A Review of Resampling Techniques for High Dimensional Random Variate Generation

Lekshmi P., Ananya Bajaj, Neha Karanjkar
School of Mathematics and Computer Science, Indian Institute of Technology Goa, Goa, INDIA

September 2024

Abstract

The efficient functioning of a digital twin requires generation of synthetic data from a random variate generation model generated after analyzing the input data and inferring the dependencies among the variates. The inflow of the input data from the real system to the digital twin will be continuous and will be in large volumes. It is important to handle this incoming high volume data and capture the underlying complex correlation to be able to produce high-quality synthetic data. The dimensionality and the correlation in this data will pose significant challenges in input modeling. This article reviews a class of non-parametric methods that use resampling techniques to generate random variates without losing the intricacies in the joint distribution.

Introduction

The process of model generation in digital twins needs to be online and automated for it to be always in sync with the real system data. As the digital twin is stochastic in nature, the input parameters will change with time and should get updated for the digital twin to reflect the real system. So it is important to have a random variate generation model integrated into the simulation model which is capable of analyzing the input data and inferring the correlation in it. However, the dimensionality and complex correlations of streaming data pose significant challenges, requiring the accurate capture of joint distributions to produce high-quality data. Copulas prove to be invaluable in these contexts, providing parametric, non-parametric, and semi-parametric methods to accurately model dependencies.

Brief review of resampling techniques used for random variate generation of high-dimensional data

Parametric techniques involve making certain assumptions about the shape or pattern of the unknown joint distribution of multivariate data. Common families used to represent the dependence structure or copula of the multivariate data include Gaussian and Gumbel. The parameters corresponding to these families are determined using techniques such as Maximum Likelihood Estimation and the Method of Moments. The parametric copula approach assumes that the probability density function (PDF) of the observed data follows a specific form and works to find the parameters of this predefined distribution function. The tools [1,2] have routines to model parametric copulas.

In contrast, nonparametric copula methods do not make any assumptions about the shape of the copula; instead, they capture it directly from the data. Common approaches in nonparametric methods involve inherent smoothing and resampling from the available data [3,4]. The class of methods involving resampling is advantageous when the underlying distribution of the data is unknown or complex, and parametric assumptions may not hold. However, these methods also have inherent flaws.

The method in [4] requires the entire input data to be stored and maintained to sample a new data point from it, which poses a data storage issue in the case of digital twins where streaming input data must be managed. Random variates are generated by using a uniformly distributed integer to randomly index into a list. New points are then generated near the observed data by sampling from a uniform distribution within the bin where the sample falls. Figures 1-8 show the results of a study evaluating the quality of the generated data from the resampling technique by varying the input size and the number of bins. When the number of bins is large, there will be gaps in the distribution if there are bins with no observed samples. And, when the number of bins is small, the distribution will be a piecewise continuous distribution due to flattening within the bins. If the input data is sparse, this method produces a joint distribution that is not smooth, due to the assumption of uniform distribution within each bin.

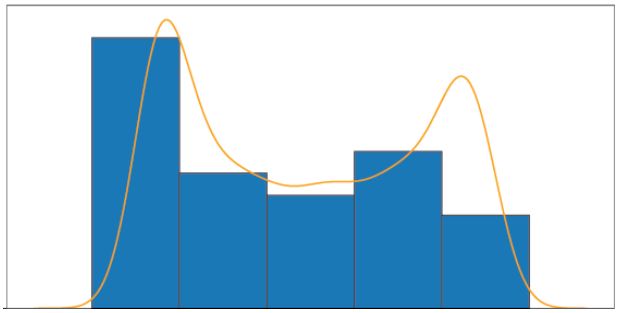

Figure 1: Bins=5; input size=10000

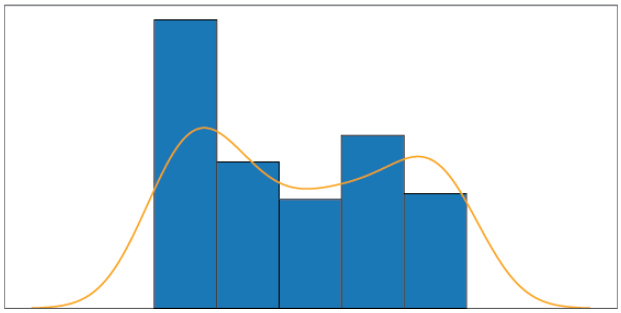

Figure 2: Bins=5; input size=100

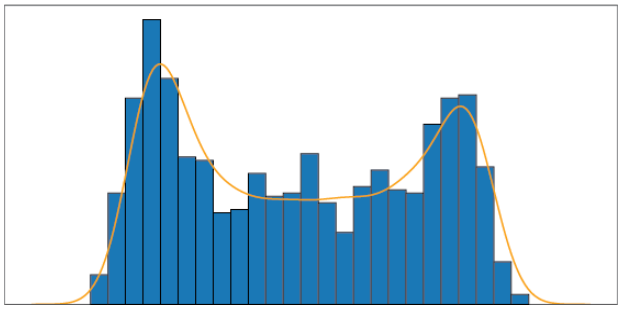

Figure 3: Bins=25; input size=10000

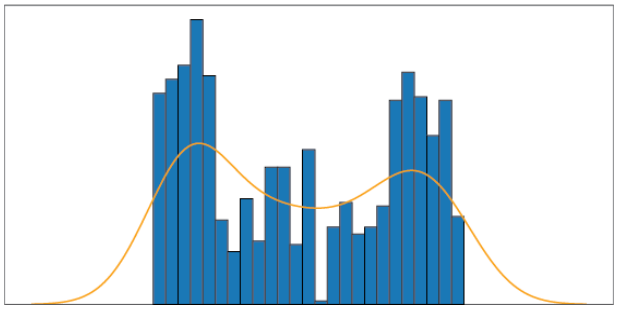

Figure 4: Bins=25; input size=100

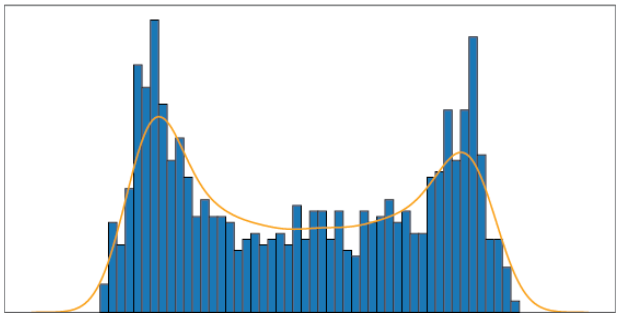

Figure 5: Bins=50; input size=10000

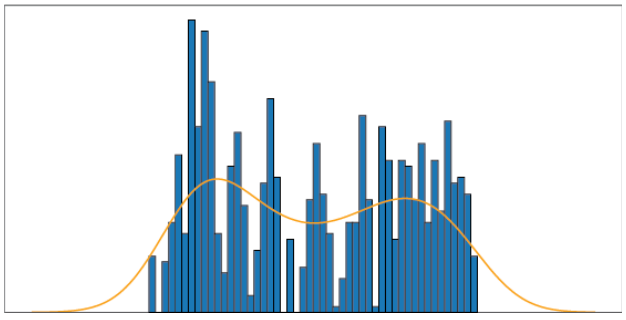

Figure 6: Bins=50; input size=100

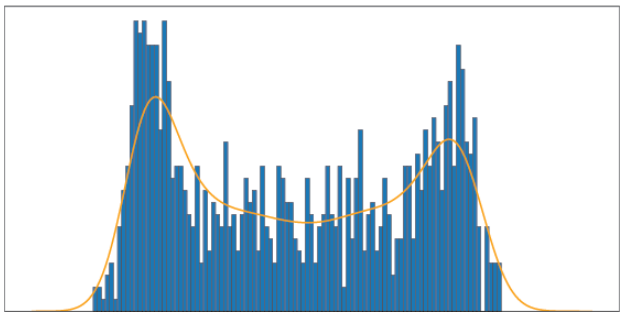

Figure 7: Bins=100; input size=10000

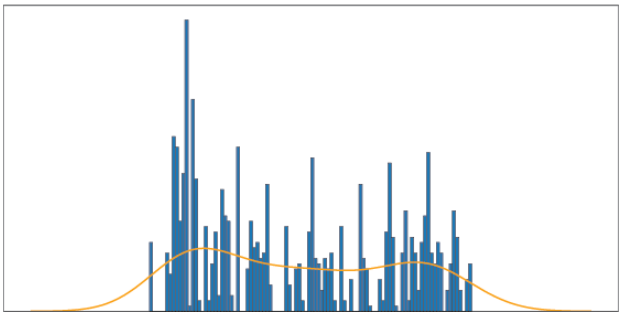

Figure 8: Bins=100; input size=100

References

[1]Maxime Jumelle, PyCopula,"PyCopula: PyCopula is an easy-to-use Python library that allows you to study random variables dependencies with copulas", Online, url:*https://pypi.org/project/copulas*, *2024.*

[2]Jacob L. Chapman, and Daniel Bok, Copulae, "Copulae:Toolbox to model copula",Online, url:*https://copulae.readthedocs.io/en/latest/getting-started.html, 2024.*

[3]R. R. Barton, and L. W. Schruben, "Resampling methods for input modeling", *Proceeding of the 2001 Winter Simulation Conference, 2001.*

[4]Juan Restrepo, Juan Rivera, Henry Laniado, and Pablo Osorio. "Nonparametric generation of synthetic data using copulas", *Electronics, 2023.*