Mixture-Based Generative Modeling for Data Imputation and Synthesis

A. Faul[†,1], D. Ginsbourger[§,1], B. Spycher[§,2],

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2022 – Sep. 2026

[1] Institute of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland
antoine.faul@unibe.ch david.ginsbourger@unibe.ch

[2] Institute of Social and Preventive Medicine, University of Bern, Switzerland
ben.spycher@unibe.ch

Abstract

In medical research, data scarcity and missing information are prevalent, posing significant challenges for statistical analysis. These issues often stem from the high cost of collecting comprehensive patient datasets, the complexity of measuring specific variables, privacy concerns, or varying levels of patient adherence.

This work addresses these challenges by developing statistical methods for data imputation and synthetic data generation using mixtures of distributions. Our motivation originates from a case study at the University Hospital of Bern, for which we have made the data publicly available [3]. The goal was to predict the 10-year risk of cardiovascular disease when some clinical inputs of the risk calculator were systematically missing. Our method involved probabilistically imputing the missing variables and propagating the resulting uncertainty into the risk calculator.

Formally, the aim of systematic imputation is to sample from the conditional distribution of a random vector $\mathbf{Y} \in \mathbb{R}^q$ given observations of $\mathbf{X} \in \mathbb{R}^p$, based on existing samples of $(\mathbf{X}, \mathbf{Y})$. We propose a generative approach which consists in estimating the joint density $f(\mathbf{x}, \mathbf{y})$ using a parametric probability density function and sampling from the conditional distribution $f(\mathbf{y}|\mathbf{x})$ through analytical formulas.

For this purpose, our work [1] introduces families of multivariate distributions stable by conditioning, including multivariate Gaussian, Student t, and skew normal distributions, but excluding, for example, multivariate q-Exponential distributions. We demonstrate that stability by conditioning of a family of trans-dimensional probability distributions can be extended to finite mixtures and marginal transformations.

This extends the applicability of analytical conditioning across a broader range of multivariate distributions. We developed an algorithm for conditional sampling using implicit copulas and latent spaces (see Figure 1).

While our initial study used Gaussian copulas [5], more complex dependence structures like the Gaussian Mixture copula model (GMCM) were employed to capture multi-modalities and tail dependencies in our latest work [1].

These generative approaches, based on copulas, can be applied for synthetic tabular data generation. They produce realistic synthetic data points and are competitive with machine learning

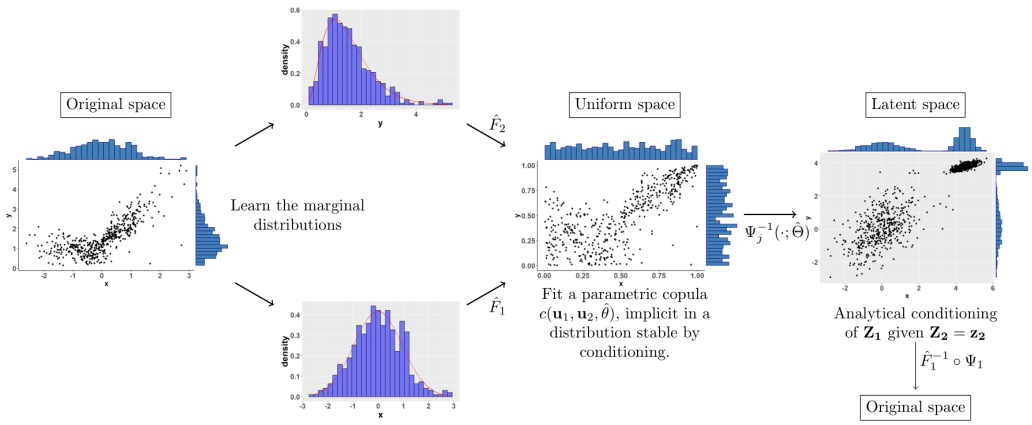

Figure 1: Workflow of the conditioning algorithm on a 2-dimensional example.

methods in moderate dimensions [2]. Evaluating data quality is crucial, yet there's no consensus on the best metrics. Therefore, we utilized a variety of existing metrics and introduced additional ones tailored to assess the statistical utility and privacy specific to our problem.

Finally, addressing complex scenarios where single generators fail to capture multi-modalities, we developed an iterative procedure inspired by the Expectation–Maximization framework to combine mixtures of diverse generators, each specializing in different data space regions [4]. We showed empirically that our approach produces high-quality synthetic data and we provide theoretical guarantees by establishing convergence rates of the mixture distribution.

Short biography (PhD student)

Antoine Faul graduated with a MSc in Engineering from Isae-Supaero and with a MSc in Statistics from Université Paris-Saclay. He started his PhD at the University of Bern in October 2022 on statistics with application to medicine. The thesis is funded by the Multidisciplinary Center for Infectious Diseases (MCID) of the University of Bern.

References

[1] Antoine Faul, David Ginsbourger, and Ben Spycher. Easy conditioning far beyond gaussian. *arXiv preprint arXiv:2409.16003*, 2024.

[2] Antoine Faul, David Ginsbourger, Ben Spycher, and Petra Stute. Copula-based synthetic data generation in medicine. 2026. Work in Progress.

[3] Antoine Faul, Philip Stange, Anja Mülhemann, Manuela Moraru, Suzanne Theis, Lena Friederichsen, Lukas Bütikofer, and Petra Stute. Menobalance: Cardiovascular risk factors from the CIMBOLIC study. *Dataset on Zenodo*, 2024.

[4] Antoine Faul, Xiao Zhou, Ossi Raisa, Mihaela Van der Shaar, and Cem Tekin. Modelling complex tabular datasets with a mixture of diverse generative models. 2026. Submitted to AISTATS 2026.

[5] Anja Mühlemann, Philip Stange, Antoine Faul, Serena Lozza-Fiacco, Rowan Iskandar, Manuela Moraru, Susanne Theis, Petra Stute, Ben D Spycher, and David Ginsbourger. Comparing imputation approaches to handle systematically missing inputs in risk calculators. *PLOS Digital Health*, 4(1):e0000712, 2025.