Kernel Density Estimation-based simulation of Monte-Carlo events at LHC

Nidhi Tripathi[1], **Salah-Eddine Dahbi,**[1] **Abhaya Kumar Swain**[1], **Xifeng Ruan** [1], **Bruce Mellado** [1,2]

[1]School of Physics and Institute for Collider Particle Physics, University of the Witwatersrand, Johannesburg, Wits 2050, South Africa
[2]iThemba LABS, National Research Foundation, PO Box 722, Somerset West 7129, South Africa

E-mail: `nidhi.tripathi@cern.ch`

Abstract. We have developed a machine learning-based generative model to estimate the kernel density of the data using the Gaussian kernel and then have generated additional samples from this distribution. This model uses scikit-learn to generate a list of particle four-momenta from the proton-proton collisions produced at the Large Hadron Collider (LHC). We demonstrate the ability of this approach to reproduce a set of kinematic features, that are used for the search for new resonances decaying to $Z\gamma$, final states at the LHC. This model is constructed to take the pre-processed $Z\gamma$ events and generate sample data with accurate statistics mimicking the original distributions and achieving better performances compared to the standard event Monte-Carlo generators.

1. Introduction

In the modern era machine learning generative models are a widely used technology around the world for synthetic data generation. A team of high energy physicist uses these techniques to increase the efficiency of data analysis and event generation at Large Hadron Collider (LHC). Variational Auto Encoders (VAE) [1, 2] and Generative Adversarial Networks (GAN) [3] are two widely used fast simulation deep generative algorithms in HEP. These two models illustrate the process of generating simulated events more rapidly and precisely reproducing the sample data distribution. However, focus of GAN and VAE is to generate data mostly in image form. These models are difficult to train and need a large amount of data for the purpose [4].

In this paper we investigate the performance of a non-deep machine learning model based on Kernel Density Estimation (KDE) [5], which is a well-known method for density estimation. This model, whose traditional name was the Parzen-Rosenblatt Windows method, learns the distribution of events in a non-parametric manner. The model estimates the density of real data distribution and generates sample data out of that distribution [6].

The largest and most potent particle accelerator and collider in the world is the LHC at the European Organization for Nuclear Research (CERN). At the LHC, the ATLAS [7] and CMS made the Higgs boson discovery in 2012 [8, 9]. Many experiments, including those at the LHC, are now devoted to search for direct evidence of physics beyond the Standard Model (SM) in the wake of this finding. The accelerators allow for a better understanding of the SM's limitations, which is a theoretical model developed to understand elementary particles and their interactions [10]. Most of the experiments at the LHC depend on Monte Carlo (MC) simulator for data generation, which is a time consuming and CPU expensive process. Machine learning based generative models has been integrated with MC to accelerate the efficiency of generating simulated events. The High-Luminosity Large Hadron Collider (HL-LHC) projects are aimed to increase the luminosity in the future. As a result, it will increase the demand of machine learning based generative models to handle the challenge.

2. Method

In this study we chose KDE for our study because it is a non-parametric methodology that necessitates no previous assumption of a distribution function for probabilities and relies on only one parameter, known as bandwidth, to accurately estimate the density. Usually small tabular data sets requires simple modeling methods based on density estimation. There are numerous density estimation methods, such as KDE [11] the Gaussian mixture model [12], and copulas [13].

2.1. Kernel density estimation

To find the shape of the estimated density function, we generate a set of points equidistant from each other and estimate the kernel density at each point. Kernel Density is a non-parametric technique for determining the probability density function of a random sample $(x_1, x_2, ..., x_n)$ from a distribution with unknown density function $f(x)$. The kernel density estimation is defined as follows:

$$p(x) = \frac{1}{nh} \sum_{j=1}^{n} K \frac{(x - x_j)}{h} \tag{1}$$

where h is the bandwidth parameter that enforces the smoothness of density estimation. The only parameter that impacts the model's accuracy is bandwidth [14, 15]. This model learns the density of the given data and generates synthetic data based on the same distribution. To generate synthetic data, we used the python libraries scikit-learn and NumPy, which implement the Ball Tree or KD Tree algorithm. KDE can be implemented in any dimension, however, its performance degrades at high dimensionality. Which is known as the curse of dimensionality. The scikit-learn library allows cross-validation tuning of the bandwidth parameter to obtain the best model and returns the parameter value that maximizes the log-likelihood of data. GridSearchCV is the function we can use to accomplish this, and it requires different bandwidth parameter values.

2.2. Dataset

The Z_γ SM MC samples used in this analysis have been generated using Madgraph5 [16] and the detector level simulation is performed using Delphes(v3) [17]. The MC-simulated Higss Boson signal used in this study for analysis purpose. The analysis focuses around the centre of mass of 150 GeV (132 GeV $< m_{ll\gamma} <$ 168 GeV). The simulated Z_γ background data was used, Higgs-like heavy scalar decaying to Z_γ ($pp \to H \to Z\gamma$) events, where $Z \to e^+e^-$ or $Z \to \mu^+\mu^-$. The kinematic features used in the study are $Z\gamma$ invariant mass $m_{\ell\ell\gamma}$, the transverse momentum, azimuthal angle, pseudo-rapidity and energy of the leading lepton, subleading lepton and photon respectively, $Pt_{\ell_1|\ell_2|\gamma}$, $\Phi_{\ell_1|\ell_2|\gamma}$, $\eta_{\ell_1|\ell_2|\gamma}$, $E_{\ell_1|\ell_2|\gamma}$, missing transverse energy E_T^{miss} and it's azimuthal angle $\Phi_{E_T^{miss}}$, the number of jets N_j, the number of central jets N_{cj}, $\Delta R_{\ell\ell}$ ($\Delta R \equiv \sqrt{(\Delta\eta_{ll})^2 + (\Delta\phi_{ll})^2}$), $Pt_{\ell\ell}/m_{\ell\ell\gamma}$, $\Delta\Phi_{\ell\ell}$, and $\Delta\Phi(E_T^{miss}, Z\gamma)$.

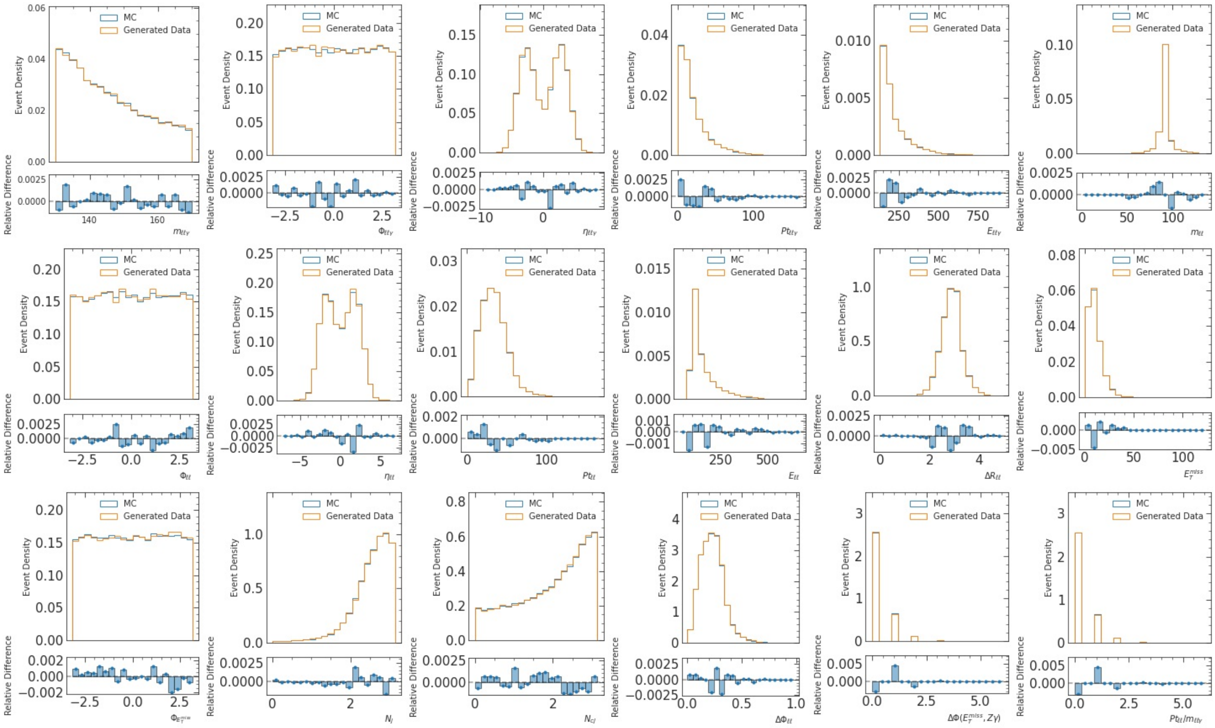

Figure 1. Plots comparing Monte Carlo (MC) and generated data from the $Z\gamma$ data set based on the best selected hyper parameter. Blue and orange plots depict MC and generated data, respectively. The blue columns beneath the data plots show the relative difference.

3. Results

Colab, a free Jupyter notebook environment, was used to run the model, which is powered by an NVIDIA Tesla K80 GPU. A generative model is built using the Scikit-learn and NumPy libraries to take pre-processed $Z\gamma$ data and generate data with accurate statistics that mimic MC data samples from the ATLAS experiment. All the features of both generated and MC data data has been compared with their corresponding local relative difference in, Figure 1. The results shows that generated model works reasonably good to reproduce sample similar to the real data. The features correlation heat-map plots for Monte Carlo, generated data, and the correlation difference between $Z\gamma$ data sets is visualised in, Figure 2.

4. Conclusion

The study presented in this paper describe the performance of Kernel Density Estimation generated synthetic data. The results show that our model generates synthetic data reasonably well. Further efforts are being made to improve the model's consistency and correlation.

References

[1] Kingma D P and Welling M 2013 *arXiv preprint arXiv:1312.6114*

[2] Rezende D J, Mohamed S and Wierstra D 2014 Stochastic backpropagation and approximate inference in deep generative models *International conference on machine learning* (PMLR) pp 1278–1286

[3] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 *Advances in neural information processing systems* **27**

[4] Gramacki A 2018 *Nonparametric kernel density estimation and its computational aspects* vol 37 (Springer)

[5] Kamalov F 2020 *Information Sciences* **512** 1192–1201

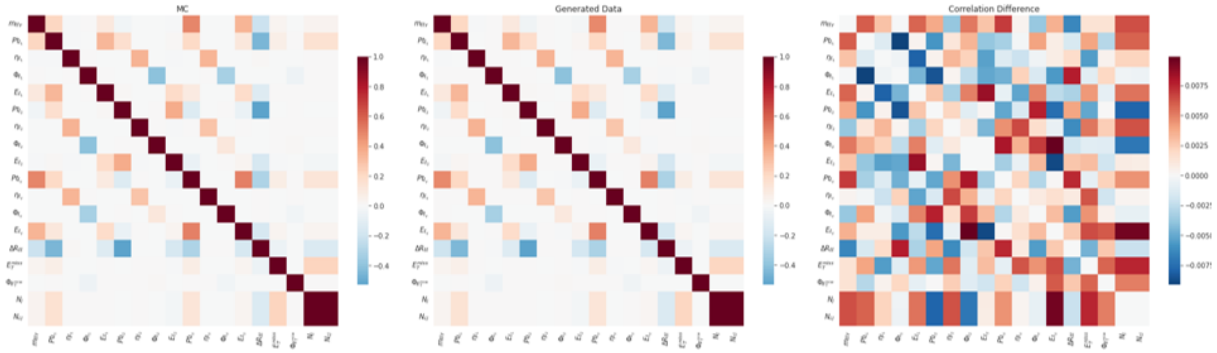

Figure 2. illustrates the comparison of correlation heatmap plots for Monte Carlo, synthetic data, and the correlation difference between $Z\gamma$ data sets. The plots indicate a good correlation.

[6] Fowler E E, Berglund A, Schell M J, Sellers T A, Eschrich S and Heine J 2020 *Journal of biomedical informatics* **105** 103408

[7] Aad G *et al.* (ATLAS) 2012 *Phys. Lett. B* **716** 1–29 (*Preprint* `1207.7214`)

[8] Chatrchyan S *et al.* (CMS) 2012 *Phys. Lett. B* **716** 30–61 (*Preprint* `1207.7235`)

[9] Higgs P W 1964 *Physical Review Letters* **13** 508

[10] Vannerem P, Müller K R, Schölkopf B, Smola A and Soldner-Rembold S 1999 *arXiv preprint hep-ex/9905027*

[11] Plesovskaya E and Ivanov S 2021 *Procedia Computer Science* **193** 442–452

[12] Chokwitthaya C, Zhu Y, Mukhopadhyay S and Jafari A 2020 Applying the gaussian mixture model to generate large synthetic data from a small data set *Construction Research Congress*

[13] Tang X S, Li D Q, Cao Z J and Phoon K K 2017 *Computers and Geotechnics* **87** 229–240

[14] Scott D W 2015 *Multivariate density estimation: theory, practice, and visualization* (John Wiley & Sons)

[15] Silverman B W 2018 *Density estimation for statistics and data analysis* (Routledge)

[16] Alwall J, Frederix R, Frixione S, Hirschi V, Maltoni F, Mattelaer O, Shao H S, Stelzer T, Torrielli P and Zaro M 2014 *Journal of High Energy Physics* **2014** 1–157

[17] Mandrik P, theFCC study group *et al.* 2019 Prospect for top quark fcnc searches at the fcc-hh *Journal of Physics: Conference Series* vol 1390 (IOP Publishing) p 012044