NeuMiss networks: differentiable programming for supervised learning with missing values

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

informatics mathematics Inria

Objectives

Supervised learning with missing values:

- both in the training set and **at prediction time** in the test set,
- under possibly **non-ignorable** missingness (Missing Not At Random).

State-of-the-art and Challenges

- A rich literature on statistical inference but few works on supervised learning with missing values.
- Most general methods for imputation are only valid if the missingness is ignorable.
- Samples are represented by varying subsets of input variables $\Rightarrow$ learn compensation mechanisms.
- The total number of possible missing data patterns is exponential in the dimension (2^d) $\Rightarrow$ keep the sample complexity polynomial.

Approach

1. Derive the **analytical expression of the optimal predictor** under various missing data mechanisms.
2. Propose a **theoretically grounded neural network architecture** (NeuMiss) designed to approximate these optimal predictors.

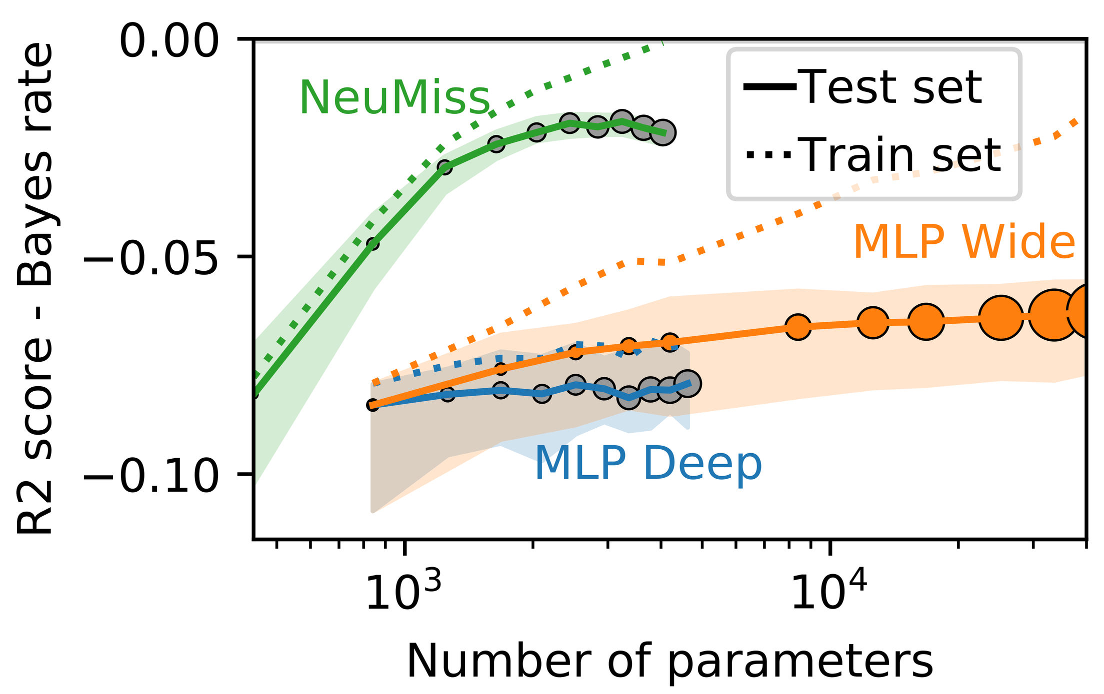

Figure 1: Comparison of NeuMiss with wide and deep neural networks with ReLU activation, fed with the data imputed by 0 and the mask.

Notations and Assumptions

Random variables

- $X \in \mathbb{R}^d$: complete data
- $\widetilde{X} \in \{\mathbb{R} \cup \{\mathtt{NA}\}\}^d$: incomplete data
- $M \in \{0,1\}^d$: mask.
- $obs(M)$: indices of the observed entries

Ex. of realizations

$x = (1.1, 2.3, 3.1, 8, 5.27)$

$\tilde{x} = (1.1, \mathtt{NA}, -3.1, 8, \mathtt{NA})$

$m = (0, 1, 0, 0, 1)$

$x_{obs(m)} = (1.1, 3.1, 8)$

Assumptions:

Linear model: $Y = \beta_0^\star + \sum_{j=1}^{d} \beta_j^\star X_j + \epsilon$

Gaussian data: $X \sim \mathcal{N}(\mu, \Sigma)$

Optimal (Bayes) predictor:

$$f^\star \in \underset{f:(\mathbb{R}\cup\{\mathtt{NA}\})^d \mapsto \mathbb{R}}{\text{argmin}} \mathbb{E}\left[\left(Y - f(\widetilde{X})\right)^2\right]$$

The optimal predictors under various missing data mechanisms

Ignorable missing data mechanisms:

- Missing Completely At Random (MCAR): $P(M = m|X) = P(M = m)$
- Missing At Random (MAR): $P(M = m|X) = P(M = m|X_{obs(m)})$

M(C)AR Bayes predictor

$$f^\star(X_{obs}, M) = \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \langle \beta_{mis}^\star, \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}) \rangle$$

Non-ignorable missing data mechansim:

Gaussian self-masking (MNAR) Bayes predictor

$$f^\star(X_{obs}, M) = \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \langle \beta_{mis}^\star, (Id + D_{mis}\Sigma_{mis|obs}^{-1})^{-1} \times (\tilde{\mu}_{mis} + D_{mis}\Sigma_{mis|obs}^{-1}(\mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}))) \rangle$$

Intuition:

The optimal predictors are linear per pattern

The slopes of the obs. variables depend on M to compensate for the missingness of other correlated variables: $f^\star(X_{obs}, M) = \beta_0(M) + \sum_{j \in \text{obs(M)}} \beta_j(M) X_j$

References

[1] S van Buuren. *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman and Hall/CRC, 2018

[2] Marine Le Morvan et al. "Linear predictor on linearly-generated data with missing values: non consistency and solutions". In: vol. 108. AISTATS. 2020, pp. 3165–3174

Approximating the Bayes predictors

Main difficulty: approx. of Σ_{obs}^{-1}, for any obs, i.e., any missing data pattern!

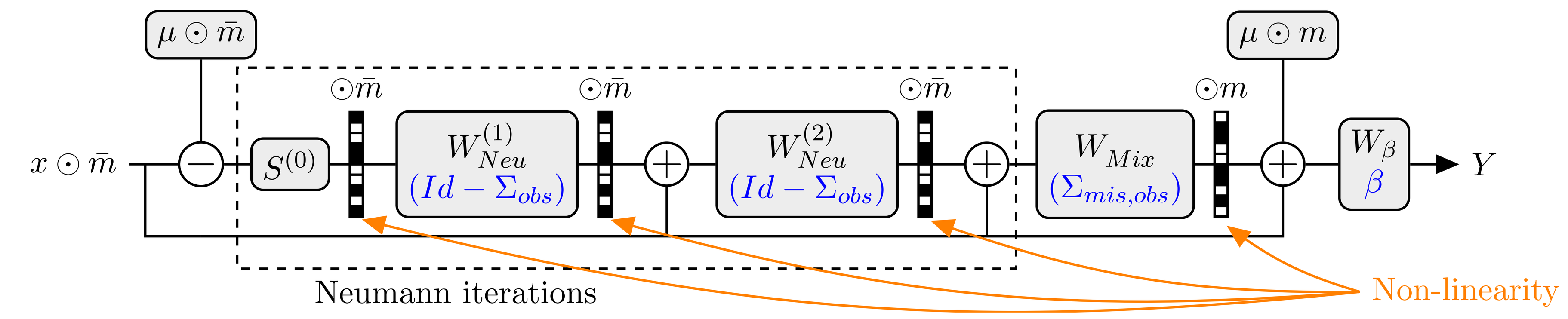

Figure 2: NeuMiss architecture for a depth of 4.

NeuMann iterations: approximate Σ_{obs}^{-1} by unrolling the order-ℓ truncation of a NeuMann series:

$$S_{obs(m)}^{(\ell)} = (Id - \Sigma_{obs(m)}) S_{obs(m)}^{(\ell-1)} + Id.$$

A new type of non-linearity: the multiplication entrywise by the mask.

Theoretically-grounded architecture: In M(C)AR and under a simplifying assumption in Gaussian self-masking, NeuMiss with a depth $\ell + 1$ can exactly compute the order-ℓ approximation of the Bayes predictor.

Experimental results

Simulated data

- Gaussian covariates
- Response is a linear model
- 50% missing values

Methods

- EM: Expectation Maximization.
- MICE [1] + LR: Conditional imputer followed by linear regression.
- MLP [2]: feedforward neural network

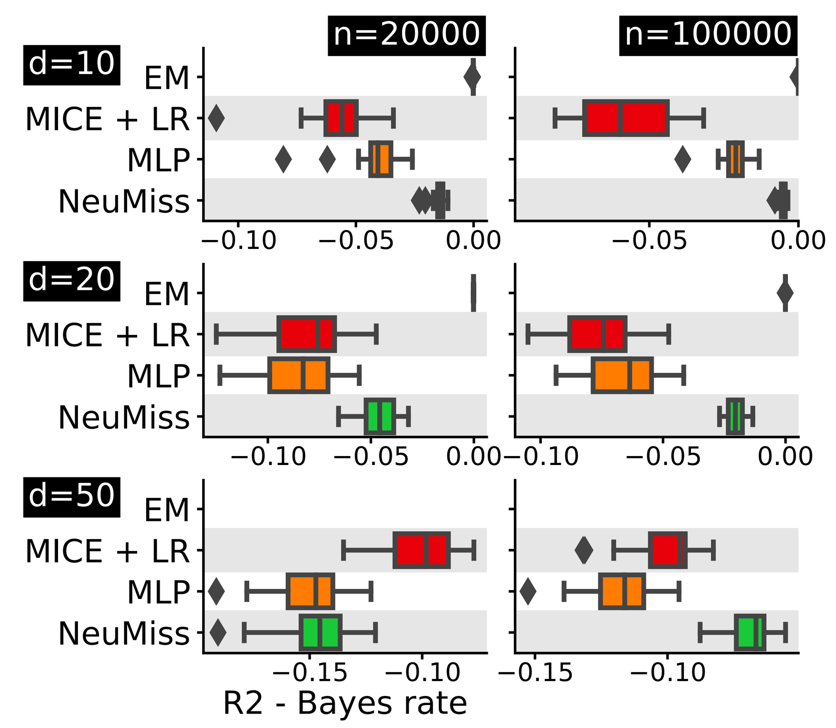

Figure 3: MCAR

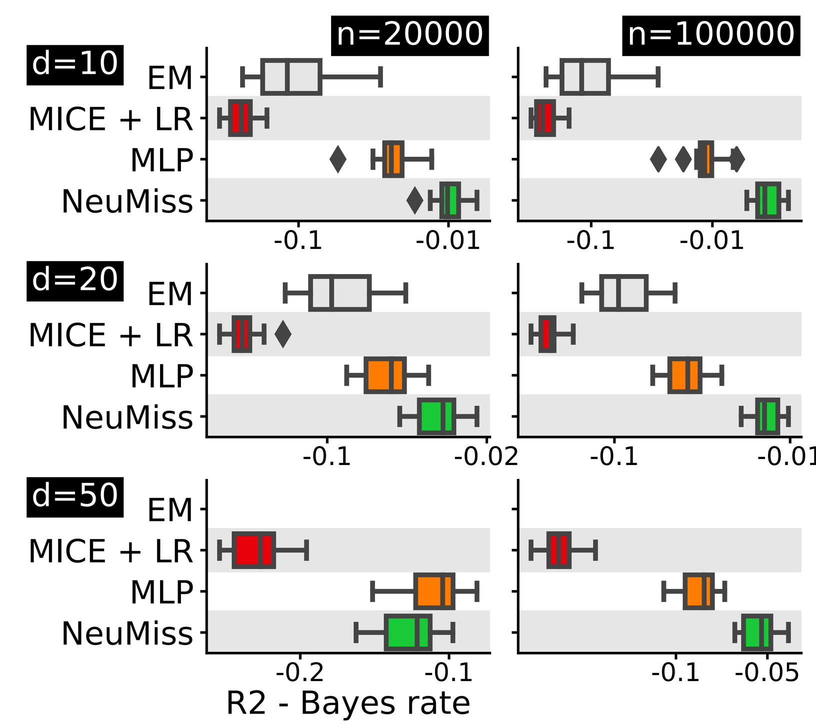

Figure 4: Gaussian self-masking

- Neumiss performances come close to the optimal performances.
- **Robustness to the missing data mechanism.**
- Suited for medium-sized datasets thanks to weights sharing across mdp.

.