Synthetic Data Generation with Nonparametric Bayesian Methods

Highlight privacy-preserving data analysis

Giovanni Mele, Chiara Noemi Nesti, Chiara Tomasini, Gianluca Villa, Andrea Violante, Stefano Zara

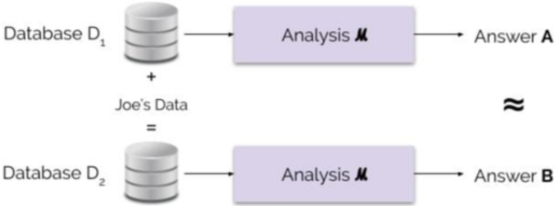
Supervised by:
Prof. Mario Beraha
PhD researcher at Politecnico di Milano

The Role of Data in Society

In today's data-driven world, sensitive information plays a crucial role in generating insights and making accurate predictions. However, its use also introduces significant privacy risks, including the potential for identity theft, highlighting the need for careful handling and protection.

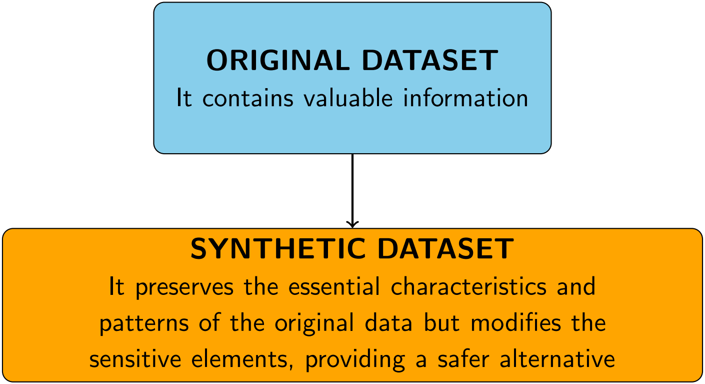
How can we harness the power of sensitive data to drive innovation and inform decisions, while ensuring privacy and ethical use?

Differential privacy



Differential privacy consists in releasing statistical information about datasets while protecting the privacy. This is done by adding some noise.

Goals:



ORIGINAL DATASET

It contains valuable information

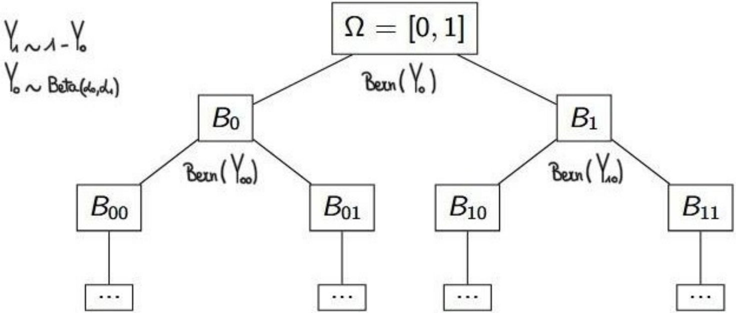
SYNTHETIC DATASET

It preserves the essential characteristics and patterns of the original data but modifies the sensitive elements, providing a safer alternative

The fundamental goal is to simulate synthetic data while preserving real patterns while minimizing privacy risks.

Pólya Tree

A Pólya tree defines a prior distribution on function spaces by recursively partitioning the interval $[0, 1]$ into subintervals. Probabilities are assigned to these subintervals using Beta-distributed random variables. This recursive partitioning process helps define a flexible and rich class of distributions, allowing for complex modeling of uncertainty in function spaces.



Pólya Tree Framework: Definitions and Parameterization

A Pólya tree is denoted by $G \sim \mathcal{PT}(G_0, \mathcal{A})$

Lavine (1992) proposed the following choice to center the $\mathcal{PT}$ around a given centering probability measure G_0.
In short, fix Π as the dyadic quantiles of G_0 and use $\alpha_{\epsilon_0} = \alpha_{\epsilon_1}$ for every level.

Lavine proposes $\alpha_{\epsilon_0 \dots \epsilon_m} = m^2$ as a canonical choice. Walker, Mallick and Paddock updated the model by defining $\alpha_{\epsilon_0 \dots \epsilon_m} = c \cdot m^2$, with $c > 0$.

The posterior distribution of a Pólya Tree remains a Pólya Tree. This is due to the conjugacy property of the model in fact the structure is preserved after updating with observed data.

Pólya Tree is a conjugate model

Letting $G_0 = \{\pi_m\}$ be the sub-interval partition and $\mathcal{A} = \{\alpha_\epsilon : \epsilon \in E\}$ the set of alphas (where $E = \bigcup_{m=0}^{\infty} \{0,1\}^m$),

suppose that $y_1, \ldots, y_n \mid G \overset{\text{i.i.d.}}{\sim} G$, and $G \sim \mathcal{PT}(G_0, \mathcal{A})$. Then

$$G \mid y \sim \mathcal{PT}(G_0, \mathcal{A}^*),$$

where the updated beta parameters $\alpha^*_\epsilon \in A$ are $\alpha_\epsilon = \alpha_\epsilon + n_\epsilon$, being n_ϵ the number of y fallen in ϵ.

Predictive Distribution

Theoretical formulation:

$$P\left(Y_{n+1} \in B_\epsilon \mid y_1, \ldots, y_n\right) = \prod_{m=1}^{M} \frac{\alpha^{*(n+1)}_{\epsilon_m(Y_{n+1})}}{\alpha^{*(n+1)}_{\epsilon_{m-1}(Y_{n+1})_0} + \alpha^{*(n+1)}_{\epsilon_{m-1}(Y_{n+1})_1}}$$

it's the **product of the expected values** of the Beta r.v. at the m-levels

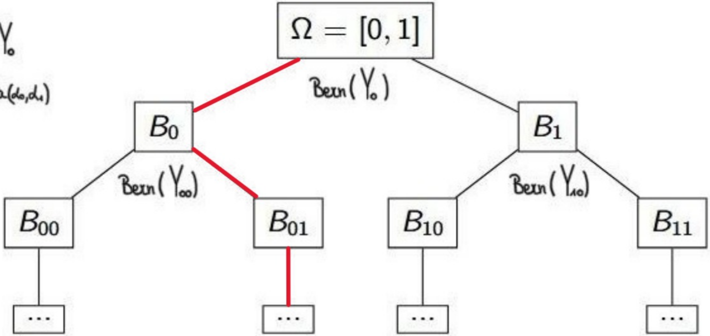
Example:
if we wanted to compute

$$P\left(Y_{n+1} \text{ falls in the bin } \epsilon = 010001011 \mid y_1, \ldots, y_n\right) =$$

$$= \frac{\alpha^*_0}{\alpha^*_0 + \alpha^*_1} \cdot \frac{\alpha^*_{01}}{\alpha^*_{00} + \alpha^*_{01}} \cdot \frac{\alpha^*_{010}}{\alpha^*_{010} + \alpha^*_{011}} \cdots \text{ up to } M = 10$$

$Y_1 \sim 1 - Y_0$

$Y_0 \sim \text{Beta}(\alpha_0, \alpha_1)$



Example: path $\epsilon = 010001011$

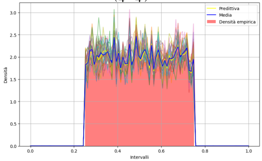
$$P(Y_{n+1} \in B_\epsilon) = \frac{\alpha_0^*}{\alpha_0^* + \alpha_1^*} \cdot \frac{\alpha_{01}^*}{\alpha_{00}^* + \alpha_{01}^*} \cdot \frac{\alpha_{010}^*}{\alpha_{010}^* + \alpha_{011}^*} \cdots \text{ up to } M = 10$$

Exploring the Sensitivity to Parameter c

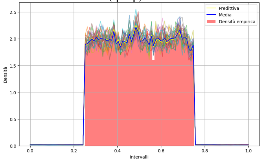
We sampled $y_1, ..., y_n \overset{iid}{\sim} U(\frac{1}{4}, \frac{3}{4})$ and then $y_1, ..., y_n \overset{iid}{\sim} Beta(2, 2)$, with n $= 10000$.

Remember that $\alpha = cm^2$ and $G_0 \sim U(0, 1)$

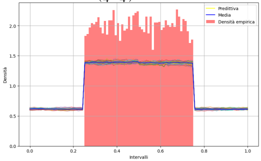
$\mathcal{U}nif(\frac{1}{4}, \frac{3}{4})$, $c = 1$



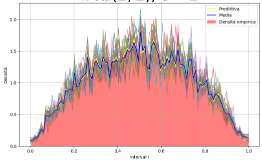
$\mathcal{U}nif(\frac{1}{4}, \frac{3}{4})$, $c = 10$



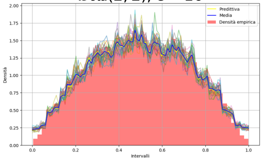
$\mathcal{U}nif(\frac{1}{4}, \frac{3}{4})$, $c = 1000$



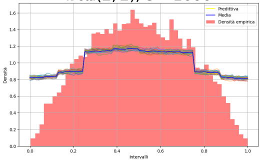
$\mathcal{B}eta(2, 2)$, $c = 1$



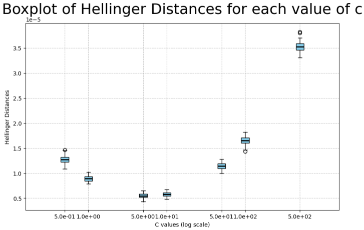
$\mathcal{B}eta(2, 2)$, $c = 10$



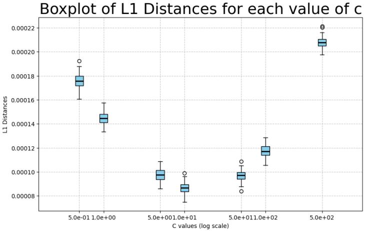
$\mathcal{B}eta(2, 2)$, $c = 1000$



Analyzing the Optimal Choice for Parameter *c*



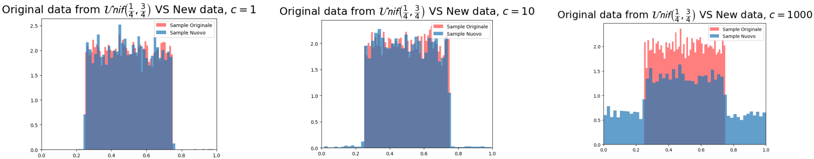
Boxplot of Hellinger Distances for each value of c



Boxplot of L1 Distances for each value of c

A higher *c* leads to greater adherence to the prior, reducing variability, while a lower *c* allows for more flexibility, enabling exploration of diverse distributions. Its choice affects the trade-off between prior belief and adaptability to observed data, particularly as the tree depth increases.

Accuracy of Our Prediction

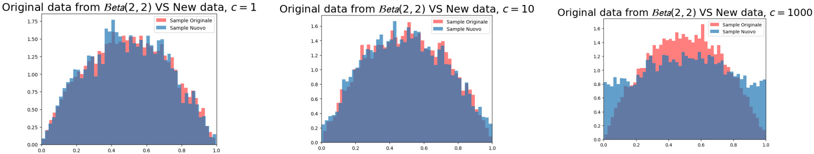


Original data from $\mathcal{U}nif(\frac{1}{4},\frac{3}{4})$ VS New data, $c = 1$

Original data from $\mathcal{U}nif(\frac{1}{4},\frac{3}{4})$ VS New data, $c = 10$

Original data from $\mathcal{U}nif(\frac{1}{4},\frac{3}{4})$ VS New data, $c = 1000$

Predictive density estimated for an $Unif(1/4, 3/4)$



Original data from $\mathcal{B}eta(2,2)$ VS New data, $c = 1$

Original data from $\mathcal{B}eta(2,2)$ VS New data, $c = 10$

Original data from $\mathcal{B}eta(2,2)$ VS New data, $c = 1000$

Predictive density estimated for a $Beta(2, 2)$