3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Postdoctoral research topic

- Title of the proposed topic: **Self-supervised missing data imputation**
- Research axis of the 3IA: **1: foundations of AI**
- **Supervisor (name, affiliation, email): Pierre-Alexandre Mattei, Inria, pierre-alexandre.mattei@inria.fr**
- Potential co-supervisor (name, affiliation): Aude Sportisse, CNRS, Grenoble
- The laboratory and/or research group: Inria Maasai, Sophia-Antipolis

---

Apply by sending an email directly to the supervisor (pierre-alexandre.mattei@inria.fr). The application will include:

- Letter of recommendation of the supervisor indicated above
- Curriculum vitæ including the list of the scientific publications
- Motivation letter
- Letter of recommendation of the thesis supervisor

⇨ **All the requested documents must be gathered and concatenated in a single PDF file named in the following format: LAST NAME of the candidate_Last Name of the supervisor_2021.pdf**

- Description of the topic:

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 [Sportisse et al., 2020], random forests [Stekhoven and Buhlmann, 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 postdoc 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].

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, John Wiley & Sons, 2019.

Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep generative modelling and imputation of incomplete data sets. ICML, 2019.

Aude Sportisse, Claire Boyer, and Julie Josse. Estimation and imputation in probabilistic principal component analysis with missing not at random data. NeurIPS, 2020

Daniel J Stekhoven and Peter Buhlmann. 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. NeurIPS, 2021