

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

Antoine Faul, David Ginsbourger,
Ben Spycher, Petra Stute.

Institute of Mathematical Statistics and Actuarial Sciences
University of Bern, Switzerland

13th October, 2026

- 1 Motivating problem
- 2 Copula-based synthetic data generation
- 3 Evaluation of synthetic data
- 4 Conclusion

Risk calculators in medicine

- Medical **risk calculators** exist for many diseases yet remain underutilized.
-  **Limitations:**
 - Require clinical measurements (e.g., blood draws).
 - Difficult for non-clinicians to use due to lack of instructional support.

Risk calculators in medicine

- Medical **risk calculators** exist for many diseases yet remain underutilized.
-  **Limitations:**
 - Require clinical measurements (e.g., blood draws).
 - Difficult for non-clinicians to use due to lack of instructional support.

Idea

Increase accessibility by substituting missing lab measurements with proxy data (e.g., BMI, wearables).

Risk calculators in medicine

- Medical **risk calculators** exist for many diseases yet remain underutilized.
-  **Limitations:**
 - Require clinical measurements (e.g., blood draws).
 - Difficult for non-clinicians to use due to lack of instructional support.

Idea

Increase accessibility by substituting missing lab measurements with proxy data (e.g., BMI, wearables).

- **Case Study:**
 - We apply this to the SCORE2 risk calculator for cardiovascular disease (CVD).
 - The approach shall be **transferable** to other risk calculators.

A motivating example

- Cohort of 443 female patients from the University hospital of Bern.

n	Age	Smoker	Diabetes	BMI	TotalC	HDLC	BP
1	58	0	0	22.15	5.93	1.85	100
2	52	0	1	30.25	3.65	1.06	120
3	45	0	0	29.75	5.07	2.03	130
4	46	1	0	21.23	NA	NA	NA
5	55	0	0	20.63	NA	NA	NA

- Predict the risk (with uncertainty) of developing a cardio-vascular disease (CVD) in the next 10 years when the clinical measurements are systematically missing.
- Data publicly available on Zenodo

Deterministic imputation

n	Age	Smoker	Diabetes	BMI	TotalC	HDLC	BP
1	58	0	0	22.15	5.93	1.85	107
2	52	0	1	30.25	3.65	1.06	122
3	45	0	0	29.75	5.07	2.03	130
4	46	1	0	21.23	4.72	1.90	105
5	55	0	0	20.63	4.20	1.74	114

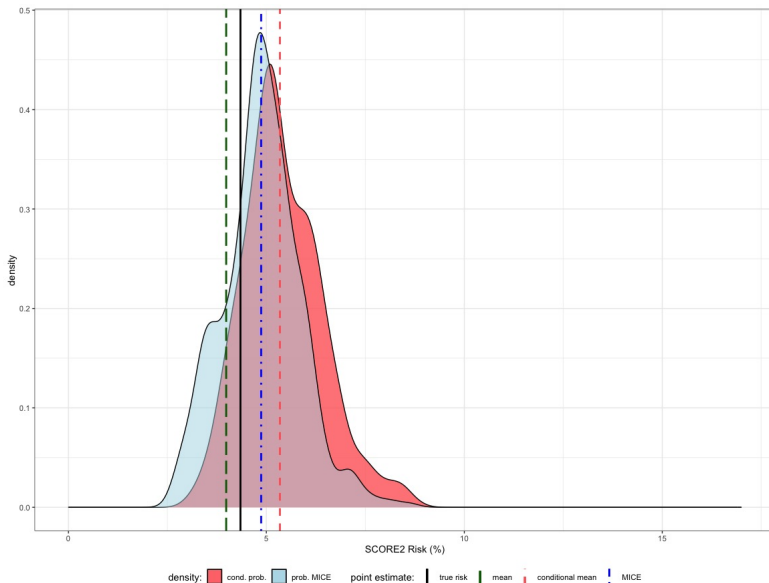
Probabilistic Imputation

n	Age	Smoker	Diabetes	BMI	TotalC	HDLC	BP
1	58	0	0	22.15	5.93	1.85	100
2	52	0	1	30.25	3.65	1.06	120
3	45	0	0	29.75	5.07	2.03	130
4	46	1	0	21.23			
5	55	0	0	20.63			

Apply SCORE2 risk calculator with imputed inputs.

→ obtain a **probability distribution** for the risk

Probabilistic vs deterministic imputation



- 1 Motivating problem
- 2 Copula-based synthetic data generation
- 3 Evaluation of synthetic data
- 4 Conclusion

Synthetic data generation

Challenges with patient data:

-  **Data scarcity:** limited number of patients in the dataset making robust statistical analysis difficult
-  **Limited availability** of similar public datasets online to benchmark or validate methods
-  **Privacy concerns:** sensitive variables — such as the SCORE2 cardiovascular risk — cannot be shared openly without risking patient re-identification

A promising solution:

- Generate **synthetic data** that preserves the key **statistical features** of the original data while preserving **privacy**
- Enable **data sharing** without compromising patient privacy and to fit statistical models to a **larger** dataset

Reminders on copulas

Copulas

A d -dimensional *copula* is a multivariate distribution function over $[0, 1]^d$ with standard uniform margins.

Sklar's theorem

For any distribution function F with margins F_j , $j = 1, \dots, d$, there exists a *copula* C such that:

$$F(x_1, \dots, x_p) = C(F_1(x_1), \dots, F_d(x_d)), \quad \forall \mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$$

A multivariate distribution function is composed by both the margins and the copula. The copula characterizes the dependence structure between components.

Examples of copulas

- The **Gaussian Copula** is the copula implicit in a multivariate Gaussian distribution. Its density can be expressed by:

$$c_P(u) = \frac{\phi_P(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))}{\prod_{i=1}^d \phi(\Phi^{-1}(u_i))}, \quad 0 \leq u_1, \dots, u_d \leq 1,$$

where ϕ and Φ are the pdf and cdf of $\mathcal{N}_d(0, I)$ standard normal distribution and ϕ_P is the pdf of $\mathcal{N}_d(0, P)$

Examples of copulas

- The **Gaussian Copula** is the copula implicit in a multivariate Gaussian distribution. Its density can be expressed by:

$$c_P(u) = \frac{\phi_P(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))}{\prod_{i=1}^d \phi(\Phi^{-1}(u_i))}, \quad 0 \leq u_1, \dots, u_d \leq 1,$$

where ϕ and Φ are the pdf and cdf of $\mathcal{N}_d(0, I)$ standard normal distribution and ϕ_P is the pdf of $\mathcal{N}_d(0, P)$

- Vine copula** models construct multivariate copulas using only bivariate copulas as building blocks. The density of a vine copula is given by:

$$\begin{aligned} f(x_1, x_2, x_3) &= c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2) \\ &\quad \times c_{32}(F_2(x_2), F_3(x_3)) \\ &\quad \times c_{12}(F_1(x_1), F_2(x_2))f_1(x_1)f_2(x_2)f_3(x_3), \end{aligned}$$

where $c_{13;2}(\cdot, \cdot; x_2)$ represents the copula density associated with the conditional distribution of (X_1, X_3) given $X_2 = x_2$, $F_{j|G}$ is the conditional distribution function of X_j given variables in G .

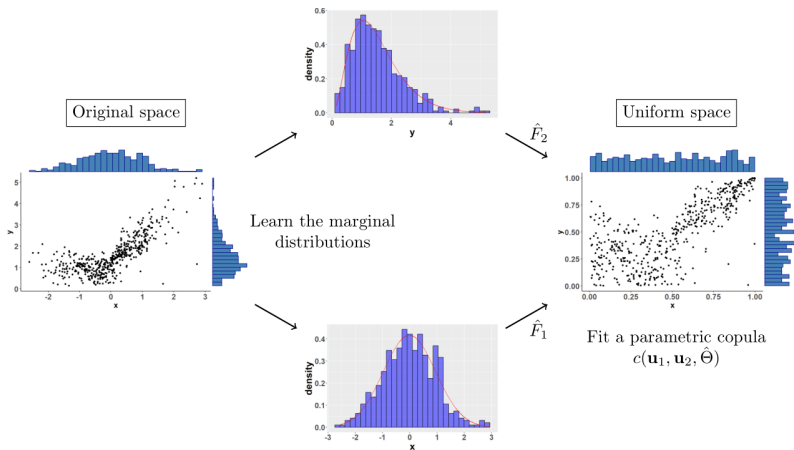
Gaussian Mixture Copula Model

- The GMCM is the copula implicit in a Gaussian mixture model, it allows to capture multi-modality in the dependence structure
- The density of a GMCM with parameters Θ is given by:

$$c_{gmc}(u_1, \dots, u_d; \Theta) = \frac{\psi(\Psi_1^{-1}(u_1), \dots, \Psi_d^{-1}(u_d))}{\prod_{i=1}^d \psi_i(\Psi_i^{-1}(u_i))}, \quad 0 \leq u_1, \dots, u_d \leq 1$$

where ψ_j and Ψ_j^{-1} denote the marginal df and inverse cdf of the GMM along the j -th dimension .

Copula-based synthetic data generation



- 1 Motivating problem
- 2 Copula-based synthetic data generation
- 3 Evaluation of synthetic data**
- 4 Conclusion

Empirical evaluation of synthetic datasets

- **Fidelity:** Measures how closely the synthetic dataset's distribution matches that of the real data.

Empirical evaluation of synthetic datasets

- **Fidelity:** Measures how closely the synthetic dataset's distribution matches that of the real data.
- **Utility:** Assessed by its effectiveness in practical applications, such as downstream regression or classification tasks.

Empirical evaluation of synthetic datasets

- **Fidelity:** Measures how closely the synthetic dataset's distribution matches that of the real data.
- **Utility:** Assessed by its effectiveness in practical applications, such as downstream regression or classification tasks.
- **Privacy:** Ensures no duplication of individual records, safeguards sensitive information, and protects against potential attacks.

Fidelity evaluation: Visualization

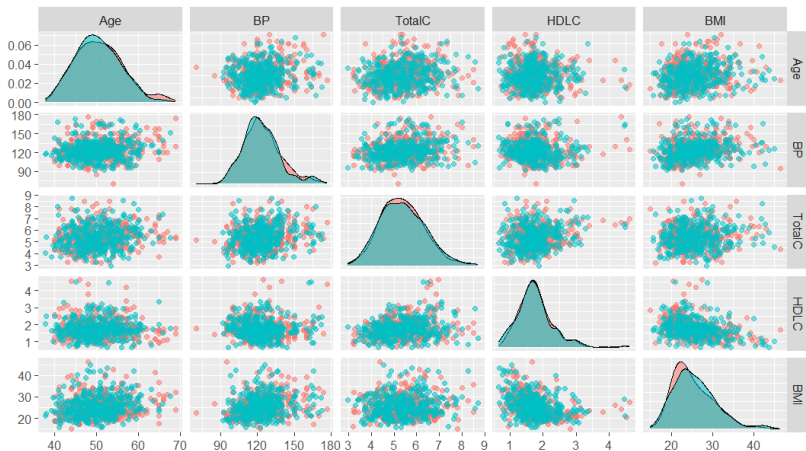


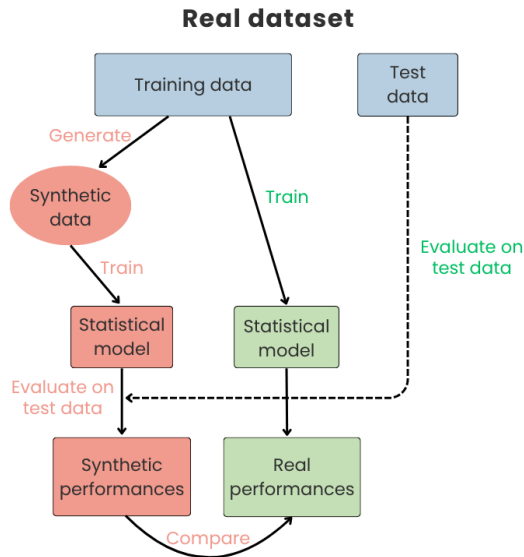
Figure 1: Synthetic data generated with the Gaussian Mixture Copula model (blue) vs original data (red) from the CIMBOLIC dataset

Fidelity evaluation: metrics

Generator Type	MMD (e-3) ↓	JSD (e-3) ↓	IKL (e-1) ↑	P-PR ↑	P-RE ↑
GC	31.74 ± 0.62	13.94 ± 0.58	7.71 ± 0.14	0.99 ± 0.003	0.88 ± 0.02
CTGAN	44.59 ± 1.37	31.60 ± 0.58	7.97 ± 0.04	0.90 ± 0.009	0.98 ± 0.006
GMCM	3.86 ± 0.41	3.38 ± 0.39	9.90 ± 0.03	0.97 ± 0.004	0.98 ± 0.005
MARG	7.34 ± 0.54	3.35 ± 0.41	9.90 ± 0.037	0.97 ± 0.005	0.98 ± 0.005
ARF	3.10 ± 0.39	3.93 ± 0.44	9.84 ± 0.07	0.97 ± 0.004	0.99 ± 0.004
Vines	3.60 ± 0.33	3.34 ± 0.44	9.91 ± 0.03	0.97 ± 0.004	0.98 ± 0.005

Table 1: Experiment Results Over the Cimboic dataset. All results are reported as the mean $\pm$ std, averaged over 100 synthetic dataset.

Utility evaluation: Train On Synthetic Test on Real (TSTR)



Utility evaluation: Results

Generator Type	CRPS↓	wCRPS (1e-2) ↓	Log-Score ↓
Real	0.340 ± 0.031	5.03 ± 2.16	0.781 ± 0.074
GMCM	0.345 ± 0.035	5.03 ± 2.41	0.804 ± 0.085
Vines	0.345 ± 0.035	5.04 ± 2.47	0.801 ± 0.089
ARF	0.348 ± 0.036	5.14 ± 2.59	0.822 ± 0.083
GC	0.358 ± 0.040	5.23 ± 2.71	0.926 ± 0.169
MARG	0.395 ± 0.044	5.97 ± 3.16	0.937 ± 0.093
CTGAN	0.550 ± 0.042	8.17 ± 2.60	1.26 ± 0.057

Table 2: Performances of SCORE-2 risk prediction with systematic missing values by using different synthetic dataset in the TSTR framework. All results are reported as the mean $\pm$ std, averaged over 100 synthetic dataset and 20 train-test splits.

Privacy Evaluation: Metrics

Three complementary privacy metrics:

- **SensAttack**: can an attacker infer the **SCORE2 risk** from synthetic data and non-sensitive attributes?
- **DCR**: how close is each synthetic record to its nearest real counterpart? (0 = privacy breach, 1 = safe)
- **Permutation test**: are real-vs-synthetic distances statistically distinguishable from within-real distances?

- 1 Motivating problem
- 2 Copula-based synthetic data generation
- 3 Evaluation of synthetic data
- 4 Conclusion**

Conclusion and Perspectives

What we showed:

- Copula-based methods as a **simple and interpretable** approach to tabular synthetic data generation
- A use-inspired evaluation of the **fidelity – utility – privacy** trade-off
- A concrete application to **systematic missing values** in a cardiovascular risk calculator (SCORE2)

Limitations and perspectives:

- Scalability: fitting copula models becomes **computationally expensive** in high dimensions
- Develop a **formal privacy framework** — e.g. membership inference attacks