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

We address the challenge of high-dimensional input modeling and random variate generation for Discrete-Event Simulation (DES) based Digital Twins (DTs). In a DES based DT, the modeler has information about the behavior and structure of the system being modeled. Oftentimes there will be dependence or correlation among the entities in the system and have to be modeled as multidimensional random variates. The process of input modeling should be online and automated for the DT to reflect the real system by leveraging the high volume, high-dimensional data from the real system. This calls for robust and efficient methods to infer correlations and fit joint distributions from data in real-time, and to generate random variates from these distributions during simulation. This paper reviews, classifies, and critiques existing methods for high-dimensional random variate generation, focusing on their suitability for DES-based DT applications.

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. This review was performed over publications from 2000-2023 in data-science and simulation-focussed venues discussing input modeling and random variate generation methods for DES based digital twin applications. The methods can be broadly classified into Statistical and Machine Learning based approaches. In most statistical techniques, the underlying dependence structure is captured by modeling the joint distribution using copulas. Copula is a set of mathematical tools that have the ability to connect two or more time-independent variables[1]. This class of techniques can be broadly classified into parametric, non-parametric and semi-parametric methods. Parametric methods are where assumptions are made about the shape of the joint distribution of the multidimensional random variates(eg, gumbel, gaussian, etc [2,4,5]. In nonparametric, the dependence structure is captured empirically from data [6-9]. Further, there are Machine learning and deep learning techniques that help in achieving this objective. Machine Learning techniques involve studying the data, analyzing its dependencies and obtaining an assumed distribution from which the new data points are then sampled[9]. This is achieved by training the models on input data. Machine Learning methods use classification and regression trees, bagging, random forests, support vector machines and Nearest-Neighbour based approaches[3,11,12,13]. Deep Learning offers promising ways to capture the dependence and correlation in multivariate data. However, training a deep learning network like VAE, LSTM, GAN, etc requires a lot

of data, time, and computation cost. Neural Input Modeling is a framework that employs a generative neural network for learning and generating multivariate data. The method in [10] uses vine copula, LSTM and reinforcement learning where copula methods and machine learning methods are combined.

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

Parametric approaches in statistical 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 [4,5] 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 [6,14]. 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 [8] 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.

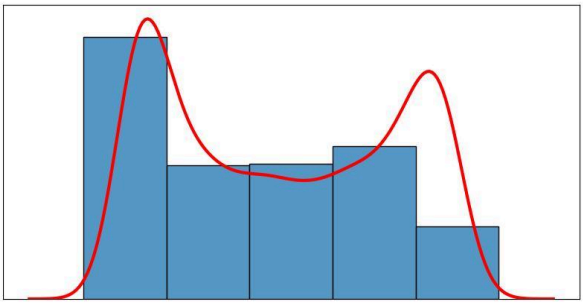

Figure 1: Bins=5; input size=10000

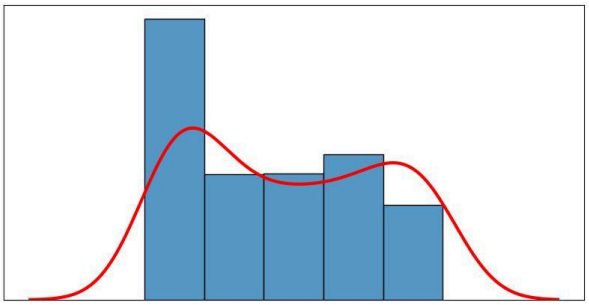

Figure 2: Bins=5; input size=100

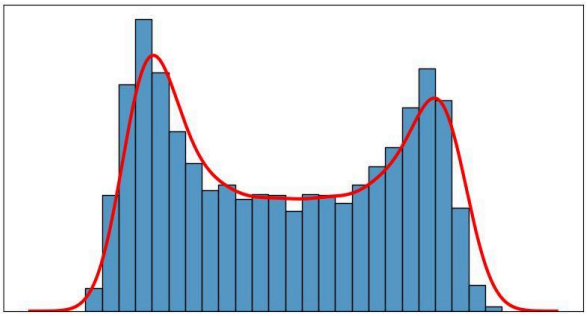

Figure 3: Bins=25; input size=10000

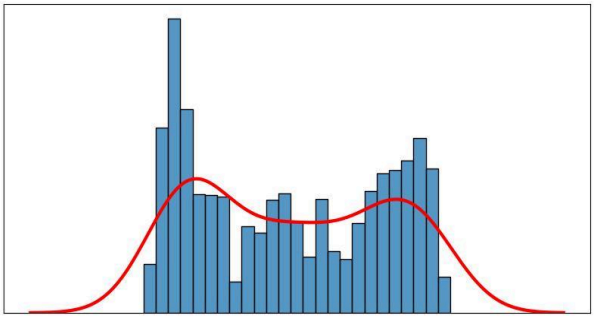

Figure 4: Bins=25; input size=100

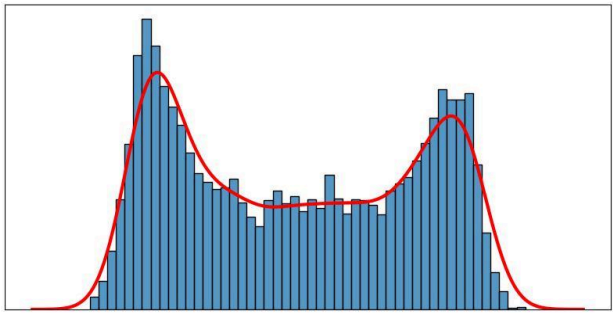

Figure 5: Bins=50; input size=10000

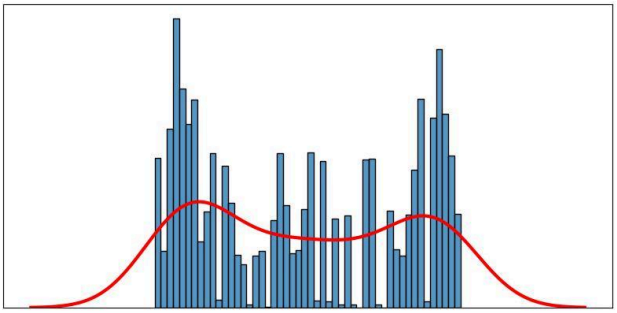

Figure 6: Bins=50; input size=100

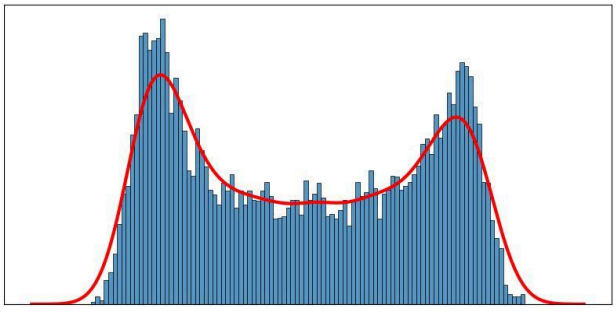

Figure 7: Bins=100; input size=10000

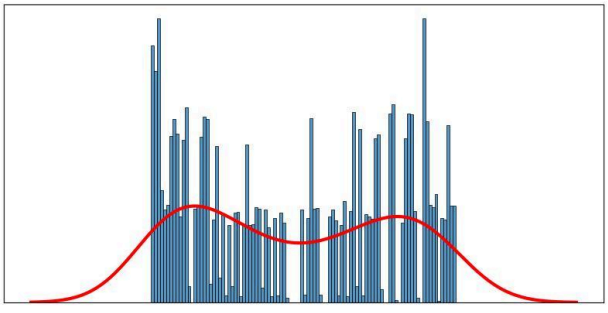

Figure 8: Bins=100; input size=100

References

[1]Roger B. Nelsen. An Introduction to Copulas (Springer Series in Statistics), *Springer*, 2006

[2]Fodil Benali, Damien Bodénès, Nicolas Labroche, Cyril de Runz. MTCopula: Synthetic Complex Data Generation Using Copula. *23rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP)*, 2021 .

[3]Beata Nowok, Gillian M. Raab, and Chris Dibben. synthpop: Be-spoke creation of synthetic data, *Journal of Statistical Software*, 74(11):1–26, 2016.

[4]Maxime Jumelle, PyCopula. PyCopula: a python library for fitting parametric copulas, *Online* url:https://pypi.org/project/copulas/, 2024

[5]Cristiano Tamborrino, Daniel Bok, Jacob Chapman Copulae. Copulae: Python Toolbox to model copula , *Online,* url: https://github.com/DanielBok/copulae, 2024

[6]Song Xi Chen and Tzee-Ming Huang. Nonparametric estimation of copula functions for dependence modelling. The *Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 35(2):265–282, 2007

[7]Y. I. Ngounou Bakam, and D. Pommeret. Nonparametric estimation of copulas and copula densities by orthogonal projections. *Econometrics and Statistics,* 2023

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

[9]Wang Cen, and Peter J. Haas. Nim: Generative neural networks for automated modeling and generation of simulation inputs. *ACM Trans. Model. Comput. Simul.,* 33(3), aug 2023

[10]Yi Sun, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Learning vine copula models for synthetic data generation, *AAAI Conference on Artificial Intelligence,* 2019

[11]Ali Furkan Kalay. Generating synthetic data with the nearest neighbors algorithm, *ArXiv* url:https://arxiv.org/abs/2210.00884, 2022

[12]Jörg Drechsler, Jerome P. Reiter. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets, *Computational Statistics & Data Analysis*, 2011

[13]Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In 2016 *IEEE International Conference on Data Science and Advanced Analytics (DSAA),* 2016

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