# Learning why missing values are missing

**Keywords** : missing-data, deep learning, evaluation of imputation methods

Several factors can contribute to missing values in a study, including data loss, sensor failures, or the aggregation of datasets from multiple sources. There is a rich literature on how to impute missing values, for example, considering the EM algorithm [Dempster et al., 1977], low rank models [Robin et al., 2019, Sportisse et al., 2020], random forests [Stekhoven and Bühlmann, 2012] or deep learning techniques with variational autoencoders [Mattei and Frellsen, 2019, Ipsen et al., 2021].

To assess the performance of imputation methods, a relevant measure is the Mean Squared Error (MSE) computed on the missing entries. The main challenge is that, in practice, we do not have access to the unobserved values. Therefore, many studies propose computing the MSE on complete datasets by introducing synthetic missing data. An R package has been developed for this purpose [Schouten et al., 2018]. In real datasets containing missing values, a common practice is to introduce additional synthetic missing values and compute the MSE only on those entries for which true values are known [Sportisse et al., 2020]. The challenge lies in simulating pertinent missing values that adhere to the same distribution as the existing ones. In particular, the manner on how values are missing must be considered: the correlations between the mask variables, which indicates where the missing values are, as well as the links between the mask and the data values. A simple example would be in cases where the data is MCAR, i.e. there is no link between the mask and the data values, and where there are no specific patterns associated with missingness (e.g., the second variable is always missing if the first is). In this case, introducing new missing values using a Bernoulli distribution can provide a suitable approximation. However, when dealing with more complex missing-data scenarios, a more sophisticated approach becomes necessary.

As far as we know, there is no existing work specifically addressing these distributional shift problems in the context of generating missing data. A primary option would be to learn the missing-data mechanism, i.e. the conditional distribution of the mask given the data values, using a variational autoencoder [Ipsen et al., 2021]. Beyond its relevance for evaluating imputation methods, the estimated mechanism could be used to devise novel imputation schemes, drawing on recent developments in the supervised paradigm [Le Morvan et al., 2020, Ipsen et al., 2022, Van Ness and Udell, 2023].

**Context of the internship** The intern will join the Maasai team of Inria Sophia-Antipolis and Université Côte d'Azur, which is composed of 25 researchers in statistical and machine learning (web: `https://team.inria.fr/maasai/`). The team is part of the Institut 3IA Côte d'Azur `https://3ia.univ-cotedazur.eu/`, which offers a lot of opportunities (thesis offers, seminars & meetings with PhD students/postdoc in machine learning).
Duration: 6 months
Salary: approx. 550€ / month
PhD opportunities within the Maasai team may be pursued after the intership, to continue this work.

**Contact** To apply, please contact Pierre-Alexandre Mattei (pierre-alexandre.mattei@inria.fr) and Aude Sportisse (aude.sportisse@inria.fr).

## References

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):

1–22, 1977.

Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. not-miwae: Deep generative modelling with missing not at random data. In *ICLR 2021-International Conference on Learning Representations*, 2021.

Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. How to deal with missing data in supervised deep learning? In *ICLR 2022-10th International Conference on Learning Representations*, 2022.

Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, and Gaël Varoquaux. Neumiss networks: differentiable programming for supervised learning with missing values. *Advances in Neural Information Processing Systems*, 33:5980–5990, 2020.

Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pages 4413–4423. PMLR, 2019.

Geneviève Robin, Olga Klopp, Julie Josse, Éric Moulines, and Robert Tibshirani. Main effects and interactions in mixed and incomplete data frames. *Journal of the American Statistical Association*, 2019.

Rianne Margaretha Schouten, Peter Lugtig, and Gerko Vink. Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930, 2018.

Aude Sportisse, Claire Boyer, and Julie Josse. Estimation and imputation in probabilistic principal component analysis with missing not at random data. *Advances in Neural Information Processing Systems*, 33:7067–7077, 2020.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Mike Van Ness and Madeleine Udell. In defense of zero imputation for tabular deep learning. In *NeurIPS 2023 Second Table Representation Learning Workshop*, 2023.