Internship offer: Self-supervised missing data imputation

Keywords : missing-data, deep learning

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].

One limitation of all these techniques is that they are all *indirect*, in the sense that the loss function that is optimised is not the imputation error. The main challenge is that, in practice, we do not have access to the unobserved values, and therefore, cannot compute this error. The goal of this internship will be to develop a *direct* method, based on self-supervised learning. The closest related works are two papers using masked generative modelling [Tashiro et al., 2021, An et al., 2024]. However, both techniques remain indirect in these sense of the previous paragraph.

An important second step would be to quantify the uncertainty of these imputations, for instance through multiple imputations [Little and Rubin, 2019] or conformal prediction [Angelopoulos et al., 2023].

This internship will be supervised by Pierre-Alexandre Mattei (Research scientist at Inria, Université Côte d'Azur). Collaborations with Aude Sportisse (Research scientist at CNRS in Grenoble) will be included. The candidate is expected to have a good knowledge of probability and statistics (limit theorems, Monte Carlo) and of supervised machine learning models (neural nets, random forests).

Context of the internship The postdoc 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.
Duration: 4 to 6 months
Salary: around 600€ gross per 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).

References

Seunghwan An, Gyeongdong Woo, Jaesung Lim, ChangHyun Kim, Sungchul Hong, and Jong-June Jeon. Masked language modeling becomes conditional density estimation for tabular data synthesis. *arXiv preprint arXiv:2405.20602*, 2024.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

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.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

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.

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.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34:24804–24816, 2021.