Bayesian Copula Models for Data Synthesis: A Love Letter and Look Forward

Joseph Feldman

joseph.feldman@duke.edu

Duke University

Coauthors

Dan Kowal, Cornell Statistics and Data Science

Jerry Reiter, Duke Statistical Science

The Outline

- Bayesian Copula Models for Data Synthesis

- Open Problems: Towards Formal Privacy

Synthetic Data for Privacy Preservation

There are many ways to create synthetic data, and I want to convince you copulas deserve a place at the table.

We generally judge synthetic data generators by the following criteria:

1. Utility
2. Privacy – either through the SDL framework or DP
3. Less talked about: Facility of synthesis

Methods for Synthetic data Generation

A brief review of popular methods

Utility requirements: mixed data types, complex dependencies, structural zeros, scalability, etc.

Three common approaches:

- Sequential synthesis (Reiter, 2005)
  - Pros: Flexible, fast, compatible with mixed data types
  - Cons: Difficult to tune in high dimensions
- Deep Learning (Eigenschink et. al, 2023)
  - Pros: Flexible, compatible with mixed data types
  - Cons: Black box, time-consuming
- Bayesian joint models* (Hu et. al 2014, Feldman and Kowal, 2022)
  - Pros: (can be) Flexible, (can be) fast, (can be) compatible with mixed data types
  - Cons: Bayesian computation is a hurdle for practitioners

My Preferred Bayesian Joint Model: The Gaussian Copula

A very simple recipe

The data-generating process for (i.i.d) p-dimensional observations under the Gaussian copula:

$$\boldsymbol{z}_i \sim N(0, C)$$

$$y_{ij} = F_j^{-1}\{\Phi(z_{ij})\}$$

Parameters:

$$(\{F_j\}_{j=1}^{p}, \boldsymbol{C})$$

Idea: Estimate the posterior distribution of copula parameters given confidential data and then generate predictive samples to build a synthetic data set

Utility and Facility Challenges

Utility challenge: Mixed data types

- Use rank and rank-probit likelihoods (Hoff 2007, Feldman and Kowal 2022)
- Software for sampling algorithms: sbgcop (CRAN), GMCImpute (https://github.com/jfeldman396/GMCImpute)

Facility challenge: prior specification for the marginals in high dimension.

- Just use ECDFs! This is a *really* good idea for synthetic data.

$$\tilde{\boldsymbol{z}} \sim N(0, \boldsymbol{C}^s)$$

$$\tilde{y}_{ij} = \hat{F}_j^{-1}\{\Phi(\tilde{z}_{ij})\}$$

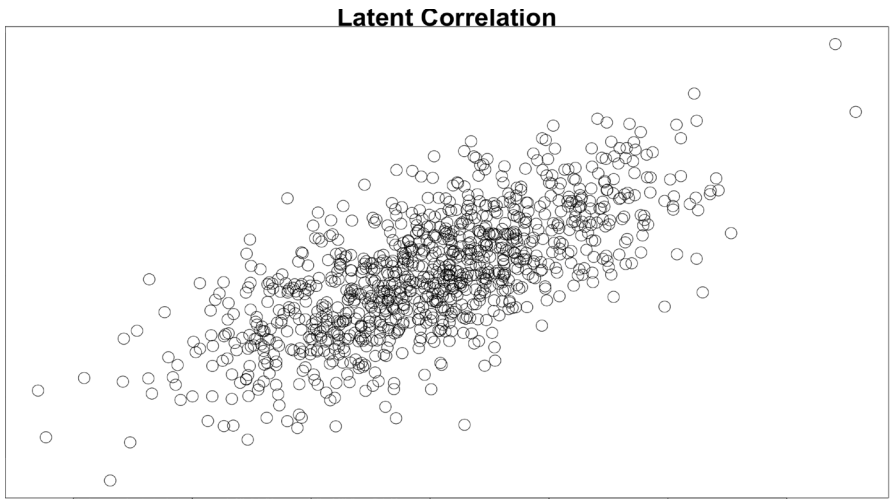
Facility challenge: Scalability of Bayesian sampling

- We've made these samplers *much* faster through coarsening (Feldman, Reiter, and Kowal 2024)

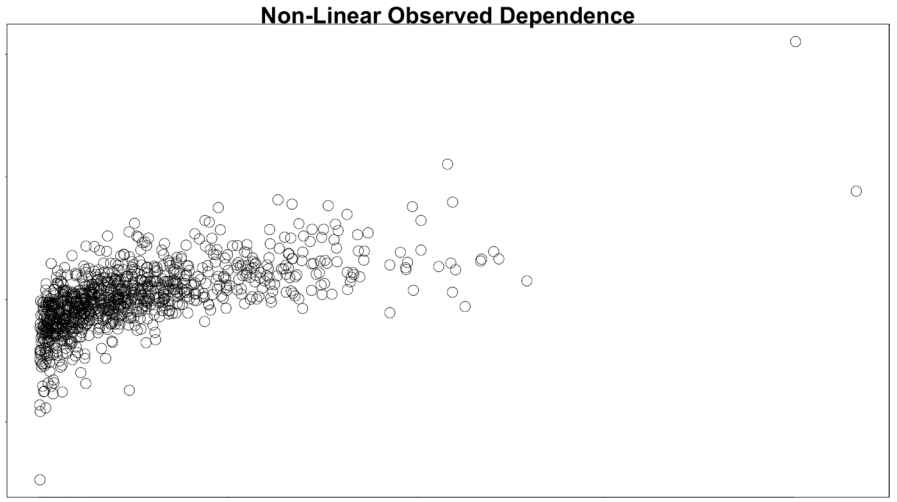
Is the juice worth the squeeze?

Question: Why do all of this work on the Gaussian copula?

Answer: Despite its simplicity, the Gaussian copula can capture complex univariate and multivariate dependencies



Latent, linear associations yield non-linearity on the observed scale!



The takeaway: there is high potential for useful synthetic data

…and we can make them even more flexible

For interactions and more complex non-linearities, we can utilize a Gaussian mixture copula

$$\boldsymbol{z}_i \sim \sum_{h=1}^{H} N(\alpha_h, \boldsymbol{C}_h)$$

$$y_{ij} = \hat{F}_j^{-1}\{\Psi_j(z_{ij})\}$$

An application to North Carolina Educational Achievement Data

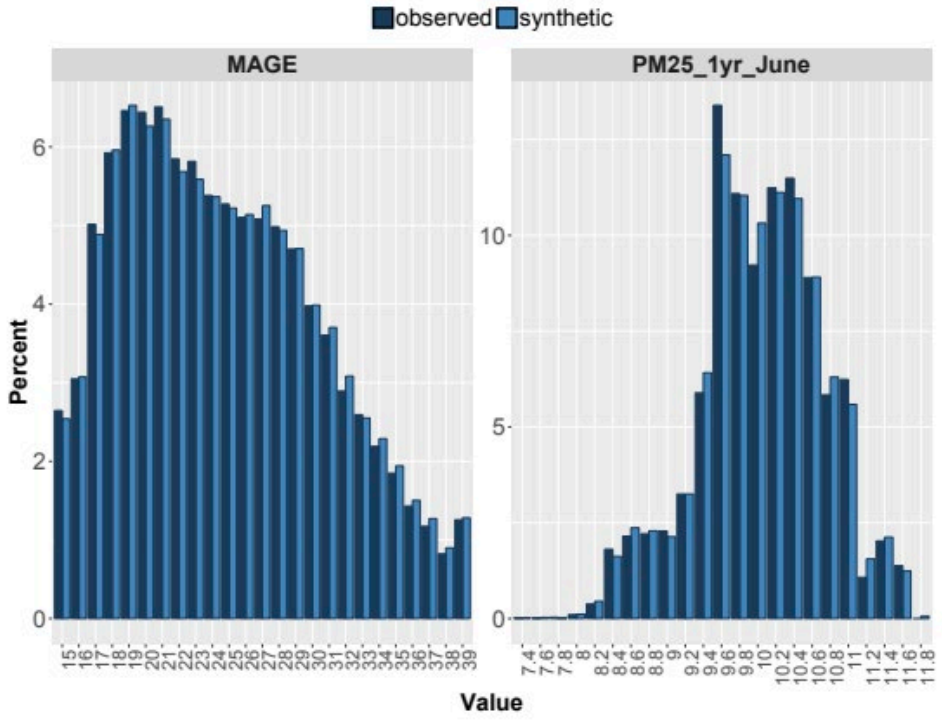
The data is comprised of health, demographic, socio-economic, and academic achievement measurements on 20,000 North Carolina children.

TABLE 1
North Carolina dataset description

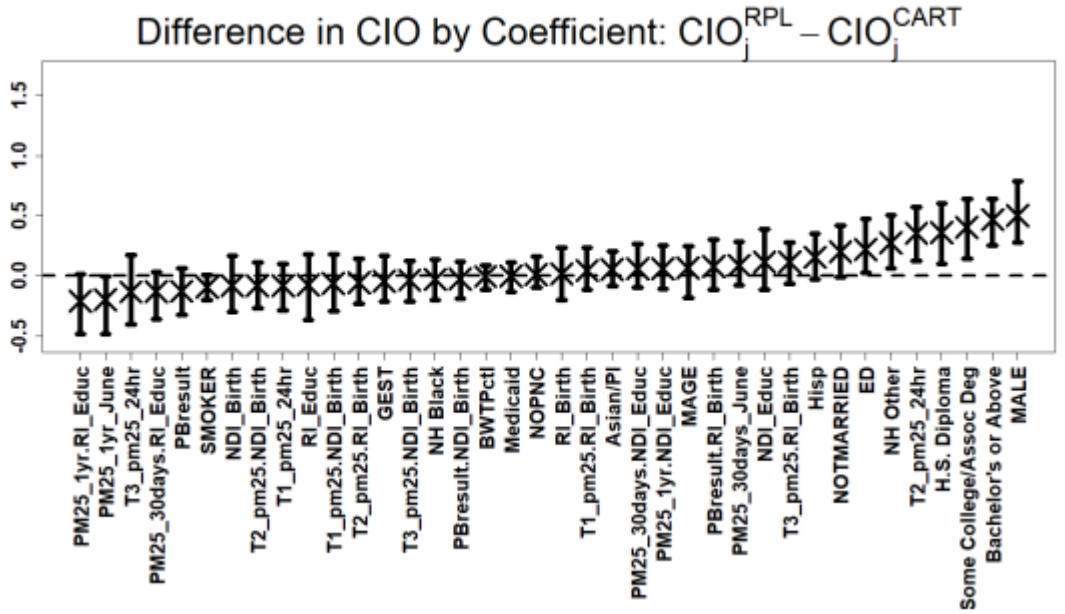
Type	Variable	Range/Values	Label
Categorical	Mother's Race	White, Black, Hispanic, Asian/Pacific Islander, Other	M_RACEGROUP
	Mother's Education	No H.S. Diploma, H.S. Diploma, Some College/Associates Degree, Bachelor's or Higher	M_EDUCGROUP
Binary	Gender	Male, Female	MALE
	Prenatal Care	Yes, No	NOPNC
	Marital Status	Married, Not Married	NOTMARRIED
	Smoker	Yes, No	SMOKER
	Econ. Disadvantaged	Yes, No	ED
	Medicaid	Yes, No	MEDICAID
Integer	EOG Reading Score	Integer 316-370	ReadScal1
	EOG Math Score	Integer 321-373	MathScal1
	Mother's Age (years)	Integer 15-40	MAGE
	Birth Weight Percentile	Integer 0-100	BWTpctl_clin
	Gestational Period (Weeks)	Integer 32-42	GEST
	Blood Lead Test Result	Integer 1-10	PBresult
Continuous	PM 2.5 (by trimester)	PPM 5.60-31.06	Ti_pm25_24hr i = 1,2,3
	Acute PM 2.5 Exposure	PPM 6.026 - 17.891	PM25_30days_June
	Chronic PM 2.5 Exposure	PPM 7.549 - 11.709	PM25_1yr_June
	Racial Isolation at Birth	0 - 1	RI_nhb_Birth
	Racial Isolation at Test	0 - 1	RI_nhb_Educ
	Neighborhood Deprivation Index at Birth	-4.5174 - 11.3888	NDI_Birth
	Neighborhood Deprivation Index at Test	-4.17036 - 10.3524	NDI_Educ

The Gaussian copula creates high utility synthetic data

Univariate



Multivariate



CIO = confidence interval overlap

What about Privacy?

By attribute disclosure risk, it does no worse than other synthesizers…

Open Questions: Towards Formal Privacy

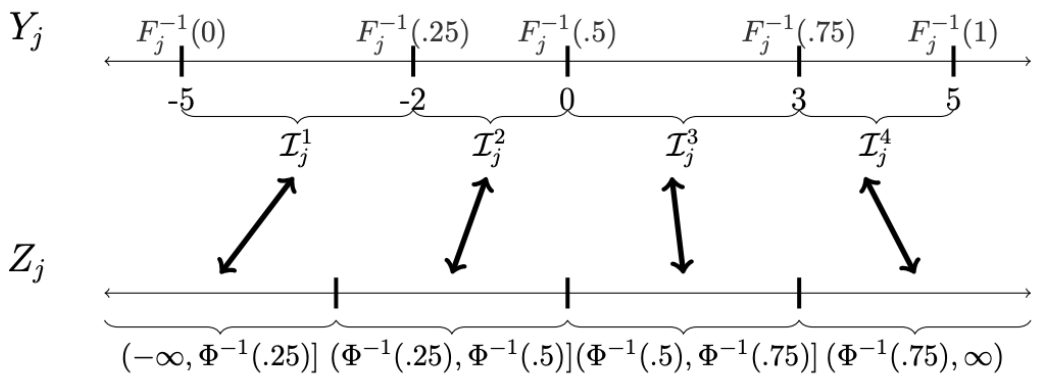
How can we estimate Bayesian models under differential privacy?

Two Ideas:

- Gibbs sampling updates = exponential mechanisms (Foulds et. al, 2021)
  - Accumulate against a privacy budget at each update
- Coarsen the observed data (Miller and Dunson, 2018)
  - Binning yields tractable likelihoods (in terms of sensitivity)

How does this work?

Idea: Marginally, condition on interval memberships $y_{ij} \in \mathcal{I}_j$



Then, the joint probability of any observation is multinomial, which has a tractable global sensitivity

Open questions

1. How do you sample from the coarsened data posterior?
2. What is the trade-off between the level of coarsening in the observed data, global sensitivity of the coarsened likelihood, and accuracy?
3. How can we modify the likelihood to account for different privacy budgets?

Questions?

A couple of our papers on Copula models:

- Bayesian Data Synthesis and the Utility-Risk Trade-Off For Mixed Epidemiological Data (AOAS, 2022)
- Nonparametric Copula Models for Multivariate, Mixed, and Missing Data (JMLR, 2024)
- Gaussian Copulas for Noningnorable Missing Data Using Auxiliary Marginal Quantiles (ArXiv, 2024)