

# Master's internship defense (MVA and PFE)

Inria - Supervised learning with missing values

Alexandre Perez

Supervised by  
Gaël Varoquaux and Marine Le Morvan  
at Inria Saclay

École Nationale des Ponts et Chaussées  
September 3, 2021



# Context

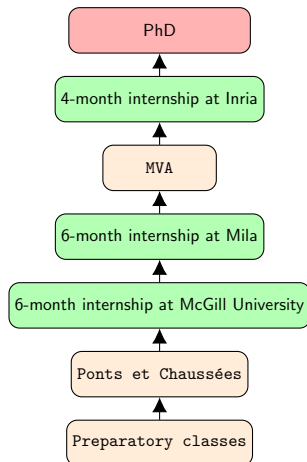
Lab: Inria Saclay

Supervisors: Gaël Varoquaux and Marine Le Morvan

Duration: 4 months, April 2021 - August 2021

Subject: Missing values in supervised learning.

- Benchmarking missing values approaches for predictive models on real data.
- Adapt NeuMiss to classification with missing values [Le Morvan et al., 2020].



# Overview

1. Introduction
2. Benchmark missing values approaches
3. Adapt NeuMiss to classification with missing values
4. Conclusion

# Introduction: the problem of missing values

- Missing values are omnipresent in real world problems
- Have long been studied in the statistical literature within the inferential framework

[Rubin, 1976] defined several missing values mechanisms:

- *Missing At Random* (MAR): the probability of a value to be missing only depends on the observed variables.
- *Missing Not At Random* (MNAR): the missingness can depend on both the observed and unobserved values.

Most missing values methods in inference rely on the MAR hypothesis since theoretical results show that the mechanism can be ignored. In practice, real data is often MNAR.

# Introduction: scope of the study

Focus on supervised learning with missing values.

Different tradeoffs: risk minimization instead of parameters estimation.

In supervised learning, most statistical models and machine learning algorithms are not designed for incomplete data.

How to deal with missing values in this framework?

- Delete samples having missing values → to avoid.
- Use imputation.
  - Constant imputation (mean, median)
  - Conditional imputation (KNN, MICE)
- Adapt or create predictive models to handle missing values natively.
  - Boosted-trees with *Missing Incorporated in Attribute* (MIA) adaptation [Twala et al., 2008].
  - NeuMiss networks in the regression setting [Le Morvan et al., 2020].

# Introduction: the two axes

## Part 1:

- How does MIA experimentally compare to imputation ?
- Contributions:
  - Add statistical tests and statistics on the database.
  - Co-wrote a 17-page manuscript submitted in the GigaScience academic journal.

## Part 2:

- Adapt NeuMiss to the classification setting.
- Contributions:
  - Derived the optimal predictors in binary classification.
  - Adapted NeuMiss.
  - Investigated calibration.
  - Experimental evaluation.
  - Submitted a 6-page abstract to ECML PKDD workshop, selected for an oral.

# Part 1: Benchmark missing values approaches

# Methods benchmarked

**Table: Methods compared in the main experiment.**

All use gradient-boosted trees as predictive model.

8 use imputation and 1 uses MIA.

In-article name	Imputer	Mask	Predictive model
MIA	-	-	Gradient-boosted trees
Mean	Mean	No	Gradient-boosted trees
Mean+mask	Mean	Yes	Gradient-boosted trees
Median	Median	No	Gradient-boosted trees
Median+mask	Median	Yes	Gradient-boosted trees
Iterative	MICE	No	Gradient-boosted trees
Iterative+mask	MICE	Yes	Gradient-boosted trees
KNN	KNN	No	Gradient-boosted trees
KNN+mask	KNN	Yes	Gradient-boosted trees



# Datasets

- Traumabase [The Traumabase Group, ], 20 000 samples.
- UK BioBank [Sudlow et al., ], 500 000 samples.
- MIMIC-III [Johnson et al., ], 60 000 samples.
- NHIS [National Center for Health Statistics, 2017], 88 000 samples.

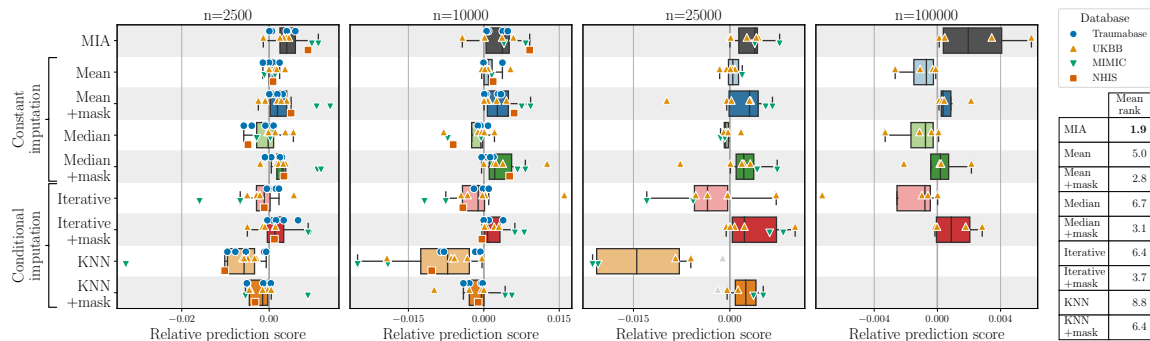
# Experimental protocol

Benchmark the 9 methods on 13 prediction tasks (10 classifications, 3 regressions).

Using cross validation and hyper-parameters tuning.

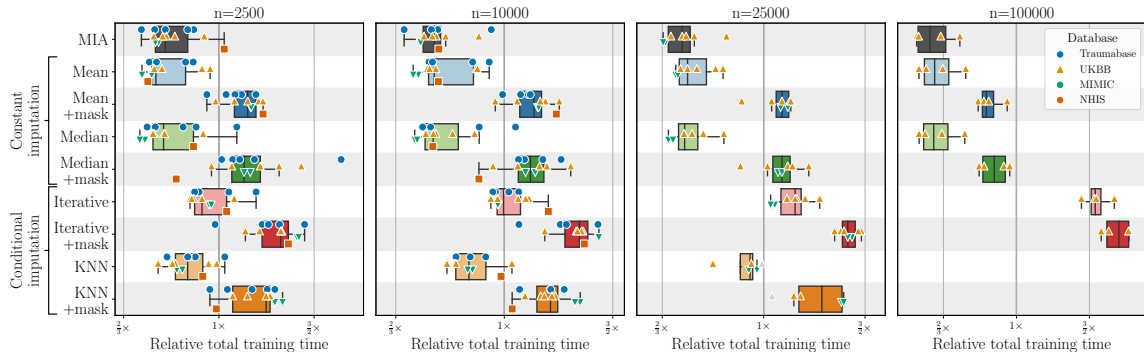
4 sub-samplings of datasets: 2 500, 10 000, 25 000 and 100 000 samples.

# Results - Prediction performance



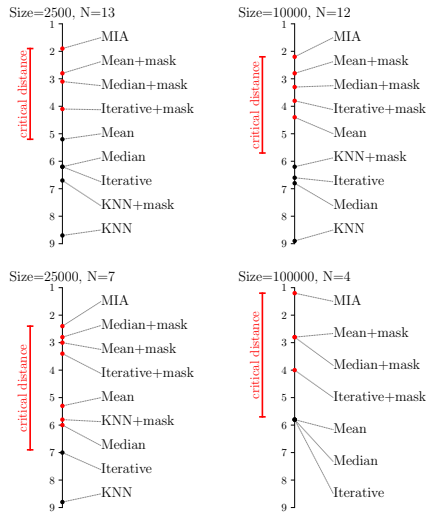
**Figure: Prediction performance.** Comparison of prediction performance and training times across the 9 methods for 13 prediction tasks spread over 4 databases, and for 4 sizes of dataset (2 500, 10 000, 25 000 and 100 000 samples).

# Results - Computational time



**Figure: Computational time.** Comparison of prediction performance and training times across the 9 methods for 13 prediction tasks spread over 4 databases, and for 4 sizes of dataset (2 500, 10 000, 25 000 and 100 000 samples).

# Results - Significance



**Figure: Mean ranks by method and by size of dataset.** The critical distance is computed using the Nemenyi test. Methods within the critical distance range do not perform significantly differently from one another.. 13 / 37

# Findings and interpretation

## Findings:

- MIA takes the lead at little cost, although not significantly.
- Adding the mask improves prediction.

## Interpretation:

- The missingness is informative (MNAR or outcome depends on missingness)  
→ imputation is not applicable.

# Strengths and limitations

## Limitations:

- Not every difference is significant.
- Would benefit having more datasets and having more datasets with large number of samples.

## Strengths of the benchmark:

- 12 000 CPU hours.
- Lots of datasets (only 6% of empirical NeurIPS articles build upon more than 10 datasets [Bouthillier and Varoquaux, 2020].)

More details in the manuscript in appendix.

## Part 2: Adapt NeuMiss to classification with missing values



# Problem setting - Notations

- A pair of random variables  $(X, Y)$ , where  $X \in \mathbb{R}^d$  and  $Y \in \mathbb{R}$  for regression or  $Y \in \{-1, 1\}$  for binary classification.
- A random vector  $M \in \{0, 1\}^d$  acting as a mask on  $X$ : for all  $1 \leq j \leq d$ ,  $M_j = 1 \Leftrightarrow X_j$  not observed.
- The incomplete feature vector  $\tilde{X} \in \mathcal{X} := (\mathbb{R} \cup \{\text{NA}\})^d$  is defined as  $\tilde{X}_j = \text{NA}$  if  $M_j = 1$  and  $\tilde{X}_j = X_j$  if  $M_j \neq 1$ .

We note *obs* (resp. *mis*) the indices of the zero (resp. non-zero) entries of  $M$ .

For a vector  $X \in \mathbb{R}^d$ , we denote by  $X_{obs}$  (resp.  $X_{mis}$ ) the observed (resp. missing) values of  $X$ .

# Problem setting - Assumptions

## Assumption (Gaussian data)

*The complete data is distributed as a multivariate Gaussian:  $X \sim \mathcal{N}(\mu, \Sigma)$ .*

# Problem setting - Assumptions

## Assumption (Missing At Random (MAR))

For all  $m \in \{0, 1\}^d$ ,  $\mathbb{P}[M = m \mid X] = \mathbb{P}[M = m \mid X_{obs}]$ .

Missing Completely At Random (MCAR) is a special case of MAR where the probability of  $M$  does not depend on the covariates.

## Assumption (Gaussian self-masking, instance of MNAR)

*The probability that a variable is missing depends on its own value through a Gaussian function:*

$$\mathbb{P}[M \mid X] = \prod_{k=1}^d \mathbb{P}[M_k \mid X_k]$$
$$\forall 1 \leq k \leq d, \quad \mathbb{P}[M_k = 1 \mid X_k] = K_k \exp\left(-\frac{(X_k - \tilde{\mu}_k)^2}{2\tilde{\sigma}_k^2}\right)$$

where  $0 < K_k < 1$ ,  $\tilde{\mu}_k \in \mathbb{R}$  and  $\tilde{\sigma}_k \in \mathbb{R}_+$ .

# NeuMiss - The regression setting

## Assumption (Linear model)

*The response  $Y \in \mathbb{R}$  is linked to the complete data  $X$  through a linear model:*

$$Y = \beta_0^* + \langle X, \beta^* \rangle + \epsilon, \quad \text{where } \beta_0^* \in \mathbb{R}, \beta^* \in \mathbb{R}^d, \text{ and } \epsilon \sim \mathcal{N}(0, \sigma^2), \sigma \in \mathbb{R}.$$

# NeuMiss - Bayes predictors in the regression setting

## Proposition (MAR Bayes predictor, [Le Morvan et al., 2020])

*For linear model, Gaussian data and in the MAR setting, the Bayes predictor reads:*

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

## Proposition (GSM Bayes predictor, [Le Morvan et al., 2020])

*For linear model, Gaussian data and in the Gaussian self-masking setting, the Bayes predictor reads:*

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

*with  $\Sigma_{mis|obs} := \Sigma_{mis} - \Sigma_{mis,obs} \Sigma_{obs}^{-1} \Sigma_{obs,mis}$  and  $D$  the diagonal matrix of entries  $(\tilde{\sigma}_1^2, \dots, \tilde{\sigma}_d^2)$ .*

# NeuMiss - The architecture

[Le Morvan et al., 2020] introduced NeuMiss, a neural network that approximates the above Bayes predictors.

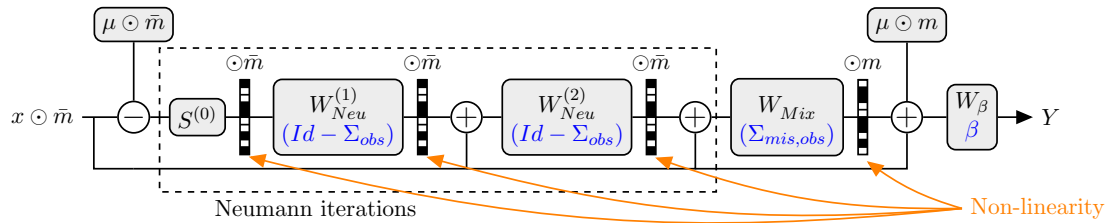
Key point: approximation of  $(\Sigma_{obs})^{-1}$  with Neumann iterates.  
 $2^d$  submatrices to approximate  $((\Sigma_{obs})^{-1} \neq (\Sigma^{-1})_{obs})$ .

Order- $l$  approximation  $S_{obs}^{(l)}$  of  $(\Sigma_{obs})^{-1}$  is defined for all  $l \geq 1$  with:

$$S_{obs}^{(l)} = (Id - \Sigma_{obs})S_{obs}^{(l-1)} + Id \quad \text{and} \quad S_{obs}^{(0)} = Id. \quad (1)$$

Practically, each layer  $i$  applies a  $x \mapsto W^{(i)}x + x$  transformation followed by a new type of non-linearity, the elementwise multiplication by the mask

# NeuMiss - The architecture



**Figure: NeuMiss network architecture with a depth of 4** —  $\bar{m} = 1 - m$ . Each weight matrix  $W^{(k)}$  corresponds to a simple transformation of the covariance matrix indicated in blue. Figure taken from [Le Morvan et al., 2020].

Reminder of the MAR Bayes predictor:

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

# Contributions



# The extended classification setting

## Assumption (Probit model)

*The response  $Y \in \{-1, 1\}$  is linked to the complete data  $X$  through the following relation:*

$$\mathbb{P}[Y = 1 \mid X, \beta_0^*, \beta^*] = \Phi(\langle X, \beta^* \rangle + \beta_0^*),$$

*where  $\Phi : \mathbb{R} \rightarrow [0, 1]$  is the cumulative distribution function of the standard normal distribution.*

# Optimal predictors in binary classification

## Proposition (MAR Bayes predictor)

*For Gaussian data generated via the probit model in the MAR setting, then:*

$$\mathbb{P}[Y = 1 \mid \tilde{X}] = \Phi \left( \frac{\nu}{(1 + \sigma^2)^{\frac{1}{2}}} \right),$$

*with:*

$$\begin{aligned} \nu &:= \beta_0^* + \langle \beta_{obs}^*, X_{obs} \rangle + \langle \beta_{mis}^*, \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}) \rangle \\ \sigma^2 &:= \beta_{mis}^{*T} \Sigma_{mis} \beta_{mis}^* - \beta_{mis}^{*T} \Sigma_{mis,obs}(\Sigma_{obs})^{-1} \Sigma_{mis,obs}^T \beta_{mis}^* \end{aligned}$$

*and the Bayes predictor can be written:*

$$f_{\tilde{X}}^*(\tilde{X}) = \text{sign}(\nu)$$

# Optimal predictors in binary classification

## Proposition (GSM Bayes predictor)

*For Gaussian data generated via the probit model in the Gaussian self-masking setting, then:*

$$\mathbb{P}[Y = 1 \mid \tilde{X}] = \Phi \left( \frac{\nu}{(1 + \sigma^2)^{\frac{1}{2}}} \right),$$

*with:*

$$\begin{aligned} \nu &= \beta_0^* + \langle \beta_{obs}^*, X_{obs} \rangle + \langle \beta_{mis}^*, (Id + D_{mis} \Sigma_{mis|obs}^{-1}) \\ &\quad \times (\tilde{\mu}_{mis} + D_{mis} \Sigma_{mis|obs}^{-1} (\mu_{mis} + \Sigma_{mis,obs} (\Sigma_{obs})^{-1} (X_{obs} - \mu_{obs}))) \rangle \\ \sigma^2 &= \beta_{mis}^{*T} \left( D_{mis}^{-1} + \Sigma_{mis|obs}^{-1} \right)^{-1} \beta_{mis}^* \end{aligned}$$

*and the Bayes predictor can be written:  $f_{\tilde{X}}^*(\tilde{X}) = \text{sign}(\nu)$*

# Considerations about calibration

## Definition (Calibration)

A classifier  $f : \mathcal{X} \rightarrow \mathbb{R}^K$  is said to be calibrated if the confidence in its prediction is equal to the probability of this prediction to be correct, i.e.:

$$\mathbb{E}[\mathbb{1}_{\hat{y}=y} \mid \hat{z}] = \hat{z}, \text{ with } \hat{y} = \operatorname{argmax}_x f(x), \hat{z} = \max_x f(x) \text{ and } x \in \mathcal{X}.$$

## Proposition

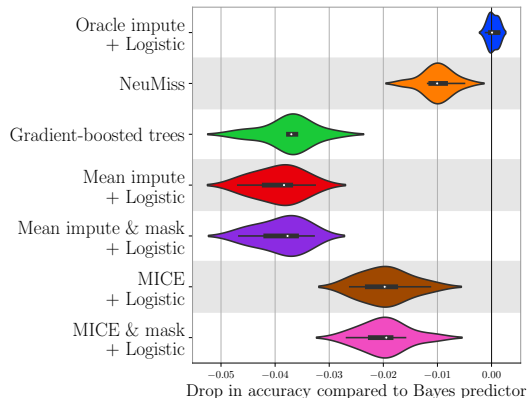
*The Bayes predictor  $f_{\tilde{X}}^*$  is calibrated.*

The calibrated confidence is  $\mathbb{P}[Y = 1 \mid \tilde{X}] = \Phi\left(\frac{\nu}{(1+\sigma^2)^{\frac{1}{2}}}\right)$ . NeuMiss computes  $\nu$ . The best guess we can provide as confidence with this adaptation is  $\Phi(\nu)$ , which is not calibrated even if it provides the optimal decision boundary.

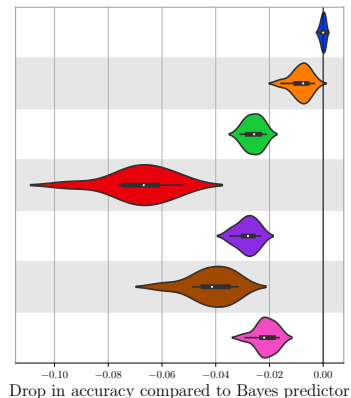
## Remark

The above adaptation of NeuMiss is not calibrated.  $\sigma$  is required to build a calibrated classifier.

# Empirical study of classification with missing values



(a) MCAR



(b) Gaussian self-masking

**Figure: Prediction benchmarks.** Accuracy of the benchmarked methods relative to the accuracy of the Bayes predictor, on 10 simulated datasets of 500 000 samples and 50 numerical features drawn from a multivariate gaussian  $\mathcal{N}(\mu, \Sigma)$  for MCAR and GSM settings. About 50% values are missing in each.

# Conclusion

# Conclusion

## Part 1:

- Using MIA provides small but systematic improvement over imputation.
- Directly handling missing values is to be considered.

## Part 2:

- Easy theoretically-grounded adaptation to binary classification.
- The adaptation is robust to missing values mechanism.
- The classifier is not calibrated → more sophisticated adaptation is required.

## General:

- Shed light on pros and cons of practices to handle missing values.
- Change habits in practice: better choices than imputation.
- Build a solid base for calibrated classification with NeuMiss.

# Conclusion - Further work

As a PhD student:

- Make a calibrated classifier.
- Relax assumptions (probit model, gaussian data, binary).
- Investigate influence of the class balance.
- Understand its training.
- Evaluate NeuMiss on *real-world* data.
- ... more to come along the way.



# Conclusion - What I learned

A very instructive internship:

- Improved technical skills:
  - Theoretical analysis, derive new analytical expressions.
  - Academic writing.
  - Python, PyTorch, scikit-learn, pandas, matplotlib,  $\text{\LaTeX}$ .
  - Designing a neural network and intensively training with HP tuning.
- Improved relational skills:
  - Communicate ideas, discuss views
  - Participate social events
  - Lightning talk and weekly meeting challenged my communication.





# Acknowledgments

Thank you for you attention!

And a particular thank to:

- Gaël Varoquaux and Marine Le Morvan
- Jean-Baptiste Poline
- Julie Josse
- Corinne Petitot

# References I

-  Bouthillier, X. and Varoquaux, G. (2020).  
Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020.  
Research report, Inria Saclay Ile de France.
-  Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M.,  
Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G.  
MIMIC-III, a freely accessible critical care database.  
3(1):160035.
-  Le Morvan, M., Josse, J., Moreau, T., Scornet, E., and Varoquaux, G. (2020).  
NeuMiss networks: differentiable programming for supervised learning with missing values.  
*Advances in Neural Information Processing Systems*, 33:5980–5990.
-  National Center for Health Statistics (2017).  
National Health Interview Survey (NHIS).

# References II



Rubin, D. B. (1976).  
Inference and missing data.  
*Biometrika*, 63(3):581–592.



Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R.  
UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age.  
12(3):e1001779.



The Traumabase Group.  
Traumabase.

# References III



Twala, B. E. T. H., Jones, M. C., and Hand, D. J. (2008).  
Good methods for coping with missing data in decision trees.  
*Pattern Recogn. Lett.*, 29:950–956.