Near-optimal rate of consistency for linear prediction with missing values

Alexis Ayme

Claire Boyer Aymeric Dieuleveut and Erwan Scornet

ICML 2022

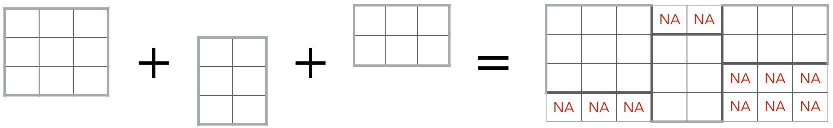
Background

- **Growing mass of data => NA (not attribut)/missing values**
- **Different sources:**

1. Bugs
2. Cost

$1	$10	$100	$0
		NA	
	NA	NA	

3. Multiplication of sources (i.e. merge)



4. Sensitive data

Age	Job	**Incomes**
		NA
		NA
NA		NA
		NA

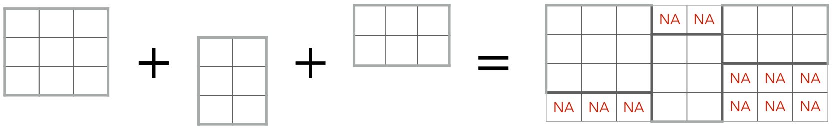
Background

- **Growing mass of data => NA (not attribut)/missing values**
- **Different sources:**

1. Bugs
2. Cost

$1	$10	$100	$0
		NA	
	NA	NA	

3. Multiplication of sources (i.e. merge)



4. Sensitive data

Age	Job	**Income**
		NA
		NA
NA		NA
		NA

- Any statistical analyses require **complete** data
  - Strategy 1: **complete** the dataset before the ML process (e.g. by collaborative filtering)
  - Strategy 2: **adapt** statistical analysis to handle missing values (e.g. EM algorithm to perform regression with NA)

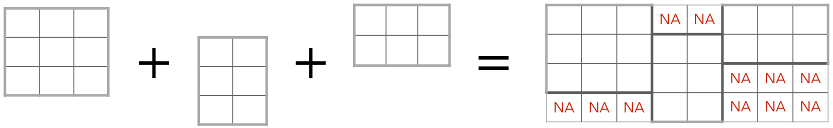
Background

- **Growing mass of data => NA (not attribut)/missing values**
- **Different sources:**

1. Bugs
2. Cost

$1	$10	$100	$0
		NA	
	NA	NA	

3. Multiplication of sources (i.e. merge)



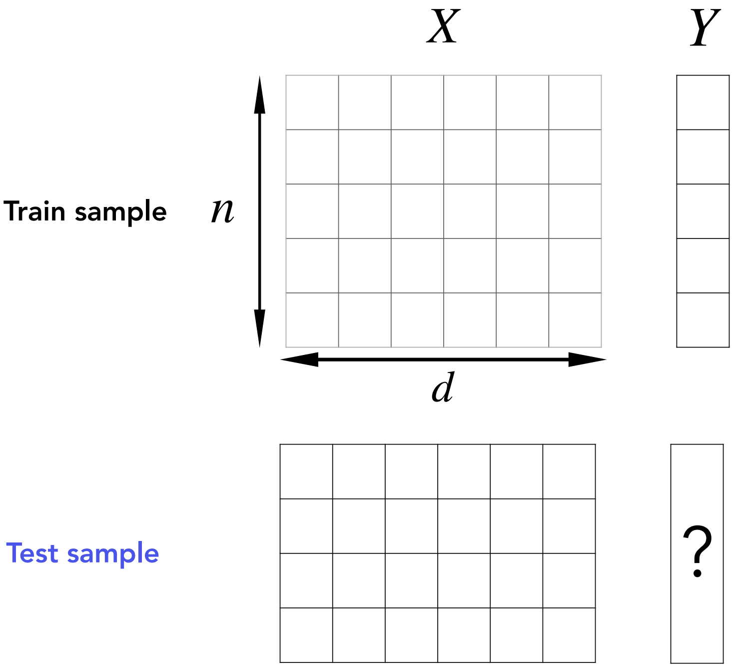
4. Sensitive data

Age	Job	**Income**
		NA
		NA
NA		NA
		NA

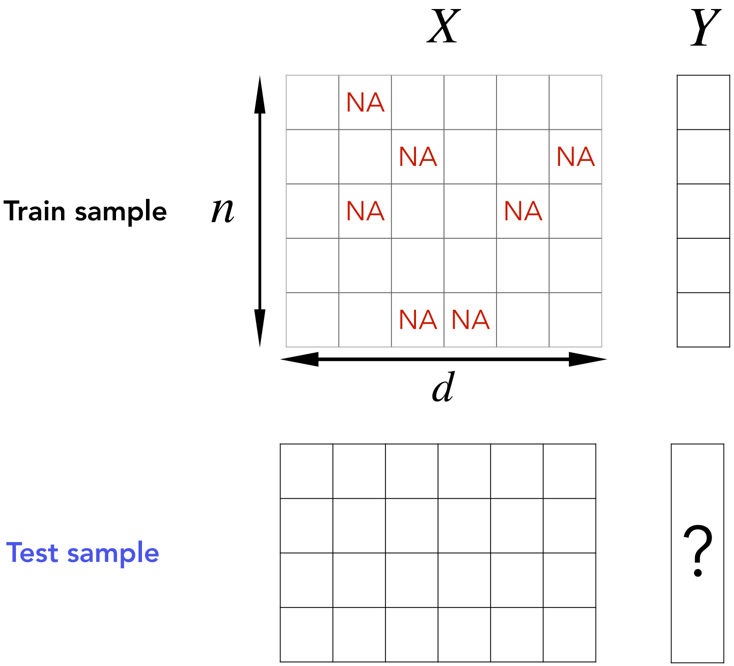
- Any statistical analyses require **complete** data
  - Strategy 1: **complete** the dataset before the ML process (e.g. by collaborative filtering)
  - Strategy 2: **adapt** statistical analysis to handle missing values (e.g. EM algorithm to perform regression with NA)

What about **supervised learning**?
i.e. prediction with NAs

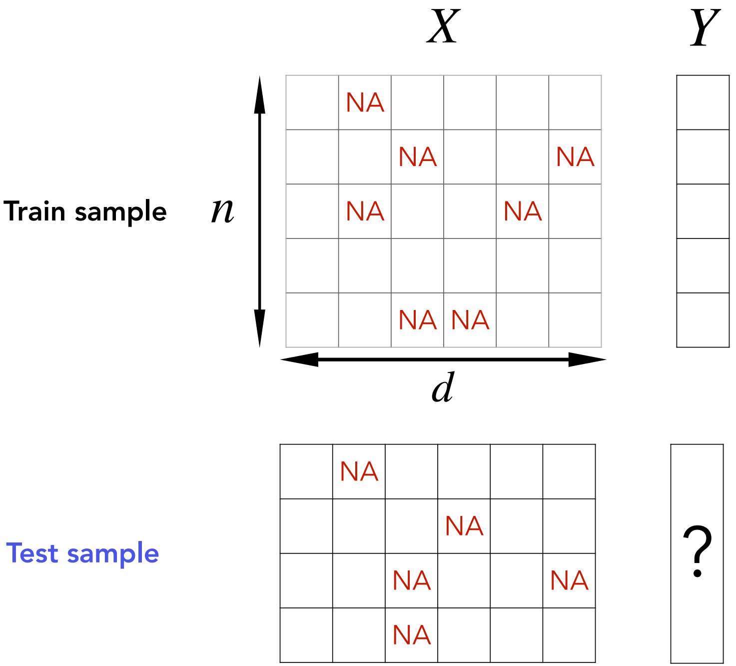
Supervised learning with missing values (NA)



Supervised learning with missing values (NA)



Supervised learning with missing values **(NA)**

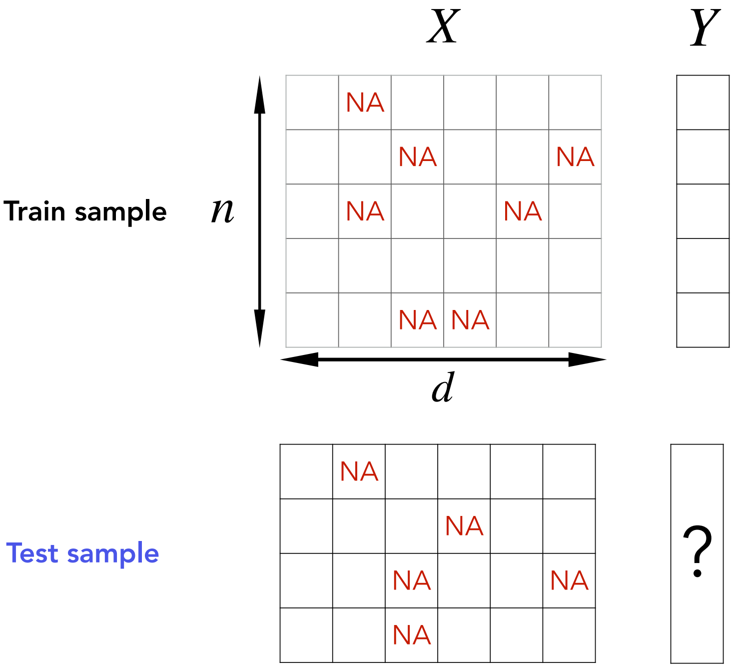


$$Y = \beta^T X + \varepsilon$$

$$\|\hat{\beta} - \beta\|_2^2$$

$$(\hat{\beta}^T X - \beta^{*} X)^2$$

Supervised learning with missing values (NA)



- **Missing pattern:** $M_i \in \{0,1\}^d$

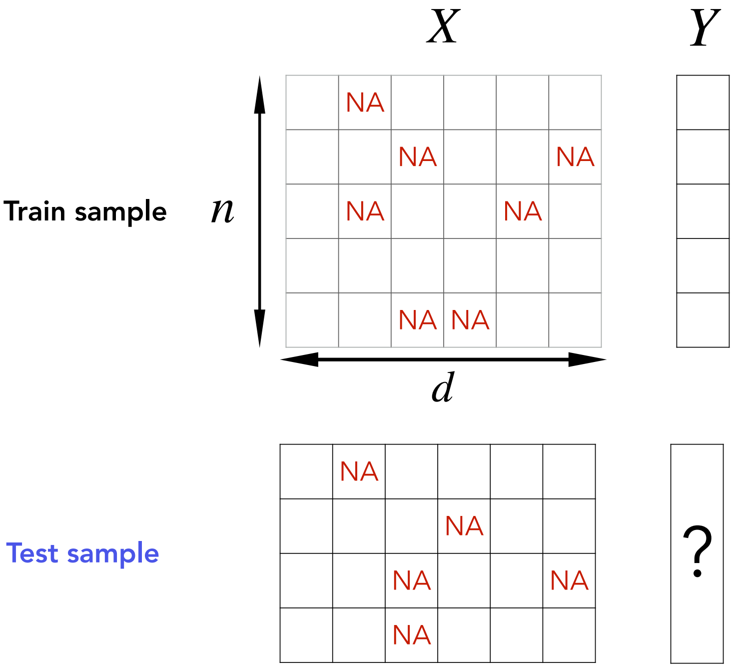
$$X_i = \begin{array}{|c|c|c|c|c|c|}\hline \text{NA} & 1 & -5 & \text{NA} & 0 & 2 \\ \hline\end{array}$$

$$\downarrow$$

$$M_i = (1,\ 0,\ 0,\ 1,\ 0,\ 0)$$

- **Input:** $Z = \left(X_{\text{obs}}, M\right)$
- **Output:** $Y \in \mathbb{R}$

Supervised learning with missing values (NA)



- **Missing pattern:** $M_i \in \{0,1\}^d$

$$X_i = \begin{array}{|c|c|c|c|c|c|}\hline \text{NA} & 1 & -5 & \text{NA} & 0 & 2 \\ \hline\end{array}$$

$$\downarrow$$

$$M_i = (1,\ 0,\ 0,\ 1,\ 0,\ 0)$$

- **Input:** $Z = (X_{\text{obs}}, M)$
- **Output:** $Y \in \mathbb{R}$

Goal: Predict on **test sample** minimizing

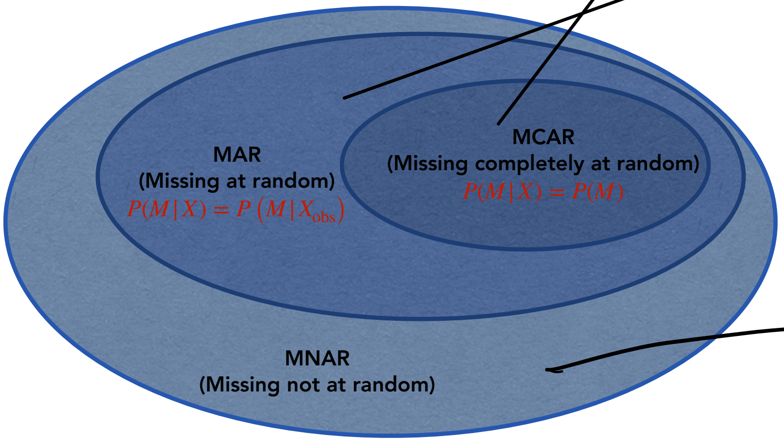
$$R(f) = \mathbb{E}_{Z,Y}\left[\left(Y - f(Z)\right)^2\right]$$

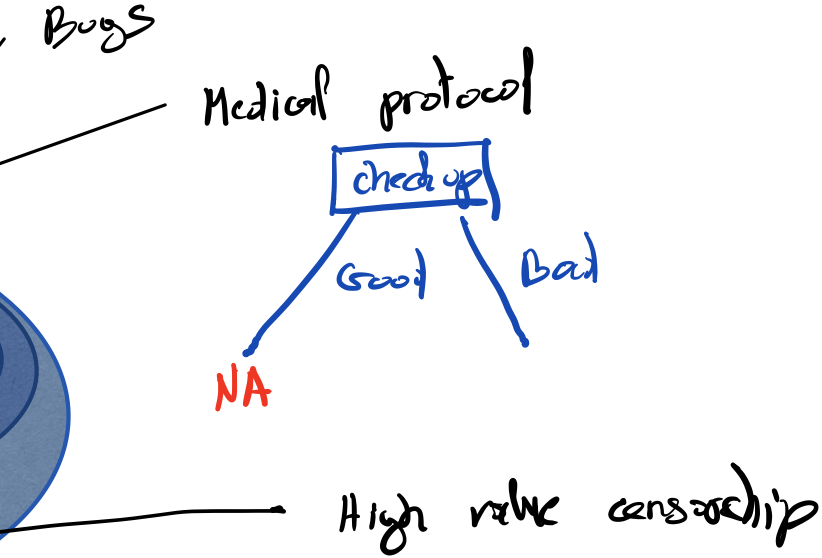
Zoo of assumptions on NA

- **First point of view :** $P(X,M) = P(M|X)P(X)$

  1) Assumption on $P(X)$:
     Example: X Gaussian Vector

  2) Assumption on $P(M|X)$:



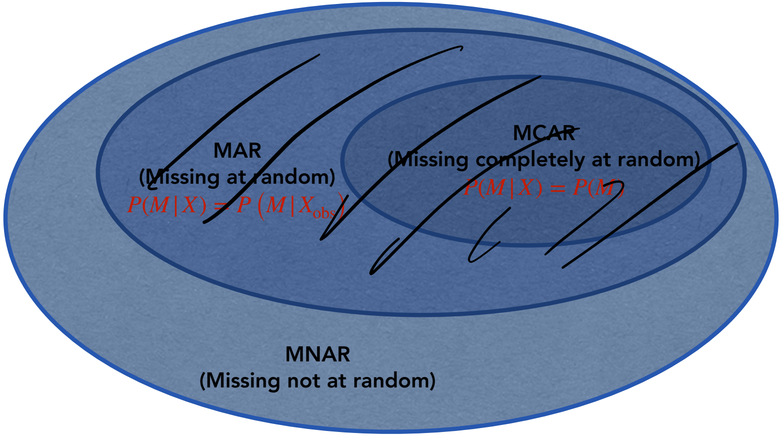


Zoo of assumptions on NA

- **First point of view :** $P(X, M) = P(M \mid X)P(X)$

  1) Assumption on $P(X)$:
     Example: X Gaussian Vector

  2) Assumption on $P(M \mid X)$:

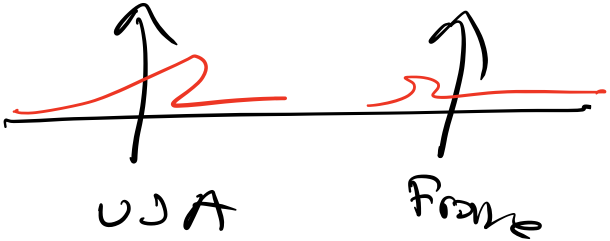


- **Second point of view :** $P(X, M) = P(M)P(X \mid M)$

  1) Assumption on $P(M)$:
     Example: $P(M = m) = p_m$

  2) Assumption on $P(X \mid M)$:

     GPMM (Gaussian pattern mixture model):
     $X \mid (M = m)$ Gaussian Vector



Pattern-by-Pattern regression

- **Assumption**: linear model for complete inputs

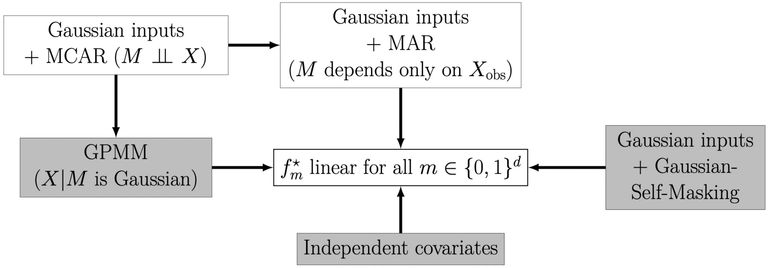
$$y_i = \beta^\top X_i + \epsilon_i$$

/!\ With NA, the Bayes predictor does *not* necessarily remain linear

- **Bayes predictor** (better prediction) decomposition

$$f^\star(Z) = \sum_{m \in \{0,1\}^d} f^\star_m(X_{obs(m)}) \mathbf{1}_{M=m}$$

Local **Bayes prediction** for the missing pattern $(M = m)$



Assumption to obtain linearity

Proposition: (Le Morvan et al. 2020)
Under **linear model** and several **missing data scenarios** (including MNAR), $f^\star_m$ are **linear**

Pattern-by-Pattern regression

- **Assumption**: linear model for complete inputs

$$y_i = \beta^\top X_i + \epsilon_i$$

/!\ With NA, the Bayes predictor does *not* necessarily remain linear

- **Bayes predictor** decomposition

$$f^\star(Z) = \sum_{m \in \{0,1\}^d} f^\star_m(X_{obs(m)}) \mathbf{1}_{M=m}$$

Local **Bayes prediction** for the missing pattern $(M = m)$

Proposition: (Le Morvan et al. 2020)
Under **linear model** and several **missing data scenarios** (including MNAR), $f^\star_m$ are **linear**

- **Pattern-by-pattern** predictor

$$\hat{f}(Z) = \sum_{m \in \{0,1\}^d} \hat{f}_m(X_{obs(m)}) \mathbf{1}_{M=m}$$

Local **Least-Square** regression on
$\{(X_{i,obs}, Y_i), M_i = m\}$

Pattern-by-Pattern regression

- **Assumption**: linear model for complete inputs

$$y_i = \beta^\top X_i + \epsilon_i$$

/!\ With NA, the Bayes predictor does *not* necessarily remain linear

- **Bayes predictor** decomposition

$$f^\star(Z) = \sum_{m \in \{0,1\}^d} f^\star_m(X_{obs(m)}) \mathbf{1}_{M=m}$$

Local **Bayes prediction** for the missing pattern $(M = m)$

Proposition: (Le Morvan et al. 2020)
Under **linear model** and several **missing data scenarios** (including MNAR), $f^\star_m$ are **linear**

- **Pattern-by-pattern** predictor

$$\hat{f}(Z) = \sum_{m \in \{0,1\}^d} \hat{f}_m(X_{obs(m)}) \mathbf{1}_{M=m}$$

Local **Least-Square** regression on $\{(X_{i,obs}, Y_i), M_i = m\}$

Theorem 1:
Under Lipschitz and sub-Gaussian assumptions,

$$\mathcal{E}(\hat{f}) := \mathbb{E}\left[\left(f^\star(Z) - \hat{f}(Z)\right)^2\right] \leq A \log(n) 2^d \frac{d}{n} + \text{Approx}$$

- Optimal for equiprobable missing patterns $\left(p_m = \frac{1}{2^d}\right)$
- Tight for the worst case of pattern-by-pattern predictors
- Sub-optimal for other distributions?

Thresholded Pattern-by-Pattern regression

- Adaptivity to the missing pattern distribution to overcome the curse of dimensionality
- Overfitting reduction

via **Thresholded P-by-P** predictor:

$\hat{p}_m$ = **frequency** of pattern m

$$\hat{f}(Z) = \sum_{m \in \{0,1\}^d} \hat{f}_m(X_{obs(m)}) \mathbf{1}_{M=m} \mathbf{1}_{\hat{p}_m > \frac{d}{n}}$$

Local **Least-Square** regression on $\left\{(X_{i,obs}, Y_i), M_i = m\right\}$

Thresholded Pattern-by-Pattern regression

- Adaptivity to the missing pattern distribution to overcome the curse of dimensionality
- Overfitting reduction

via **Thresholded P-by-P** predictor:

$\hat{p}_m$ = **frequency** of pattern m

$$\hat{f}(Z) = \sum_{m \in \{0,1\}^d} \hat{f}_m(X_{obs(m)}) \mathbf{1}_{M=m} \mathbf{1}_{\hat{p}_m > \frac{d}{n}}$$

Local **Least-Square** regression on $\{(X_{i,obs}, Y_i), M_i = m\}$

- Definition: missing pattern **complexity**

$$\mathfrak{C}_p\left(\frac{d}{n}\right) = \sum_{m \in \{0,1\}^d} p_m \wedge \frac{d}{n}$$

Theorem 2: (Main result)
Under Lipschitz and Sub-Gaussian assumptions

$$\mathscr{E}(\hat{f}) \leq A \log(n) \mathfrak{C}_p\left(\frac{d}{n}\right) + \text{Approx}$$

Thresholded Pattern-by-Pattern regression

- Adaptivity to the missing pattern distribution to overcome the curse of dimensionality
- Overfitting reduction

via **Thresholded P-by-P** predictor:

$\hat{p}_m$ = **frequency** of pattern m

$$\hat{f}(Z) = \sum_{m \in \{0,1\}^d} \hat{f}_m(X_{obs(m)}) \mathbf{1}_{M=m} \mathbf{1}_{\hat{p}_m > \frac{d}{n}}$$

Local **Least-Square** regression on $\{(X_{i,obs}, Y_i), M_i = m\}$

- Definition: missing pattern **complexity**

$$\mathfrak{C}_p\left(\frac{d}{n}\right) = \sum_{m \in \{0,1\}^d} p_m \wedge \frac{d}{n}$$

Theorem 2: (Main result)
Under Lipschitz and Sub-Gaussian assumptions

$$\mathscr{E}(\hat{f}) \leq A \log(n) \mathfrak{C}_p\left(\frac{d}{n}\right) + \text{Approx}$$

Examples:

1. **Uniform** distribution: $\mathfrak{C}_p\left(\frac{d}{n}\right) = 2^d \frac{d}{n}$
2. **Bernoulli** distribution: $M_j \sim \mathscr{B}(\epsilon)$ and $\epsilon \leq \frac{d}{n}$

$$\mathfrak{C}_p\left(\frac{d}{n}\right) \leq \frac{d^2}{n}$$

The thresholded P-by-P predictor is near-optimal

- **Minimax risk**

Worst case on a class of problem $\mathscr{P}_p$

$$\mathscr{E}_{\text{mini}}(p) = \inf_{\tilde{f}} \sup_{\mathbb{P} \in \mathscr{P}_p} \mathbb{E}_{\mathbb{P}} \left[\left(\tilde{f}(Z) - f^\star(Z) \right)^2 \right]$$

Best algorithm

where $\mathscr{P}_p$ represents a class of data distributions

- for which the missing pattern distribution is p
- under Lipschitz and Sub-Gaussian assumptions

The thresholded P-by-P predictor is near-optimal

- **Minimax risk**

Worst case on a class of problem $\mathscr{P}_p$

$$\mathscr{E}_{\text{mini}}(p) = \inf_{\tilde{f}} \sup_{\mathbb{P}\in\mathscr{P}_p} \mathbb{E}_{\mathbb{P}}\left[\left(\tilde{f}(Z) - f^\star(Z)\right)^2\right]$$

Best algorithm

where $\mathscr{P}_p$ represents a class of data distributions
- for which the missing pattern distribution is p
- under Lipschitz and Sub-Gaussian assumptions

Theorem 3:

$$\sigma^2 \mathfrak{C}_p\left(\frac{1}{n}\right) \lesssim \mathscr{E}_{\text{mini}}(p) \underbrace{\leq \mathscr{E}(\hat{f}) \leq A\log(n)\mathfrak{C}_p\left(\frac{d}{n}\right)}_{\text{Theorem 2}}$$

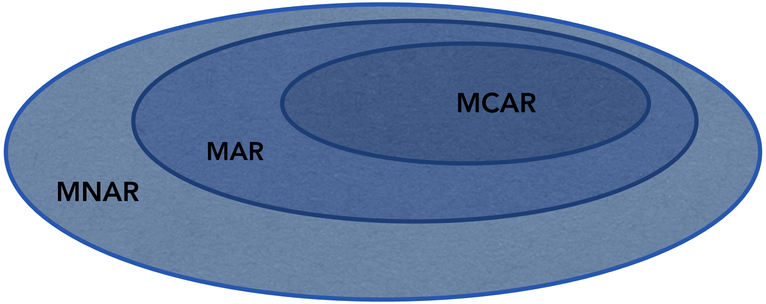
- Lower bound still holds when $\mathscr{P}_p$ includes **MAR** missing values

Examples

1. **Uniform** distribution: $\mathfrak{C}_p\left(\frac{1}{n}\right) = \frac{2^d}{n}$, $\mathfrak{C}_p\left(\frac{d}{n}\right) = 2^d\frac{d}{n}$
2. **Bernoulli** distribution: $\mathfrak{C}_p\left(\frac{1}{n}\right) = \frac{d}{n}$, $\mathfrak{C}_p\left(\frac{d}{n}\right) = \frac{d^2}{n}$

The thresholded P-by-P predictor is near-optimal

- Inference POV



Theorem 3:

$$\sigma^2 \mathfrak{C}_p\left(\frac{1}{n}\right) \lesssim \mathscr{E}_{\text{mini}}(p) \underbrace{\leq \mathscr{E}(\hat{f}) \leq A \log(n) \mathfrak{C}_p\left(\frac{d}{n}\right)}_{\text{Theorem 2}}$$

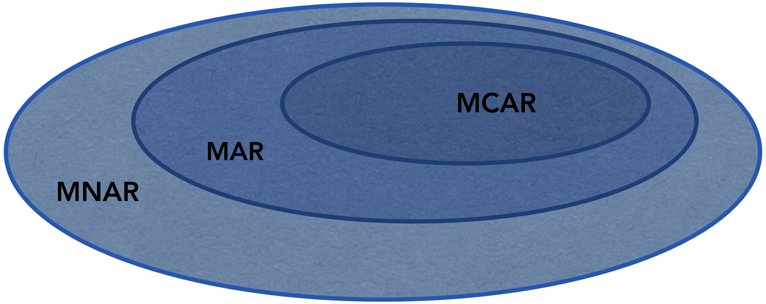
- Lower bound still holds when $\mathscr{P}_p$ includes **MAR** missing values

Examples

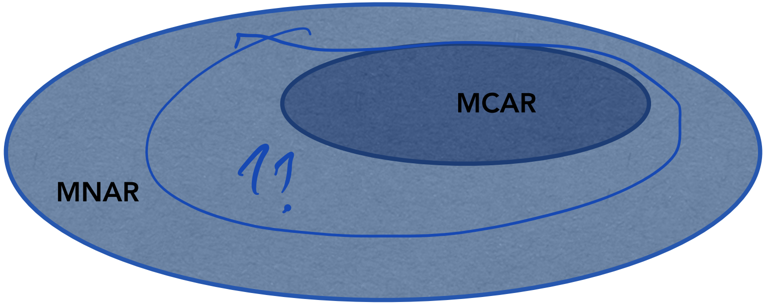
1. **Uniform** distribution: $\mathfrak{C}_p\left(\frac{1}{n}\right) = \frac{2^d}{n}, \mathfrak{C}_p\left(\frac{d}{n}\right) = 2^d \frac{d}{n}$
2. **Bernoulli** distribution: $\mathfrak{C}_p\left(\frac{1}{n}\right) = \frac{d}{n}, \mathfrak{C}_p\left(\frac{d}{n}\right) = \frac{d^2}{n}$

The thresholded P-by-P predictor is near-optimal

- Inference POV



- Supervised learning POV



Theorem 3:

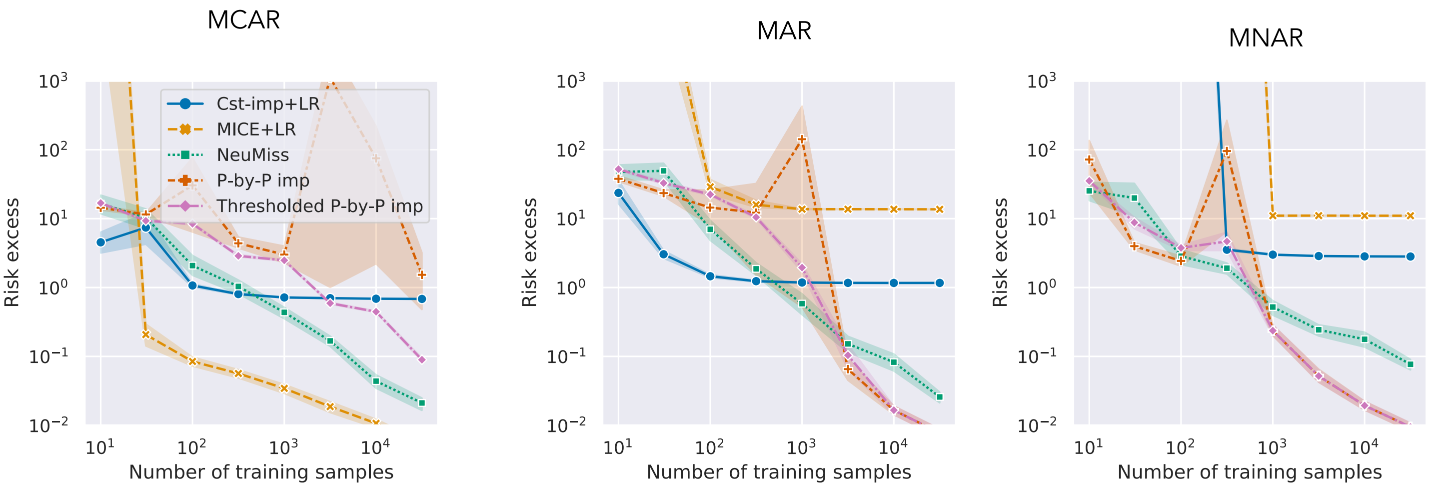
$$\sigma^2 \mathfrak{C}_p\left(\frac{1}{n}\right) \lesssim \mathscr{E}_{\text{mini}}(p) \underbrace{\le \mathscr{E}(\hat{f}) \le A \log(n) \mathfrak{C}_p\left(\frac{d}{n}\right)}_{\text{Theorem 2}}$$

- Lower bound still holds when $\mathscr{P}_p$ includes **MAR** missing values

Examples

1. **Uniform** distribution: $\mathfrak{C}_p\left(\frac{1}{n}\right) = \frac{2^d}{n}$, $\mathfrak{C}_p\left(\frac{d}{n}\right) = 2^d \frac{d}{n}$
2. **Bernoulli** distribution: $\mathfrak{C}_p\left(\frac{1}{n}\right) = \frac{d}{n}$, $\mathfrak{C}_p\left(\frac{d}{n}\right) = \frac{d^2}{n}$

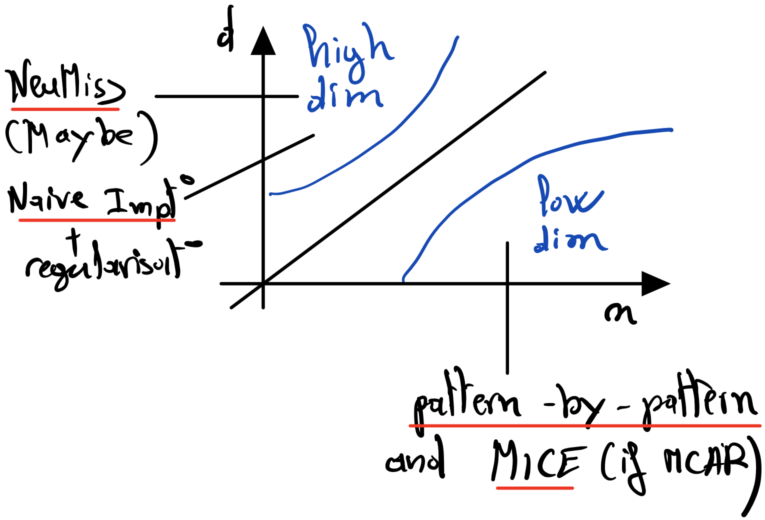
Numerical experiments



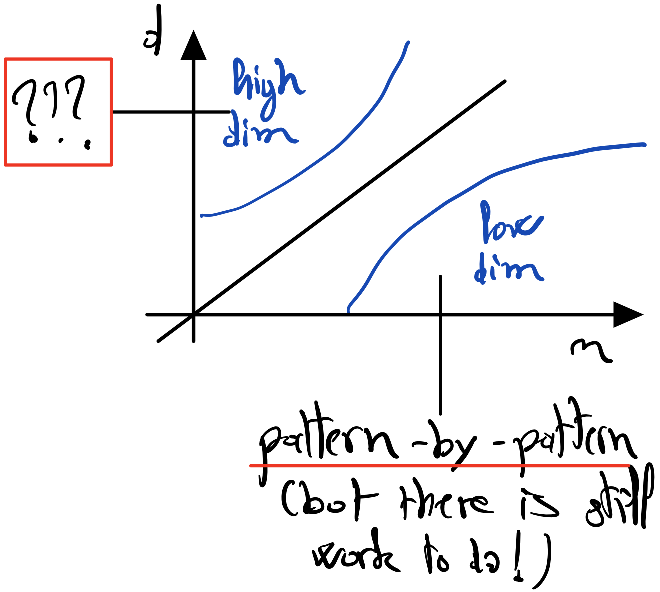
Excess risk w.r.t. n with $d = 8$

What's work ?

In practice



In theory



Conclusion

Theoretical contributions

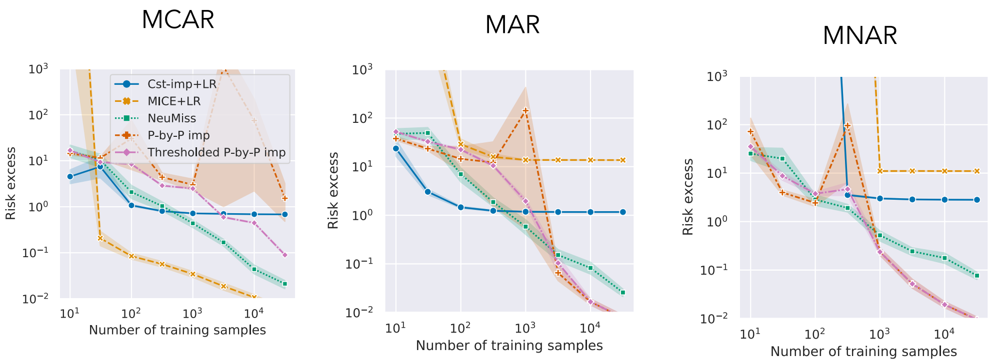
- New thresholded predictor
- Adaptative upper bound
- Near optimal

$$\hat{f}(Z) = \sum_{m \in \{0,1\}^d} \hat{f}_m(X_{obs(m)}) \mathbf{1}_{M=m} \mathbf{1}_{\hat{p}_m > \frac{d}{n}}$$

$$\sigma^2 \mathfrak{C}_p\left(\frac{1}{n}\right) \lesssim \mathscr{E}_{\text{mini}}(\mathscr{P}) \leq A \log(n) \mathfrak{C}_p\left(\frac{d}{n}\right)$$

Numerical experiments

- Thresholded P-by-P predictor:
  - reduced variance
  - consistent regardless of the missing scenario



Excess risk w.r.t. n with $d = 8$

Near-optimal rate of consistency for linear prediction with missing values

Corresponding Author: alexis.ayme@sorbonne-universite.fr

SORBONNE UNIVERSITÉ

l'X ÉCOLE POLYTECHNIQUE UNIVERSITÉ PARIS-SACLAY

Inria

Special thanks to the Paris City Council for the financial support

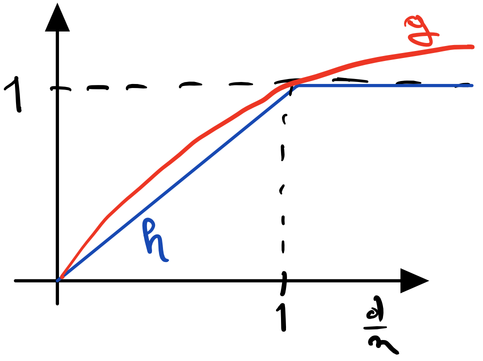
Some theoritical details:

Why $\mathfrak{C}_p$ is adaptive?

$$\mathfrak{C}_p\left(\frac{d}{n}\right) = \sum_{m \in \{0,1\}^d} p_m \wedge \frac{d}{n}$$

$$= \sum_m p_m \underbrace{\left(1 \wedge \frac{d}{n p_m}\right)}_{h\left(\frac{d}{n p_m}\right)}$$

$$\leq \sum_m p_m \, g\left(\frac{d}{n p_m}\right)$$



Two examples:

1) $g = id \implies \mathfrak{C}_p\left(\frac{d}{n}\right) \leq \sum_m p_m \frac{d}{n p_m} = 2^d \frac{d}{n}$

2) $\left| \begin{array}{l} g(x) = x^\alpha \\ \alpha \in]0,1[\end{array} \right. \implies \mathfrak{C}_p\left(\frac{d}{n}\right) \leq \sum_m p_m^{1-\alpha} \left(\frac{d}{n}\right)^\alpha = \left(\frac{d}{n}\right)^\alpha e^{\alpha H_{1-\alpha}(p)}$

where $H_{1-\alpha}(p)$ = Renyi Entropy