Linear prediction with NA, Imputation versus specific methods

Alexis Ayme

Under the supervision of:
Claire Boyer, Aymeric Dieuleveut and Erwan Scornet

Background

- **Growing mass of data => NA (not attributed)/missing values**

- **Different sources:**

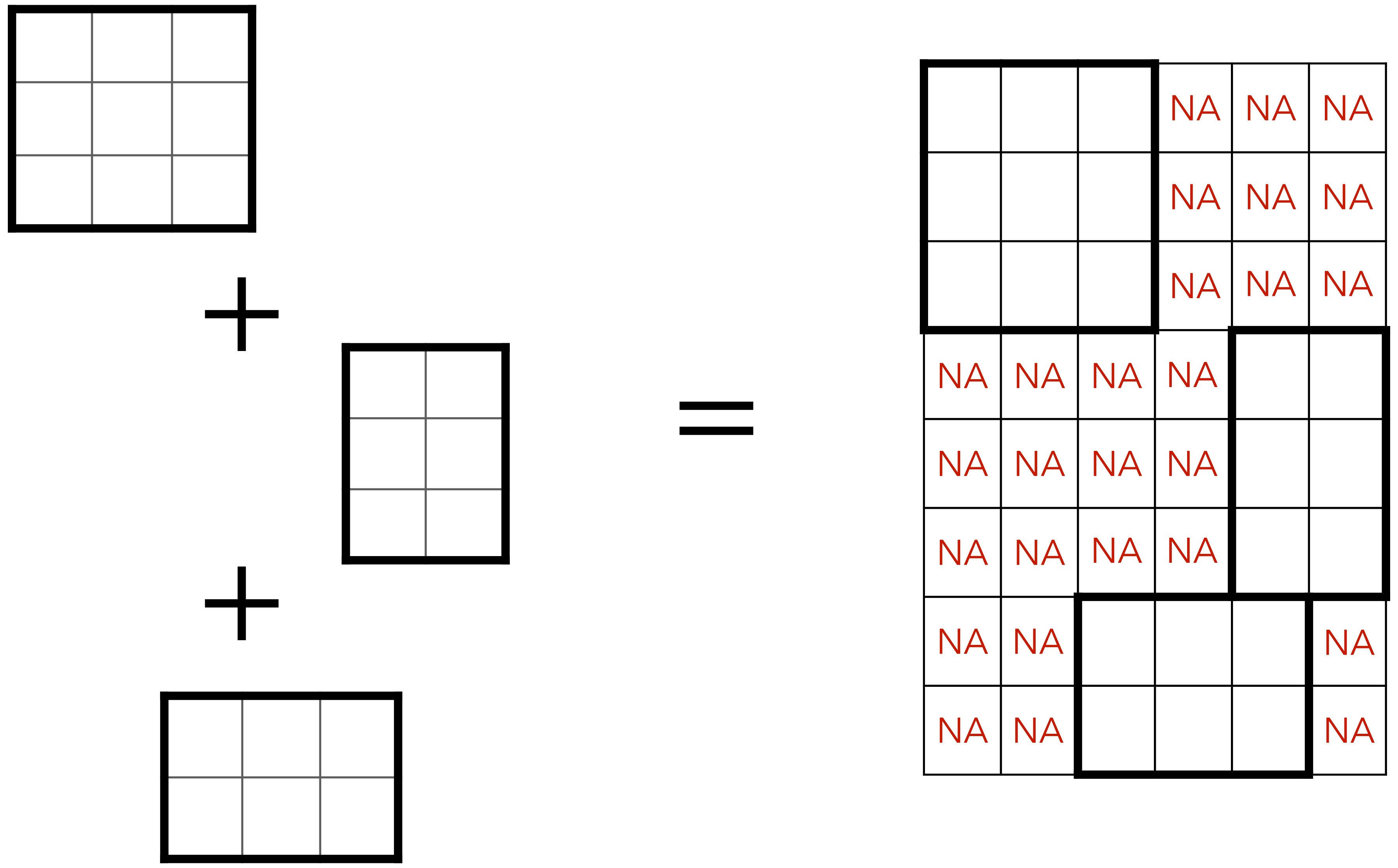
1. Bugs
2. Cost
3. Multiplication of sources (i.e. merging)
4. Sensitive data

Age	Job	**Income**
		NA
		NA
NA		NA

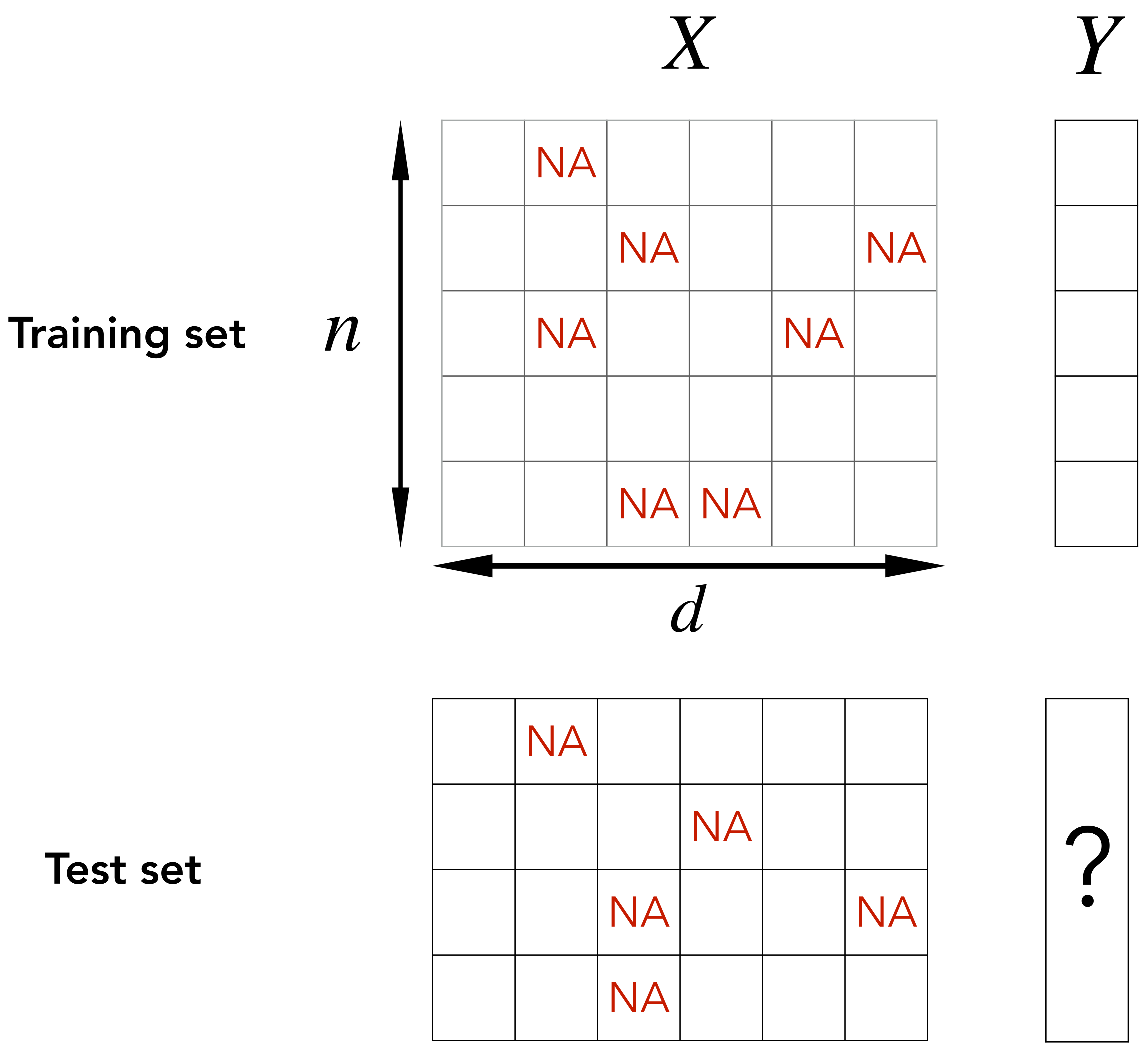
$1	$10	$100	$0
		NA	
	NA	NA	
		NA	
	NA	NA	

- **Growing mass of data => High-dimensional dataset**

1. Cost
2. Multiplication of sources (i.e. merging)
3. Genotype, text



Introduction: Supervised learning with missing values (NA)

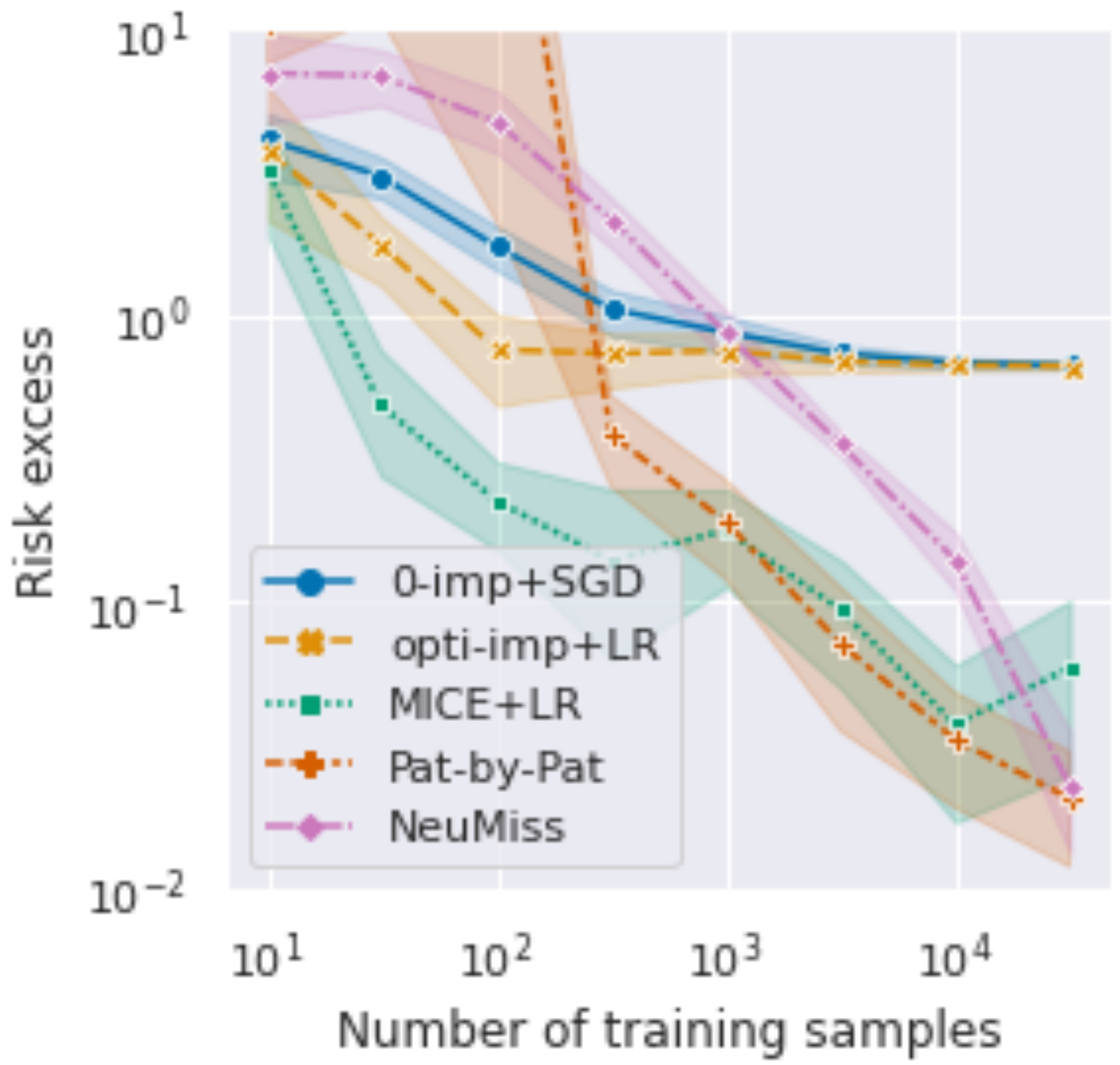


Introduction: Supervised learning with missing values (NA)

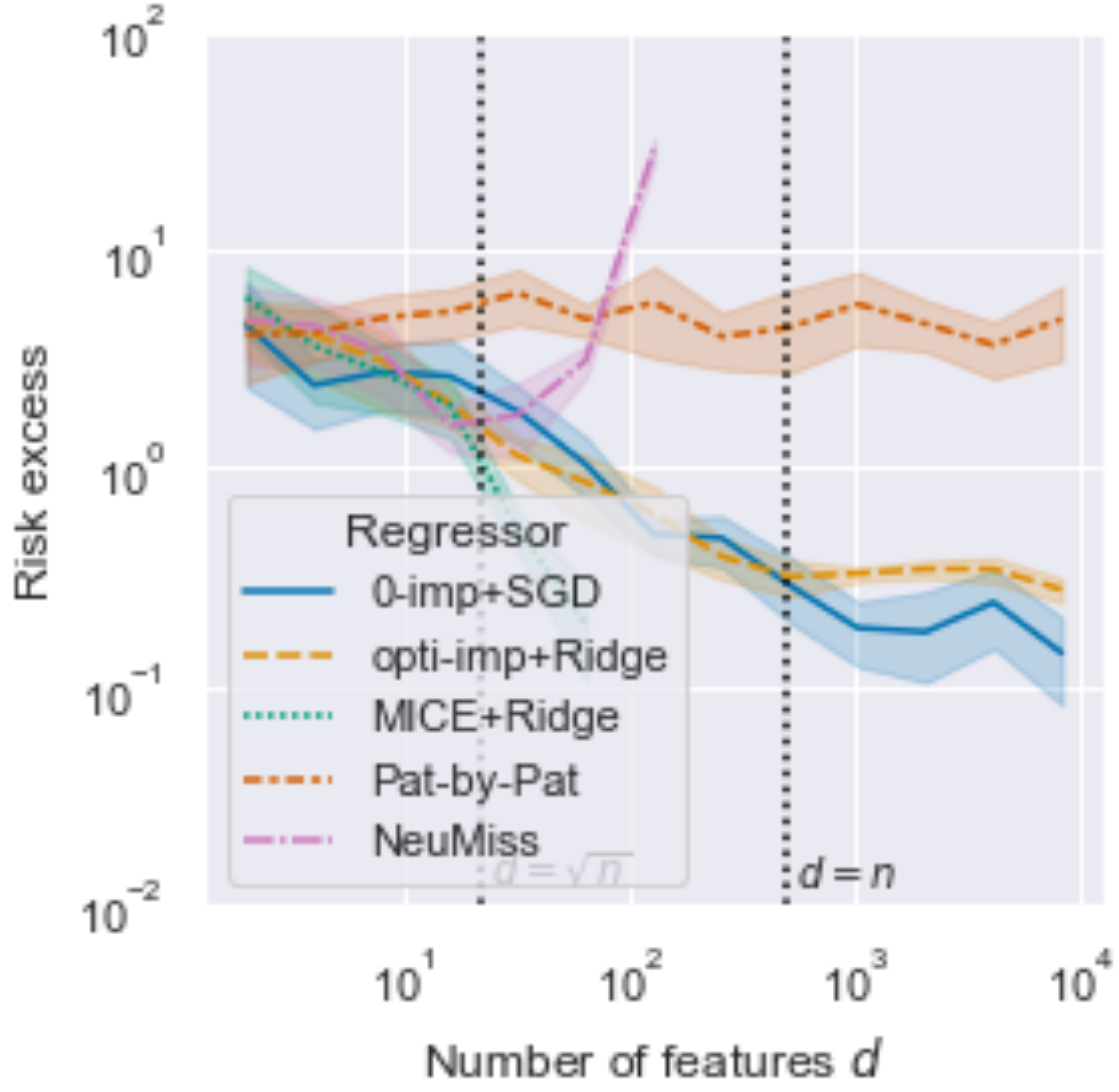
- **Handle missing values with:**

1. Impute then regress procedure
2. Specific method

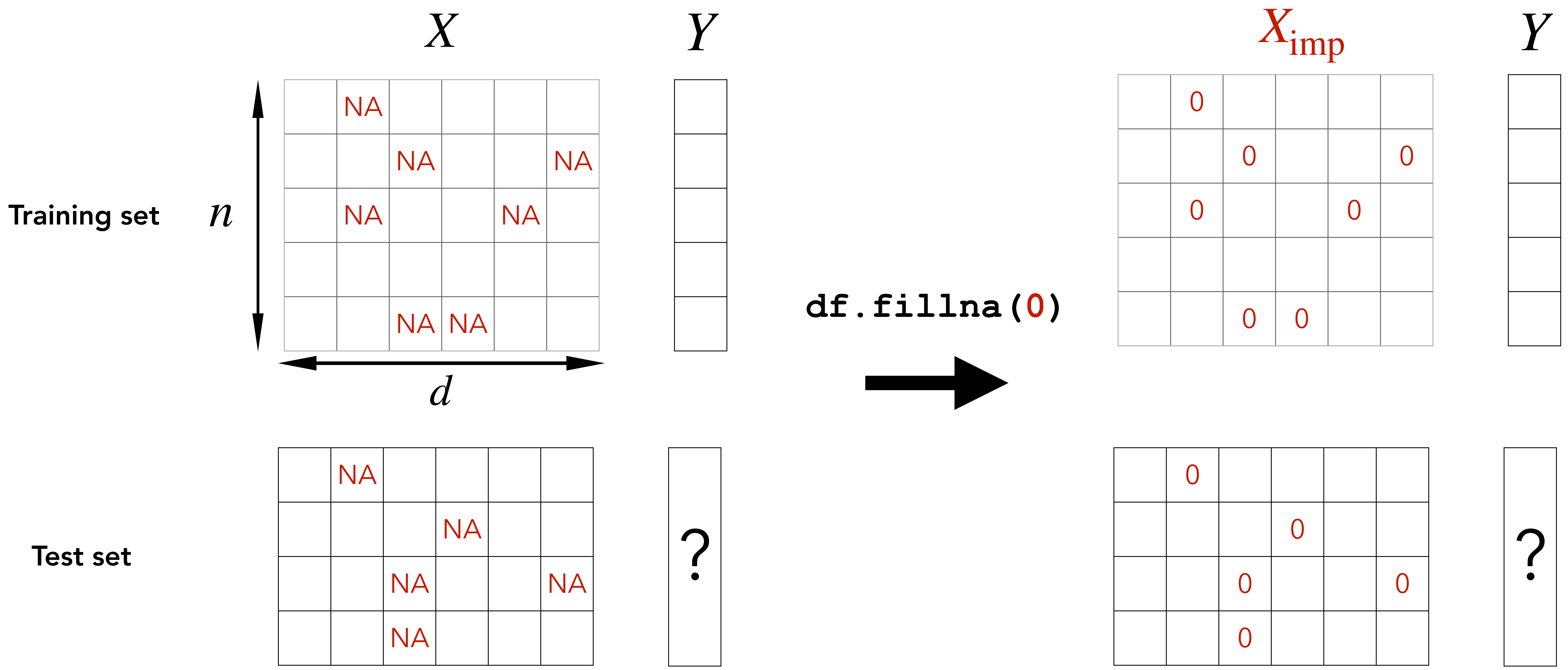
- **Low dimension $n \to +\infty$**



- **High dimension $d \to +\infty$**

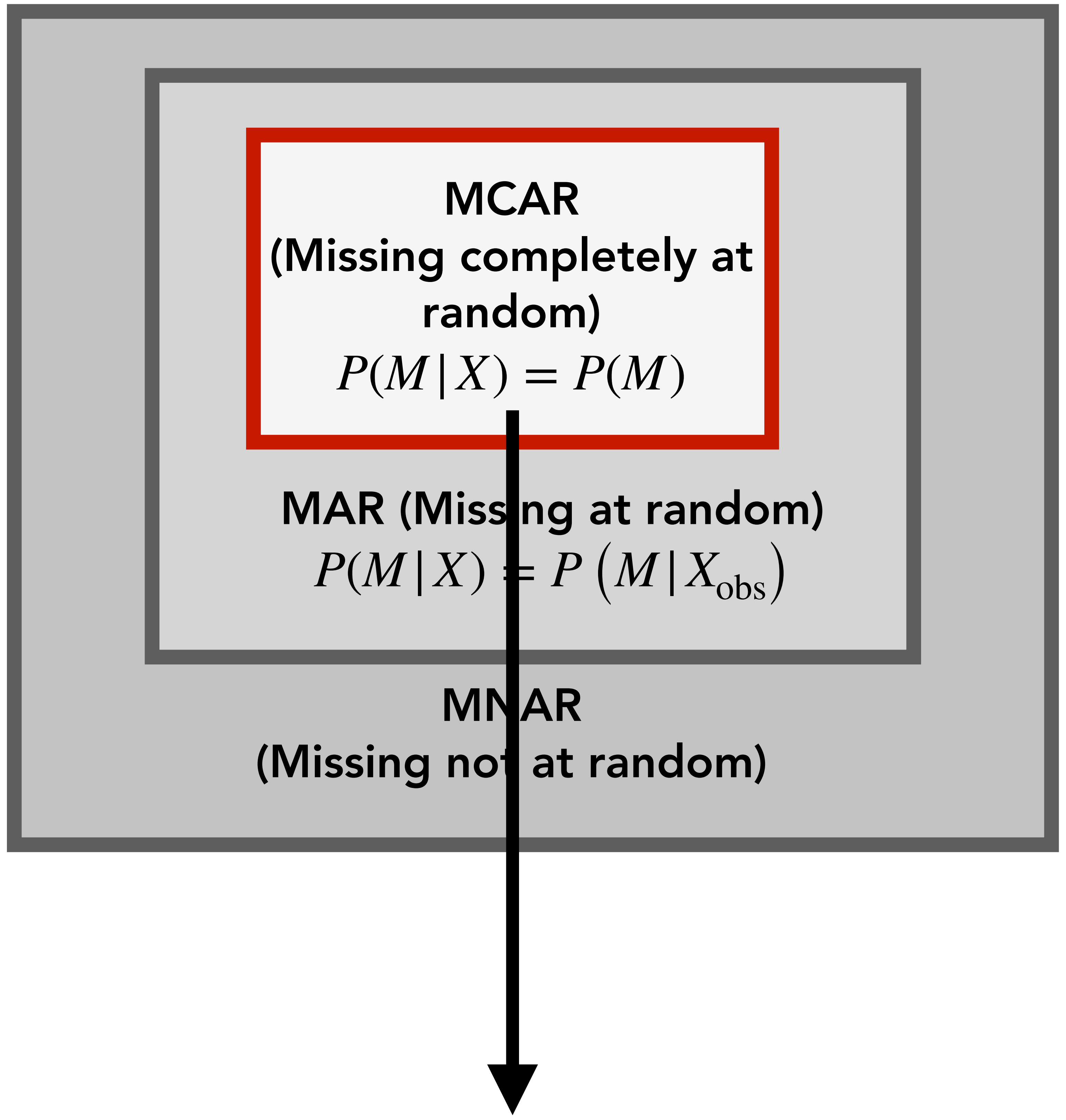


Introduction: Naive imputation



Introduction

- **Missing values**



Bernoulli Model: Missing values
i.i.d $M_1, \dots, M_d \sim \mathscr{B}(1-\rho)$

- **In this talk:**

1) High-dimensional dataset

$$d \gg n$$

2) Predict on **imputed data**

$$R_{\mathrm{imp}}(f) = \mathbb{E}_{X,Y}\left[\left(Y - f(X_{\mathrm{imp}})\right)^2\right]$$

3) Comparison with the complete case

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

Definition

- **Linear model** (well/miss-specified): $\beta \in \mathbb{R}^d, \mathbb{E}[\epsilon X] = 0$

$$Y = \theta_\star^\top X + \epsilon$$

- **Bayes risk**:

$$R^\star = \inf_\theta R(\theta)$$

$$R^\star_{\text{imp}} = \inf_\theta R_{\text{imp}}(\theta)$$

Imputation bias:

$$B_{\text{imp}} = R^\star_{\text{imp}} - R^\star$$

Imputation by 0 = implicit ridge?

- **Ridge penalization**

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

Theorem 2: Under Bernoulli model and $\Sigma_{j,j} = 1$ for all $j \in [d]$,

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

1. Optimal predictor has a small norm
2. **TAKE AT HOME: Imputation induce a Ridge penalization**

- **Ridge bias**

$$B_{\text{ridge},\lambda} = \inf_\theta\{R(\theta) - R(\theta_\star) + \lambda\|\theta\|_2^2\}$$

Theorem 2: Under Bernoulli model and $\Sigma_{j,j} = 1$ for all $j \in [d]$,

$$B_{\text{imp}} = B_{\text{ridge},\lambda_{\text{imp}}}$$

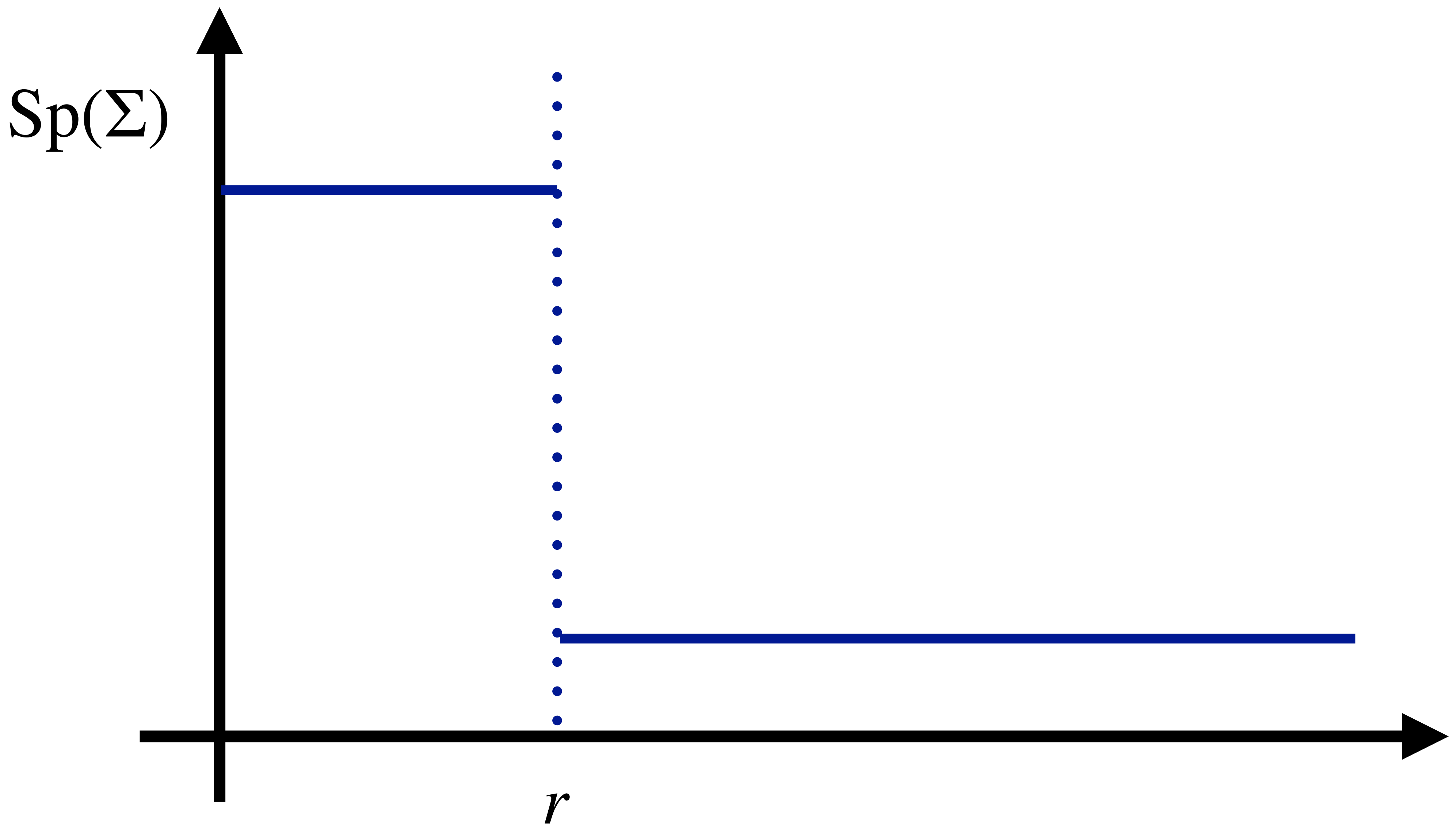
where $\lambda_{\text{imp}} = \dfrac{\rho}{1-\rho}$

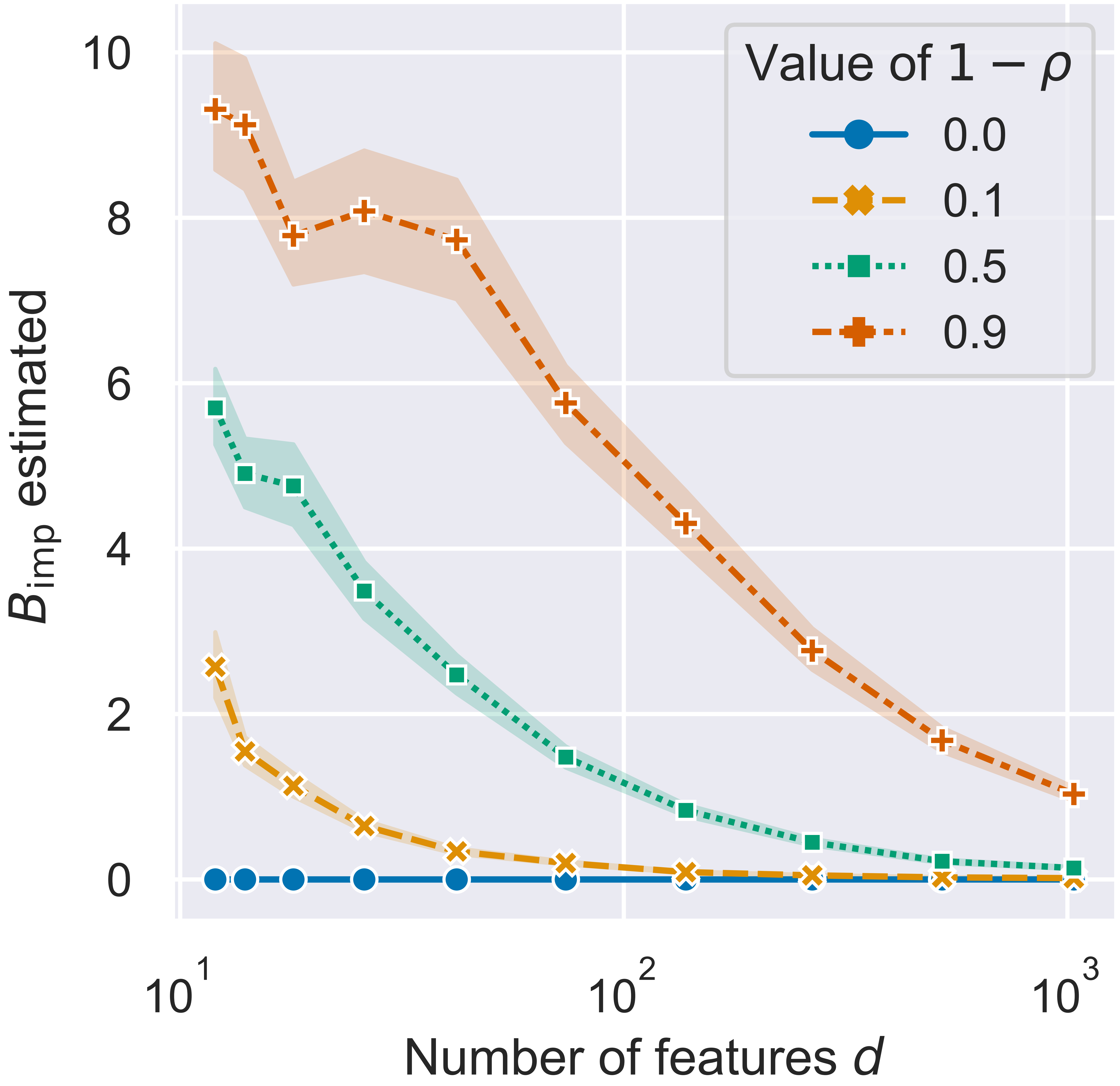
1. No strong assumptions on X
2. λ_{imp} depends only on $1-\rho$ the proportion of missing values.
3. Available for all MCAR setting with another λ_{imp}
4. Bias decreases with the dimension
5. **TAKE AT HOME: MCAR missing values seem to be at the same price of Ridge penalization**

Illustration: on low rank data

- **Low rank data (or Spiked):** $\mathrm{rank}(\Sigma) \approx r$

$$B_{\mathrm{imp}} \lesssim \frac{r}{d}\mathbb{E}Y^2$$





Learn imputed data with SGD

- **SGD recursion:** with learning rate $\gamma = \dfrac{1}{d\sqrt{n}}$

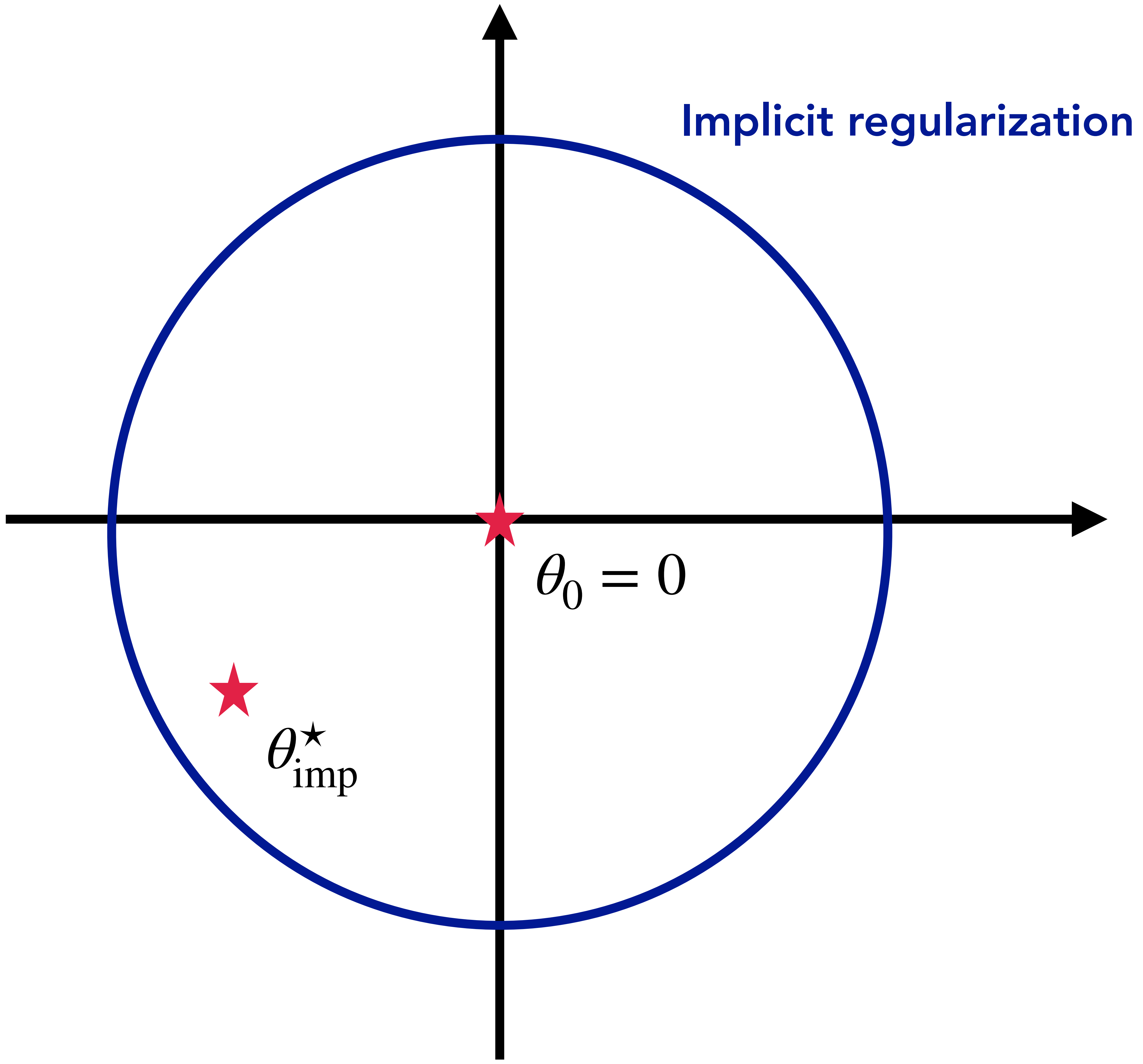
$$\begin{cases} \theta_0 = 0 \\ \theta_{\mathrm{imp},t} = \left[I - \gamma X_{\mathrm{imp},t} X_{\mathrm{imp},t}^\top\right] \theta_{\mathrm{imp},t-1} + \gamma Y_t X_{\mathrm{imp},t} \end{cases}$$

- **Polyak Ruppert average:**

$$\bar{\theta}_{\mathrm{imp},n} = \frac{1}{n+1} \sum_{t=1}^{n} \theta_{\mathrm{imp},t}$$

Theorem 2: Under classical SGD assumptions,

$$\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp,n}}\right)\right] - R^\star \leq B_{\mathrm{imp}} + \frac{d}{\sqrt{n}} \|\theta_{\mathrm{imp}}^2\|_2^2 + \frac{\sigma^2}{\sqrt{n}}$$



Learn imputed data with SGD

- **SGD recursion:**

$$\begin{cases} \theta_0 = 0 \\ \theta_{\mathrm{imp},t} = \left[I - \gamma X_{\mathrm{imp},t} X_{\mathrm{imp},t}^\top\right] \theta_{\mathrm{imp},t-1} + \gamma Y_t X_{\mathrm{imp},t} \end{cases}$$

- **Polyak Ruppert average:**

$$\bar{\theta}_{\mathrm{imp},n} = \frac{1}{n+1} \sum_{t=1}^{n} \theta_{\mathrm{imp},t}$$

Theorem 2: Under classical SGD assumptions,

$$\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp,n}}\right)\right] - R^\star \leq B_{\mathrm{imp}} + \frac{d}{\sqrt{n}} \|\theta_{\mathrm{imp}}^2\|_2^2 + \frac{\sigma^2}{\sqrt{n}}$$

- **Illustration on low rank:**

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

1. We leverage on implicit regularization.
2. Streaming online (one passe)
3. Trade-off between imputation bias and initial condition.
4. **TAKE AT HOME: Imputation bias vanishes for** $d \gg \sqrt{n}$

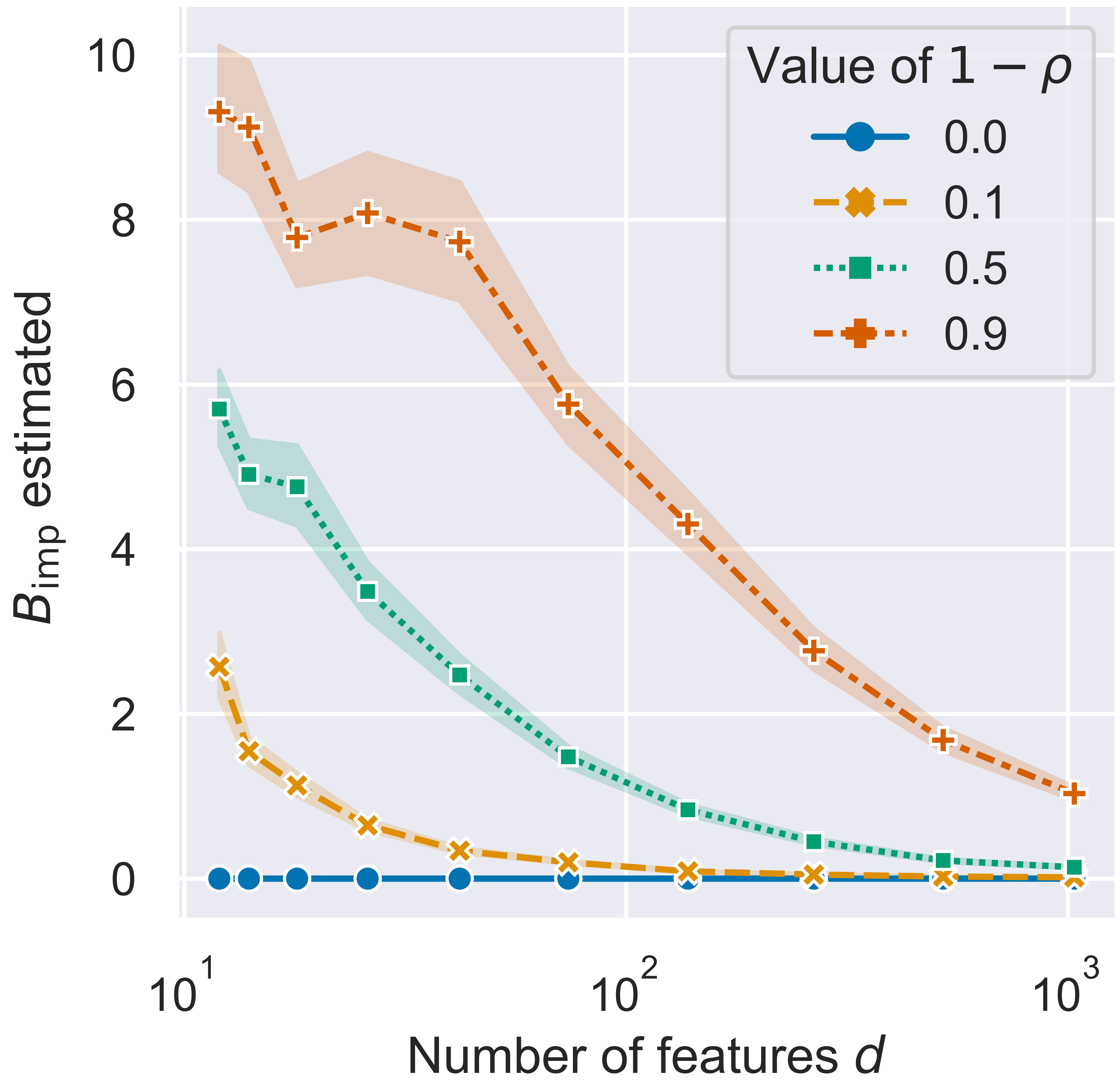
Conclusion

1. **Imputation (even very cheap) out-performs specific method to handle missing values in high-dimension.**

2. **Imputation by 0 induce a Ridge penalization.**

3. **Imputation bias vanishes with dimension as a consequence missing values are not an issue in high dimension.**

4. $d \gg \sqrt{n}$ **regime leads to slow rates of consistency.**



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

Toy example

- **Complete Model:**

$$Y = X_1 \; .$$

$$X = (X_1, X_1, \ldots, X_1)$$

$$M_1, \ldots, M_d \sim \mathscr{B}(1/2).$$

$$R^\star = 0$$

- **With imputed missing values:**

$$\theta_1 = (1,0,\ldots,0)^\top$$

$$\downarrow$$

$$\theta_1^\top X_{\mathrm{imp}} = X_1 M_1$$

$$\downarrow$$

$$R(\theta_1) = \frac{1}{2}\mathbb{E}[X_1^2]$$

$$\theta_2 = 2(1/d, 1/d, \ldots, 1/d)^\top$$

$$\downarrow$$

$$\theta_2^\top X_{\mathrm{imp}} = \frac{2X_1}{d}\sum_j M_J$$

$$\downarrow$$

$$R(\theta_2) = \frac{1}{d}\mathbb{E}[X_1^2]$$

$$\downarrow$$

$$\boxed{B_{\mathrm{imp}} = R^\star - R_0^\star \leq \frac{1}{d}\mathbb{E}[X_1^2]}$$