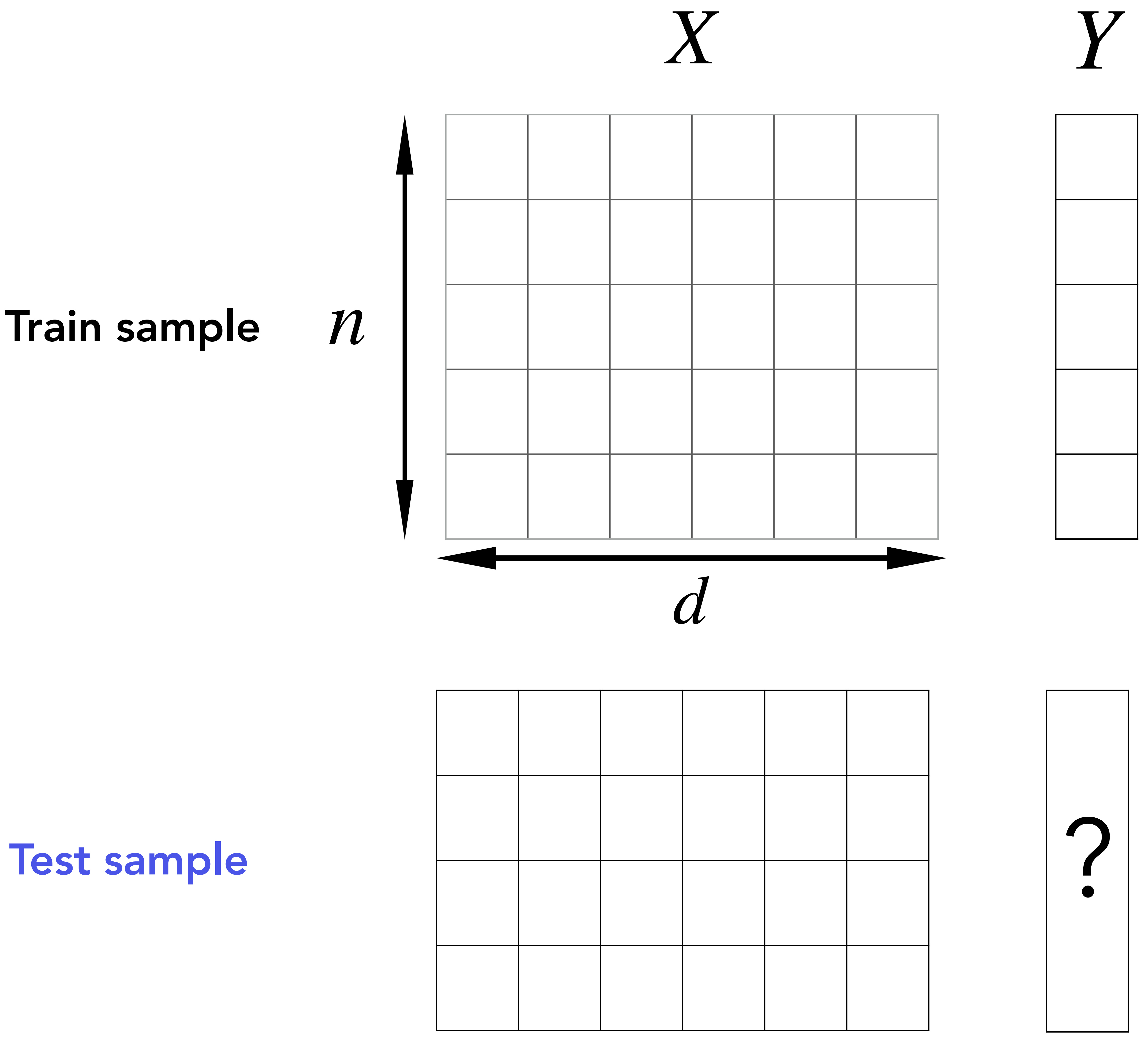
Near-optimal rate of consistency for linear prediction with missing values

Alexis Ayme

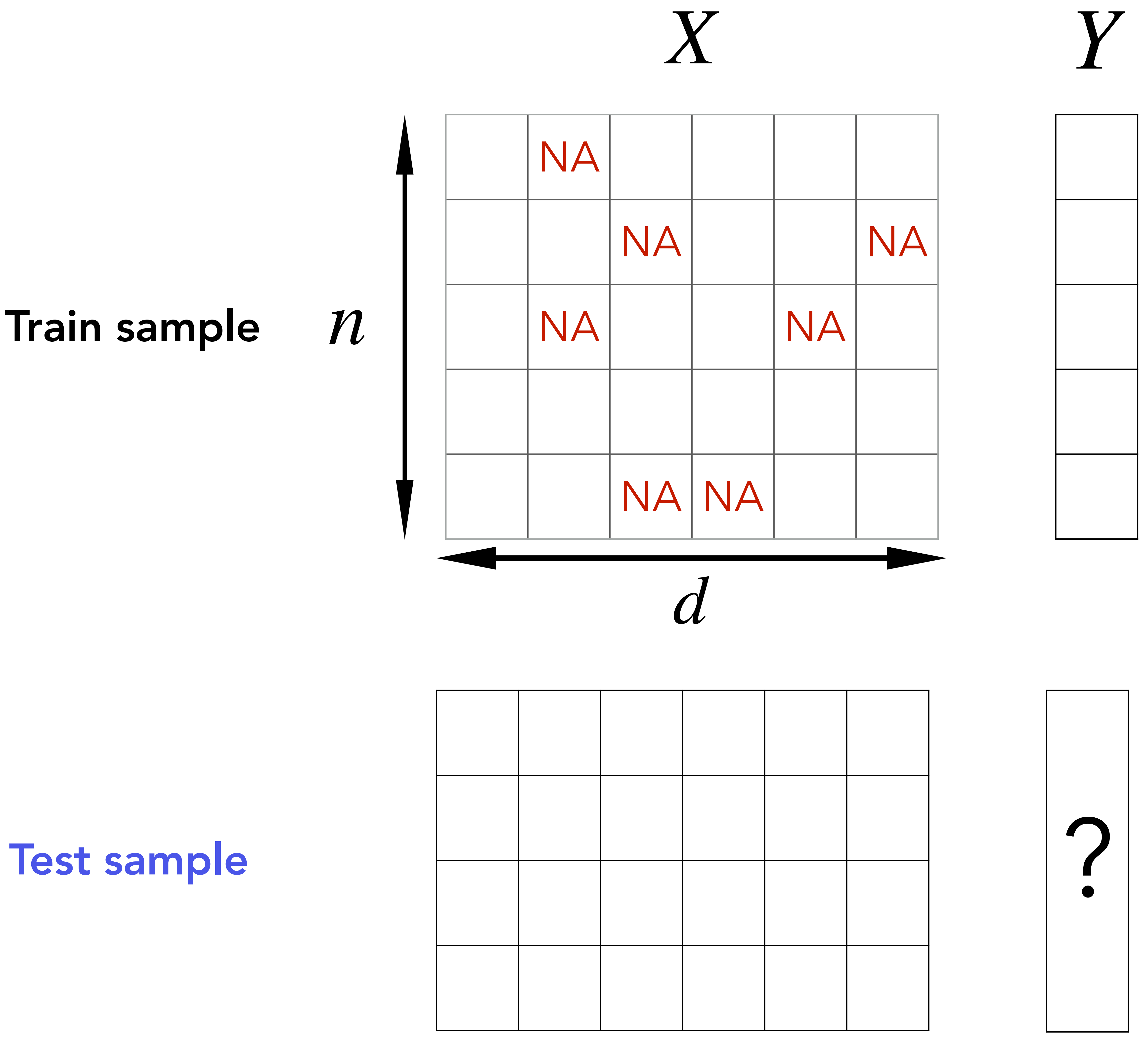
Claire Boyer Aymeric Dieuleveut and Erwan Scornet

ICML 2022

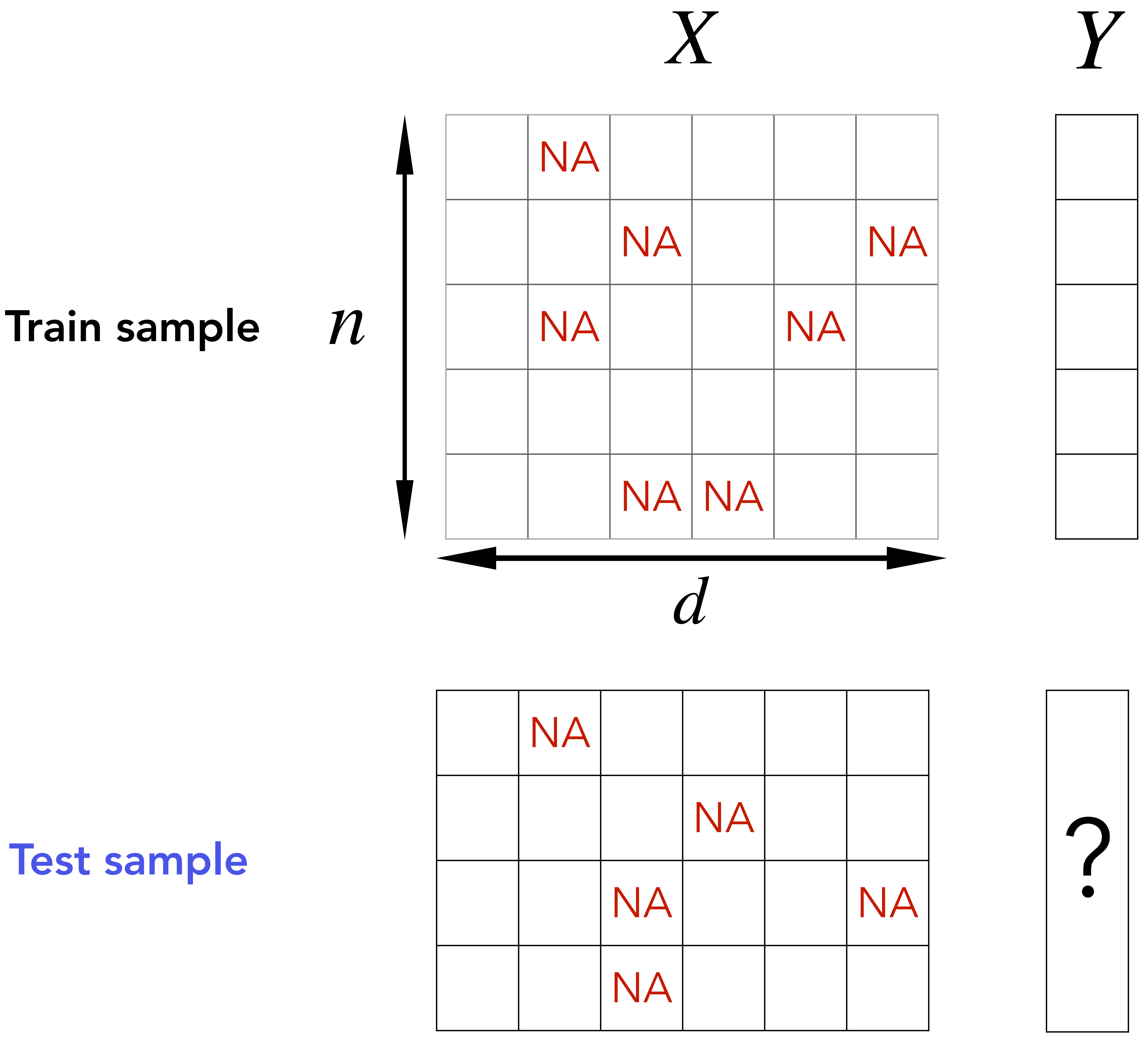
Supervised learning with missing values (NA)



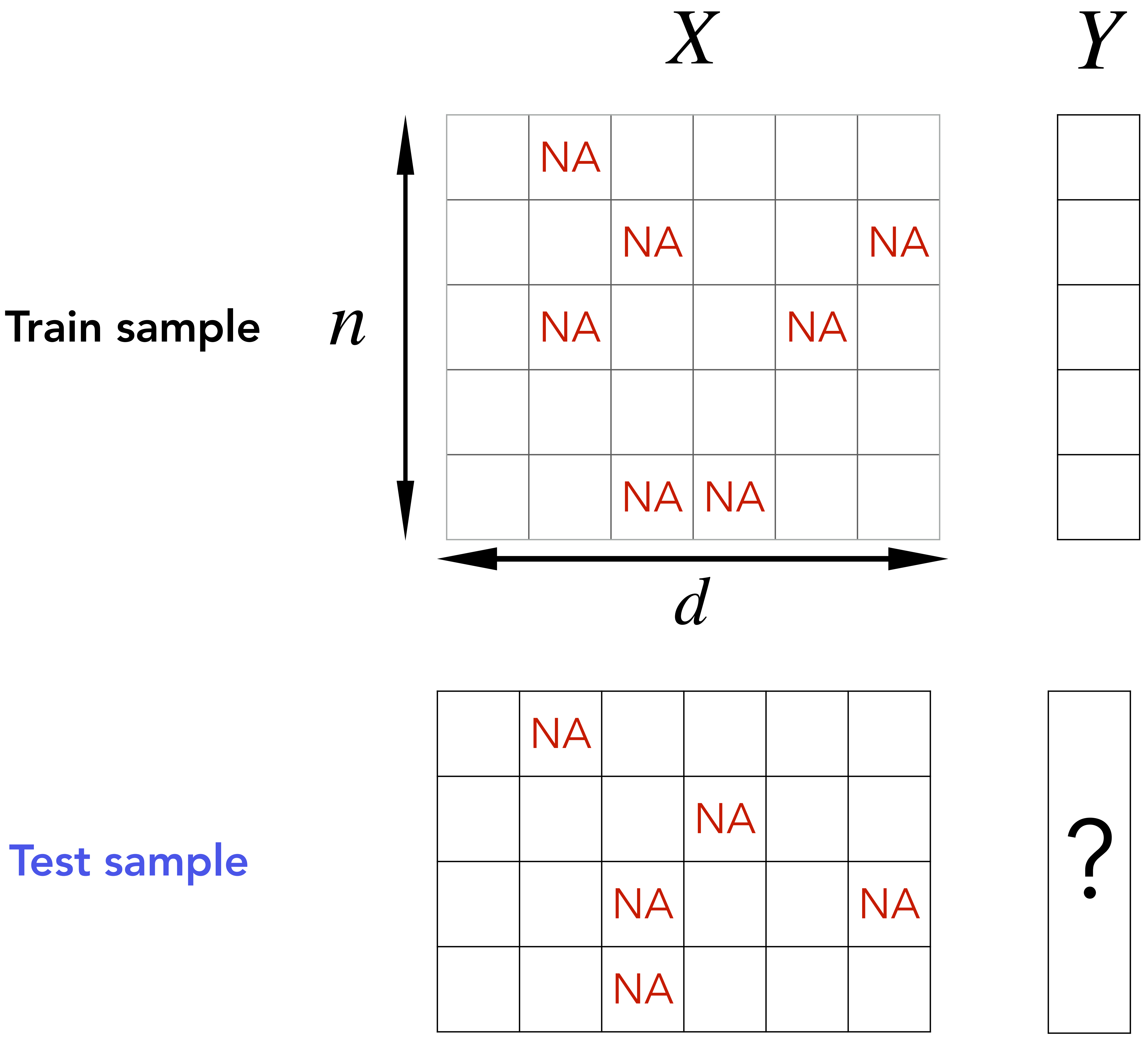
Supervised learning with missing values (NA)



Supervised learning with missing values (NA)



Supervised learning with missing values (NA)



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

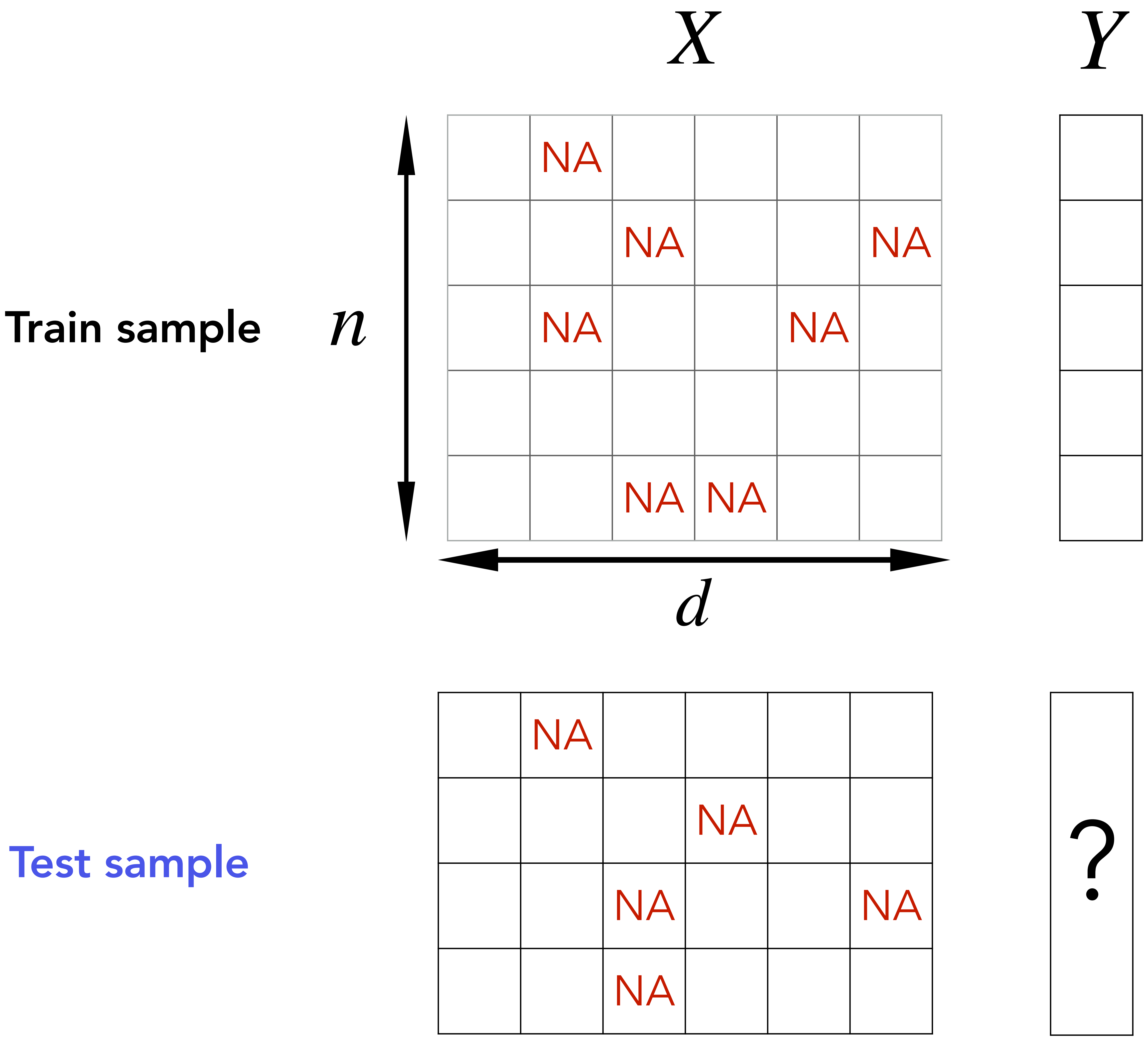
$$X_i = \begin{array}{|c|c|c|c|c|c|}\hline \text{NA} & 1 & -5 & \text{NA} & 0 & 2 \\ \hline\end{array}$$

$$\downarrow$$

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

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

Goal: Predict on **test sample** minimizing

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

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 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,

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

- 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

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

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

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

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

Conclusion

Theoretical contributions

- New thresholded predictor
- Adaptive upper bound
- Near optimal

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

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

Conclusion

Theoretical contributions

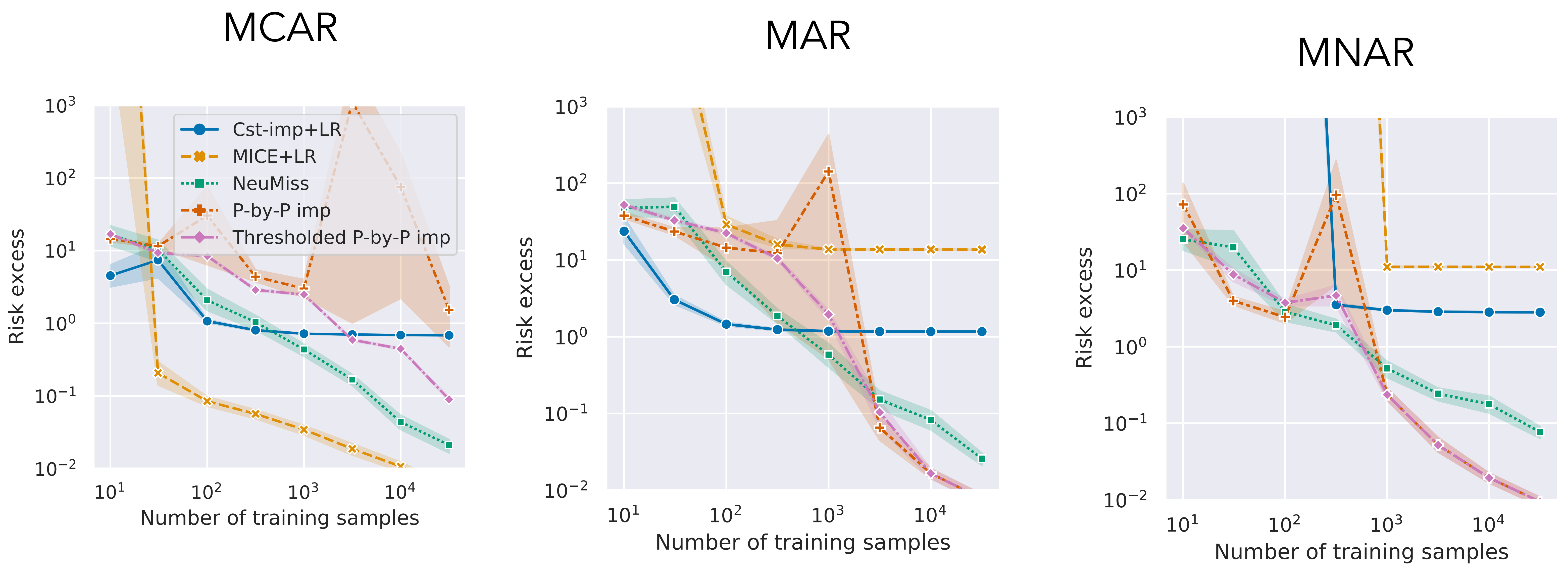
- New thresholded predictor
- Adaptative upper bound
- Near optimal

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

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

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

Special thanks to the Paris City Council for the financial support