Copula based tabular synthetic data generation: a sound trade-off

A. Faul[†,1], D. Ginsbourger[§,1], B. Spycher[§,2], P. Stute[3]

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

[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

[3] Department of Obstetrics and Gynecology University Women's Hospital, Bern, Switzerland
petra.stute@insel.ch

Abstract

The field of synthetic data generation has expanded rapidly in recent years, leading to the development of numerous machine learning methods for tabular data synthesis. These models offer significant potential but may often be challenging to interpret and requiring large training datasets. Furthermore, they are susceptible to overfitting, which can compromise privacy by inadvertently replicating specific patient details instead of capturing general patterns.

While increasing the complexity of generators usually enhances synthetic data *fidelity* - how closely the synthetic data statistically resembles the real data - and *utility* - the usefulness of synthetic data for a downstream application - it tends to deteriorate *privacy* - how much information the synthetic data discloses about the original data. Managing this trade-off is central to the challenge of synthetic data generation.

Copulas offer a natural approach to learn joint distributions from data by decoupling the modelling of marginal distributions versus of the dependence structure. This allows modelling multivariate data with various marginal distributions, as typical in medicine, and results in a flexible, interpretable method for synthetic data generation.

In the present work, we present a comprehensive framework for synthetic data generation using copulas. The approach is particularly valuable for datasets of moderate dimensions and a limited number of observations, a common scenario in medical research. While the framework is primarily aimed at continuous variables, it can also be used in the case of discrete and mixed variables. We introduce various types of copulas to model complex dependencies, including multimodal dependencies with the Gaussian Mixture Copula Model (GMCM). As main example, we consider a dataset from the Gynecology Department at University Hospital (Inselspital) in Bern, containing both personal and clinical information for 359 female patients aged 40 to 69. This dataset has been collected in the framework of the CIMBOLIC study.

Furthermore, using evaluation metrics —both established and newly developed towards the main objectives of *fidelity*, *utility*, and *privacy*-, we demonstrate that copula-based methods inherently achieve an effective trade-off by capturing the distributions of tabular datasets without overfitting. This results in high-quality, privacy-enhanced synthetic data.

In particular, we propose an approach based on permutation tests to assess the privacy of a synthetic dataset.