Learning with missing values: from estimation to prediction

Erwan Scornet
Lecturer at Sorbonne University

October 2025

SCIENCES SORBONNE UNIVERSITÉ

Collaborators

- Alexis Ayme, Post-doc ENS Ulm, Paris. Linear models, Optim.
- Claire Boyer, Professor Paris-Saclay. Signal, Optim.
- Aymeric Dieuleveut, Professor at IPP, Paris. Optim.
- Julie Josse, Senior researcher, INRIA, Montpellier. Causality
- Marine Le Morvan, Junior researcher, INRIA, Paris. Supervised learning
- Christophe Muller, PhD student, Oxford
- Jeffrey Naf, Assist. Professor, GSEM, Geneva. Distributional Prediction
- Angel Reyero Lobo, PhD Student, INRIA, Toulouse. Variable Importance
- Gael Varoquaux, Senior researcher, INRIA, Paris. ML, Scikit-learn









Thanks for the slides too!

Traumabase: an observational French registry[2]

- 40000 trauma patients
- 300 heterogeneous features from pre-hospital and in-hospital settings
- 40 trauma centers, 4000 new patients per year

Center	Accident	Age	Sex	Lactate	Blood Pres.	Shock	Platelet	...
Beaujon	fall	54	m	NM	180	yes	292000	
Pitie	gun	26	m	NA	131	no	323000	
Beaujon	moto	63	m	3.9	NR	yes	318000	
Pitie	moto	30	w	Imp	107	no	211000	
⋮								⋱

⇒ **Explain and Predict** hemorrhagic shock, need for neurosurgery and need for a trauma center given pre-hospital features.

Ex: logistic regression/ random forests + **Quantify uncertainty**[1]

[1] Zaffran, **J.**, Dieuleveut, Romano. Conformal Prediction with Missing Values. *ICML 2023.*

[2] `www.traumabase.eu - https://www.traumatrix.fr/`

Missing values[3], [4], [5]

Missing values are everywhere: unanswered questions in a survey, lost data, damaged plants, machines that fail...

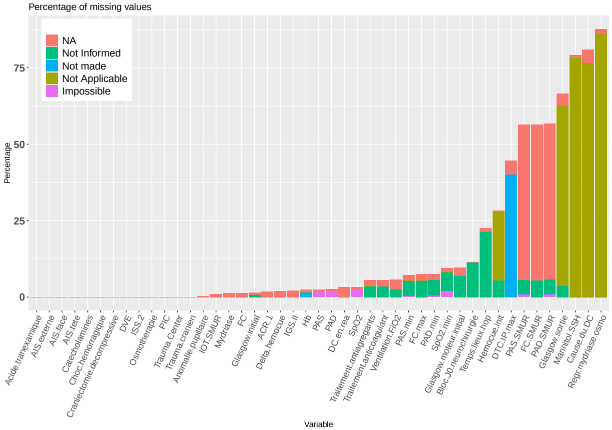
"The best thing to do with missing values is not to have any"

Gertrude Mary Cox (1900-1978)

⇒ Still an issue in the "big data" area (data from different sources)

[3] Little & Rubin (2019). Statistical Analysis with Missing Data, Third Edition, Wiley.
[4] Van Buuren (2018). Flexible Imputation of Data. Second Edition, Chapman & Hall.
[5] Schafer (1997). Analysis of Incomplete Multivariate Data, Chapman & Hall.

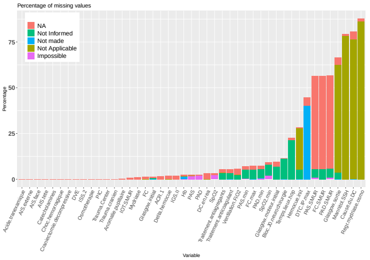
Missing data: important bottleneck in statistical practice



Different types of missing values

- ▷ Not informed: not recorded
- ▷ Not made: possibly due to patient status
- ▷ Not applicable: not supposed to be measured
- ▷ Impossible
- ▷ NA: unkown

Missing data: important bottleneck in statistical practice

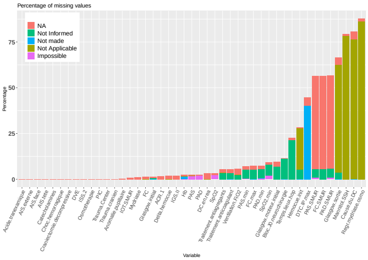


"One of the ironies of Big Data is that missing data play an ever more significant role"[6]

---

[6] Zhu, Wang, Samworth. High-dimensional PCA with heterogeneous missingness. *JRSSB*. 2022.

Missing data: important bottleneck in statistical practice



"One of the ironies of Big Data is that missing data play an ever more significant role"[6]

Complete case analysis: delete incomplete samples

- **Bias**: Resulting sample not representative of the target population
- **Information loss**: Take a matrix with d features where each entry is missing with probability $1/100$, remove a row (of length d) when one entry is missing

$$d = 5 \quad \Longrightarrow \quad \approx 95\% \text{ of rows kept}$$
$$d = 300 \quad \Longrightarrow \quad \approx 5\% \text{ of rows kept}$$

[6] Zhu, Wang, Samworth. High-dimensional PCA with heterogeneous missingness. *JRSSB*. 2022.

Linear models

Linear model

$$Y = X^T \beta^\star + \text{noise}$$

- ▷ $Y \in \mathbb{R}$ (regression) outcome is always observed
- ▷ $X \in \mathbb{R}^d$ contains missing values!

Linear models

Linear model

$$Y = X^T \beta^\star + \text{noise}$$

- ▷ $Y \in \mathbb{R}$ (regression) outcome is always observed
- ▷ $X \in \mathbb{R}^d$ contains missing values!

Three different tasks: imputation, estimation, prediction.

1. **Imputation** - Replace missing values to obtain a complete data set, on which any classical analysis can be performed.
2. **Estimation** - Provide an estimate of $\beta^\star$ - allows predicting outputs of complete data.

Linear models

Linear model

$$Y = X^T \beta^\star + \text{noise}$$

- ▷ $Y \in \mathbb{R}$ (regression) outcome is always observed
- ▷ $X \in \mathbb{R}^d$ contains missing values!

Three different tasks: imputation, estimation, prediction.

1. **Imputation** - Replace missing values to obtain a complete data set, on which any classical analysis can be performed.
2. **Estimation** - Provide an estimate of $\beta^\star$ - allows predicting outputs of complete data.
3. **Prediction** - Predict Y for a new X with missing entries

Warning: A good estimate of $\beta^\star$ does not lead to a prediction of Y

$$X = (\mathrm{na}, 5, \mathrm{na}, -6) \qquad X^\top \beta^\star = ??$$

Solutions to handle missing values in the covariates

Abundant literature: Creation of Rmistatic platform[7] (> 150 packages)

- **Imputation**: (Single/Multiple) imputation to get a/several complete data set(s). Ex: (M)ICE
- **Estimation**: Modify the estimation process to deal with missing values - Maximum likelihood inference: Expectation Maximization algorithms[8]
- **Prediction**: Predict an outcome with missing data in covariates[9][10]. Solutions: using deterministic (e.g. constant) imputation or Missing Incorporated in Attributes for trees based methods (`grf package`)

[7] Mayer, **J.** et al. A unified platform for missing values methods and workflows. *R journal*. 2022.

[8] Jiang, **J.** et al. Logistic Regression with Missing Covariates *CSDA*. 2019. - `misaem package`

[9] **J.** et al. Consistency of supervised learning with missing values. *Stats papers*. 2018-2024.

[10] Le morvan, **J.** et al. What's a good imputation to predict with missing values? *Neurips2021*.

Outline

Summary

Missing values mechanism: Rubin's taxonomy[11,12]

- Random Variables:
  - ▷ $X^\star \in \mathbb{R}^d$: complete unavailable data, $X \in \mathbb{R}^d$: observed data with NA
  - ▷ $M \in \{0,1\}^d$: missing pattern, or mask, $M_j = 1$ if and only if X_j is missing
- Realizations: For a pattern m, $o(x,m) = (x_j)_{j\in\{1,\dots,d\}:m_j=0}$ the observed elements of x and while $o^c(x,m) = (x_j)_{j\in\{1,\dots,d\}:m_j=1}$, the missing elements.

$$
\begin{aligned}
x^\star &= (1,2,3,8,5)\\
x &= (1,\mathrm{NA},3,8,\mathrm{NA})\\
m &= (0,1,0,0,1)\\
o(x,m) &= (1,3,8), \qquad o^c(x^\star,m) = (2,5)
\end{aligned}
$$

[11] Rubin. Inference and missing data. *Biometrika*. 1976.

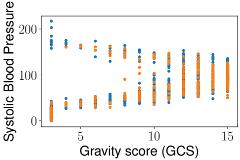
[12] What Is Meant by "Missing at Random"? Seaman, et al. *Statistical Science*. 2013.

Missing values mechanism: Rubin's taxonomy[11],[12]

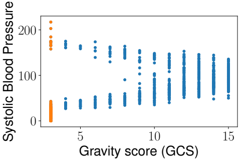
- Random Variables:
  - ▷ $X^\star \in \mathbb{R}^d$: complete unavailable data, $X \in \mathbb{R}^d$: observed data with NA
  - ▷ $M \in \{0,1\}^d$: missing pattern, or mask, $M_j = 1$ if and only if X_j is missing

For a pattern m, $o(x,m) = (x_j)_{j \in \{1,\ldots,d\}: m_j = 0}$ the observed elements of x and while $o^c(x,m) = (x_j)_{j \in \{1,\ldots,d\}: m_j = 1}$, the missing elements.

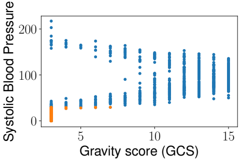
Ex: Simulated missing values according to the 3 mechanisms (Orange points will be missing) in Systolic Blood Pressure - GCS is always observed



Missing Completely at Random (MCAR)
$m \in \mathcal{M}, x \in \mathcal{X}$,
$\mathbb{P}(M = m | x) = \mathbb{P}(M = m)$



Missing at Random (MAR)
$\forall m \in \mathcal{M}, x \in \mathcal{X}$
$\mathbb{P}(M = m | x)$
$= \mathbb{P}(M = m | o(x,m))$



Missing Not At Random (MNAR)
If not MAR: it is MNAR

[11] Rubin. Inference and missing data. *Biometrika*. 1976.

[12] What Is Meant by "Missing at Random"? Seaman, et al. *Statistical Science*. 2013.

Two views to model the joint distribution of (X, M)

▷ Selection Model[13]: $p^*(M = m, x) = \mathbb{P}(M = m \mid x)p^*(x)$

Definition: SM-MAR

$$\mathbb{P}(M = m|x) = \mathbb{P}(M = m|o(x, m)) \text{ for all } m \in \mathcal{M}, x \in \mathcal{X}.$$

The proba. of any m occurring only depends on the obs part of x.

▷ Pattern Mixture Model[14]: $p^*(M = m, x) = p^*(x \mid M = m)\mathbb{P}(M = m)$

Definition: PMM-MAR

$$p^*(o^c(x, m) \mid o(x, m), M = m) = p^*(o^c(x, m) \mid o(x, m)).$$

for all $m \in \mathcal{M}, x \in \mathcal{X}$. The conditional distrib. of missing given obs. in pattern m is equal to the unconditional one.[a]

[a] Molenberghs et al. Every MNAR model has a MAR counterpart with equal fit. *JRSSB*. 2008

- Proposition: SM-MAR is equivalent to PMM-MAR

[13] Heckman. Sample selection bias as a specification error. *Econometrica*. 1979

[14] Little. Pattern-mixture models for multivariate incomplete data. *JASA*. 1993

Testing the missing values mechanism

- Can we observe the missing value mechanism from the sample?

Unfortunately, the general answer is no

[15] Little. *A Test of Missing Completely at Random for Multivariate Data with Missing Values*. 1988

[16] Michel, **Naf**, Spohn, Meinshausen. PKLM: a flexible MCAR test using classification, Psychometrika. 2025

[17] Berrett, Samworth. *Optimal nonparametric testing of missing completely at random and its connections to compatibility*, AoS. 2023

Testing the missing values mechanism

- Can we observe the missing value mechanism from the sample?

Unfortunately, the general answer is no

MCAR vs MAR in Gaussian setting

- If we assume MAR is true we can test H_0 : MCAR vs H_A : MAR.
- A classical test is the Little test[15] that operates under the assumption of Gaussianity.

[15] Little. *A Test of Missing Completely at Random for Multivariate Data with Missing Values*. 1988

[16] Michel, **Naf**, Spohn, Meinshausen. PKLM: a flexible MCAR test using classification, Psychometrika. 2025

[17] Berrett, Samworth. *Optimal nonparametric testing of missing completely at random and its connections to compatibility*, AoS. 2023

Testing the missing values mechanism

- Can we observe the missing value mechanism from the sample?

Unfortunately, the general answer is no

MCAR vs MAR in Gaussian setting

- If we assume MAR is true we can test H_0 : MCAR vs H_A : MAR.
- A classical test is the Little test[15] that operates under the assumption of Gaussianity.

Nonparametric tests

- One of the very few (if not only) useable nonparametric test is our PKLMTest[16]
- There is also interesting theoretical work[17]

[15] Little. *A Test of Missing Completely at Random for Multivariate Data with Missing Values*. 1988

[16] Michel, **Naf**, Spohn, Meinshausen. PKLM: a flexible MCAR test using classification, Psychometrika. 2025

[17] Berrett, Samworth. *Optimal nonparametric testing of missing completely at random and its connections to compatibility*, AoS. 2023

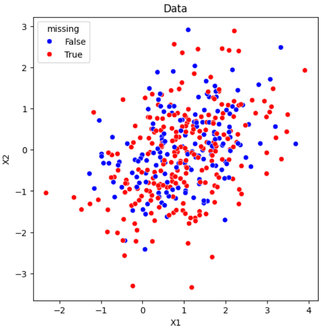
Summary

Single imputation[18]

Generative setting

- $(X_1, X_2) \sim \mathcal{N}((\mu_{x_1}, \mu_{x_2}), \Sigma)$; $n = 400$
- $(\mu_{x_1}, \mu_{x_2}) = (1, 0)$ and $\Sigma = ((1, 0.3), (0.3, 1))$
- MCAR missing values on X_2 only with probability $p = 0.6$.

Discard incomplete observations and then estimate parameters



$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.3$

$\hat{\mu}_{x_2} = 0.043$
$\hat{\sigma}_{x_2} = 0.926$
$\hat{\rho} = 0.368$

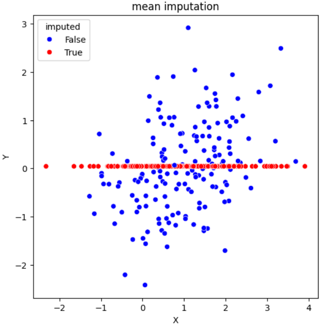
[18]The code to reproduce the plots is available in Rmistastic

Single imputation

Generative setting

- ▷ $(X_1, X_2) \sim \mathcal{N}((\mu_{x_1}, \mu_{x_2}), \Sigma)$; $n = 400$
- ▷ $(\mu_{x_1}, \mu_{x_2}) = (1, 0)$ and $\Sigma = ((1, 0.3), (0.3, 1))$
- ▷ MCAR missing values on X_2 only with probability $p = 0.6$.

Impute by the mean and then estimate parameters



$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.3$

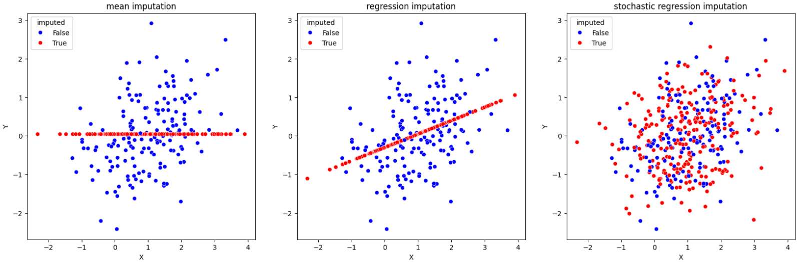
$\hat{\mu}_{x_2} = 0.043$
$\hat{\sigma}_{x_2} = 0.586$
$\hat{\rho} = 0.227$

Mean imputation deforms joint and marginal distributions

Objective: to impute while preserving distribution

Assuming a bivariate gaussian distribution $x_{i2} = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

- ▷ Regression imputation: Estimate β (here with complete data) and impute $\hat{x}_{i2} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} \Rightarrow$ variance underestimated and correlation overestimated
- ▷ Stochastic reg. imputation: Estimate β and σ - impute from the predictive $\hat{x}_{i2} \sim \mathcal{N}\left(\beta_0 + \hat{\beta}_1 x_{i1}, \hat{\sigma}^2\right) \Rightarrow$ preserve distributions



$\mu_{x_2} = 0$
$\sigma_{x_2} = 1$
$\rho = 0.3$

0.043	0.038	0.037
0.926	0.647	0.909
0.368	0.539	0.275

Impute while preserving distribution. Multivariate case

- ▷ Assuming a joint distribution
  - ◇ Gaussian model $x_i \sim \mathcal{N}(\mu, \Sigma)$
  - ◇ Low rank : $X_{n \times d} = \mu_{n \times d} + \varepsilon \; \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with μ of low rank
    - $\Rightarrow$ Different regularization depending on noise regime [18]
    - $\Rightarrow$ Count data[19], ordinal data, categorical data, blocks/multilevel data
  - ◇ Optimal transport [20], deep generative models: GAIN[21], MIWAE [22], etc. [23] [24]
- ▷ Iterating conditional models (joint distribution implicitly defined)
  - ◇ with parametric regression (M)ICE: (Multiple) Imput. by Chained Equations [25]
  - ◇ iterative imputation of each variable by random forests [26]

---

[18] **J.** & Wager. Stable autoencoding for regularized low-rank matrix estimation. *JMLR*. 2016.
[19] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects & interac. in mixed data. *JASA*. 2019.
[20] Muzelec, Cuturi, Boyer, **J.** Missing Data Imputation using Optimal Transport. *ICML*. 2020.
[21] Yoon et al. GAIN: Missing data imputation using generative adversarial nets. *ICML*. 2018.
[22] Mattei & Frellsen. Miwae: Deep generative model & imput. of inc. data. *ICML*. 2018.
[23] Deng et al. Extended missing data imput. via gans. *DMKD*. 2022.
[24] Fang, Bao. Fragmgan: gan for fragmentary data imputation. *STRF* 2023.
[25] van Buuren, S. Flexible Imputation of Missing Data. Chapman & Hall/CRC Press. 2018.
[26] Stekhoven, Bühlmann. MissForest–non-parametric imputation for mixed data. *Bioinfo. 2012.*

Imputation by Chained Equations (ICE)

Init.

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F

Imputation by Chained Equations (ICE)

Impute via mean/mode

Init.

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F

Age	Inc.	Gen.
34	13	F
18	12	F
26	14	M
26	13	F

Imputation by Chained Equations (ICE)

Impute via mean/mode

Init.

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F

Age	Inc.	Gen.
34	13	F
18	12	F
26	14	M
26	13	F

Set values of Age originally missing as unknown

1^{st} step Age

Age	Inc.	Gen.
34	13	F
18	12	F
?	14	M
?	13	F

Imputation by Chained Equations (ICE)

Impute via mean/mode

Init.

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F

Age	Inc.	Gen.
34	13	F
18	12	F
26	14	M
26	13	F

Set values of Age originally missing as unknown

Fit a predictive model on complete observation to predict Age

1^{st} step Age

Age	Inc.	Gen.
34	13	F
18	12	F
?	14	M
?	13	F

Age	Inc.	Gen.
34	13	F
18	12	F
?	14	M
?	13	F

Imputation by Chained Equations (ICE)

Impute via mean/mode

Init.

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F

Age	Inc.	Gen.
34	13	F
18	12	F
26	14	M
26	13	F

Set values of Age originally missing as unknown

Fit a predictive model on complete observation to predict Age

Use the fitted model to impute ?

1^{st} step Age

Age	Inc.	Gen.
34	13	F
18	12	F
?	14	M
?	13	F

Age	Inc.	Gen.
34	13	F
18	12	F
?	14	M
?	13	F

Age	Inc.	Gen.
34	13	F
18	12	F
50	14	M
34	13	F

Imputation by Chained Equations (ICE)

Impute via mean/mode

Init.

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F

Age	Inc.	Gen.
34	13	F
18	12	F
26	14	M
26	13	F

Set values of Inc.
originally missing
as unknown

'Inc.' step

Age	Inc.	Gen.
34	?	F
18	12	F
50	14	M
34	?	F

Imputation by Chained Equations (ICE)

Impute via mean/mode

Init.

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F

Age	Inc.	Gen.
34	13	F
18	12	F
26	14	M
26	13	F

Set values of Inc. originally missing as unknown

Fit a predictive model on complete observation to predict Inc.

'Inc.' step

Age	Inc.	Gen.
34	?	F
18	12	F
50	14	M
34	?	F

Age	Inc.	Gen.
34	?	F
18	12	F
50	14	M
34	?	F

Imputation by Chained Equations (ICE)

Impute via mean/mode

Init.

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F

Age	Inc.	Gen.
34	13	F
18	12	F
26	14	M
26	13	F

Set values of Inc. originally missing as unknown

Fit a predictive model on complete observation to predict Inc.

Use the fitted model to impute ?

'Inc.' step

Age	Inc.	Gen.
34	?	F
18	12	F
50	14	M
34	?	F

Age	Inc.	Gen.
34	?	F
18	12	F
50	14	M
34	?	F

Age	Inc.	Gen.
34	12	F
18	12	F
50	14	M
34	12	F

Imputation by Chained Equations (ICE)

Impute via mean/mode

Init.

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F

Age	Inc.	Gen.
34	13	F
18	12	F
26	14	M
26	13	F

Set values of Gen.
originally missing
as unknown

'Gen.' step

Age	Inc.	Gen.
34	12	F
18	12	?
50	14	M
34	12	F

Imputation by Chained Equations (ICE)

Impute via mean/mode

Init.

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F

Age	Inc.	Gen.
34	13	F
18	12	F
26	14	M
26	13	F

Set values of Gen. originally missing as unknown

Fit a predictive model on complete observation to predict Gen.

'Gen.' step

Age	Inc.	Gen.
34	12	F
18	12	?
50	14	M
34	12	F

Age	Inc.	Gen.
34	12	F
18	12	?
50	14	M
34	12	F

Imputation by Chained Equations (ICE)

Impute via mean/mode

Init.

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F

Age	Inc.	Gen.
34	13	F
18	12	F
26	14	M
26	13	F

Set values of Gen. originally missing as unknown

Fit a predictive model on complete observation to predict Gen.

Use the fitted model to impute ?

'Gen.' step

Age	Inc.	Gen.
34	12	F
18	12	?
50	14	M
34	12	F

Age	Inc.	Gen.
34	12	F
18	12	?
50	14	M
34	12	F

Age	Inc.	Gen.
34	12	F
18	12	F
50	14	M
34	12	F

Imputation by Chained Equations (ICE)

Init.

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F

Age	Inc.	Gen.
34	13	F
18	12	F
26	14	M
26	13	F

'Age' step

Age	Inc.	Gen.
34	13	F
18	12	F
?	14	M
?	13	F

Age	Inc.	Gen.
34	13	F
18	12	F
?	14	M
?	13	F

Age	Inc.	Gen.
34	13	F
18	12	F
50	14	M
34	13	F

'Inc.' step

Age	Inc.	Gen.
34	?	F
18	12	F
50	14	M
34	?	F

Age	Inc.	Gen.
34	?	F
18	12	F
50	14	M
34	?	F

Age	Inc.	Gen.
34	12	F
18	12	F
50	14	M
34	12	F

'Gen.' step

Age	Inc.	Gen.
34	12	F
18	12	?
50	14	M
34	12	F

Age	Inc.	Gen.
34	12	F
18	12	?
50	14	M
34	12	F

Age	Inc.	Gen.
34	12	F
18	12	F
50	14	M
34	12	F

Hyperparameters - R implementation

- Initialization
- Number of cycles
- Ordering of variables: same order, random order...

Hyperparameters - R implementation

- Initialization
- Number of cycles
- Ordering of variables: same order, random order...
- Predictive models
  - Predictive mean matching (numeric data)[27]
  - Logistic regression imputation (binary data)[28]
  - Multinomial regression imputation (unordered categorical data)
  - Proportional odds model (ordered categorical data) [29]

[27] https://stefvanbuuren.name/fimd/sec-pmm.html

[28] https://www.rdocumentation.org/packages/mice/versions/3.17.0/topics/mice.impute.logreg

[29] https://online.stat.psu.edu/stat504/lesson/8/8.4

Hyperparameters - R implementation

- Initialization
- Number of cycles
- Ordering of variables: same order, random order...
- Predictive models
  - Predictive mean matching (numeric data)
  - Logistic regression imputation (binary data)
  - Multinomial regression imputation (unordered categorical data)
  - Proportional odds model (ordered categorical data)

Logistic regression imputation - Bayesian logistic regression

- Fit a logistic model on the data
- Construct $\hat{\beta}$ and an estimation of its covariance matrix $\hat{\Sigma}$.
- Draw $\tilde{\beta} \sim \mathcal{N}(\hat{\beta}, \hat{\Sigma})$.
- Compute the predicted score as $\sigma(X^\top \tilde{\beta})$.
- Impute by drawing a Bernoulli with parameter $\sigma(X^\top \tilde{\beta})$.

Hyperparameters - R implementation

- Initialization
- Number of cycles
- Ordering of variables: same order, random order...
- Predictive models
  - Predictive mean matching (numeric data)
  - Logistic regression imputation (binary data)
  - Multinomial regression imputation (unordered categorical data)
  - Proportional odds model (ordered categorical data)

Predictive mean matching

- Fit a linear model on the data
- Construct $\hat{\beta}$ and an estimation of its covariance matrix $\hat{\Sigma}$.
- Draw $\tilde{\beta} \sim \mathcal{N}(\hat{\beta}, \hat{\Sigma})$.
- Compute the predicted scores as $X^\top \tilde{\beta}$.
- Find the $k = 5$ observations for which $X_i^\top \hat{\beta}$ is the closest to $X^\top \tilde{\beta}$
- Impute by drawing uniformly at random one observations among the k selected observations.

Hyperparameters - R implementation

- Initialization
- Number of cycles
- Ordering of variables: same order, random order...
- Predictive models
  - Predictive mean matching (numeric data)
  - Logistic regression imputation (binary data)
  - Multinomial regression imputation (unordered categorical data)
  - Proportional odds model (ordered categorical data)

Random forests - Mice.RF

- Fit a random forest on the data
- For a given 'missing' observation, put it down each tree and collect all observations in all leaves
- Impute by drawing at random an observation among the previous set

Hyperparameters - R implementation

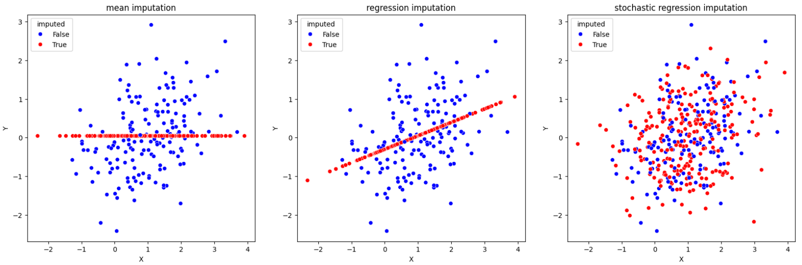
- Initialization
- Number of cycles
- Ordering of variables: same order, random order...
- Predictive models
  - Predictive mean matching (numeric data)
  - Logistic regression imputation (binary data)
  - Multinomial regression imputation (unordered categorical data)
  - Proportional odds model (ordered categorical data)

Random forests - MissForest

- Fit a random forest on the data
- Impute by predicting the value output by the RF

Summary

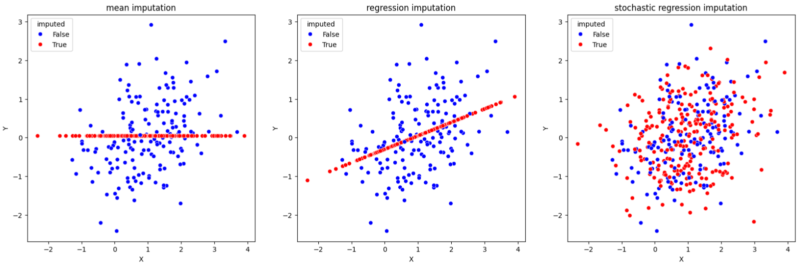
Single imputation methods



$\mu_y = 0$
$\sigma_y = 1$
$\rho = 0.3$

0.043	0.038	0.037
0.586	0.647	0.909
0.227	0.539	0.275

Single imputation methods



$\mu_y = 0$	0.043	0.038	0.037
$\sigma_y = 1$	0.586	0.647	0.909
$\rho = 0.3$	0.227	0.539	0.275

How to build confidence intervals for μ_y?

Confidence interval for a mean

Let $Y = (Y_1, \ldots, Y_n)'$ be i.i.d. independent Gaussian $\mathcal{N}(\mu_y, \sigma_y^2)$.

- Unknown variance:

$$\frac{\hat{\mu}_y - \mu_y}{\hat{\sigma}_{\hat{\mu}_y}} \sim T(n-1)$$

- Unknown variance:

$$\sqrt{n}\left(\frac{\hat{\mu}_y - \mu_y}{\hat{\sigma}_y}\right) \sim T(n-1)$$

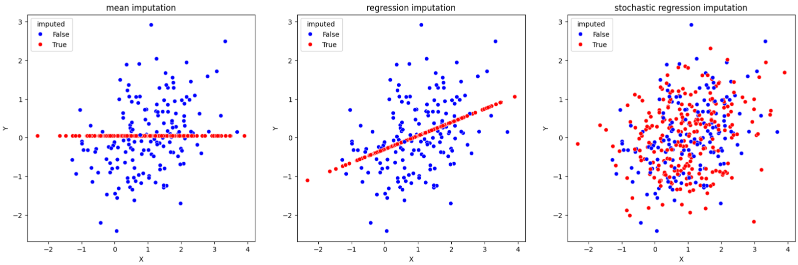
- CI for μ_y at level α: $\left[\hat{\mu}_y - \frac{\widehat{\sigma}_y}{\sqrt{n}} qt_{1-\alpha/2}(n-1)\ ,\ \hat{\mu}_y + \frac{\hat{\sigma}_y}{\sqrt{n}} qt_{1-\alpha/2}(n-1)\right]$

Simulation - Computing coverage

1. Generate bivariate Gaussian data $(\mu_y = 0, \sigma_y = 1, \rho = 0.6)$
2. Put MCAR missing values on y and impute missing entries
3. Compute the confidence interval of μ_y
4. Count if the true value $\mu_y = 0$ is in the confidence interval
5. Repeat the steps 1-4, 10000 times

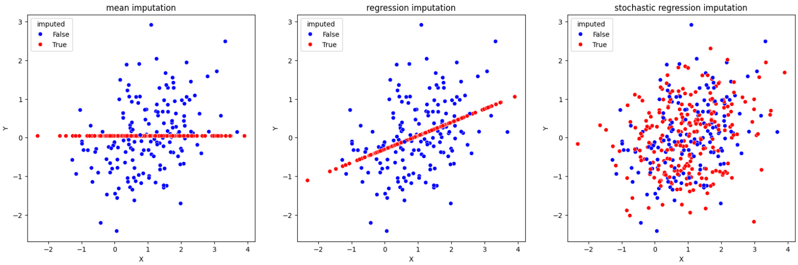
Code available on Rmistatic.

Single imputation methods: Danger!



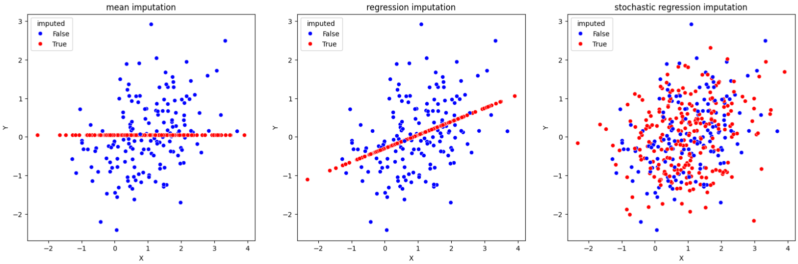
$\mu_y = 0$	-0.005	-0.004	-0.004
$\sigma_y = 1$	0.629	0.673	0.996
$\rho = 0.6$	0.189	0.443	0.301
$CI \mu_y 95\%$			

Single imputation methods: Danger!



	mean imputation	regression imputation	stochastic regression imputation
$\mu_y = 0$	-0.005	-0.004	-0.004
$\sigma_y = 1$	0.629	0.673	0.996
$\rho = 0.6$	0.189	0.443	0.301
$CI\mu_y 95\%$	56.0	57.7	73.4

Single imputation methods: Danger!



$\mu_y = 0$	-0.005	-0.004	-0.004
$\sigma_y = 1$	0.629	0.673	0.996
$\rho = 0.6$	0.189	0.443	0.301
$CI \mu_y 95\%$	56.0	57.7	73.4

$\Rightarrow$ Standard errors $\hat{\sigma}_{\hat{\mu}_y}$ based on the imputed data set are underestimated

The idea of imputation is both seductive and dangerous *(Dempster and Rubin, 1983)*

Asymptotic theory

Asymptotic confidence interval for μ_y: $\left[\hat{\mu}_y - z_{\alpha/2}\frac{\hat{\sigma}_y}{\sqrt{n}}; \hat{\mu}_y - z_{1-\alpha/2}\frac{\hat{\sigma}_y}{\sqrt{n}}\right]$

Consider MCAR values and

- Impute missing values on via (stochastic) linear regression
- $\hat{\mu}_y$ is the average of y computed on the imputed data set

Asymptotic variance (Little & Rubin, 2019. p158)

$$Var[\hat{\mu}_y - \mu_y] \simeq \frac{\hat{\sigma}_y^2}{n_{full}}\left(1 - \hat{\rho}^2\frac{n - n_{full}}{n}\right),$$

where $\hat{\sigma}_y$ is estimated on the complete observations only and n_{full} the number of complete observations.

- If there are few missing data ($n_{full} \sim (n)$), then $Var[\hat{\mu}_y - \mu_y] \sim \hat{\sigma}_y^2/n$, the ACI has the correct asymptotic coverage (Idem if $\rho = 1$).
- But, in general, **coverage of single imputation is too low**: need to take into account the uncertainty associated to the predictions.

Multiple imputation: correct standard errors

1) Generate M plausible values for each missing value

X_1	X_2	X_3	Y
3	20	10	s
-6	45	6	s
0	4	30	no s
-4	32	35	s
1	63	40	s
-2	15	12	no s

X_1	X_2	X_3	Y
-7	20	10	s
-6	45	9	s
0	12	30	no s
13	32	35	s
1	63	40	s
-2	10	12	no s

X_1	X_2	X_3	Y
7	20	10	s
-6	45	12	s
0	-5	30	no s
2	32	35	s
1	63	40	s
-2	20	12	no s

2) Perform the analysis on each imputed data set: $\hat{\beta}_m, \widehat{Var}\left(\hat{\beta}_m\right)$

3) Combine the results (Rubin's rules)[27]:

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

$$T = \underbrace{\frac{1}{M}\sum_{m=1}^{M}\widehat{Var}\left(\hat{\beta}_m\right)}_{\textbf{Within}\text{-imputation variance}} + \underbrace{\left(1+\tfrac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_m - \hat{\beta}\right)^2}_{\textbf{Between}\text{-imputation variance}}$$

[27]see Chapter 14 of Semiparametric Theory and Missing Data. A.A. Tsiatis. 2006.

Multiple imputation: bivariate case[28]

MI based on stochastic regression

1. Generate M imputed data sets: for $m = 1, ..., M$,
   - draw $\hat{y}_i$ from $\mathcal{N}(x_i\hat{\beta}, \hat{\sigma}^2)$
2. Performe the analysis on each imputed data set
3. Compute the variance (= within + between imputation variance)

[28] Code available on Rmistatic.

Multiple imputation: bivariate case[28]

MI based on stochastic regression

1. Generate M imputed data sets: for $m = 1, ..., M$,
   - draw $\hat{y}_i$ from $\mathcal{N}(x_i\hat{\beta}, \hat{\sigma}^2)$
2. Performe the analysis on each imputed data set
3. Compute the variance ($=$ within $+$ between imputation variance)

	$M = 1$	$M = 50$
$\mu_y = 0$	-0.004	-0.004
$\sigma_y = 1$	0.996	0.996
$\rho = 0.3$	0.301	0.301
$CI\mu_y 95\%$	73.4	92

[28] Code available on Rmistatic.

Multiple imputation: bivariate case[28]

MI based on stochastic regression

1. Generate M imputed data sets: for $m = 1, ..., M$,
   - draw $\hat{y}_i$ from $\mathcal{N}(x_i\hat{\beta}, \hat{\sigma}^2)$
2. Performe the analysis on each imputed data set
3. Compute the variance (= within + between imputation variance)

	$M = 1$	$M = 50$
$\mu_y = 0$	-0.004	-0.004
$\sigma_y = 1$	0.996	0.996
$\rho = 0.3$	0.301	0.301
$CI\,\mu_y 95\%$	73.4	92

- Variability of the parameters is missing: "improper" imputation
- Prediction variance = estimation variance plus noise

[28] Code available on Rmistatic.

Multiple imputation: bivariate case[28]

MI based on stochastic regression

1. Generate M imputed data sets: for $m = 1, ..., M$,
   - ▷ Generate $\hat{\beta}^1, ..., \hat{\beta}^M$ by bootstrap or via posterior distribution (Data Augmentation, Tanner & Wong, 1987))
   - ▷ Impute missing values $\hat{y}_i^m$ by drawing $\mathcal{N}(x_i\hat{\beta}^m, (\hat{\sigma}^2)^m)$
2. Performe the analysis on each imputed data set
3. Compute the variance (= within + between imputation variance)

	$M = 1$	$M = 50$	$M = 50$ with boot.
$\mu_y = 0$	-0.004	-0.004	-0.004
$\sigma_y = 1$	0.996	0.996	0.996
$\rho = 0.3$	0.301	0.301	0.301
$CI\mu_y 95\%$	73.4	92	96

[28] Code available on Rmistatic.

Multiple imputation[29]

⇒ Aim: provide an estimation of all parameters with their estimated variance.

Parametric Multiple imputation

1. Generating M imputed data sets, taking into account:
   - ▷ structural noise (e.g. σ^2 via stochastic regression)
   - ▷ parameter variance (e.g. via bootstrapping)
2. Performing the analysis on each imputed data set[a],
3. Compute the variance (= within + between imputation variance)

$$\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m \quad T = \frac{1}{M}\sum \widehat{Var}\left(\hat{\beta}_m\right) + \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum\left(\hat{\beta}_m - \hat{\beta}\right)^2$$

[a] The analysis model may be "in agreement" with the imputation model: congeniality.

[29] Little & Rubin. 2019. Statistical Analysis with Missing Data, 3rd Edition. Wiley

Multiple imputation[29]

⇒ Aim: provide an estimation of all parameters with their estimated variance.

NonParametric Multiple imputation

1. Generating M imputed data sets, taking into account:
    - ▷ structural noise (e.g. σ^2 via stochastic regression)
    - ▷ parameter variance (e.g. via bootstrapping)
2. Performing the analysis on each imputed data set[a],
3. Aggregate the result of each analysis (e.g. taking the mean of predicted output values)

[a] The analysis model may be "in agreement" with the imputation model: congeniality.

[29] Little & Rubin. 2019. Statistical Analysis with Missing Data, 3rd Edition. Wiley

Multiple Imputation with joint modeling

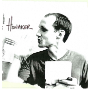
$\Rightarrow$ Hypothesis $x_i \sim \mathcal{N}(\mu, \Sigma)$

Expectation Maximization Bootstrap

1. Bootstrap rows: $X^1, \ldots, X^M$
2. EM algorithm: $(\hat{\mu}^1, \hat{\Sigma}^1), \ldots, (\hat{\mu}^M, \hat{\Sigma}^M)$
3. Imputation: $\hat{x}^m_{i,miss}$ drawn from $\mathcal{N}\left(\hat{\mu}^m_{miss|obs}, \hat{\Sigma}^m_{miss|obs}\right)$

Easy to parallelized. Implemented in `Amelia` (website)

Amelia Earhart



James Honaker

Gary King

Matt Blackwell

Multiple imputation by chained equations or FCS[32]

- Impute variables 1 by 1 using all other variables as inputs (round-robin)
- One model/variable: flexible for different types of variables
- Cycle through variables: iteratively refining imputations

MICE

1. Initial imputation: mean imputation
2. For a variable j
   - $(\hat{\beta}_{-j}, \hat{\sigma}_{-j})$ drawn from a Bootstrap: $(\hat{\beta}^1_{-j}, \hat{\sigma}^1_{-j}), \ldots, (\hat{\beta}^M_{-j}, \hat{\sigma}^M_{-j})$
   - Impute X_j^m via stochastic regression $\mathcal{N}\left((x_{i,-j})'\hat{\beta}^m_{-j}, \hat{\sigma}^m_{-j}\right)$
3. Cycling through variables

$\Rightarrow$ With continuous variables & regression/variable: gibbs $\mathcal{N}(\mu, \Sigma)$ [30] [31]

"*There is no clear-cut method for determining whether MICE has converged*"
Implemented in R package `mice` & IterativeImputer from `scikitlearn` (default iterative ridge regression)

Stef van Buuren

[30] Monte Carlo statistical methods (Robert, Casella, 2004) (p344),
[31] The EM algorithm and extensions (McLachlan, et al. 1998) (p243)
[32] van Buuren. 2018. Flexible Imputation of Missing Data. Second Edition. CRC Press

Joint versus Conditional modeling

Conditional modeling takes the lead?

- ▷ Flexible: one model/variable. Easy to deal with interactions and variables of different nature (binary, ordinal, categorical...)
- ▷ Many statistical models are conditional models
- ▷ Tailor to your data - Super powerful in practice

⇒ Drawbacks: one model/variable. **Computational costly**[a]

[a] Improvement on mice pmm for large sample size, see mice github repo - still costly for large d

What to do with high correlation or when $n < p$

- ▷ JM shrinks the covariance $\Sigma + k\mathbb{I}$ (selection of k?)
- ▷ CM: ridge regression or predictors selection/variable

Challenges with multiple imputation

- ▷ MI in high dimension? Theory with small n, large p?
- ▷ Aggregating lasso regressions? clustering?

Summary

How to evaluate imputation quality?

- Aim: imputed data must resemble complete data.

Original data set

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F
34	NA	M
22	28	F
29	10	NA
34	NA	F
80	NA	NA
68	15	F

How to evaluate imputation quality?

- Aim: imputed data must resemble complete data.

Original data set

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F
34	NA	M
22	28	F
29	10	NA
34	NA	F
80	NA	NA
68	15	F

Imputed data set

Age	Inc.	Gen.
34	13	F
18	12	F
30	14	M
30	13	F
34	13	M
22	28	F
29	10	F
34	13	F
80	13	F
68	15	F

How to evaluate imputation quality?

▷ Aim: imputed data must resemble complete data.

Original data set

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F
34	NA	M
22	28	F
29	10	NA
34	NA	F
80	NA	NA
68	15	F

Imputed data set

Age	Inc.	Gen.
34	13	F
18	12	F
30	14	M
30	13	F
34	13	M
22	28	F
29	10	F
34	13	F
80	13	F
68	15	F

What is the quality of data imputation?

How to evaluate imputation quality?

▷ Aim: imputed data must resemble complete data.

Original data set

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F
34	NA	M
22	28	F
29	10	NA
34	NA	F
80	NA	NA
68	15	F

How to evaluate imputation quality?

▷ Aim: imputed data must resemble complete data.

Original data set

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F
34	NA	M
22	28	F
29	10	NA
34	NA	F
80	NA	NA
68	15	F

How to evaluate imputation quality?

- Aim: imputed data must resemble complete data.

Original data set

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F
34	NA	M
22	28	F
29	10	NA
34	NA	F
80	NA	NA
68	15	F

Additional missing values

Age	Inc.	Gen.
34	NA	F
NA	NA	NA
NA	14	NA
NA	NA	F
34	NA	M
22	NA	F
NA	10	NA
34	NA	F
80	NA	NA
68	NA	NA

How to evaluate imputation quality?

▷ Aim: imputed data must resemble complete data.

Original data set

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F
34	NA	M
22	28	F
29	10	NA
34	NA	F
80	NA	NA
68	15	F

Additional missing values

Age	Inc.	Gen.
34	NA	F
NA	NA	NA
NA	14	NA
NA	NA	F
34	NA	M
22	NA	F
NA	10	NA
34	NA	F
80	NA	NA
68	NA	NA

Imputed missing values

Age	Inc.	Gen.
34	12	F
46	12	F
46	14	F
46	12	F
34	12	M
22	12	F
46	10	F
34	12	F
80	12	F
68	12	F

How to evaluate imputation quality?

- Aim: imputed data must resemble complete data.

Original data set

Age	Inc.	Gen.
34	NA	F
18	12	NA
NA	14	M
NA	NA	F
34	NA	M
22	28	F
29	10	NA
34	NA	F
80	NA	NA
68	15	F

Additional missing values

Age	Inc.	Gen.
34	NA	F
NA	NA	NA
NA	14	NA
NA	NA	F
34	NA	M
22	NA	F
NA	10	NA
34	NA	F
80	NA	NA
68	NA	NA

Imputed missing values

Age	Inc.	Gen.
34	12	F
46	12	F
46	14	F
46	12	F
34	12	M
22	12	F
46	10	F
34	12	F
80	12	F
68	12	F

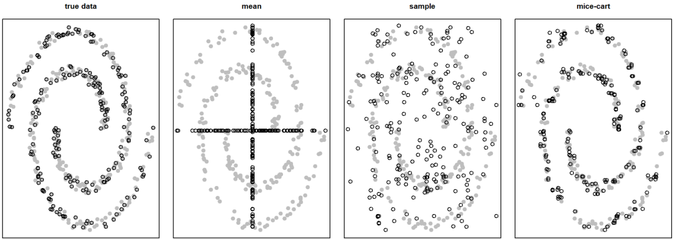
Compared initial vs imputed values via predictive metrics (MSE, MAE...)

Measures related to imputation quality

Pointwise predictive measure such as MSE rank highest imputation close to the conditional expectation

▷ Favor imputation with small variability

Imputation is a distributional task so one should use distributional measures[33][34] to assess its quality.



[33] Székely & Rizzo. Energy statistics *Journal of stat. planning & inference*. 2013

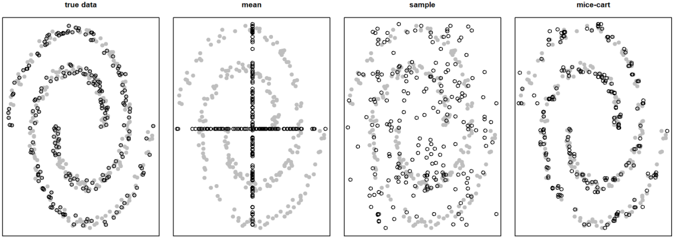
[34] Gneiting, Raftery, Strictly Proper Scoring Rules, Prediction, and Estimation, *JASA*, 2007

Measures related to imputation quality

Pointwise predictive measure such as MSE rank highest imputation close to the conditional expectation

▷ Favor imputation with small variability

Imputation is a distributional task so one should use distributional measures[33][34] to assess its quality.



Imputation method	Mean	Sample	Mice-CART
Renormalized RMSE	0	-0.18	-0.22

[33] Székely & Rizzo. Energy statistics *Journal of stat. planning & inference*. 2013
[34] Gneiting, Raftery, Strictly Proper Scoring Rules, Prediction, and Estimation, *JASA*, 2007

Distributional measures

- Energy score (distribution vs a point)

$$es(H, x) = \frac{1}{2}\mathbb{E}_{X,X' \sim H}[\|X - X'\|_{\mathbb{R}^d}] - \mathbb{E}_{X \sim H}[\|X - x\|_{\mathbb{R}^d}]$$

- The **energy** score can be used to score **distributional prediction/imputation**

Controlled simulation setting

- Generate complete data
- Mask some data according to MCAR/MAR/MNAR mechanism
- Learn a distributional imputation method H
- For any $x \in \mathbb{R}^d$, sample imputed values from H to estimate $es(H, x)$
- Average over $X \sim P^*$ (complete data distribution) to estimate

$$S(H, P^*) := \mathbb{E}_{Y \sim P^*}[es(H, Y)]$$

- The question of how to evaluate imputation methods becomes much harder when the **true underlying values are not available**.

A new procedure

- ▷ Consider a distribution κ on the subsets of $\{1, \ldots, d\}$
- ▷ For each $A \subset \{1, \ldots, d\}$, we let P_A^M be the marginal distribution of M on A. We denote $M_A \sim P_A^M$.
- ▷ We also let $H_A | M_A = m_A$, i.e. the distribution of an imputation H, given the missingness pattern m_A on the projection A.

A new procedure

- Consider a distribution κ on the subsets of $\{1, \ldots, d\}$
- For each $A \subset \{1, \ldots, d\}$, we let P_A^M be the marginal distribution of M on A. We denote $M_A \sim P_A^M$.
- We also let $H_A | M_A = m_A$, i.e. the distribution of an imputation H, given the missingness pattern m_A on the projection A.

Imputation score of imputation H

$$S_{NA}^{\star}(H, P) = \mathbb{E}_{A \sim \kappa, M_A \sim P_A^M, X_A \sim H_{M_A}} \left[\log \left(\frac{p_A(X_A | M_A = \mathbf{0})}{h_{M_A}(X_A)} \right) \right].$$

A new procedure

Group observations into J groups according to their missing data pattern $M_1, \ldots, M_J$.

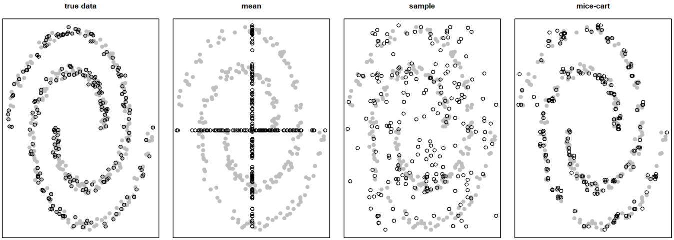
Procedure

For each missing pattern m among $M_1, \ldots, M_J$

1. Choose $\texttt{num.proj}$ projections on $\{1, \ldots, d\}$ such that each projection contains at least one observed and one missing component.
2. Obtain the imputed data from pattern m, denoted by $\hat{X}_i$. Split them into two halves $\hat{X}_i^0$ and $\hat{X}_i^1$
3. For each projection A_k ($k = 1, \ldots, \texttt{num.proj}$),
   a) Get the complete data $X_{A_k}^{comp}$ from the projected data X_{A_k}
   b) Get the projected imputed data $\hat{X}_{i,A_k}^0$
   c) Fit a forest with $\texttt{num.trees.per.proj}$ to discriminate $X_{A_k}^{comp}$ from $\hat{X}_{i,A_k}^0$ (ensuring balanced classes).
4. Aggregate all forests and let $\hat{g}_A(x)$ be the probability output by the forest at x.
5. Compute the individual scores $\log \hat{g}_A(x)$ for $x \in \hat{X}_i^1$
6. Average all scores across all observations, missing patterns and imputed data sets (multiple imputation) to get the final imputation score.

Measures related to imputation quality

Imputation is a distributional task so one should use distributional measures[35][36] to assess its quality.

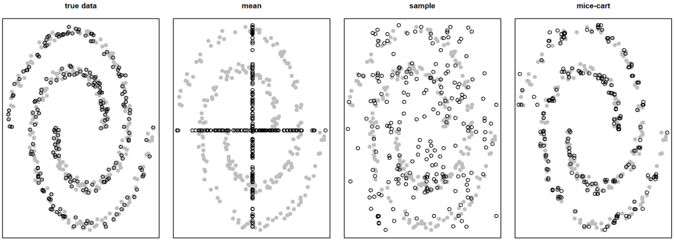


Imputation method	Mean	Sample	Mice-CART
Renormalized RMSE	0	-0.18	-0.22
Renormalized Energy score			

[35] Székely & Rizzo. Energy statistics *Journal of stat. planning & inference*. 2013
[36] Gneiting, Raftery, Strictly Proper Scoring Rules, Prediction, and Estimation, *JASA*, 2007

Measures related to imputation quality

Imputation is a distributional task so one should use distributional measures[35][36] to assess its quality.



Imputation method	Mean	Sample	Mice-CART
Renormalized RMSE	0	-0.18	-0.22
Renormalized Energy score	-22.4	-1.39	0

[35] Székely & Rizzo. Energy statistics *Journal of stat. planning & inference*. 2013

[36] Gneiting, Raftery, Strictly Proper Scoring Rules, Prediction, and Estimation, *JASA*, 2007

Major characteristics of imputations

Imputation should

(1) be a distributional regression method,
(2) be able to capture nonlinearities in the data,
(3) be able to deal with distributional shifts in the observed variables,

- Conditional and marginal **distribution shifts** can occur for different patterns under MAR
- Conditional shifts are handled with FCS

Method	(1)	(2)	(3)
missForest (Stekhoven & Bühlmann, 2011)		✓	
mice-cart (Burgette & Reiter, 2010)	✓	✓	
mice-RF (Doove et al., 2014)	✓	✓	
mice-DRF (Näf et al., 2024)	✓	✓	
mice-norm.nob (Gaussian)	✓		✓
mice-norm.predict (Regression)			✓

MAR with shift in cond. distribution between patterns

- Example: two patterns $m_1 = (0, 0)$ and $m_2 = (1, 0)$, with $\Sigma = ((2, 1), (1, 1)))$ and **a shift**:

$$X \mid M = m_1 \sim N\left((0,0), \Sigma\right)$$
$$X \mid M = m_2 \sim N\left((5,5), \Sigma\right).$$

MAR with shift in cond. distribution between patterns

• Example: two patterns $m_1 = (0,0)$ and $m_2 = (1,0)$, with $\Sigma = ((2,1),(1,1)))$ and **a shift**:

$$X \mid M = m_1 \sim N\left((0,0),\Sigma\right)$$
$$X \mid M = m_2 \sim N\left((5,5),\Sigma\right).$$

• A special case of MAR: conditional distributions are the same across patterns:

$$X_1|X_2, M = m_1 = X_1|X_2, M = m_2.$$

Definition (Conditional indep. MAR - CIMAR)

For all $m, m' \in \mathcal{M}, x \in \mathcal{X}$,

$$p^*(o^c(x,m) \mid o(x,m), M = m') = p^*(o^c(x,m) \mid o(x,m)).$$

MAR with shift in cond. distribution between patterns

• Example: two patterns $m_1 = (0,0)$ and $m_2 = (1,0)$, with $\Sigma = ((2,1),(1,1)))$ and **a shift**:

$$X \mid M = m_1 \sim N\left((0,0), \Sigma\right)$$
$$X \mid M = m_2 \sim N\left((5,5), \Sigma\right).$$

• A special case of MAR: conditional distributions are the same across patterns:

$$X_1|X_2, M = m_1 = X_1|X_2, M = m_2.$$

Definition (Conditional indep. MAR - CIMAR)

For all $m, m' \in \mathcal{M}, x \in \mathcal{X}$,

$$p^*(o^c(x,m) \mid o(x,m), M = m') = p^*(o^c(x,m) \mid o(x,m)).$$

Beware! Even in this case, the joint distribution varies across pattern, since the marginal distribution of X_2 changes

Forests generalize poorly outside of the training set

- Example: two patterns $m_1 = (0, 0)$ and $m_2 = (1, 0)$, with $\Sigma = ((2, 1), (1, 1))$ and **a shift** $X \mid M = m_1 \sim N((0, 0), \Sigma))$, $X \mid M = m_2 \sim N((5, 5), \Sigma))$.

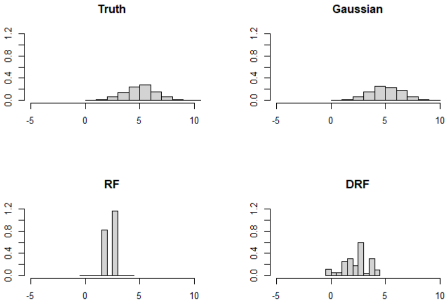


Figure: True distribution against a draw from different imputation methods.

DRF, a distributional method, fails to deal with covariate shift

▷ Imputation should be centered around 5.

MAR with shifts in cond. distribution between patterns

Consider $X \in \mathbb{R}^3$ with three different missing patterns:

$$m_1 = (0,0,0), \quad m_2 = (1,0,0) \quad \text{and} \quad m_3 = (1,1,0).$$

MCAR: No change allowed.

For all $m, m' \in \mathcal{M}, x \in \mathcal{X}$, $p^*(x) = p^*(x \mid M = m) = p^*(x \mid M = m')$

CIMAR: No conditional changes allowed

$p^*(x_1, x_2 \mid x_3, M = m_1) = p^*(x_1, x_2 \mid x_3, M = m_2) = p^*(x_1, x_2 \mid x_3, M = m_3) =$ $p^*(x_1, x_2 \mid x_3)$

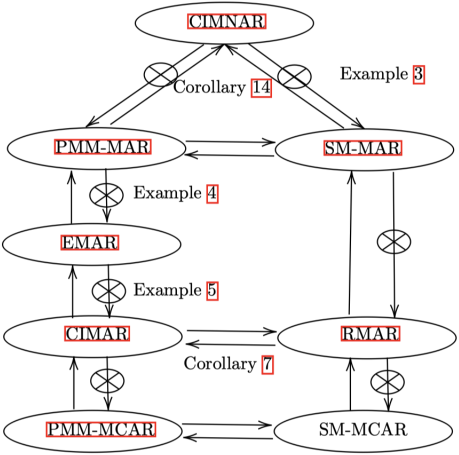
Distrib. of $X_1, X_2 \mid X_3$ is not allowed to change from one pattern to another, though the marginal distrib. of X_3 can change.

PMM-MAR: many changes allowed

$p^*(x_1, x_2 \mid x_3, M = m_3) = p^*(x_1, x_2 \mid x_3)$

Both distrib. of observed variables and conditional ones can change from pattern to pattern.

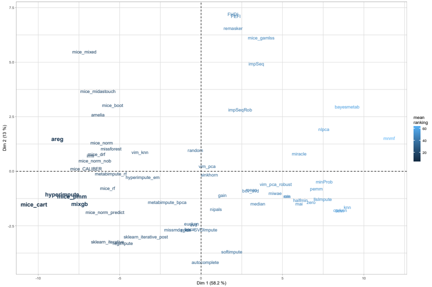
Relationships between the M(N)AR conditions[37]



[37] **Naf**, Scornet **J.**. (2024). What is a good imputation under MAR. *Submitted.*

Benchmarking imputation methods

- ▷ 65 methods (R & Python)
- ▷ 14 datasets: 100-50000 observations and 3-400 features
- ▷ 10-30 % NA MCAR, MAR, Standardized energy distance



- ▷ Mice-cart[38], aregImpute (close to mice+splines+pmm)[39], Hyperimpute (mice + model selection RF, XGBoost, Logistic Reg., etc)[40], Mice mixed[41]

[38] Buuren & Groothuis-O. (2011). Multivariate imputation by chained equations in R. *JSS*.
[39] Harrell & Dupont (2018). Hmisc: Harrell miscellaneous. R package. Stat. Comput.
[40] Jarrett et al. (2022). Hyperimpute: Gen. iter. imput. with automatic model selection. *ICML*.
[41] Varga (2020). missCompare: Intuitive Missing Data Imputation. R package. Stat. Comput.

Take home message on inference & imputation

- Different missing data scenario designed for likelihood inference (e.g. EM algorithm) but that can be very complex (distribution shift in MAR).
- Use single imputation only for point estimates
- In general, look for an imputation that preserve the joint distribution of the data
- Compare imputation methods with distributional metrics like energy distance
- **Multiple imputation** aims at estimating the parameters and their variability taking into account **the uncertainty of the missing values**
- Use Multiple imputation to get confidence intervals
- mice-DRF promising (code available) - mice-Engression[42]

[42] Shen & Meinshausen (2024). Engression: extrapolation through the lens of distributional regression. *JRSS B.*

Summary

Formalizing the problem

- **Assumption** - The response Y is a function of the (unavailable) complete data plus some noise:

$$Y = f^\star(X) + \varepsilon, \quad X \in \mathbb{R}^d,\ Y \in \mathbb{R}.$$

- Optimization problem:

$$\min_{f:(\mathbb{R}\cup\{\mathtt{NA}\})^d \mapsto \mathbb{R}} \mathcal{R}(f) := \mathbb{E}\left[\left(Y - f(\widetilde{X})\right)^2\right]$$

- A Bayes predictor is a minimizer of the risk. It is given by:

$$\tilde{f}^\star(\widetilde{X}) := \mathbb{E}\left[Y | X_{obs(M)}, M\right] = \mathbb{E}\left[f(X) | X_{obs(M)}, M\right]$$

  where $M \in \{0,1\}^d$ is the missingness indicator.

- The Bayes rate $\mathcal{R}^\star$ is the risk of the Bayes predictor: $\mathcal{R}^\star = \mathcal{R}(\tilde{f}^\star)$.
- A Bayes optimal function f achieves the Bayes rate, i.e, $\mathcal{R}(f) = \mathcal{R}^\star$.

Supervised learning with missing values

$\tilde{X} = X \odot (1 - M) + \texttt{NA} \odot M$. New feature space is $\widetilde{\mathbb{R}}^d = (\mathbb{R} \cup \{\texttt{NA}\})^d$.

$$Y = \begin{pmatrix} 4.6 \\ 7.9 \\ 8.3 \\ 4.6 \end{pmatrix} \quad \tilde{X} = \begin{pmatrix} 9.1 & \texttt{NA} & 1 \\ 2.1 & \texttt{NA} & 3 \\ \texttt{NA} & 9.6 & 2 \\ \texttt{NA} & 5.5 & 6 \end{pmatrix} \quad X = \begin{pmatrix} 9.1 & 8.5 & 1 \\ 2.1 & 3.5 & 3 \\ 6.7 & 9.6 & 2 \\ 4.2 & 5.5 & 6 \end{pmatrix} \quad M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Supervised learning with missing values

$\tilde{X} = X \odot (1 - M) + \texttt{NA} \odot M$. New feature space is $\widetilde{\mathbb{R}}^d = (\mathbb{R} \cup \{\texttt{NA}\})^d$.

$$Y = \begin{pmatrix} 4.6 \\ 7.9 \\ 8.3 \\ 4.6 \end{pmatrix} \quad \tilde{X} = \begin{pmatrix} 9.1 & \texttt{NA} & 1 \\ 2.1 & \texttt{NA} & 3 \\ \texttt{NA} & 9.6 & 2 \\ \texttt{NA} & 5.5 & 6 \end{pmatrix} \quad X = \begin{pmatrix} 9.1 & 8.5 & 1 \\ 2.1 & 3.5 & 3 \\ 6.7 & 9.6 & 2 \\ 4.2 & 5.5 & 6 \end{pmatrix} \quad M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Finding the Bayes predictor.

$$f^\star \in \underset{f:\ \widetilde{\mathbb{R}}^d \to \mathbb{R}}{\operatorname{argmin}} \ \mathbb{E}\left[\left(Y - f(\tilde{X})\right)^2\right].$$

$$f^\star(\tilde{X}) = \sum_{m \in \{0,1\}^d} \mathbb{E}\left[Y | X_{obs(m)}, M = m\right] \ \mathbb{1}_{M=m}$$

$\Rightarrow$ One model per pattern (2^d) (Rubin, 1984, generalized propensity score)

Make prediction with missing data great again

Bayes predictor.

$$f^\star(\tilde{X}) = \sum_{m\in\{0,1\}^d} \mathbb{E}\left[Y|X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$$

- ▷ Difficulty due to the half nature of the input space
- ▷ Worst case: 2^d models to learn

Two common strategies:

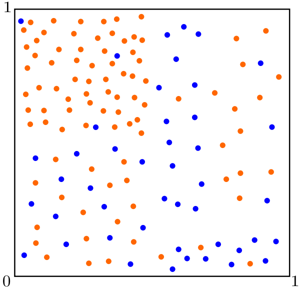
- ▷ **Impute-then-regress strategies** - impute the data then learn on the imputed data set
  - ◇ Computationally efficient but possibly inconsistent
- ▷ **Pattern-by-pattern strategies** - use a different predictor for each missing pattern
  - ◇ Consistent by design but intractable in most situations

CART (Classification And Regression Tree, 1984)

Built by recursively splitting cells until some stopping criterion is satisfied.

Find the feature $j^\star$, the threshold $z^\star$ which minimises the loss

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



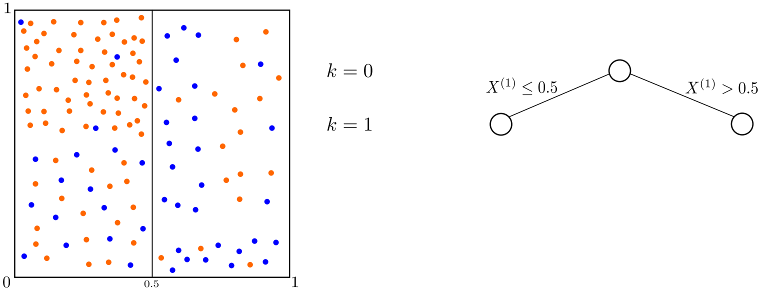
$k = 0$

CART (Classification And Regression Tree, 1984)

Built by recursively splitting cells until some stopping criterion is satisfied.

Find the feature $j^\star$, the threshold $z^\star$ which minimises the loss

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

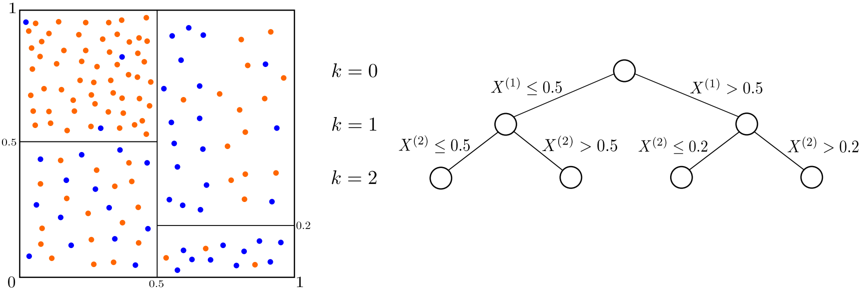


CART (Classification And Regression Tree, 1984)

Built by recursively splitting cells until some stopping criterion is satisfied.

Find the feature $j^\star$, the threshold $z^\star$ which minimises the loss

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

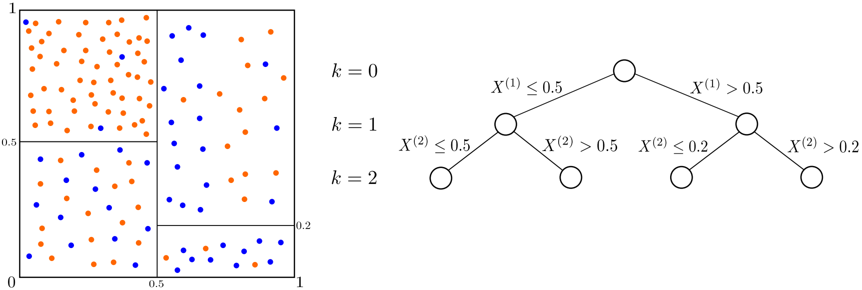


CART (Classification And Regression Tree, 1984)

Built by recursively splitting cells until some stopping criterion is satisfied.

Find the feature $j^\star$, the threshold $z^\star$ which minimises the loss

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



Two difficulties with missing data

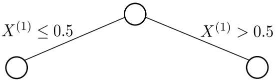
- ▷ How to find the best split?
- ▷ How to propagate missing data down the tree?

CART with missing values

	X_1	X_2	Y
1			
2	NA		
3	NA		
4			

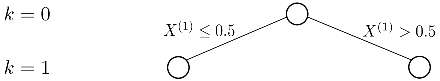
$k = 0$

$k = 1$



CART with missing values

	X_1	X_2	Y
1			
2	NA		
3	NA		
4			



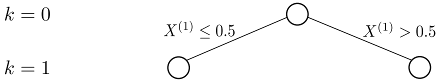
Two steps:

1. For each variable, compute the splitting criterion on observed values only (e.g., 1 & 4 for X_1)

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

CART with missing values

	X_1	X_2	Y
1			
2	NA		
3	NA		
4			



Two steps:

1. For each variable, compute the splitting criterion on observed values only (e.g., 1 & 4 for X_1)

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

2. Propagate observations (2 & 3) with missing values?
   - ▷ Probabilistic split: $\mathcal{B}ernoulli(\#L/(\#L + \#R))$ (`C4.5`)
   - ▷ Block: Send all to a side by minimizing the error (`lightgbm`)
   - ▷ Surrogate split: Search another variable that gives a close partition (rpart)

Missing incorporated in attribute (MIA)[43]

One step: select the variable, the threshold and propagate missing values

1. $\{\widetilde{X}_j \leqslant z \text{ or } \widetilde{X}_j = \texttt{NA}\}$ vs $\{\widetilde{X}_j > z\}$
2. $\{\widetilde{X}_j \leqslant z\}$ vs $\{\widetilde{X}_j > z \text{ or } \widetilde{X}_j = \texttt{NA}\}$
3. $\{\widetilde{X}_j \neq \texttt{NA}\}$ vs $\{\widetilde{X}_j = \texttt{NA}\}$.

- The splitting location z depends on the missing values
- **Missing values treated like a category** (well to handle $\mathbb{R} \cup \texttt{NA}$)
- Good for informative pattern, target one model per pattern:

$$\mathbb{E}\left[Y \middle| \tilde{X}\right] = \sum_{m \in \{0,1\}^d} \mathbb{E}\left[Y | X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$$

- Implementations `grf/partykit package`, `XGBoost`
- Extremely **good performances** in practice **for any mechanism**

[43] Twala et al. (2008). Methods for coping with missing data in decision trees. *Pattern Recog.*

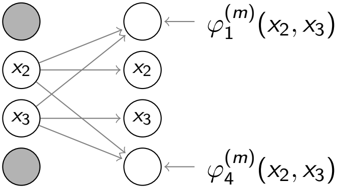
Impute-then-Regress procedures

- Impute-then-Regress procedures consist in
  1. Impute missing values
  2. train a supervised learning algorithm on the imputed data set.

Impute-then-Regress procedures

- Impute-then-Regress procedures consist in
  1. Impute missing values
  2. train a supervised learning algorithm on the imputed data set.
- More formally, define Impute-then-Regress procedures as functions of the form:

$$g \circ \Phi, \text{ where } \Phi \in \mathcal{F}^I, \ g : \mathbb{R}^d \mapsto \mathbb{R}.$$



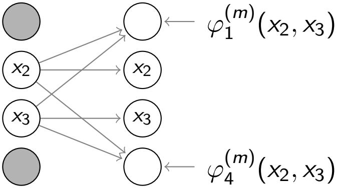
where imputation functions $\Phi \in \mathcal{F}^I$ are of the form:

Impute-then-Regress procedures

- Impute-then-Regress procedures consist in
  1. Impute missing values
  2. train a supervised learning algorithm on the imputed data set.
- More formally, define Impute-then-Regress procedures as functions of the form:

$$g \circ \Phi, \text{ where } \Phi \in \mathcal{F}^I, \ g : \mathbb{R}^d \mapsto \mathbb{R}.$$

where imputation functions $\Phi \in \mathcal{F}^I$ are of the form:



Can Impute-then-Regress procedures be Bayes optimal?

Impute-then-Regress procedures are Bayes optimal

Given an imputation function Φ, we define $g^\star_\Phi$ the minimizer of the population risk on imputed data as

$$g^\star_\Phi \in \underset{g:\mathbb{R}^d \mapsto \mathbb{R}}{\operatorname{argmin}} \quad \mathbb{E}\left[\left(Y - g \circ \Phi(\widetilde{X})\right)^2\right].$$

Impute-then-Regress procedures are Bayes optimal

Given an imputation function Φ, we define $g^\star_\Phi$ the minimizer of the population risk on imputed data as

$$g^\star_\Phi \in \underset{g:\mathbb{R}^d \mapsto \mathbb{R}}{\text{argmin}} \; \mathbb{E}\left[\left(Y - g \circ \Phi(\widetilde{X})\right)^2\right].$$

Theorem (Le Morvan et al., 2021)

Assume that X admits a density, the response Y is generated as $Y = f^\star(X) + \varepsilon$ and $\Phi \in \mathcal{F}^I_\infty$ (C^∞ imputation functions). Then,

- *for all missing data mechanisms,*
- *and for almost all imputation functions,*

$$g^\star_\Phi \circ \Phi \text{ is Bayes optimal.}$$

Impute-then-Regress procedures are Bayes optimal

Given an imputation function Φ, we define $g^\star_\Phi$ the minimizer of the population risk on imputed data as

$$g^\star_\Phi \in \underset{g:\mathbb{R}^d \mapsto \mathbb{R}}{\operatorname{argmin}} \quad \mathbb{E}\left[\left(Y - g \circ \Phi(\widetilde{X})\right)^2\right].$$

Theorem (Le Morvan et al., 2021)

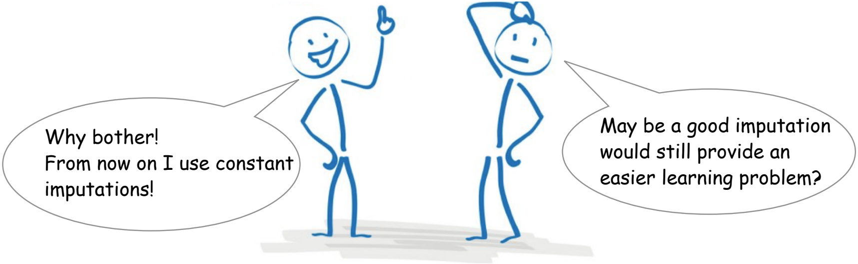
Assume that X admits a density, the response Y is generated as $Y = f^\star(X) + \varepsilon$ and $\Phi \in \mathcal{F}^I_\infty$ (C^∞ imputation functions). Then,

- *for all missing data mechanisms,*
- *and for almost all imputation functions,*

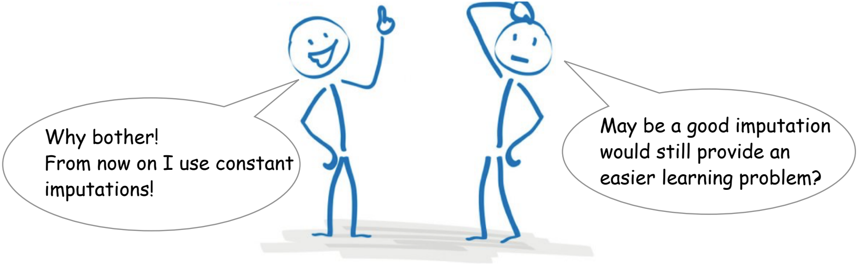
$$g^\star_\Phi \circ \Phi \text{ is Bayes optimal.}$$

For almost all imputation functions, and all missing data mechanisms, a universally consistent algorithm trained on the imputed data is a consistent procedure.

Which imputation function should one choose?



Which imputation function should one choose?



Question *Are there continuous Impute-then-Regress decompositions of Bayes predictors?*

From now on, we suppose $f^\star$ (Byes predictor with complete data) is smooth and consider the conditional expectation Φ^{CI}.

Learning on conditionally imputed data

Question *What can we say about the optimal predictor on the conditionally imputed data:* $g^{\star}_{\Phi^{CI}} \circ \Phi^{CI}$*?*

Learning on conditionally imputed data

Question *What can we say about the optimal predictor on the conditionally imputed data:* $g^{\star}_{\Phi^{CI}} \circ \Phi^{CI}$ *?*

Theorem (Le Morvan et al., 2021)

Suppose that $f^{\star} \circ \Phi^{CI}$ *is not Bayes optimal, and that the probability of observing all variables is strictly positive, i.e.,* $P(M = \mathbf{0}, X = x) > 0$, *for all* x. *Then there is no continuous function* g *such that* $g \circ \Phi^{CI}$ *is Bayes optimal.*

- In the above setting, $g^{\star}_{\Phi^{CI}}$ is not continuous. Thus, imputing via conditional expectation leads to a difficult learning problem.
- Almost all imputations lead to consistent estimators but some ease the training of the supervised learning algorithm.

Imputation-then-regress: does imputation matter?

Adding the mask to the input (one mask per feature):

$$\begin{matrix} X_1 & X_2 \end{matrix} \qquad\qquad \begin{matrix} X_1 & X_2 & M_1 & M_2 \end{matrix}$$

$$\begin{pmatrix} 1 & 2 \\ 3 & \mathbf{NA} \\ \mathbf{NA} & 4 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 0 & 0 \\ 3 & \mathbf{NA} & 0 & \mathbf{1} \\ \mathbf{NA} & 4 & \mathbf{1} & 0 \end{pmatrix}$$

From an empirical study over 19 datasets[44]:

[44] M. Le Morvan, G. Varoquaux, Imp. for pred.: beware of diminish. returns. (ICLR2025)

[45] Mike et al. (2023). The Missing Indicator Method: From Low to High Dimensions. SIGKDD.

Imputation-then-regress: does imputation matter?

Adding the mask to the input (one mask per feature):

$$\begin{pmatrix} X_1 & X_2 \\ 1 & 2 \\ 3 & \mathbf{NA} \\ \mathbf{NA} & 4 \end{pmatrix} \rightarrow \begin{pmatrix} X_1 & X_2 & M_1 & M_2 \\ 1 & 2 & 0 & 0 \\ 3 & \mathbf{NA} & 0 & \mathbf{1} \\ \mathbf{NA} & 4 & \mathbf{1} & 0 \end{pmatrix}$$

From an empirical study over 19 datasets[44]:

- Imputation accuracy matters less when using expressive models or when incorporating the mask as complementary inputs[45]
- Imputation accuracy matters much more for generated linear outcomes than for real-data outcome
- Adding the mask as input is beneficial for prediction performances even for MCAR settings, where missingness is uninformative.

[44] M. Le Morvan, G. Varoquaux, Imp. for pred.: beware of diminish. returns. (ICLR2025)
[45] Mike et al. (2023). The Missing Indicator Method: From Low to High Dimensions. SIGKDD.

Imputation-then-regress: does imputation matter?

Adding the mask to the input (one mask per feature):

$$\begin{array}{cc} X_1 & X_2 \end{array} \qquad \begin{array}{cccc} X_1 & X_2 & M_1 & M_2 \end{array}$$
$$\begin{pmatrix} 1 & 2 \\ 3 & \mathbf{NA} \\ \mathbf{NA} & 4 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 0 & 0 \\ 3 & \mathbf{NA} & 0 & \mathbf{1} \\ \mathbf{NA} & 4 & \mathbf{1} & 0 \end{pmatrix}$$

From an empirical study over 19 datasets[44]:

- Imputation accuracy matters less when using expressive models or when incorporating the mask as complementary inputs[45]
- Imputation accuracy matters much more for generated linear outcomes than for real-data outcome
- Adding the mask as input is beneficial for prediction performances even for MCAR settings, where missingness is uninformative.

Investing in more flexible models is more efficient than investing in more complex imputations.

[44] M. Le Morvan, G. Varoquaux, Imp. for pred.: beware of diminish. returns. (ICLR2025)
[45] Mike et al. (2023). The Missing Indicator Method: From Low to High Dimensions. SIGKDD.

Summary so far

Bayes predictor

$$f^{\star}(\tilde{X}) = \sum_{m \in \{0,1\}^d} \mathbb{E}\left[Y | X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$$

Two common strategies:

- Impute-then-regress strategies - impute the data then learn on the imputed data set
  - Computationally efficient but possibly inconsistent
  - Consistent if used with a non-parametric learning algorithm (forests, tree boosting, nearest neighbor...)
- Pattern-by-pattern strategies - use a different predictor for each missing pattern
  - Consistent by design but intractable in most situations

Summary

Missing data and linear models

Our aim

Predict on new data, which may contain missing entries.

MCAR
(missing completely at random)
$\mathbb{P}(M|X) = \mathbb{P}(M)$

MAR (missing at random)
$\mathbb{P}(M|X) = \mathbb{P}(M|X^{(obs)})$

MNAR (missing not at random)

Linear model

$$Y = X^T \beta^\star + \text{noise}$$

- $Y \in \mathbb{R}$ (regression) outcome is always observed
- $X \in \mathbb{R}^d$ contains missing values!
- $\beta^\star$ model parameter

Linear models do not remain linear

Let

$$Y = X_1 + X_2 + \varepsilon,$$

where $X_2 = \exp(X_1) + \varepsilon_1$. Now, assume that only X_1 is observed. Then, the model can be rewritten as

$$Y = X_1 + \exp(X_1) + \varepsilon + \varepsilon_1,$$

where $f(X_1) = X_1 + \exp(X_1)$ is the Bayes predictor.

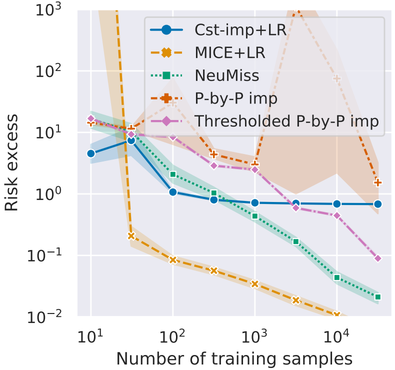
Here, the submodel for which only X_1 is observed is not linear.

$\Rightarrow$ There exists a large variety of submodels for a same linear model.
$\Rightarrow$ Submodel natures depend on the structure of X and on the missing-value mechanism.

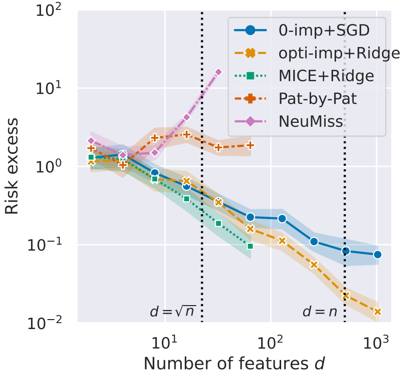
Handling missing values in linear models for prediction

2 possible approaches

- ▷ Patter-by-pattern methods
- ▷ Impute-then-regress procedures

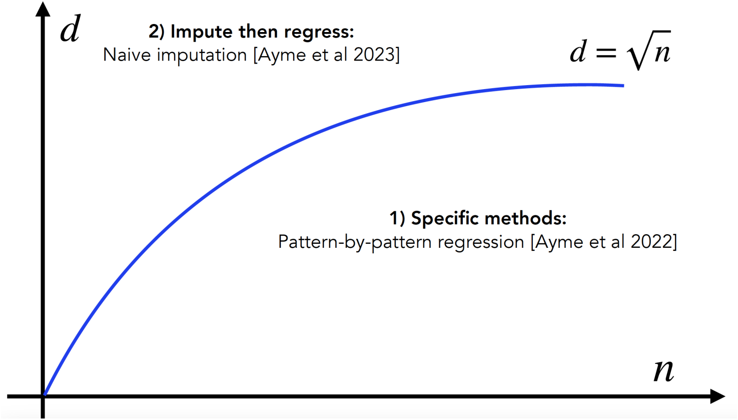


Fixed dimension



Fixed sample size

Different strategies for prediction



Specific methods: formalization

- Dataset $\mathcal{D}_n = \{(Z_i, Y_i), i \in [n]\}$ where

$$Z_i = (X_{obs(M_i)}, M_i).$$

- New test point $Z = (X_{obs(M)}, M)$ (with unknown target Y).

Goal in prediction

Find a linear function $\widehat{f}$ that minimizes the risk

$$R_{\text{miss}}(\widehat{f}) = \mathbb{E}\left[\left(\widehat{f}(Z) - Y\right)^2\right].$$

Pattern-by-pattern Bayes predictor

Consider either

▷ $X \sim \mathcal{N}(\mu, \Sigma)$ — Gaussian (G)

or,

▷ $X|(M = m) \sim \mathcal{N}(\mu^m, \Sigma^m)$ — Gaussian pattern mixture model (GPMM)

Decompose the Bayes predictor

$$f^\star(Z) = \sum_{m \in \mathcal{M}} f^\star_m(X_{obs(m)}) \mathbb{1}_{M=m},$$

with $f^\star_m$ the Bayes predictor conditionally on the event $(M = m)$.

Proposition [Le Morvan et al 2020]

If [(MCAR or MAR) and G] or GPMM then, for all $m \in \mathcal{M}$,

$f^\star_m$ is linear.

A missing-distribution-free upper bound

Predictor $\widehat{f}(Z) = \sum_{m \in \mathcal{M}} \widehat{f}_m(X_{obs(m)}) \mathbb{1}_{M=m}$ (pattern-by-pattern OLS)
where $\widehat{f}_m$ is a modified least-square regression rule trained on

$$\mathcal{D}_m = \left\{(X_{i,obs(m)}, Y_i), M_i = m\right\}.$$

Theorem (simplified) [Le Morvan et al. 2020] [Ayme, Boyer, Dieuleveut, S. 2022]

If [(MCAR or MAR) and G] or GPMM then

$$\mathbb{E}\left[\left(\widehat{f}(Z) - f^{\star}(Z)\right)^2\right] \lesssim \log(n) 2^d \frac{d}{n}$$

where the constant depends on the level of noise.

A missing-distribution-free upper bound

Predictor $\widehat{f}(Z) = \sum_{m \in \mathcal{M}} \widehat{f}_m(X_{obs(m)}) \mathbb{1}_{M=m}$ (pattern-by-pattern OLS)
where $\widehat{f}_m$ is a modified least-square regression rule trained on

$$\mathcal{D}_m = \left\{ (X_{i,obs(m)}, Y_i), M_i = m \right\}.$$

Theorem (simplified) [Le Morvan et al. 2020] [Ayme, Boyer, Dieuleveut, S. 2022]

If [(MCAR or MAR) and G] or GPMM then

$$\mathbb{E}\left[\left(\widehat{f}(Z) - f^\star(Z) \right)^2 \right] \lesssim \log(n) 2^d \frac{d}{n}$$

where the constant depends on the level of noise.

- This result does not depend on the distribution of missing patterns.
- Number of parameters is $p := d2^d$. This result suffers from the curse of dimensionality even with small d.

A missing pattern distribution adaptive bound

Idea: Regression only on high frequency missing patterns

$$\widehat{f}(Z) = \sum_{m \in \mathcal{M}} \widehat{f}_m(X_{obs(m)}) \mathbb{1}_{M=m} \mathbb{1}_{|\mathcal{D}_m| \geqslant d}.$$

A missing pattern distribution adaptive bound

Idea: Regression only on high frequency missing patterns

$$\widehat{f}(Z) = \sum_{m \in \mathcal{M}} \widehat{f}_m(X_{obs(m)}) \mathbb{1}_{M=m} \mathbb{1}_{|\mathcal{D}_m| \geqslant d}.$$

Theorem [Ayme, Boyer, Dieuleveut, S. 2022]

$$\mathbb{E}\left[\left(\widehat{f}(Z) - f^\star(Z)\right)^2\right] \lesssim \log(n)\mathcal{E}_p\left(d/n\right),$$

with $\mathcal{E}_p\left(d/n\right) := \sum_m \min(p_m, d/n)$.

- ▷ Valid for MCAR, MAR and MNAR settings.
- ▷ Adaptive to missing data distribution via $\mathcal{E}_p\left(d/n\right) \leqslant \mathrm{Card}(\mathcal{M})(d/n)$.

Examples

1. Uniform distribution: $\mathcal{E}_p\left(\frac{d}{n}\right) = 2^d d/n$
2. Bernoulli distribution: $M_j \sim \mathcal{B}(\varepsilon)$ with $\varepsilon \leqslant d/n$: $\mathcal{E}_p\left(\frac{d}{n}\right) = d^2/n$

A lower bound

Let $\mathcal{P}_p$ be a class of data distributions $\begin{cases} X|(M=m) \sim \mathcal{N}(\mu^m, \Sigma^m) \\ \text{Linear model} \\ \mathbb{P}[M=m] = p_m \end{cases}$

$$\underset{\text{error}}{\text{Minimax}}(p) = \underbrace{\min_{\tilde{f}}}_{\text{Best algo}} \underbrace{\max_{\mathbb{P}\in\mathcal{P}_p}}_{\text{Worst case on a class } \mathcal{P}_p \text{ of problems}} \mathbb{E}_{\mathbb{P}}\left[(\tilde{f}(Z) - f^\star(Z))^2\right]$$

A lower bound

Let $\mathcal{P}_p$ be a class of data distributions $\begin{cases} X|(M=m) \sim \mathcal{N}(\mu^m, \Sigma^m) \\ \text{Linear model} \\ \mathbb{P}[M=m] = p_m \end{cases}$

$$\underset{\text{error}}{\text{Minimax}}(p) = \underbrace{\min_{\tilde{f}}}_{\text{Best algo}} \quad \underbrace{\max_{\mathbb{P}\in\mathcal{P}_p}}_{\text{Worst case on a class } \mathcal{P}_p \text{ of problems}} \quad \mathbb{E}_{\mathbb{P}}\left[(\tilde{f}(Z) - f^\star(Z))^2\right]$$

Theorem [Ayme, Boyer, Dieuleveut, S. 2022]

$$\sigma^2 \mathcal{E}_p\left(\frac{1}{n}\right) \lesssim \underset{\text{error}}{\text{Minimax}}(p) \leqslant \mathbb{E}\left[\left(\widehat{f}(Z) - f^\star(Z)\right)^2\right] \lesssim \log(n)\mathcal{E}_p\left(\frac{d}{n}\right)$$

A lower bound

Let $\mathcal{P}_p$ be a class of data distributions $\begin{cases} X|(M=m) \sim \mathcal{N}(\mu^m, \Sigma^m) \\ \text{Linear model} \\ \mathbb{P}[M=m] = p_m \end{cases}$

$$\underset{\text{error}}{\text{Minimax}}(p) = \underbrace{\min_{\tilde{f}}}_{\text{Best algo}} \underbrace{\max_{\mathbb{P} \in \mathcal{P}_p}}_{\text{Worst case on a class } \mathcal{P}_p \text{ of problems}} \mathbb{E}_{\mathbb{P}}\left[(\tilde{f}(Z) - f^\star(Z))^2\right]$$

Theorem [Ayme, Boyer, Dieuleveut, S. 2022]

$$\sigma^2 \mathcal{E}_p\left(\frac{1}{n}\right) \lesssim \underset{\text{error}}{\text{Minimax}}(p) \leqslant \mathbb{E}\left[\left(\widehat{f}(Z) - f^\star(Z)\right)^2\right] \lesssim \log(n)\mathcal{E}_p\left(\frac{d}{n}\right)$$

Examples

- Uniform distribution $\quad \mathcal{E}_p\left(\frac{1}{n}\right) = 2^d/n \qquad \mathcal{E}_p\left(\frac{d}{n}\right) = 2^d d/n$
- Bernoulli distribution $M_j \sim \mathcal{B}(\varepsilon)$ with $\varepsilon \leqslant d/n$ $\quad \mathcal{E}_p\left(\frac{1}{n}\right) = d/n \qquad \mathcal{E}_p\left(\frac{d}{n}\right) = d^2/n$

Take-home messages

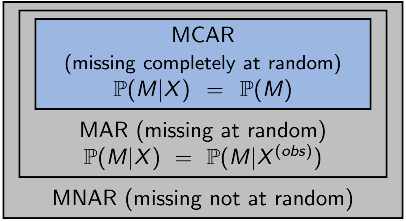
- ☞ For data regimes where n is large, several problems can be learned, even for MNAR.
- ☞ The procedure can be modified to adapt to the distribution of missing patterns.
- ☞ **The dimension is an issue**, even under the classical assumptions (MAR)

Impute-then-regress?

- Impute-then-regress method
  1. Impute the missing values by 0 to get X_{imp} (e.g., via `df.fillna(0)`)
  2. Perform a SGD regression

Impute-then-regress?

- Impute-then-regress method
  1. Impute the missing values by 0 to get X_{imp} (e.g., via `df.fillna(0)`)
  2. Perform a SGD regression
- Focus on MCAR values: $M_1, \ldots, M_d \sim \mathcal{B}(\rho)$
  $\rho =$ probability to be observed



impute by 0= doesn't exploit observed values?

Risk decomposition

- $R^\star$ = optimal risk without missing data
- $R^\star_{\mathrm{miss}}$ = optimal risk with missing data

$$\Delta_{\mathrm{miss}} := R^\star_{\mathrm{miss}} - R^\star \qquad \text{(missing data error)}$$

- $R_{\mathsf{imp}}(\theta)$ = the risk of $f_\theta(X_{\mathrm{obs}}, M) = \theta^\top X_{\mathsf{imp}}$
- $R_{\mathrm{imp}}(\theta^\star_{\mathsf{imp}})$ = optimal risk of linear prediction after imputation by 0

$$\Delta_{\mathrm{imp/miss}} := R_{\mathrm{imp}}(\theta^\star_{\mathsf{imp}}) - R^\star_{\mathrm{miss}} \qquad \text{(imputation error)}$$

- Risk decomposition:

$$R_{\mathrm{miss}}(f_\theta) = R^\star + \underbrace{\Delta_{\mathrm{miss}} + \Delta_{\mathrm{imp/miss}}}_{\text{missing data and imputation error}} + \underbrace{R_{\mathrm{miss}}(f_\theta) - R_{\mathrm{imp}}(\theta^\star_{\mathsf{imp}})}_{\text{estimation/optimization error}}$$

Toy example: how imputed inputs disturb learning

- Complete model
  - $Y = X_1$
  - $X = (X_1, \ldots, X_1)$
  - $R^\star = 0$
  - $M_1, \ldots, M_d \sim \mathcal{B}(1/2)$

Toy example: how imputed inputs disturb learning

- Complete model
  - $Y = X_1$
  - $X = (X_1, \dots, X_1)$
  - $R^\star = 0$
  - $M_1, \dots, M_d \sim \mathcal{B}(1/2)$

- With imputed inputs and $\theta_1 = (1, 0, \dots, 0)^\top$
  - $X_{\text{imp}}^\top \theta_1 = X_1 M_1$
  - $R_{\text{imp}}(\theta_1) = \frac{1}{2}\mathbb{E}\left[Y^2\right]$

- With imputed inputs and $\theta_2 = 2(1/d, 1/d, \dots, 1/d)^\top$
  - $X_{\text{imp}}^\top \theta_2 = \frac{2}{d} X_1 \sum_j M_j$
  - $R_{\text{imp}}(\theta_2) = \frac{1}{d}\mathbb{E}\left[X_1^2\right]$
  - $\Delta_{\text{miss}} + \Delta_{\text{imp/miss}} \leqslant R_{\text{imp}}(\theta_2) - R^\star \leqslant \frac{1}{d}\mathbb{E}\left[Y^2\right]$

Toy example: how imputed inputs disturb learning

- ▷ Complete model
  - ◇ $Y = X_1$
  - ◇ $X = (X_1, \ldots, X_1)$
  - ◇ $R^\star = 0$
  - ◇ $M_1, \ldots, M_d \sim \mathcal{B}(1/2)$

- ▷ With imputed inputs and $\theta_1 = (1, 0, \ldots, 0)^\top$
  - ◇ $X_{\text{imp}}^\top \theta_1 = X_1 M_1$
  - ◇ $R_{\text{imp}}(\theta_1) = \frac{1}{2}\mathbb{E}\left[Y^2\right]$

- ▷ With imputed inputs and $\theta_2 = 2(1/d, 1/d, \ldots, 1/d)^\top$
  - ◇ $X_{\text{imp}}^\top \theta_2 = \frac{2}{d} X_1 \sum_j M_j$
  - ◇ $R_{\text{imp}}(\theta_2) = \frac{1}{d}\mathbb{E}\left[X_1^2\right]$
  - ◇ $\Delta_{\text{miss}} + \Delta_{\text{imp/miss}} \leqslant R_{\text{imp}}(\theta_2) - R^\star \leqslant \frac{1}{d}\mathbb{E}\left[Y^2\right]$

correlation ⇒ low imputation/missing values error ?

Learning w/ imputed-by-0 data = ridge reg?

▷ Ridge-regularized risk with complete data

$$R_\lambda(\theta) = R(\theta) + \lambda\|\theta\|_2^2$$

▷ **Standard in high-dimension settings**

Theorem [Ayme, Boyer, Dieuleveut, S. 2023]

Under the MCAR Bernoulli model of probability ρ of observation and $Var(X_j) = 1\ \forall j$,

$$R_{\text{imp}}(\theta) = R(\rho\theta) + \rho(1-\rho)\|\theta\|_2^2$$

Consequences

1. $\Delta_{\text{miss}} + \Delta_{\text{imp/miss}}$ = ridge bias for $\lambda = \frac{1-\rho}{\rho}$
2. $\theta^\star_{\text{imp}}$ on a small ball around 0 (implicit regularization)

☞ Imputed MCAR missing values seem to be at the same price of ridge regularization

Learning with low-rank and imputed-by-0 data

- **Low-rank data**: covariance matrix $\Sigma = [XX^\top]$ is

$$\Sigma = \sum_{j=1}^{r} \lambda_j v_j v_j^\top,$$

with $\lambda_1 = \cdots = \lambda_r$ and $r \ll d$.

- Bias on low-rank data:

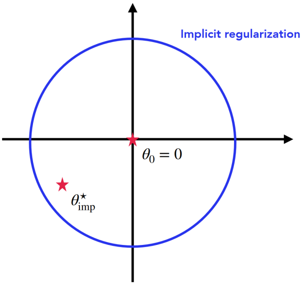
$$\Delta_{\mathrm{miss}} + \Delta_{\mathrm{imp/miss}} \lesssim \frac{1-\rho}{\rho} \frac{r}{d} \mathbb{E}[Y^2]$$

correlation $\Rightarrow$ low imputation/missing values error !

Learning with imputed-by-0 data via SGD

- Averaged SGD iterates:

$$\begin{cases} \theta_{\text{imp},t} & = \left[I - \gamma X_{\text{imp},t} X_{\text{imp},t}^\top\right] \theta_{\text{imp},t-1} + \gamma Y_t X_{\text{imp},t} \\ \bar{\theta}_{\text{imp},n} & = \frac{1}{n+1} \sum_{t=1}^{n} \theta_{\text{imp},t} \end{cases}$$



- Why use SGD ?
  1. Streaming online (one pass only)
  2. Minimizes directly the generalization risk R
  3. Friendly assumptions
  4. Leverage the implicit regularization of naive imputations choosing $\theta_{\text{imp},0} = 0$ and $\gamma = 1/d\sqrt{n}$.

Learning with imputed-by-0 data via SGD

Theorem [Ayme, Boyer, Dieuleveut, S. 2023]

Under classical assumptions for SGD,

$$\mathbb{E}\left[R_{\mathsf{imp}}(\bar{\theta}_{\mathsf{imp},n})\right] - R^{\star} \leqslant \Delta_{\mathrm{miss}} + \Delta_{\mathrm{imp/miss}} + \frac{d}{\sqrt{n}}\|\theta^{\star}_{\mathsf{imp}}\|_2^2 + \frac{\text{noise variance}}{\sqrt{n}}$$

Learning with imputed-by-0 data via SGD

Theorem [Ayme, Boyer, Dieuleveut, S. 2023]

Under classical assumptions for SGD,

$$\mathbb{E}\left[R_{\mathrm{imp}}(\bar{\theta}_{\mathrm{imp},n})\right] - R^\star \leqslant \Delta_{\mathrm{miss}} + \Delta_{\mathrm{imp/miss}} + \frac{d}{\sqrt{n}}\|\theta^\star_{\mathrm{imp}}\|_2^2 + \frac{\text{noise variance}}{\sqrt{n}}$$

▷ Example: low-rank setting

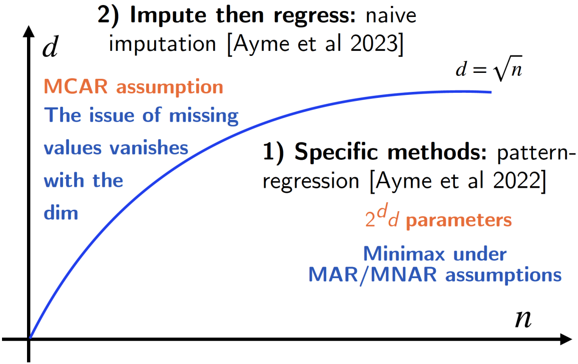
$$\mathbb{E}\left[R_{\mathrm{imp}}(\bar{\theta}_{\mathrm{imp},n})\right] - R^\star \lesssim \left(\frac{1}{\rho\sqrt{n}} + \frac{1-\rho}{d}\right)\frac{r}{d}\mathbb{E}Y^2 + \frac{\text{noise variance}}{\sqrt{n}}$$

▶ Imputation bias vanishes for $d \gg \sqrt{n}$

Naive imputation implicitly regularizes HD linear models

- ▷ MCAR inputs (observation rate=ρ)
- ▷ All in all

Performing standard linear regression on imputed-by-0 data	=	Adding a ridge regularization w/ parameter $\lambda = \frac{\text{1-observation rate}}{\text{observation rate}}$



Linear Discriminant Analysis [46]

LDA

Let $\mathbb{P}(Y = 1) = 0.5$ and $\forall k \in \{-1, 1\}$, $X|Y = k \sim \mathcal{N}(\mu_k, \Sigma)$.

Bayes predictor for the complete case:

$$h^{\star}_{\mathrm{comp}}(x) := \mathrm{sign}\left(\left(\mu_1 - \mu_{-1}\right)^{\top} \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_{-1}}{2}\right)\right).$$

[46] A primer on linear classification with missing data A.D.R. Lobo, A. Ayme, C. Boyer, E. Scornet

Linear Discriminant Analysis [46]

Bayes predictor for the complete case:

$$h^\star_{\mathrm{comp}}(x) := \mathrm{sign}\left((\mu_1 - \mu_{-1})^\top \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_{-1}}{2} \right) \right).$$

Proposition: Bayes predictor for LDA+MCAR

Assume LDA + MCAR. Then the PbP Bayes classifier satisfies

$$h^\star_m(x_{\mathrm{obs}(m)}) = \mathrm{sign}\Big(\big(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\big)^\top \Sigma^{-1}_{\mathrm{obs}(m)} \times \left(x_{\mathrm{obs}(m)} - \frac{\mu_{1,\mathrm{obs}(m)} + \mu_{-1,\mathrm{obs}(m)}}{2} \right) \Big).$$

[46] A primer on linear classification with missing data A.D.R. Lobo, A. Ayme, C. Boyer, E. Scornet

Linear Discriminant Analysis [46]

Bayes predictor for the complete case:

$$h^\star_{\mathrm{comp}}(x) := \mathrm{sign}\left(\left(\mu_1 - \mu_{-1}\right)^\top \Sigma^{-1}\left(x - \frac{\mu_1 + \mu_{-1}}{2}\right)\right).$$

Proposition: Bayes predictor for LDA+MCAR

Assume LDA + MCAR. Then the PbP Bayes classifier satisfies

$$h^\star_m(x_{\mathrm{obs}(m)}) = \mathrm{sign}\Big(\left(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\right)^\top \Sigma^{-1}_{\mathrm{obs}(m)} \times \left(x_{\mathrm{obs}(m)} - \frac{\mu_{1,\mathrm{obs}(m)} + \mu_{-1,\mathrm{obs}(m)}}{2}\right)\Big).$$

- PbP strategy is Bayes optimal
- Constant imputation is not optimal (if Σ is nondiagonal)
- Extension to MNAR scenarios (GPMM)

[46] A primer on linear classification with missing data A.D.R. Lobo, A. Ayme, C. Boyer, E. Scornet

Logistic Model

Logistic model

$\mathbb{P}[Y=1|X] = \sigma(\beta_0^\star + \sum_j \beta_j^\star X_j)$ with $\sigma(t) = 1/(1+e^{-t})$.

Bayes classifier: $g^\star(\widetilde{x}) = \mathbb{1}_{\eta^\star(\widetilde{x})>0.5}$ with $\eta^\star(\widetilde{x}) = \mathbb{E}[Y|\widetilde{X}=\widetilde{x}]$.

Ill-specified PbP logistic regression

Assume MCAR data in a logistic model for complete data with $X_1,\dots,X_d$ independent Gaussian random variables. Let $m \in \{0,1\}^d$ and assume that there exists a vector $\beta_m^\star \in \mathbb{R}^{|\mathrm{obs}(m)|+1}$ such that

$$\mathbb{P}\left(Y=1|X_{\mathrm{obs}(m)}, M=m\right) = \sigma\left(\beta_{0,m}^\star + \sum_{j\in\mathrm{obs}(m)} \beta_{j,m}^\star X_j\right).$$

Then, for all $j \in \mathrm{mis}(m)$, $\beta_j^\star = 0$

Logistic Model

Ill-specified PbP logistic regression

Assume MCAR data in a logistic model for complete data with $X_1, \ldots, X_d$ independent Gaussian random variables. Let $m \in \{0,1\}^d$ and assume that there exists a vector $\beta^\star_m \in \mathbb{R}^{|\mathrm{obs}(m)|+1}$ such that

$$\mathbb{P}\left(Y = 1 | X_{\mathrm{obs}(m)}, M = m\right) = \sigma\left(\beta^\star_{0,m} + \sum_{j \in \mathrm{obs}(m)} \beta^\star_{j,m} X_j\right).$$

Then, for all $j \in \mathrm{mis}(m)$, $\beta^\star_j = 0$

- ▷ Logistic model cannot hold on complete data AND on data with a given missing pattern
- ▷ Constant imputation Impute-then-Logistic-Regression is ill specified

$$\mathbb{E}[Y|X_{\mathrm{obs}(M)}, M = m] = \mathbb{E}\Big[\mathbb{E}[Y|X]|X_{\mathrm{obs}(M)}\Big] = \mathbb{E}\Big[\sigma(X)|X_{\mathrm{obs}(M)}\Big]$$
$$\neq \sigma\left(\mathbb{E}[\beta^\star_0 + \sum_{j=1}^{d} \beta^\star_j X_j | X_{\mathrm{obs}(M)}]\right).$$

Theoretical Results[48]

Denote $\Phi(t)$ the probit function: $\Phi(t) = (2\pi)^{-1/2} \int_{-\infty}^{t} e^{-t^2/2} dt$,

Theorem

Assume a logistic model on complete data and a GPMM: $X|M = m \sim \mathcal{N}(\mu_m, \Sigma_m)$. Then, for all m, the Bayes probability on pattern m, $\eta^\star_m$, satisfies for all $x \in \mathbb{R}^{|obs(m)|}$,

$$\left| \eta^\star_m(x) - \sigma\left(\frac{\alpha_{0,m} + \alpha_m^\top x}{\sqrt{1 + (\pi/8)\tilde{\sigma}_m^2}} \right) \right| \leqslant 2\|\varepsilon\|_\infty \approx 0.036,$$

where $\varepsilon(t) = \Phi(t) - \sigma(t\sqrt{8/\pi})$, and $\alpha_{0,m}, \alpha_m, \tilde{\sigma}_m^2$.

Theoretical ground for understanding why PbP logistic regression performs well in practice while being ill-specified.

See also[47]

[47] K.A. Verchand, A. Montanari, High-dimensional logistic regression with missing data: Imputation, regularization, and universality

[48] C. Muller, E. Scornet, J. Josse, When Pattern-by-Pattern Works: Theoretical and Empirical Insights for Logistic Models with Missing Values

Methods evaluated

- Pattern-by-pattern (`PbP`): Logistic regression on each pattern

Methods evaluated

- ▷ Pattern-by-pattern (`PbP`)
- ▷ Mean imputation (`Mean.IMP`): Mean per covariate

Methods evaluated

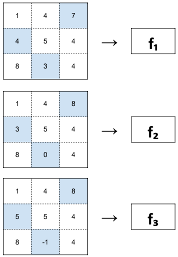
- Pattern-by-pattern (`PbP`)
- Mean imputation (`Mean.IMP`)
- Fully specified (`SAEM`): Fully parametrized model, assuming normal covariates + logistic regression, optimized by Iterative EM

Methods evaluated

- Pattern-by-pattern (`PbP`)
- Mean imputation (`Mean.IMP`)
- Fully specified (`SAEM`)
- Imputation by MICE (`MICE.IMP`): Iterative imputation by iterative MICE algorithm

Methods evaluated

- ▷ Pattern-by-pattern (`PbP`)
- ▷ Mean imputation (`Mean.IMP`)
- ▷ Fully specified (`SAEM`)
- ▷ Imputation by MICE (`MICE.IMP`)
  - → Allow multiple imputations (`MICE.`K`.IMP`): Fit logistic on each dataset, average predictions



Methods evaluated

- ▷ Pattern-by-pattern (`PbP`)
- ▷ Mean imputation (`Mean.IMP`)
- ▷ Fully specified (`SAEM`)
- ▷ Imputation by MICE (`MICE.IMP`)
  - → Allow multiple imputations (MICE.K.IMP)
  - → Add M during imputation process (`MICE.M.IMP`)

X'

1	4	NA	0	0	1
NA	5	4	1	0	0
8	NA	4	0	1	0
NA	5	4	1	0	0
1	3	1	0	0	0
NA	1	4	1	0	0
7	8	4	0	0	0
4	5	NA	0	0	1

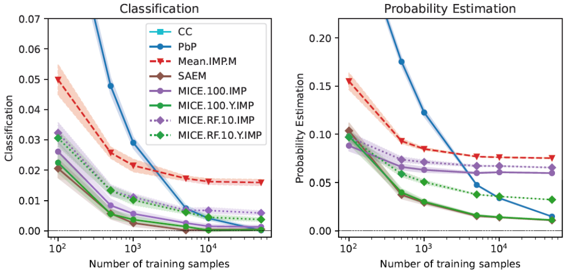
Methods evaluated

- ▷ Pattern-by-pattern (PbP)
- ▷ Mean imputation (Mean.IMP)
- ▷ Fully specified (SAEM)
- ▷ Imputation by MICE (MICE.IMP)
  - → Allow multiple imputations (MICE.K.IMP)
  - → Add M during imputation process (MICE.M.IMP)
  - → Add Y during training of imputation process (MICE.Y.IMP)

X'

1	4	NA	0
NA	5	4	0
8	NA	4	1
NA	5	4	1
1	3	1	1
NA	1	4	0
7	8	4	0
4	5	NA	1

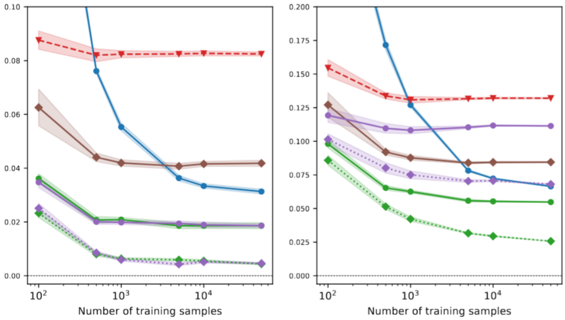
Gaussian features (MCAR)



- ▷ $X \sim \mathcal{N}(0, \Sigma)$
- ▷ 5 dimensions
- ▷ 10 replicates
- ▷ Toeplitz correlation matrix (0.65 corr.)
- ▷ MCAR with prob. 0.25

$\rightarrow$ PbP approaching the Bayes prob. (large training set)
$\rightarrow$ Necessary to use multiple imputations with MICE
$\rightarrow$ Necessary to add Y to MICE imputation
$\rightarrow$ SAEM and MICE.100.Y.IMP best overall

Non-linear features (MCAR)



- ▷ X non-linear transformation of $\mathcal{N}(0, \Sigma)$
- ▷ 5 dimensions
- ▷ 10 replicates
- ▷ Σ Toeplitz matrix (0.65)
- ▷ MCAR with prob 0.25

$\rightarrow$ No method can estimate Bayes probabilities
$\rightarrow$ SAEM suffers from misspecification
$\rightarrow$ PbP not approaching Bayes, coherent with our Theorem

Non-linear features (MCAR): details per pattern

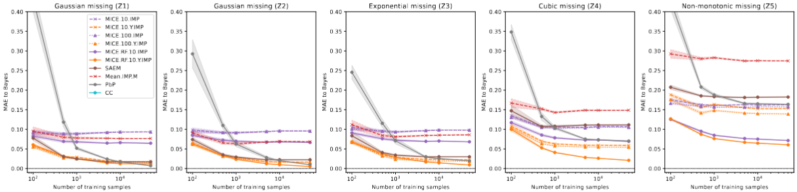
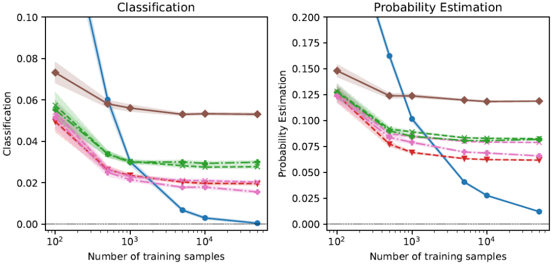


Figure C.7: Performances of selected procedures in terms of MAE from Bayes probabilities. The results are displayed by missing pattern in the test set (with one missing index: [1,0,0,0,0], ..., [0,0,0,0,1]). Means and standard errors over 10 replicates of non-linear features with MCAR missingness are displayed (see Section 5.1). The curves from **MICE.10.IMP** and **MICE.100.IMP** overlap in the first 4 plots.

Mixture of Gaussian (MNAR)



- ▷ $X|M = m \sim \mathcal{N}(\mu_m, \Sigma_m)$
- ▷ 5 dimensions
- ▷ 10 replicates
- ▷ Σ Toeplitz matrix
- ▷ MCAR with prob 0.25

$\rightarrow$ Only the PbP strategy performs well
$\rightarrow$ Coherent with theory

Conclusion

- Theoretically, altough misspecified, Pattern-by-pattern performs well under gaussian covariates (MCAR or Pattern Mixture Model)
- Confirmed experimentally: in GPMM-MNAR, PbP is one of the most competitive methods.

Empirically,

- MICE imputation consistently performing well in MCAR setting
    1. With the use of multiple imputations
    2. With the inclusion of Y in covariates
    3. But needs non-linear inner regressor for non-linear covariates
- M(N)AR settings are more tricky

Summary

Take home messages on Imputation

Missing mechanisms

- Different missing data scenario (MCAR, MAR, MNAR).
- Both % of NA & structure matter (5% of NA can be an issue)
- MAR was designed for likelihood inference (e.g. EM algorithm) but can hide many complex distributions (distribution shift in MAR).
- Few implementations of EM strategies.

Take home messages on Imputation

Missing mechanisms

- ▷ Different missing data scenario (MCAR, MAR, MNAR).
- ▷ Both % of NA & structure matter (5% of NA can be an issue)
- ▷ MAR was designed for likelihood inference (e.g. EM algorithm) but can hide many complex distributions (distribution shift in MAR).
- ▷ Few implementations of EM strategies.

Imputation

- ▷ Results in a complete data set, on which any method can be applied.
- ▷ *Imputation is both seductive & dangerous* (Dempster & Rubin, 1983).
  - ◇ Seductive: *can lull the user into the pleasant state of* **believing that the data are complete**
  - ◇ Dangerous: *it lumps together situations where the problem is minor enough to be handled in this way & situations where estimators applied to the imputed data have* **substantial biases**.

Take home messages on Imputation

Missing mechanisms

- ▷ Different missing data scenario (MCAR, MAR, MNAR).
- ▷ Both % of NA & structure matter (5% of NA can be an issue)
- ▷ MAR was designed for likelihood inference (e.g. EM algorithm) but can hide many complex distributions (distribution shift in MAR).
- ▷ Few implementations of EM strategies.

Imputation

- ▷ Results in a complete data set, on which any method can be applied.
- ▷ *Imputation is both seductive & dangerous* (Dempster & Rubin, 1983).

Single imputation

- ▷ From simple (mean imputation) to more complex strategies (MissForest)
- ▷ Useful for point estimates
- ▷ Distort the marginal and joint distributions
- ▷ Lead to confidence interval with poor coverage

Take home messages on Imputation

Single imputation

- From simple (mean imputation) to more complex strategies (MissForest)
- Useful for point estimates
- Distort the marginal and joint distributions
- Lead to confidence interval with poor coverage

Take home messages on Imputation

Single imputation

- From simple (mean imputation) to more complex strategies (MissForest)
- Useful for point estimates
- Distort the marginal and joint distributions
- Lead to confidence interval with poor coverage

Multiple imputation

- Look for an imputation that preserve the joint distribution of the data
- MI aims at estimating the parameters and their variability taking into account the uncertainty of the missing values
- Useful for confidence intervals
- Compare imputations with distributional metrics like energy distance
- mice-DRF promising (code available) - mice-Engression[a]

[a] Shen & Meinshausen (2024). Engression: extrapolation through the lens of distributional regression. *JRSS B.*

Take home messages - Supervised learning

Aim

Estimating the Bayes predictor in presence of missing values

$$f^\star(\tilde{X}) = \sum_{m \in \{0,1\}^d} \mathbb{E}\left[Y | X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$$

Two common strategies:

- ▷ **Impute-then-regress strategies** - impute the data then learn on the imputed data set
  - ◇ Computationally efficient but possibly inconsistent
- ▷ **Pattern-by-pattern strategies** - use a different predictor for each missing pattern
  - ◇ Consistent by design but intractable in most situations

Take-home messages - Supervised learning

Decision trees

- Decision trees are among the few methods able to natively handle missing values (MIA)
- Amounts to PbP strategies with a data-driven selection of relevant patterns

Impute-then-regress

- Consistent for any imputation method when the predictor is universally consistent
- Use the same imputation for train and test sets
- In finite sample, some imputation may ease the training of the predictor (e.g., Conditional Imputation is not well-suited in general)
- Rethinking imputation: a good imputation is the one that makes the prediction easy

Take-home messages - Linear models

Pattern-by-Pattern

- Rate in $2^d/n$ in the worst case
- Improved by performing regressions on the most frequent patterns only
- Rate in d^2/n for MCAR Bernoulli, with a probability of missingness small enough
- MNAR/MAR is not suited for prediction (GPMM)

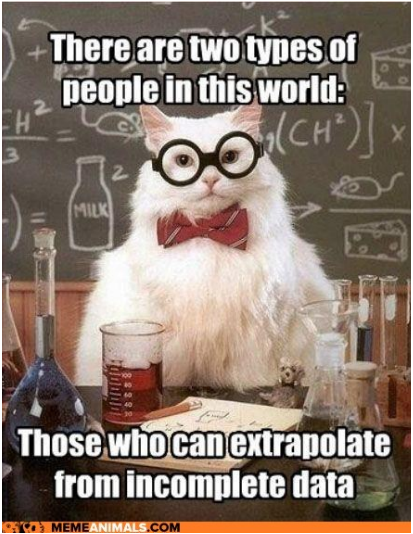
Impute-then-Regress

- Inconsistent in fixed dimension
- Consistent in high dimensions with a slow rate $n^{-1/2}$
- Imputation by zero amounts to a ridge regularization with a strength depending on the missing probability

Logistic regression model

- PbP and constant imputation result in inconsistent predictor
- But in presence of Gaussian features, Bayes probabilities are correctly estimated by PbP
- PbP competitive in GPMM-MNAR scenario but deteriorates when input distribution is not Gaussian

Thank you![49]



[49] More ressources: `https://rmisstastic.netlify.app/`