What's a good imputation to predict with missing values?

Marine Le Morvan, Julie Josse, Erwan Scornet, Gaël Varoquaux

Objectives

- **Supervised learning with missing values** poses different challenges compared to inference with missing values or imputation.

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

- In practice, **Impute-then-Regress procedures** are widely used, but there is **very little theoretical grounding** supporting their use.
- We thus ask the following questions:
  - » **Can Impute-then-Regress procedures be Bayes optimal?**
  - » **How should we choose the imputation function?**
  - » **What if the data is Missing Not At Random (MNAR)?**

Bayes optimality

Theorem

Let $g^\star_\Phi$ be the minimizer of the risk on the data imputed by Φ. Assume that $\Phi \in \mathcal{F}^I_\infty$, and that the response Y satisfies $Y = f^\star(X)+\epsilon$. Then, for all missing data mechanisms and almost all imputation functions, $g^\star_\Phi \circ \Phi$ is Bayes optimal.

In other words:

- For almost all imputation functions $\Phi \in \mathcal{F}^I_\infty$, a universally consistent algorithm trained on the imputed data $\Phi(\widetilde{X})$ is Bayes consistent.

$\Longrightarrow$ Asymptotically, it is not necessary to impute well to predict well.

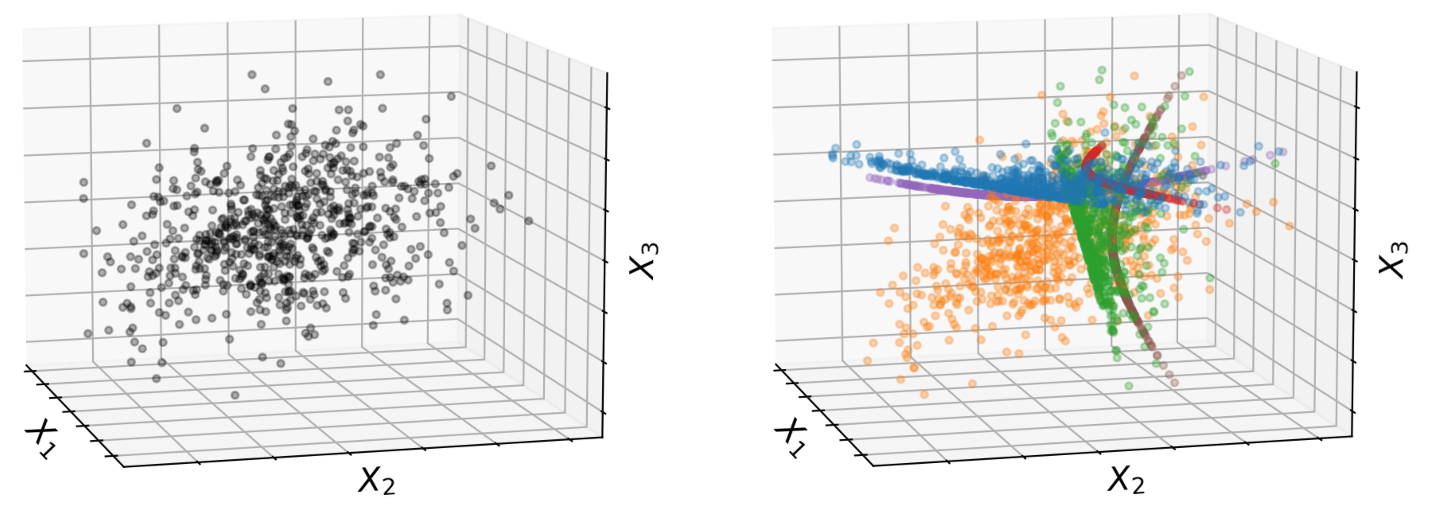
Predictive modeling calls for different imputation strategies.

Sketch of the proof

❶ All data points with a missing data pattern m are mapped to a manifold $\mathcal{M}^{(m)}$ of dimension $|obs(m)|$ (**Preimage Theorem**).

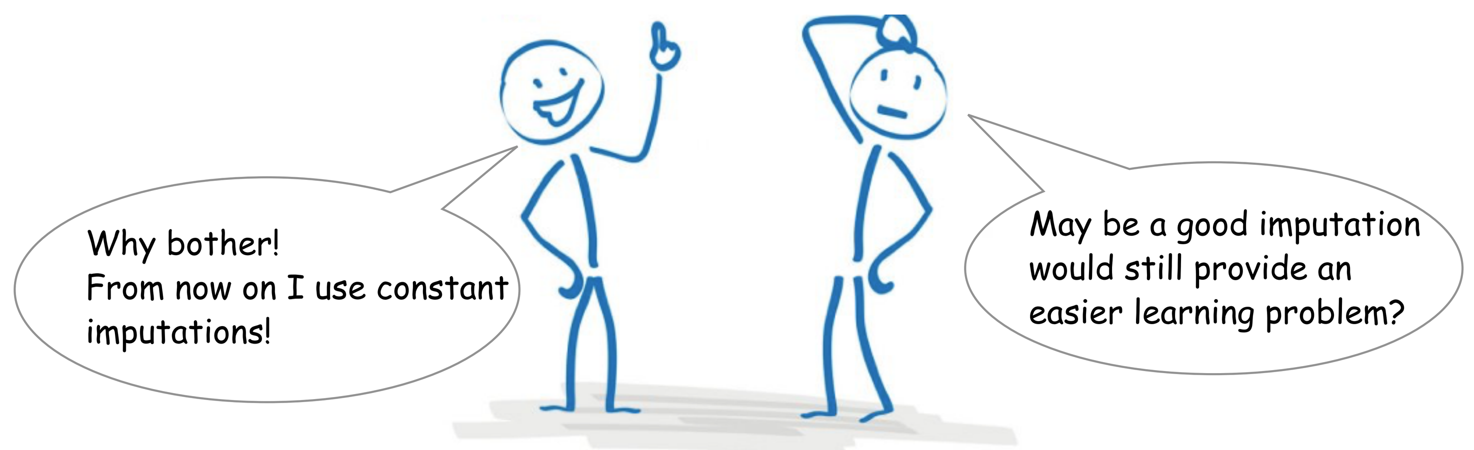
❷ The missing data patterns of imputed data points can almost surely be de-identified (**Thom transversality Theorem**).

❸ Given 2), we can build prediction functions, independent of m, that are Bayes optimal for all missing data patterns.



Complete data — Imputed data (manifolds)

Continuous decompositions



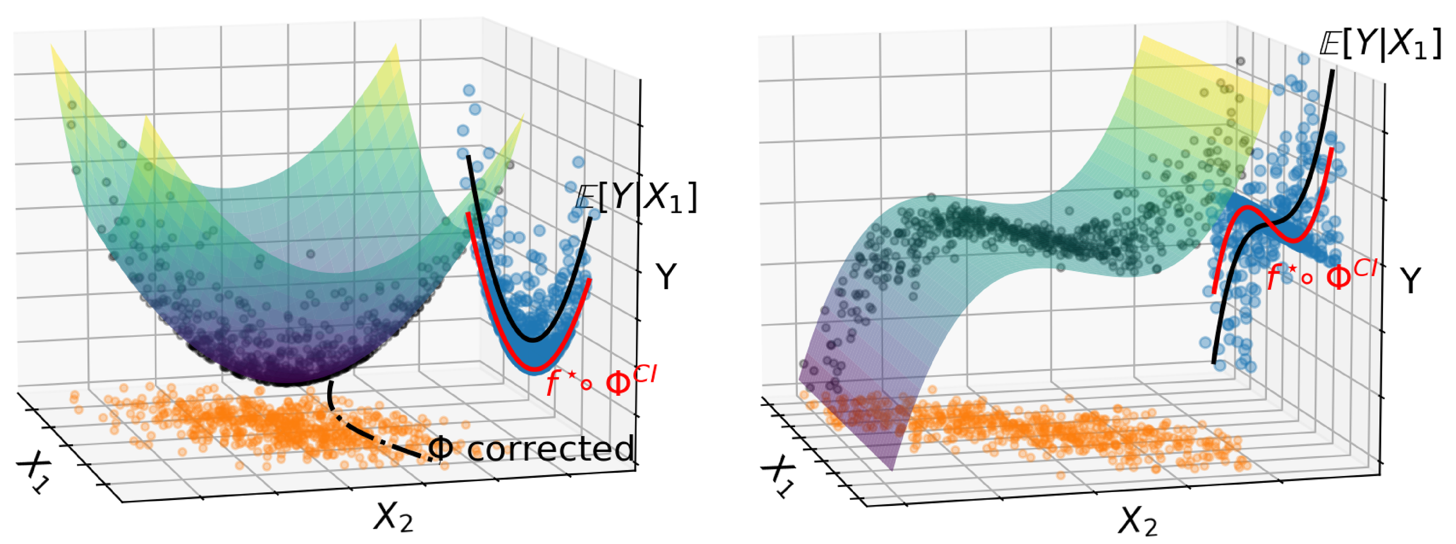
Can we find **continuous** Impute-then-Regress decompositions of the Bayes predictor?

Q1 - What is the risk of chaining oracles: $f^\star \odot \Phi^{CI}$?
where Φ^{CI} is the oracle imputation $\mathbb{E}\left[X_{mis}|X_{obs}\right]$. The excess risk is small whenever there is no direction in which both 1) the curvature of $f^\star$ is high and 2) the variance of the missing data given the observed one is high.

Q2 - Can we find a continuous function g s.t. $g \odot \Phi^{CI}$ is Bayes optimal?
Suppose that the probability of observing all variables is strictly positive. Then there is **no continuous prediction function** g such that $g \odot \Phi^{CI}$ is Bayes optimal, unless it is $f^\star$.

Q3 - Keeping the regression function fixed as $f^\star$, can we find a continuous imputation function Φ so that $f^\star \odot \Phi$ is Bayes optimal?

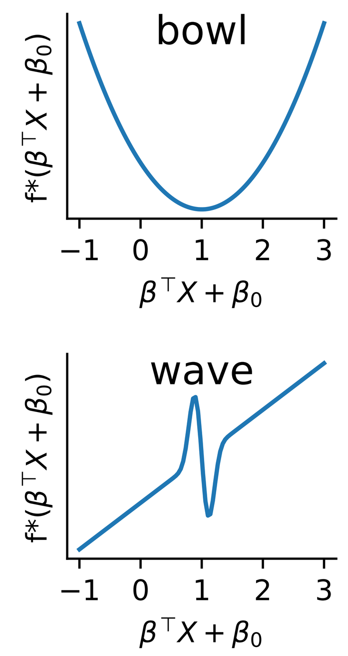
Sometimes yes! — But not always...



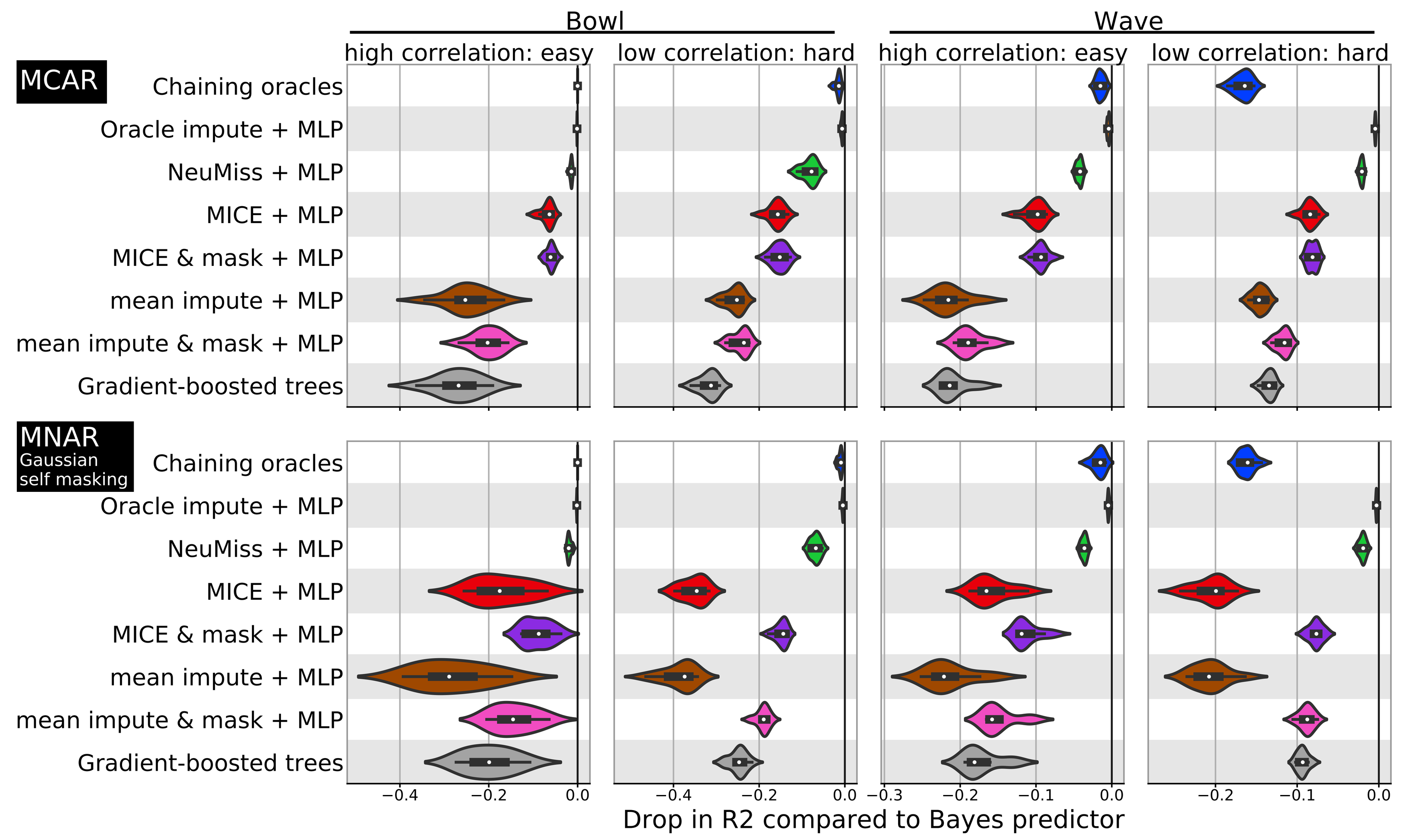
Experimental results

Data simulations

- Gaussian data: "high" and "low' covariance settings.
- $Y = f^\star(X) + \epsilon$
- 50% missing values with 2 mechanisms:
  ✓ MCAR
  ✓ Gaussian self-masking (MNAR)
- n=100,000 and d=50.

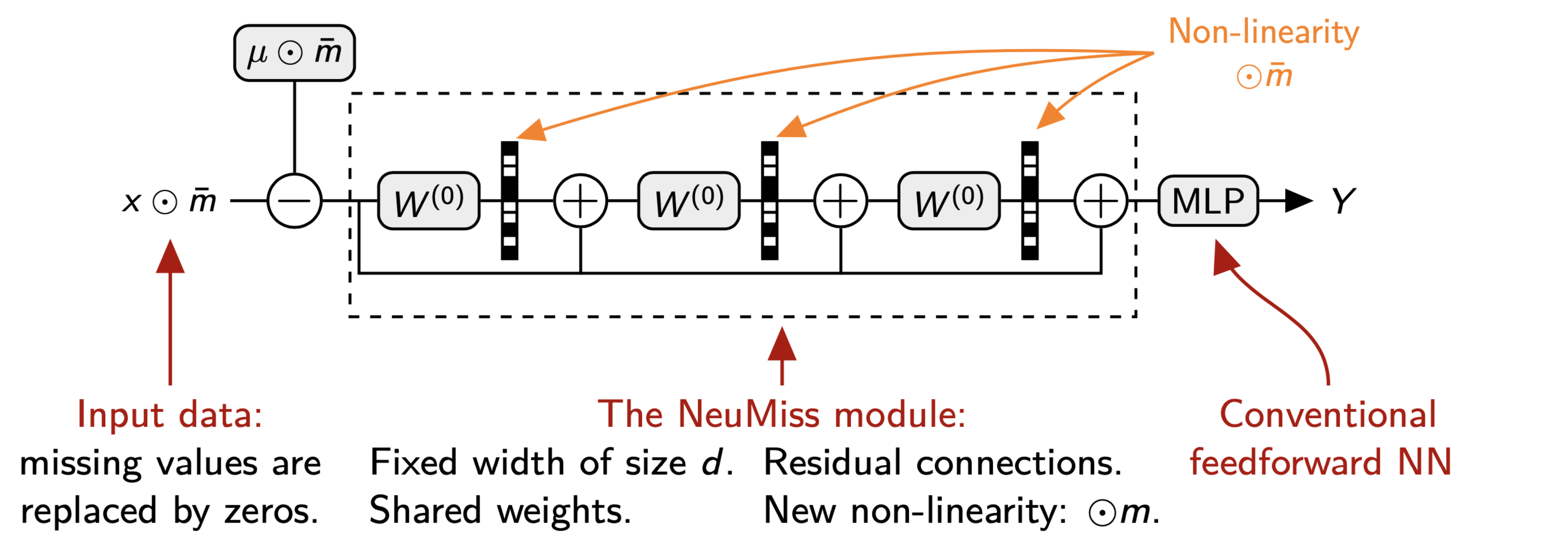


Impute-then-Regress benchmark



Zoom on NeuMiss [1] + MLP: an architecture for missing values

- Can be seen as an implicit and jointly learned Impute-then-Regress architecture for learning with missing values.
- Theoretically grounded: differentiable approximation of the conditional expectation.



[1] Marine Le Morvan et al. "NeuMiss networks: differentiable programming for supervised learning with missing values.". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 5980–5990