal{C}(X_j) = \mathcal{I} - \mathcal{I}_L^{best} - \mathcal{I}_R^{best}$$

Splitting criterion:

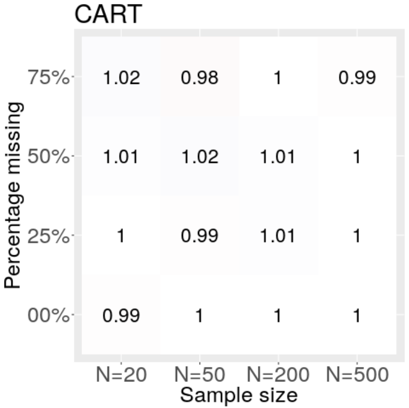
$$\mathcal{C}(X_j) \propto T(X_j)$$

With missing values: sums over available points.

2 Choice of splitting variable

Balanced setting $Y = X_1 + X_2 + \varepsilon$.

The ratio $\mathcal{C}(X_1)/\mathcal{C}(X_2)$ should be close to 1.

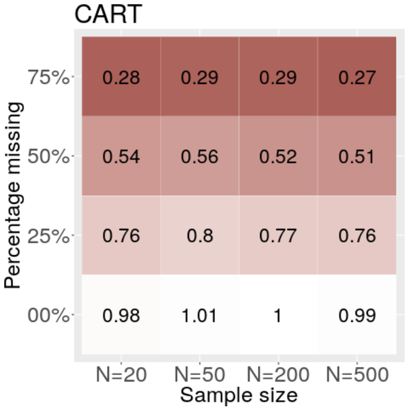


Missing at random on all variables

2 Choice of splitting variable

Balanced setting $Y = X_1 + X_2 + \varepsilon$.

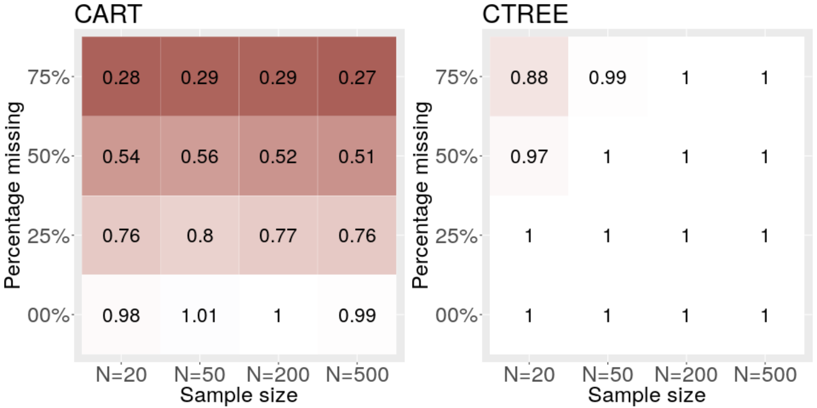
The ratio $\mathcal{C}(X_1)/\mathcal{C}(X_2)$ should be close to 1.



Missing at random on X_1

2 Choice of splitting variable

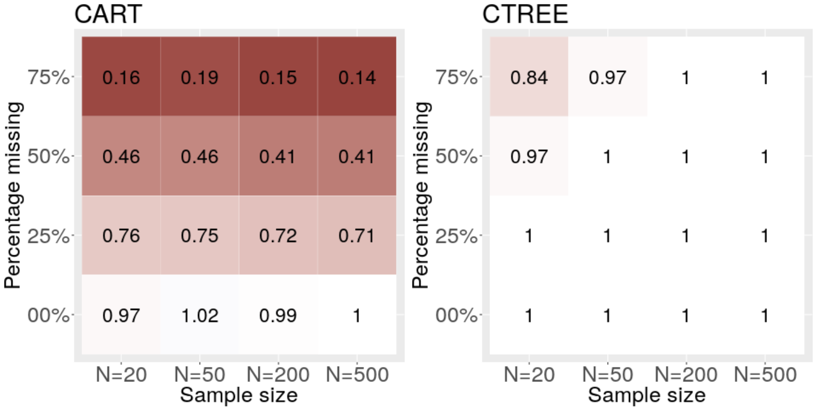
Balanced setting $Y = X_1 + X_2 + \varepsilon$.
The ratio $\mathcal{C}(X_1)/\mathcal{C}(X_2)$ should be close to 1.



Missing at random on X_1

2 Choice of splitting variable

Balanced setting $Y = X_1 + X_2 + \varepsilon$.
The ratio $\mathcal{C}(X_1)/\mathcal{C}(X_2)$ should be close to 1.

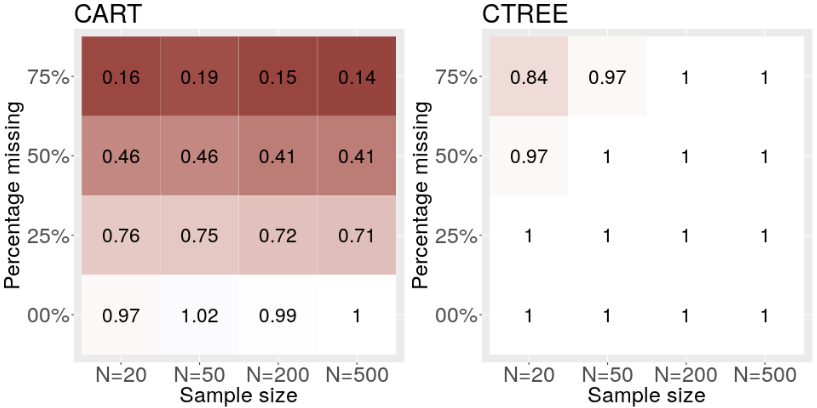


Missing on X_1 depending on the value of Y

2 Choice of splitting variable

Balanced setting $Y = X_1 + X_2 + \varepsilon$.

The ratio $\mathcal{C}(X_1)/\mathcal{C}(X_2)$ should be close to 1.



Missing on X_1 depending on the value of Y

$\Rightarrow$ Conditional trees show negligible bias.

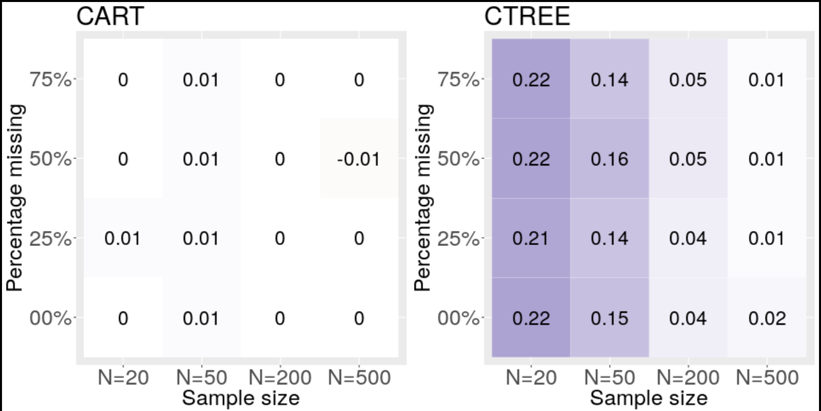
2 Prediction error

Same setting: $Y = X_1 + X_2 + \varepsilon$.

Metric: systematic bias on the prediction of Y.

Same setting: $Y = X_1 + X_2 + \varepsilon$.

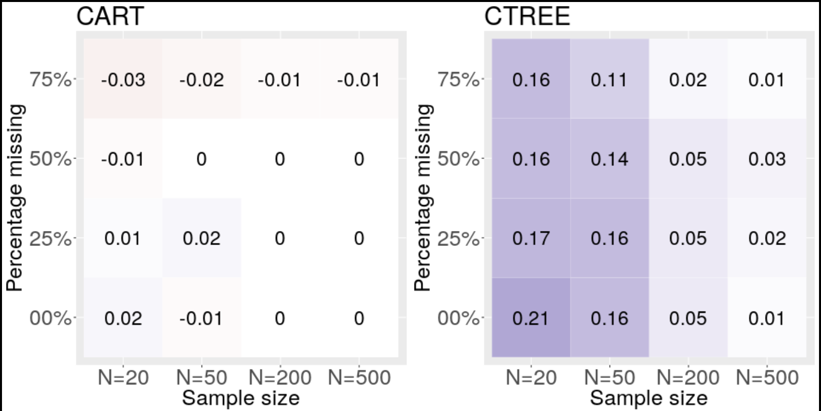
Metric: systematic bias on the prediction of Y.



Missing at random on all variables

Same setting: $Y = X_1 + X_2 + \varepsilon$.

Metric: systematic bias on the prediction of Y.

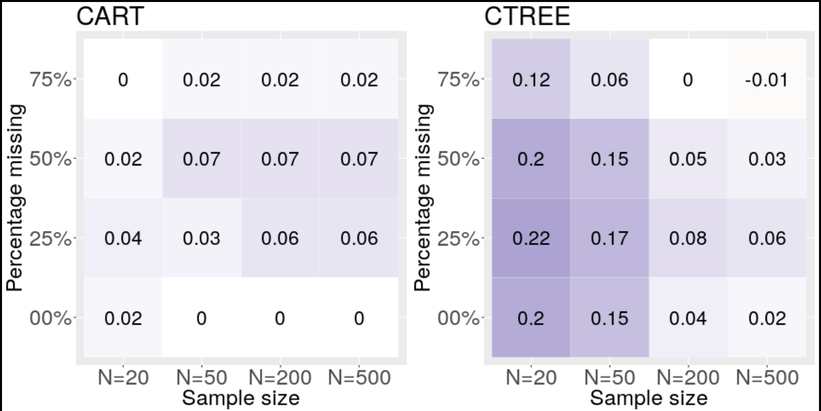


Missing at random on X_1

2 Prediction error

Same setting: $Y = X_1 + X_2 + \varepsilon$.

Metric: systematic bias on the prediction of Y.



Missing on X_1 depending on the value of Y

2 Conclusion on random forest with missing data

Inference $\neq$ prediction.

- Conditional trees correct the bias in inference of parameters.
- `CART` is more robust in prediction than in inference.
- Prediction seems easier and more useful to us.

MissingBigData

- Missing data is ubiquitus in big data
- Dependence between missingness & effect breaks analysis

$\Rightarrow$ Models that capture dependences

Compensating biases

- Missingness can appear as selection bias: causal literature
- Modeling of missingness to correct causal interpretations
- Inverse probability weighting: prediction problem

Random forests with missing data

- Uncontrolled variance in split criteria biases selections
- Prediction is more robust

3 References I

T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. 2006.

J. Josse, F. Husson, and V. Audigier. Mimca: Multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing*, 27: 501–518, 2016.

K. Mohan and J. Pearl. Graphical models for processing missing data. *arXiv:1801.03583*, 2018.

M. Udell and A. Townsend. Nice latent variable models have log-rank. *arXiv:1705.07474*, 2017.