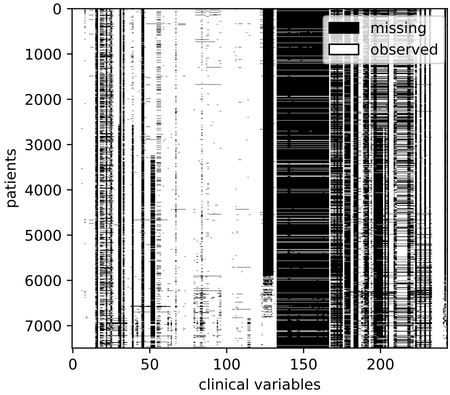
Going beyond the fear of emptyness to gain consistency

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, Julie Josse, Marine Le Morvan, **Erwan Scornet**, Gael Varoquaux

LABORATOIRE DE PROBABILITÉS STATISTIQUE & MODÉLISATION

Incomplete data is ubiquitous in many fields

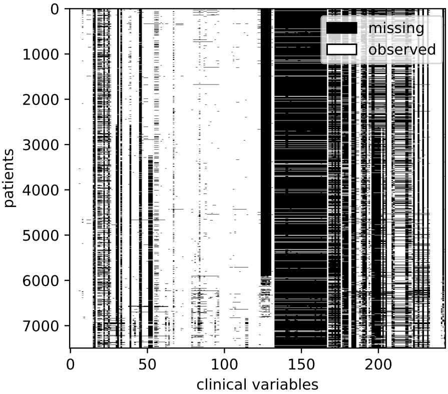


Traumabase clinical records.

Sources of missingness:

- Survey nonresponse.
- Sensor failure.
- Changing data gathering procedure.
- Database join.
- ...

Incomplete data is ubiquitous in many fields



Traumabase clinical records.

Sources of missingness:

- Survey nonresponse.
- Sensor failure.
- Changing data gathering procedure.
- Database join.
- ...

An $n \times p$ matrix, each entry is missing with probability 0.01

- $p = 5 \implies \approx 95\%$ of rows kept;
- $p = 300 \implies \approx 5\%$ of rows kept.

Missing data and linear models

- Classic literature focuses on estimation and imputation (Rubin 76) via
  - Likelihood based methods under MAR.
  - Multiple imputation under MAR.

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

Estimation vs prediction: what is the difference?

1. **Estimation**:
   - provide an estimate of $\beta^\star$

$\rightarrow$ Inference, and prediction with complete data.

Estimation vs prediction: what is the difference?

1. **Estimation**:
   - provide an estimate of $\beta^\star$

   $\rightarrow$ Inference, and prediction with complete data.

2. **Prediction**:
   - We want to predict Y for a new X with missing entries

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

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

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:

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

- 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

Summary

Impute-then-Regress procedures

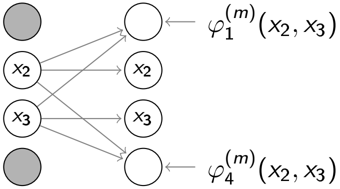
- Impute-then-Regress procedures consist in
  1. Impute missing values
  2. train a supervised learning algorithm on the imputed data set.

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:

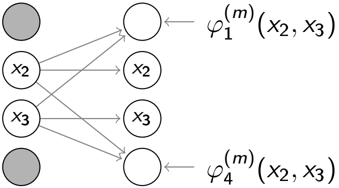


Impute-then-Regress procedures

- 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

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

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

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

Theorem (Le Morvan et al., 2021)

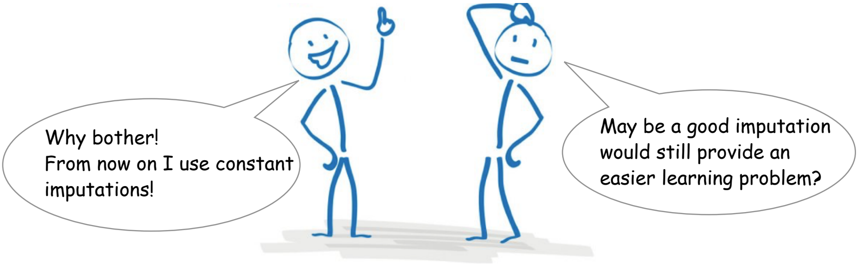
Assume that X admits a density, the response Y is generated as $Y = f^\star(X) + \varepsilon$ and $\Phi \in \mathcal{F}^I_\infty$ (C^∞ imputation functions). Then,

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

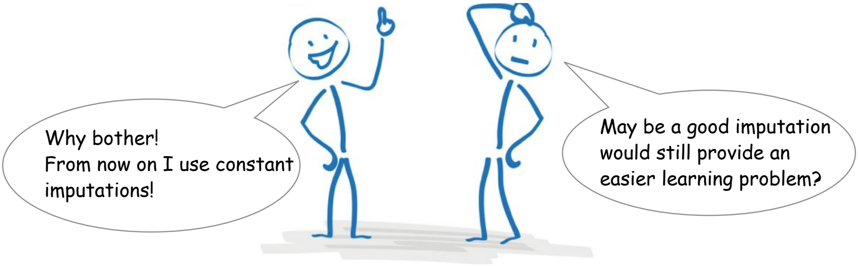
$g^\star_\Phi \circ \Phi$ *is Bayes optimal.*

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.,* $P(M = 0, X = x) > 0$*, 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.

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.

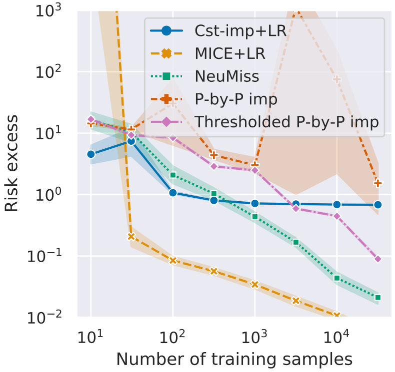
Here, the submodel for which only X_1 is observed is not linear.

$\Rightarrow$ 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.

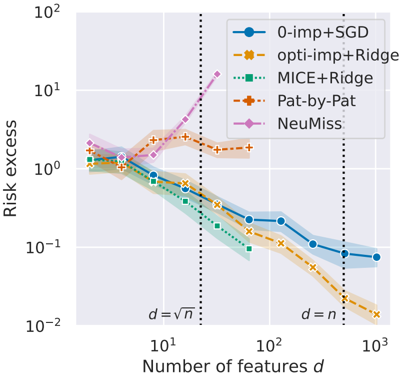
Handling missing values in linear models for prediction

2 possible approaches

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

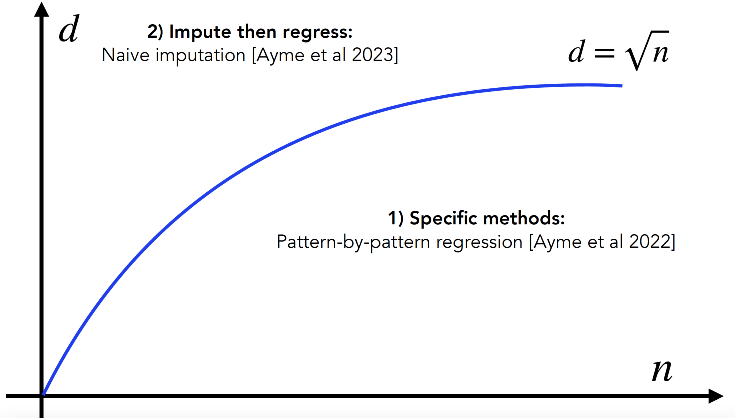


Fixed dimension



Fixed sample size

Different strategies for prediction



Summary

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 $\begin{cases} X|(M=m) \sim \mathcal{N}(\mu^m, \Sigma^m) \\ \text{Linear model} \\ \mathbb{P}[M=m] = p_m \end{cases}$

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

A lower bound

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

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

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

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

A lower bound

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

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

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

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

Examples

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

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)

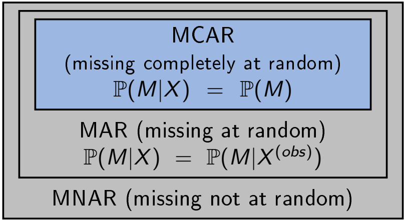
Summary

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

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

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

correlation $\Rightarrow$ 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:

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

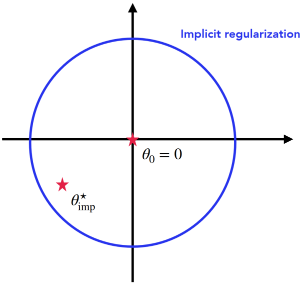
correlation $\Rightarrow$ low imputation/missing values error !

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,

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

- Example: low-rank setting

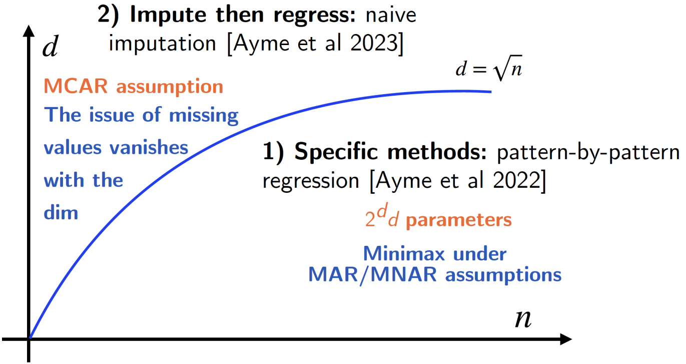
$$\mathbb{E}\left[R_{\text{imp}}(\bar{\theta}_{\text{imp},n})\right] - R^\star \lesssim \left(\frac{1}{\rho\sqrt{n}} + \frac{1-\rho}{d}\right)\frac{r}{d}\mathbb{E}Y^2 + \frac{\text{noise variance}}{\sqrt{n}}$$

- 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 $\lambda = \frac{\text{1-observation rate}}{\text{observation rate}}$



Summary

Toy example of Random features

- Latent observations (hidden) $Z \in \mathbb{R}^p$ with $p = 4$:

$$Z = (\text{age, weight, height, hair color})$$

- Target: $Y = \beta^\top Z +$ noise
- We take **randomly** d features of Z to obtain X:
  - Low dimension $d = 2$:

$$X = (\text{age, height})$$

uncorrelated regime

  - High dimension $d = 10$:

$$X = (\text{age, height, height, age, weight, hair color, weight, age, height})$$

correlated regime

First random features models

Gaussian random features:

- Input: $X_{i,j} = Z_i^\top W_j$

  Latent variables $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} \mathcal{N}(0, I_p)$

  Random weights $W_1, \ldots, W_d \overset{i.i.d.}{\sim} \mathcal{U}(\mathbb{S}^{p-1})$
- Output: $Y_i = Z_i^\top \beta^\star +$ noise of variance σ^2

First random features models

Gaussian random features:

- Input: $X_{i,j} = Z_i^\top W_j$

  Latent variables $Z_1, \dots, Z_n \overset{i.i.d.}{\sim} \mathcal{N}(0, I_p)$

  Random weights $W_1, \dots, W_d \overset{i.i.d.}{\sim} \mathcal{U}(\mathbb{S}^{p-1})$
- Output: $Y_i = Z_i^\top \beta^\star +$ noise of variance σ^2

Key quantities:

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

$$\Delta_{\mathrm{miss}}(d) := [R^\star_{\mathrm{miss}}(d) - R^\star(d)]$$

- $R^\star_{\mathrm{imp}}(d)$ = optimal risk of linear prediction after imputation by 0

$$\Delta_{\mathrm{imp/miss}}(d) := \left[R^\star_{\mathrm{imp}}(d) - R^\star_{\mathrm{miss}}(d)\right]$$

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

The story of naive imputation and missing values

Theorem [Ayme, Boyer, Dieuleveut, Scornet 2024]

Under MCAR assumptions,

- Optimal risk without missing data

$$[R^\star(d)] = \begin{cases} \sigma^2 + \frac{p-d}{p}\|\beta^\star\|_2^2, & \text{when } d < p \\ \sigma^2 & \text{when } d \geqslant p \end{cases}$$

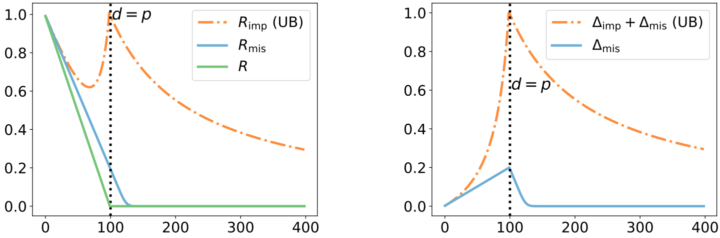
- Error due to missing data

$$\begin{cases} \Delta_{\text{miss}}(d) = (1-\rho)\frac{d}{p}\|\beta^\star\|_2^2 & \text{when } d < p \\ \Delta_{\text{miss}}(d) \leqslant c_{\rho,p}^d \|\beta^\star\|_2^2, & \text{when } d \geqslant p \quad (\text{with } c_{\rho,p} < 1) \end{cases}$$

- Error due to linear prediction on imputed data

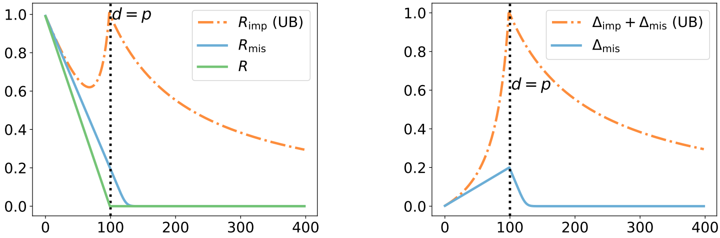
$$\begin{cases} \Delta_{\text{imp/miss}}(d) \leqslant \frac{\rho(d-1)}{p-\rho(d-1)-2}\Delta_{\text{miss}}(d) & \text{when } d < p \\ \Delta_{\text{imp/miss}}(d) + \Delta_{\text{miss}}(d) \leqslant \frac{p}{\rho d + (1-\rho)p}\|\beta^\star\|_2^2 & \text{when } d \geqslant p \end{cases}$$

The story of naive imputation and missing values



- Low dimensions (uncorrelated regime):
  - Missing values error represents $1 - \rho$ of the explained variance without missing values: missing features are lost
  - Error due to imputation is negligible: imputation is optimal

The story of naive imputation and missing values



- Low dimensions (uncorrelated regime):
  - Missing values error represents $1 - \rho$ of the explained variance without missing values: missing features are lost
  - Error due to imputation is negligible: imputation is optimal
- High dimensions (correlated regime):
  - Error due to missing values error decreases exponentially fast: missing features can be retrieve from the others
  - Extension of the low rank setting for the imputation bias: correlation $\Rightarrow$ low imputation bias

Extension of the high dimensions result

$\lim_d \Delta_{\mathrm{imp/miss}}(d) + \Delta_{\mathrm{miss}}(d) = 0$ **still holds**, for instance when

- General random features:
  - **Non-linear** inputs: $X_{i,j} = \psi(Z_i, W_j)$
  - **Non-linear** output: $Y = f^\star(Z) + \varepsilon$ with $f^\star$ continuous

  Ex: Random Fourier features (RFF)

$$W_j = (A_j, B_j) \sim \mathcal{N}(0, I) \otimes \mathcal{U}([0, 2\pi])$$
$$X_{i,j} = \cos(A_j^\top Z_i + B_j)$$

- **Non-MCAR** missing values:

  Ex: Logistic model on the latent covariate:

$$\mathbb{P}\left(M_j = 1 | Z\right) = \frac{1}{1 + e^{w'_{0j} + w_j'^\top Z}}$$

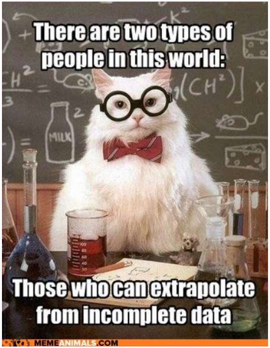
Conclusion

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
  - Linear models - Zero imputation is inconsistent but converges in high-dimensional settings (rate of $\sqrt{d/n}$)
- Pattern-by-pattern strategies - use a different predictor for each missing pattern
  - Consistent by design but intractable in most situations
  - Linear models - Rate of consistency of d^2/n for independent Bernoulli missing indicators **but** $2^d/n$ in general (not improvable)

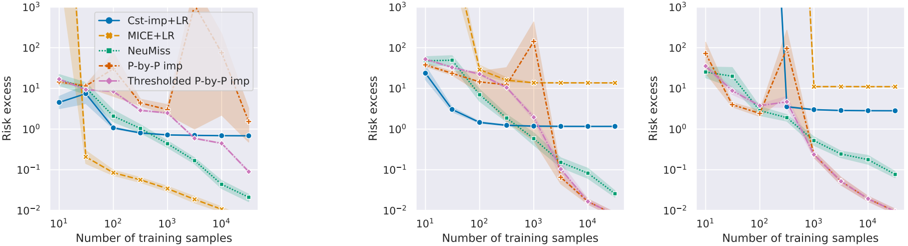
Conclusion



Thank you!

- *Near-optimal rate of consistency for linear models with missing values.* A. Ayme, C. Boyer, A. Dieuleveut, E. Scornet. ICML 2022.
- *Naive imputation implicitly regularizes high-dimensional linear models.* A. Ayme, C. Boyer, A. Dieuleveut, E. Scornet. ICML 2023.
- *Random features models: a way to study the success of naive imputation.* A. Ayme, C. Boyer, A. Dieuleveut, E. Scornet. ICML 2024.

Numerical XP for prediction



	MCAR		MAR		MNAR	
Regressors	Unbiased	Rate	Unbiased	Rate	Unbiased	Rate
Cst-imp+LR	No	Fast	No	Fast	No	Fast
MICE+LR	Yes	Fast	No	Fast	No	Fast
NeuMiss	Yes	Fast	Yes	Slow	Yes	Slow
P-by-P	Yes	Slow	Yes	Slow	Yes	Slow
Tresh. P-by-P	Yes	Slow	Yes	Fast	Yes	Fast