NeuMiss networks: differentiable programming for supervised learning with missing values

Marine Le Morvan[1,2]

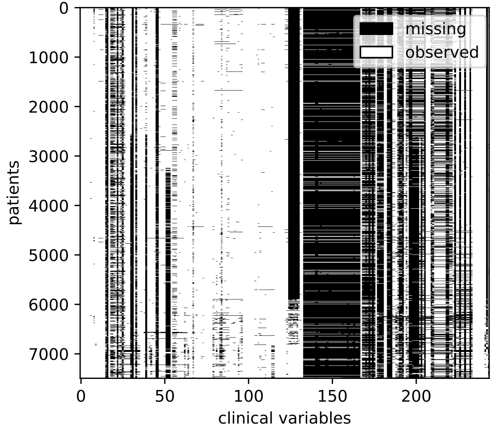
Julie Josse[1,3] **Thomas Moreau**[1]
Erwan Scornet[3] **Gaël Varoquaux**[1, 4]

[1] Université Paris-Saclay, Inria, CEA, Palaiseau, 91120, France
[2] Université Paris-Saclay, CNRS/IN2P3, IJCLab, 91405 Orsay, France
[3] CMAP, UMR7641, Ecole Polytechnique, IP Paris, 91128 Palaiseau, France
[4] Mila, McGill University, Montréal, Canada

Montpellier - December 2020

Incomplete data is ubiquitous in many fields



Traumabase clinical health records.

Sources of missingness:

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

Missing data is frequent in economics, social, political or health sciences.

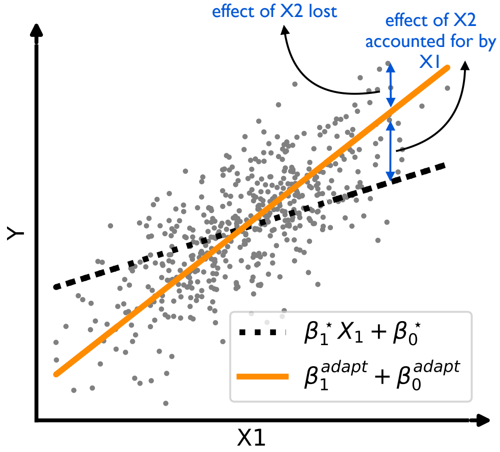
The classical literature on missing values

Since the 70s, an abundant literature on missing data has flourished.

- Missing data mechanisms are usually divided into 3 categories:
  - MCAR (Missing Completely at Random)
  - MAR (Missing at Random)
  - MNAR (Missing Non At Random)
- The literature has been mainly focused on **inference** and **imputation** tasks:
  - Likelihood based methods under MAR.
  - Multiple imputation under MAR.
  - Inverse probability weighting under MAR.

But very few works have addressed **supervised learning** with missing values, whatever the missing data mechanism.

Intuition: linear regression with missing values



$$Y = \beta_1^\star X_1 + \beta_2^\star X_2 + \beta_0^\star$$
$$\mathrm{cor}(X_1, X_2) = 0.5.$$

If X_2 is missing, the coefficient of X_1 should **compensate for the missingness of X_2**.

The difficulty of supervised learning with missing values is to handle **up to** 2^d missing data patterns (i.e. 2^d possible inputs of varying length).

The literature on supervised learning with missing values

- Some recent works:
  - Josse et al. 2019: Imputation by a constant is Bayes consistent, but the function to be learned can be overly complex (hyp: MAR).
  - Le Morvan et al. 2020: In the simple case of linear regression, a single layer MLP is Bayes consistent, but provided 2^d hidden units.
  - Many adaptations of neural networks to missing values, often involving imputing by 0 and concatenating with the mask, but no underlying theory.

What architecture should we use to handle missing values? How complex should it be? What would be a good architecture design?

The NeuMiss network

For the case of linear regression under various missing data mechanisms:

- We propose a **theoretically grounded neural network architecture**, designed to approximate the Bayes predictor.
- The complexity of the architecture stays small thanks to the **sharing of parameters across missing data patterns**.
- Its originality and strength comes from the use of a **new type of non-linearity**: the multiplication by the missingness indicator.
- It is **robust to the missing data mechanism**, including difficult MNAR settings such as self-masking.

Content

1. Optimal predictors in the presence of missing values
2. NeuMiss networks: learning by approximating the Bayes predictor
3. Empirical results

Outline

1. Optimal predictors in the presence of missing values
2. NeuMiss networks: learning by approximating the Bayes predictor
3. Empirical results

Notations and assumptions

Random variables

- $X \in \mathbb{R}^d$: complete data (unavailable)
- $\widetilde{X} \in \{\mathbb{R} \cup \{\texttt{NA}\}\}^d$: incomplete data (available)
- $M \in \{0,1\}^d$: mask.

$obs(M)$ (resp. $mis(M)$) are the indices of the observed (resp. missing) entries.

Notation abuse: $A_{obs(m),obs(m)} = A_{obs(m)}$

Ex. of realizations

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

$\widetilde{x} = (1.1, \texttt{NA}, -3.1, 8, \texttt{NA})$

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

$x_{obs(m)} = (1.1, 3.1, 8)$,
$x_{mis(m)} = (2.3, 5.27)$

Assumptions:

linear model + Gaussian data:

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

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

Bayes predictor:

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

The Bayes predictor under M(C)AR

- **MCAR**: For all $m \in \{0,1\}^d$, $P(M = m|X) = P(M = m)$.
- **MAR**: For all $m \in \{0,1\}^d$, $P(M = m|X) = P(M = m|X_{obs(m)})$.

Proposition (M(C)AR Bayes predictor)

Under the linear model and Gaussian data assumptions, and a MCAR or MAR missing data mechanism, the Bayes predictor $f^\star$ takes the form:

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

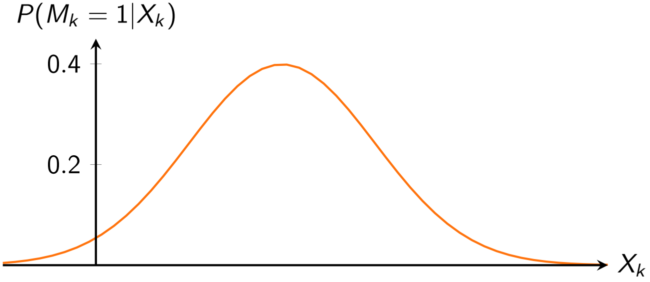
General idea of the proof:

$$\begin{aligned} f^\star(X_{obs}, M) &= \mathbb{E}[Y|X_{obs(M)}, M] \\ &= \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \langle \beta_{mis}^\star, \mathbb{E}[X_{mis}|X_{obs}, M] \rangle \\ &= \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \langle \beta_{mis}^\star, \mathbb{E}[X_{mis}|X_{obs}] \rangle \end{aligned}$$

The Bayes predictor under Gaussian self-masking (MNAR)

- **Gaussian self-masking (MNAR)**: The missing data mechanism is self-masked with $P(M|X) = \prod_{k=1}^{d} P(M_k|X_k)$ and $\forall k \in [\![1, d]\!]$,

$$P(M_k = 1|X_k) = K_k \exp\left(-\frac{1}{2}\frac{(X_k - \widetilde{\mu}_k)^2}{\widetilde{\sigma}_k^2}\right) \qquad \text{with } 0 < K_k < 1.$$



The Bayes predictor under Gaussian self-masking (MNAR)

- **Gaussian self-masking (MNAR)**: The missing data mechanism is self-masked with $P(M|X) = \prod_{k=1}^{d} P(M_k|X_k)$ and $\forall k \in [\![1, d]\!]$,

$$P(M_k = 1|X_k) = K_k \exp\left(-\frac{1}{2}\frac{(X_k - \widetilde{\mu}_k)^2}{\widetilde{\sigma}_k^2}\right) \qquad \text{with } 0 < K_k < 1.$$

Proposition (Gaussian self-masking (MNAR) Bayes predictor)

Under the linear model and Gaussian data assumptions, and a Gaussian self-masking (MNAR) missing data mechanism, the Bayes predictor $f^\star$ takes the form:

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

where $\Sigma_{mis|obs} = \Sigma_{mis,mis} - \Sigma_{mis,obs}\Sigma_{obs}^{-1}\Sigma_{obs,mis}$ *and* $D = \mathrm{diag}(\widetilde{\sigma}_1^2, \ldots, \widetilde{\sigma}_d^2)$.

Outline

1. Optimal predictors in the presence of missing values
2. NeuMiss networks: learning by approximating the Bayes predictor
3. Empirical results

How to approximate the Bayes predictors?

M(C)AR Bayes predictor:

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

nb of parameters

How to approximate the Bayes predictors?

M(C)AR Bayes predictor:

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

nb of parameters

Expectation-Maximisation

$O(d^2)$ parameters

No robustness to the missing data mech.

High computational complexity!!!
(untractable when d reaches a few dozens)

How to approximate the Bayes predictors?

M(C)AR Bayes predictor:

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

nb of parameters

Expectation-Maximisation	MLP
$O(d^2)$ parameters	$O(2^d)$ parameters
No robustness to the missing data mech.	Largely over-parametrized.
High computational complexity!!! (untractable when d reaches a few dozens)	

How to approximate the Bayes predictors?

M(C)AR Bayes predictor:

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

nb of parameters

Expectation-Maximisation	NeuMiss networks	MLP
$O(d^2)$ parameters	$O(d^2)$ parameters	$O(2^d)$ parameters
No robustness to the missing data mech.	Sharing parameters across missing data patterns	Largely over-parametrized.
High computational complexity!!! (untractable when d reaches a few dozens)	$O(d^2)$ computational complexity.	

Differentiable approximations of the inverse covariances

- We propose to approximate $(\Sigma_{obs(m)})^{-1}$, for any m, by an order-ℓ approximation $S^{(\ell)}_{obs(m)}$, defined recursively as:

$$S^{(\ell)}_{obs(m)} = (Id_{obs(m)} - \frac{1}{L}\Sigma_{obs(m)})S^{(\ell-1)}_{obs(m)} + \frac{1}{L}Id.$$

  where $L \in \mathbb{R}^+$ is greater than the largest eigenvalue of $\Sigma_{obs(m)}$.

- The iterates converge linearly to $(\Sigma_{obs(m)})^{-1}$.

- Note: the iterates can be expressed as a series, corresponding to a Neumann series if $S^{(0)} = Id$ and $\ell = \infty$, i.e,

$$(\Sigma_{obs(m)})^{-1} = \frac{1}{L}\sum_{k=0}^{\infty}(Id_{obs(m)} - \frac{1}{L}\Sigma_{obs(m)})^k$$

Differentiable approximations of the inverse covariances

Define the order-ℓ approximation of the Bayes predictor in M(C)AR settings

$$f_\ell^\star(X_{obs}, M) = \langle \beta_{obs}^\star, X_{obs} \rangle + \langle \beta_{mis}^\star, \mu_{mis} + \Sigma_{mis,obs} S_{obs(m)}^{(\ell)} (X_{obs} - \mu_{obs}) \rangle.$$

Proposition (Risk of the order-ℓ approximation)

Suppose that the spectral radius of Σ is strictly smaller than one. Then under the linear model and Gaussian data assumptions, and a MCAR or MAR missing data mechanism, for all $\ell \geq 1$,

$$\mathbb{E}\left[\left(f_\ell^\star(X_{obs}, M) - f^\star(X_{obs}, M)\right)^2\right] \leq \frac{(1-\nu)^{2\ell} \|\beta^\star\|_2^2}{\nu} \mathbb{E}\left[\left\| Id - S_{obs(M)}^{(0)} \Sigma_{obs(M)} \right\|_2^2\right]$$

where ν is the smallest eigenvalue of Σ.

NeuMiss network architecture

- M(C)AR Bayes predictor:

$$f^{\star}_{\ell}(X_{obs}, M) \approx \beta^{\star}_{0} + \langle \beta^{\star}_{obs}, X_{obs} \rangle + \langle \beta^{\star}_{mis}, \mu_{mis} + \Sigma_{mis,obs} S^{(\ell)}_{obs}(X_{obs} - \mu_{obs}) \rangle$$

- Approximation of $(\Sigma_{obs})^{-1}$: $S^{(\ell)}_{obs(m)} = (Id_{obs(m)} - \frac{1}{L}\Sigma_{obs(m)}) S^{(\ell-1)}_{obs(m)} + \frac{1}{L} Id.$

- NeuMiss network architecture (illustrated with a depth of 4):

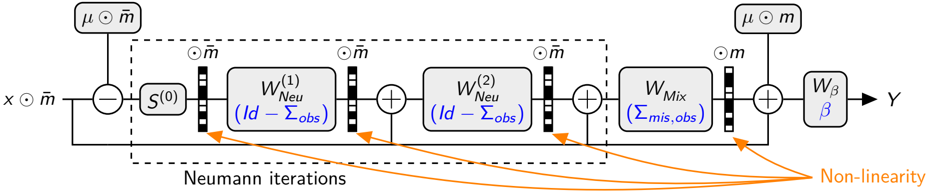


Figure: $\bar{m} = 1 - m$. Each weight matrix $W^{(k)}_{Neu}$ corresponds to a simple transformation of the covariance matrix indicated in blue.

NeuMiss network and Gaussian self-masking (MNAR)

- M(C)AR Bayes predictor:

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

- Suppose that $D_{mis}\Sigma_{mis|obs}^{-1} \approx \hat{D}_{mis}$ where $\hat{D}$ is a diagonal matrix. Then the Gaussian self-masking Bayes predicor is:

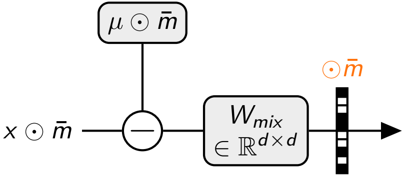
$$\begin{aligned} f^\star(X_{obs}, M) \approx \beta_0^\star + \left\langle \beta_{obs}^\star, X_{obs} \right\rangle + \langle \beta_{mis}^\star, (Id_{mis} + \hat{D}_{mis})^{-1}(\tilde{\mu}_{mis} + \hat{D}_{mis}\mu_{mis}) \\ + (Id_{mis} + \hat{D}_{mis})^{-1}\hat{D}_{mis}\Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}) \big\rangle \end{aligned}$$

The self-masking Bayes predictor can be well approximated by **adjusting the values learned for the params μ and W_{mix}** if $D_{mis}\Sigma_{mis|obs}^{-1}$ are close to diagonal.

Link with the feedforward network

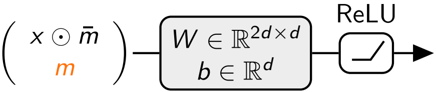
NeuMiss depth-1 layer

$\mathcal{H}_{\odot m} : \mathbb{R}^d \mapsto \mathbb{R}^d$



Feedforward layer (d hidden units)

$\mathcal{H}_{ReLU} : \mathbb{R}^d \times \{0, 1\} \mapsto \mathbb{R}^d$



Proposition (equivalence MLP - depth-1 NeuMiss network)

Denote by h_k^{ReLU} and $h_k^{\odot m}$ the output of the k^{th} hidden units of each layer. Then there exists a configuration of the weights of $\mathcal{H}_{ReLU}$ such that

$$\forall k,\ \forall (m, x_{obs(m)}), \quad h_k^{ReLU}(x_{obs}, m) = h^k_{\odot m}(x_{obs}, m) + c_k, \quad c_k \in \mathbb{R}$$

Outline

1 Optimal predictors in the presence of missing values

2 NeuMiss networks: learning by approximating the Bayes predictor

3 Empirical results

The $\odot m$ nonlinearity is crucial to the performance

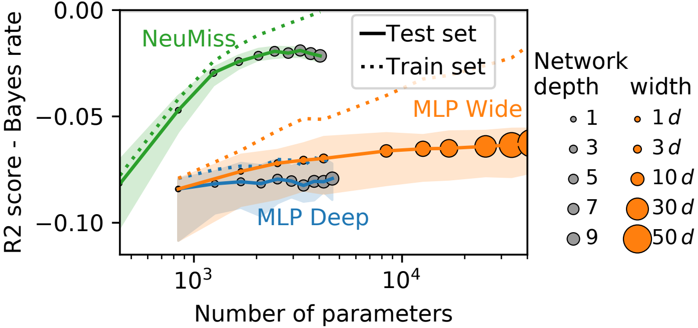


Figure: **Performance as a function of capacity across architectures** — Data are generated under a linear model with Gaussian covariates in a MCAR setting (50% missing values, $n = 10^5$, $d = 20$).

Comparison of performances with competitors

- Data
  - linear model
  - Gaussian data
  - SNR $= 10$
- Missing data mechanisms (50% missing values)
  - MCAR
  - MAR
  - Gaussian self-masking (MNAR)
  - Probit self-msking (MNAR)
- Methods
  - **EM**: Expectation-Maximisation.
  - **Mice + LR**: Conditional imputation followed by linear regression.
  - **MLP**: 1 hidden layer with varied nb of hidden units (between d and $100d$), ReLU nonlinearties, data imputed by 0 concatenated with the mask as input, ADAM, adaptative learning rate.
  - **NeuMiss** The NeuMiss architecture, depth varied between 0 and 10, SGD, adaptative learning rate.

Comparison of performances with competitors

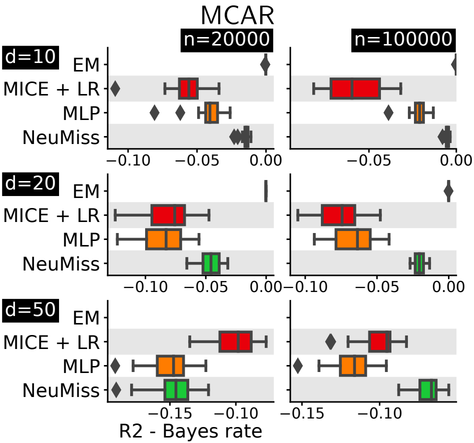


Figure: **Predictive performances in various scenarios** — varying missing-value mechanisms, number of samples n, and number of features d. All experiments are repeated 20 times. For self-masking settings, the x-axis is in log scale, to accommodate the large difference between methods.

Comparison of performances with competitors

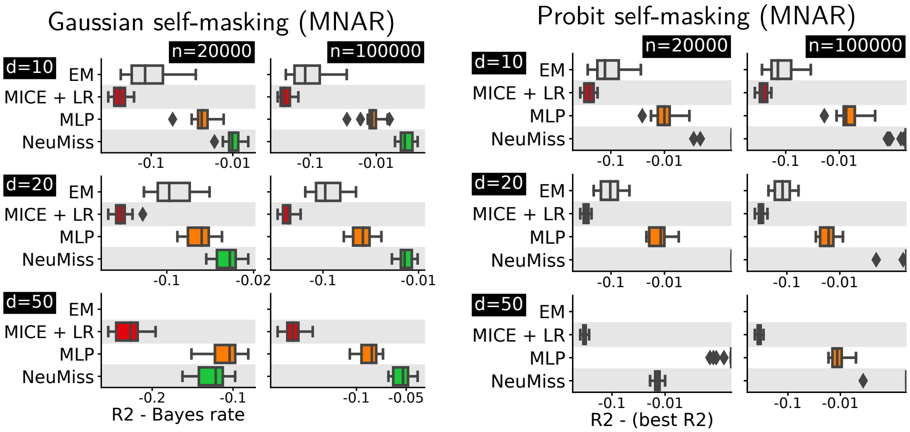
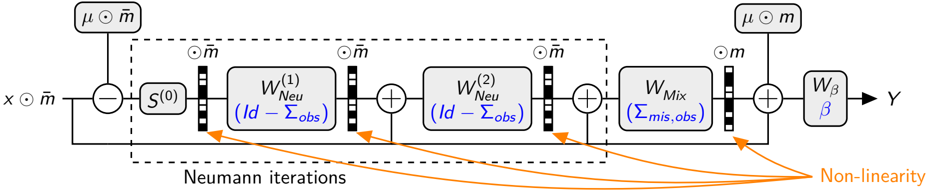


Figure: **Predictive performances in various scenarios** — varying missing-value mechanisms, number of samples *n*, and number of features *d*. All experiments are repeated 20 times. For self-masking settings, the x-xaxis is in log scale, to accommodate the large difference between methods.

Take away



- **Theoretically-grounded** architecture,
- with a new type of non-linearity: **the $\odot$ non-linearity**.
- **Robustness to the missing data mechanism**.
- **Suited for medium-sized datasets** thanks to weight sharing across missing data patterns.

Thank you for your attention!

References I

Josse, Julie et al. (2019). "On the consistency of supervised learning with missing values". In: *arXiv preprint arXiv:1902.06931*.

Le Morvan, Marine et al. (2020). "Linear predictor on linearly-generated data with missing values: non consistency and solutions". In: *arXiv preprint arXiv:2002.00658*.