An information-theoretic evaluation of vine copula models for high-dimensional covariate distributions

Niklas Hartung · Aleksandra Khatova

Institute of Mathematics, University of Potsdam, Germany

Motivation

Virtual populations are used to assess the impact of inter-individual variability on drug exposure and effect. While mechanistic approaches exist for some variables (e.g., lean body weight-based scaling [1]), empirical methods are required in general. Vine copulas can be used as a modelling tool to capture nonlinear relationships and which works for high-dimensional data [2,3].

Open question: Quantitative goodness-of-fit evaluation.

In the here-presented work, we evaluate vine copula models fitted to two high-dimensional clinical datasets in terms of Kullback-Leibler divergence.

Datasets for evaluation

General population: NHANES 2009-2012 [4]

- $d = 10$ continuous physical measurements / health variables
- $n = 1776$ individuals (selection of adults and further processing)

Critically ill population: MIMIC-IV [5]

- $d = 30$ continuous physical measurements / health variables
- $n = 4799$ individuals (selection of adults and further processing)
- Heterogeneous, admitted to the ICU with different conditions

Vine copulas [2]

Copulas separate marginals and dependencies

$$\underbrace{p(x_1, \dots, x_d)}_{\text{multivariate density}} = \underbrace{p(x_1) \cdot \dots \cdot p(x_d)}_{\text{marginals}} \cdot \underbrace{c(F(x_1), \dots, F(x_d))}_{\text{copula}}$$

- Two-stage modelling procedure: first fit marginals, then a copula
- Multivariate copula (e.g. Gaussian) or pair copula decomposition (vines)

Vines organize pair copula decompositions

$$c(u_1, \dots, u_d) = \underbrace{\prod_{e \in E(\mathcal{V})}}_{\text{edge in vine } \mathcal{V}} \underbrace{c_{i_e, j_e | D_e}(u_{i_e}, u_{j_e} | u_{D_e})}_{\text{conditional pair copula}}$$

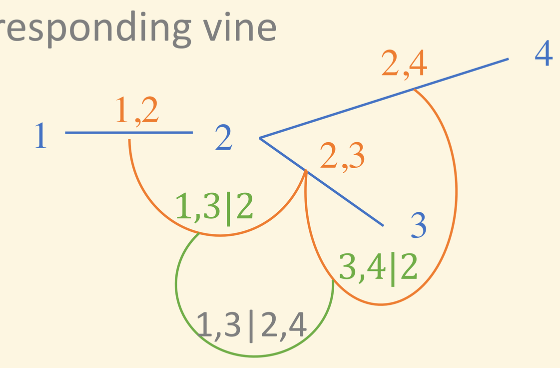
- Pair copulas from different parametric families can be combined
- Selection of vine structure based on pairwise concordance
- Efficient estimation of and simulation from vine copula models [6]

Example ($d = 4$):

Pair copula decomposition

$$c(u_1, \dots, u_4) = c_{1,2}(u_1, u_2) c_{2,3}(u_2, u_3) c_{2,4}(u_2, u_4) \cdot c_{1,3|2}(u_1, u_3 | u_2) c_{3,4|2}(u_3, u_4 | u_2) \cdot c_{1,3|2,4}(u_1, u_3 | u_2, u_4)$$

Corresponding vine



Kullback-Leibler (KL) divergence [7]

$$D_{KL}(p||q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

$p \mathrel{\hat{=}}$ unknown data-generating density
$q \mathrel{\hat{=}}$ density of surrogate model

$D_{KL}(p||q) \mathrel{\hat{=}}$ how much information is lost when using q instead of p

Estimation of KL divergence

(given samples $x^{(1)}, \dots, x^{(n)} \sim p$ and $y^{(1)}, \dots, y^{(m)} \sim q$)

chosen in this work

Density-based:

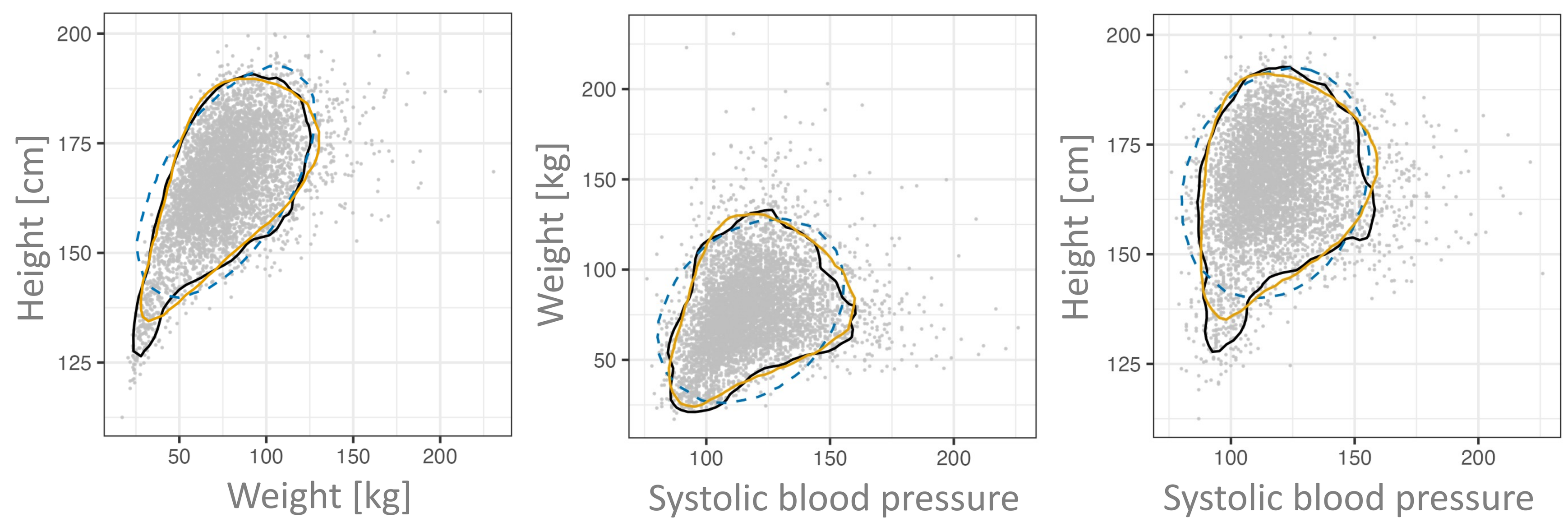
- Accurate when applicable
- Unfeasible for dimensions $d \gg 5$

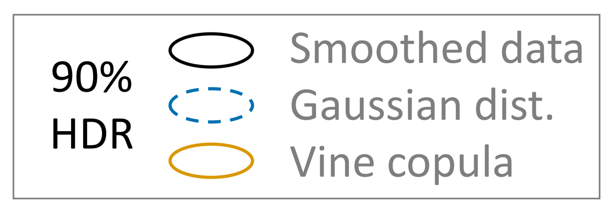
Nearest-neighbour-based:

- Better scaling with dimension d
- Need to correct finite sample bias [8]

Results

Vine copula fits capture nonlinear dependencies (3D example)

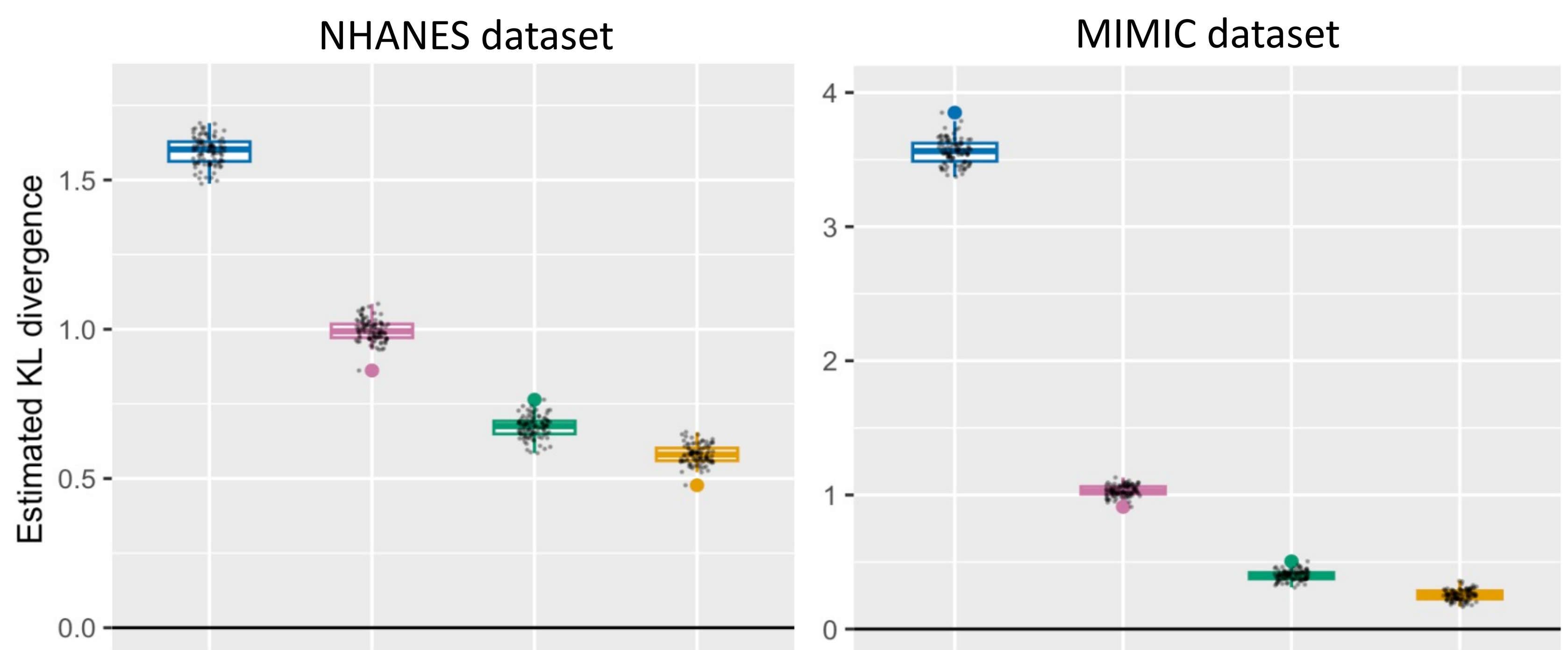


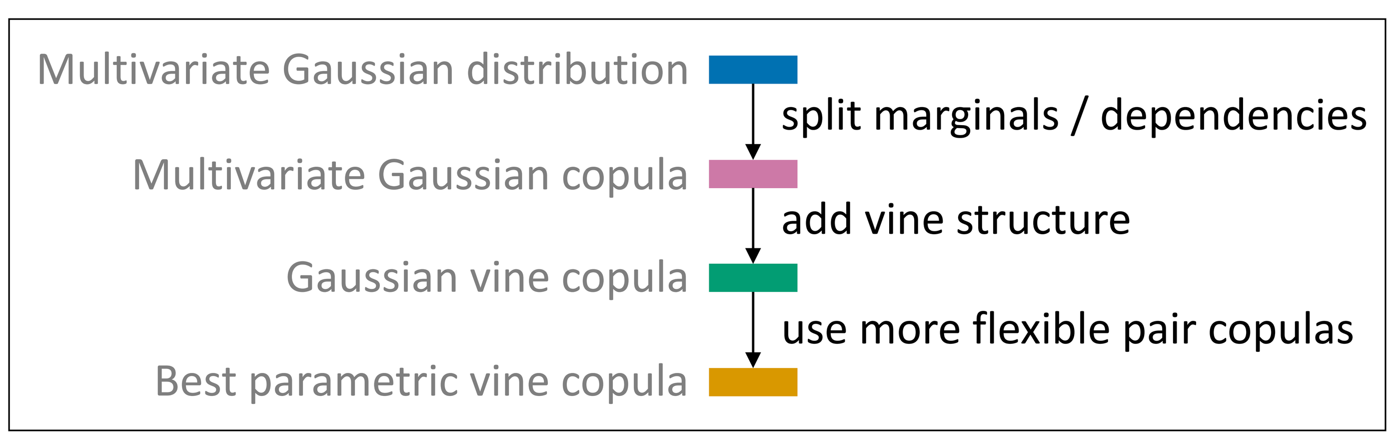


$\widehat{D_{KL}}(p_{\text{Data}}||q_{\text{Gauss}}) = 0.25$

$\widehat{D_{KL}}(p_{\text{Data}}||q_{\text{Vine}}) = 0.04$

Vine copula models have lowest KL divergence





Conclusion

- Vine copulas are more accurate than multivariate Gaussian distributions or copulas
- Lower D_{KL} for MIMIC compared to NHANES (hypothesis: more variability smoothens distributions)
- Categorical data are challenging (no functional R implementation available)

References

[1] Huisinga, W. et al. CPT:PSP (2012) 1:e4
[2] Czado, C. Analyzing Dependent Data with Vine Copulas. Lect. Notes Stat. 222 (Springer, 2019)
[3] Zwep, L. et al. PAGE 30 (2022) Abstr 10099
[4] https:// cran.r-project.org/package=NHANES
[5] Johnson, A. et al., MIMIC-IV (version 2.2). PhysioNet (2023). https://doi.org/10.13026/6mm1-ek67
[6] https://cran.r-project.org/package=rvinecopulib
[7] Kullback, S., Leibler, R.A., Ann Math Stat (1951), 22 (1): 79-86
[8] Wang, Q. et al., IEEE Trans Inf Theory (2009) 55(5): 2392-2405

Contact

For additional information, please contact Niklas Hartung,
niklas.hartung@uni-potsdam.de