Auto-Encoder based algorithms for anomaly detection

Ioan Dinu	IFNN-HH/LPC
Louis Vaslin	LPC
Julien Donini	LPC

IN2P3/IRFU ML Workshop – 16-17 March 2021

Introduction

- Anomaly detection

Objective : Identification of **new physics** signals without having a priori knowledge on them

Motive : Targeted search (supervised) are biased to find one specific supposed signal
Though we don't know what actual BSM physics looks like

Growing interest of HEP community this last years

Many attempts and challenges on this topic (e.g. LHC Olympics Challenge)

Introduction

- LHC Olympics (January and July 2020) arXiv:2101.08320 [hep-ph]

Challenge : develop **model-independent** Machine Learning anomaly detection methods for BSM searches

Data format : 4-vector particle flow information of multijet events simulated with Pythia and Delphes

Feature Extraction : Jet kinematics, substructure variables or any other observables need to be computed and extracted by applying clustering algorithms

Datasets available :

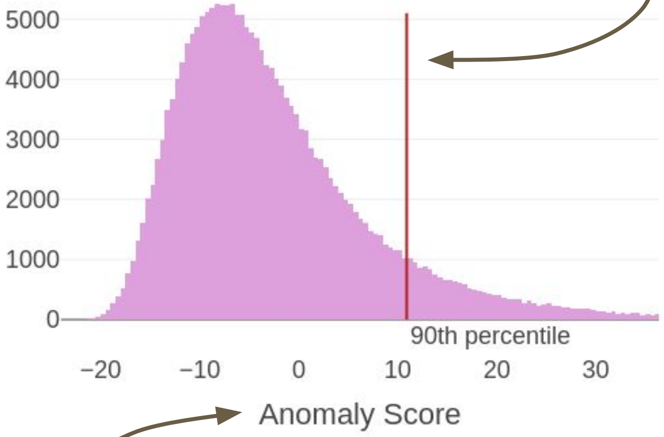
RnD dataset: *QCD background* (1M), *dijet signal* (100k) and *trijet signal* (100k) sample
Background-only training set (1M)
3 different black-boxes with potential signal (1M each)

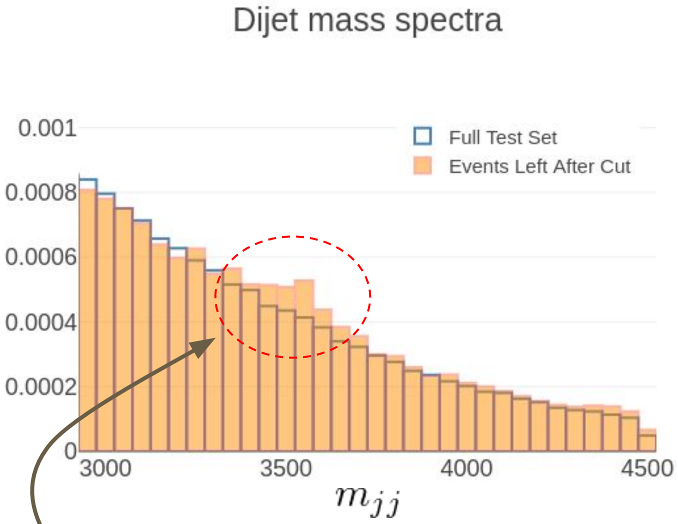
BB1 : 3.8 TeV Z' decaying in dijet with 834 signal event
BB2 : QCD background only
BB3 : 4.2 gKK decaying in dijet and trijet (BR trijet = 0.625)

Note: background is modeled differently across all datasets

Strategy

2. Cut at a threshold





3. Compare spectra

1. Neural Network based anomaly score

Methods

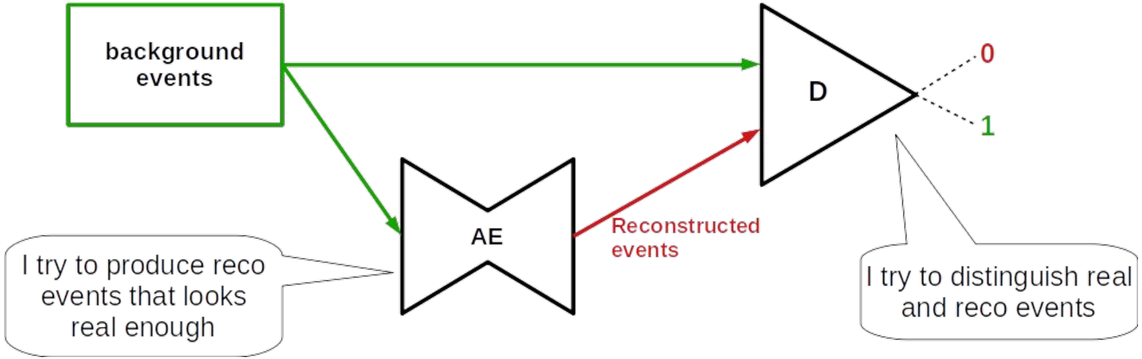
- GAN-AE

Inspired by the principle of GANs

AE and D are trained together with opposite objectives

Goal :

Train the AE using information that don't only comes from reconstruction error


background events
D
0
1
AE
Reconstructed events
I try to produce reco events that looks real enough
I try to distinguish real and reco events

Loss functions :

For D : Binary Crossentropy (BC) trained on a labeled mixture of true and reco events

For AE : BC + ε x Mean Euclidean Distance (MED) + α x DisCo using "wrong" labels for D

see slide 7

Methods

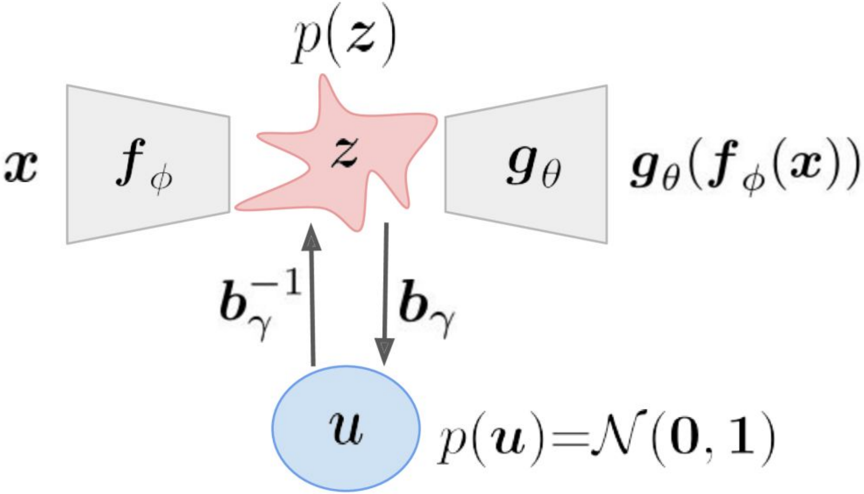
- Probabilistic Autoencoder

  Ref: arXiv:2006.05479 [cs.LG]

Autoencoder: Learns to encode and reconstruct events from a latent representation

Normalizing Flow: Learns a bijective mapping from the latent space to a multivariate normal space.

Both reconstruction error and density of the latent representation are used in order to compute an **anomaly score:**


$p(\boldsymbol{z})$
$\boldsymbol{x}$
$\boldsymbol{f}_\phi$
$\boldsymbol{z}$
$\boldsymbol{g}_\theta$
$\boldsymbol{g}_\theta(\boldsymbol{f}_\phi(\boldsymbol{x}))$
$\boldsymbol{b}_\gamma^{-1}$
$\boldsymbol{b}_\gamma$
$\boldsymbol{u}$
$p(\boldsymbol{u})=\mathcal{N}(\boldsymbol{0},\boldsymbol{1})$

$$\ln p(\vec{x}) \approx -\frac{1}{2}(\vec{x}-\vec{x}')^2 \cdot \vec{\sigma}^{\circ -2} - \frac{1}{2} b_\gamma(\vec{z})^2 + \ln|\det \mathcal{J}_\gamma|$$

Mass Decorrelation Techniques

- Distance Correlation (DisCo)

Inspired by arXiv:2001.05310 $\mathrm{dCorr}^2(\mathrm{X},\mathrm{Y})=\frac{\mathrm{dCov}(\mathrm{X},\mathrm{Y})}{\mathrm{dCov}(\mathrm{X},\mathrm{X})\mathrm{dCov}(\mathrm{Y},\mathrm{Y})}$ with dCov the distance covariance
Act as a regularization term pushing X (ED) and Y (mjj) to be decorrelated

- Sample reweighting

Define sample weights to be applied during training based on their dijet mass
Objective : Make the dijet mass distribution appears “flat” to help with decorrelation

- Quantile transformer

Makes each training feature to be uniformly distributed by applying a different transformation to every quantile in order to mitigate any potential bias.

BumpHunter

- Principle

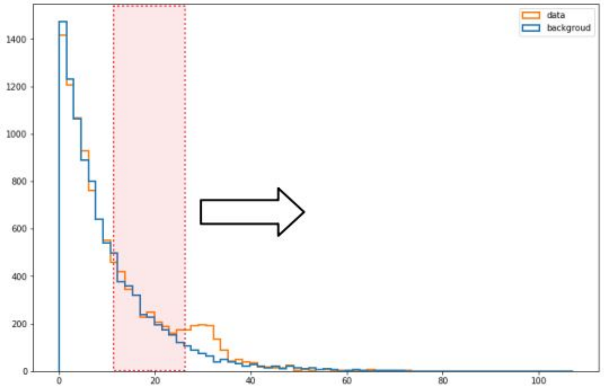
Scan the data histogram and compare it to a reference background

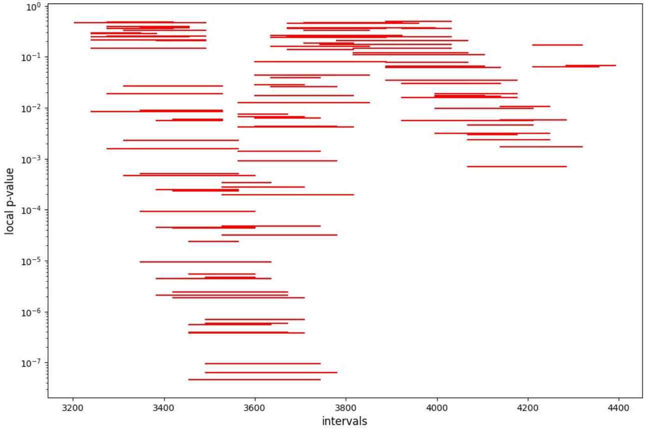
Compute the local and global p-value of the most significant excess in data

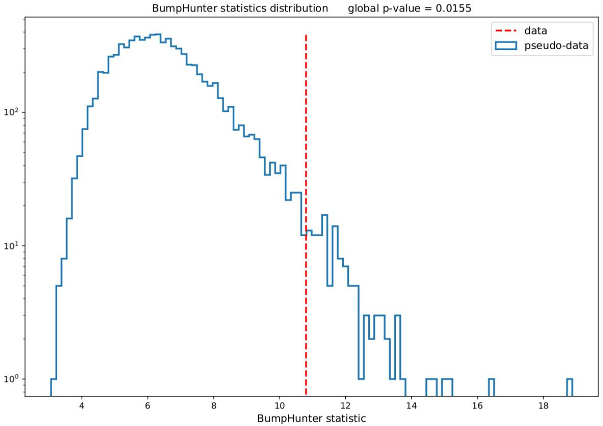
Background shape can be data driven (unsupervised search)

Use side-band normalization to enhance significance

Package available for python3
https://github.com/lovaslin/pyBumpHunter







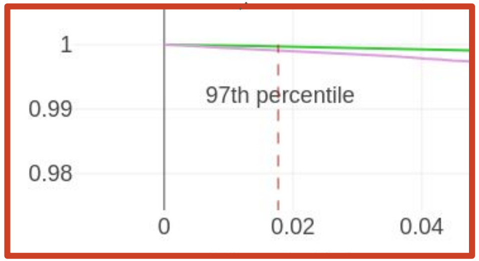
Test on RnD Data

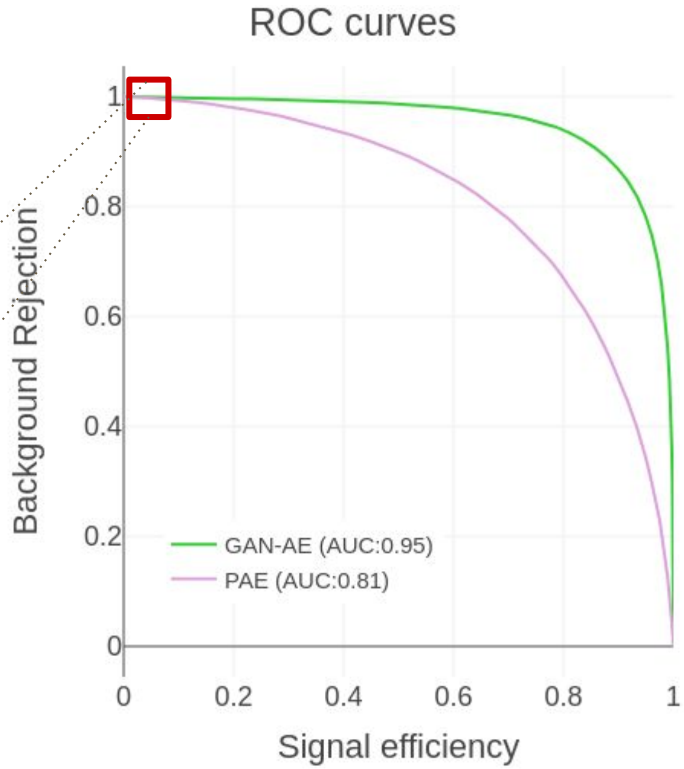
- Balanced test dataset

Testing on a balanced dataset with equal dijet signal and QCD background events from the R&D dataset:

GAN-AE model shows a much more impressive classification performance overall.

For small signal fractions where high anomaly score threshold need to be applied, the differences are not that stark:



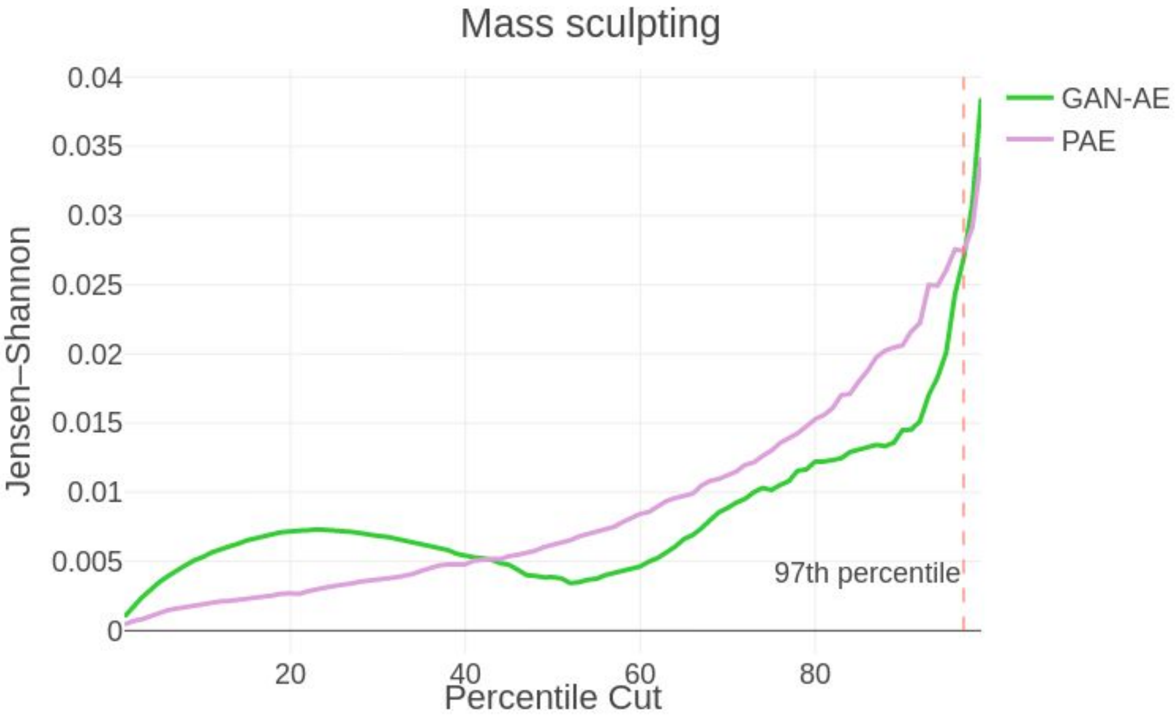


Test on RnD Data

- Mass sculpting

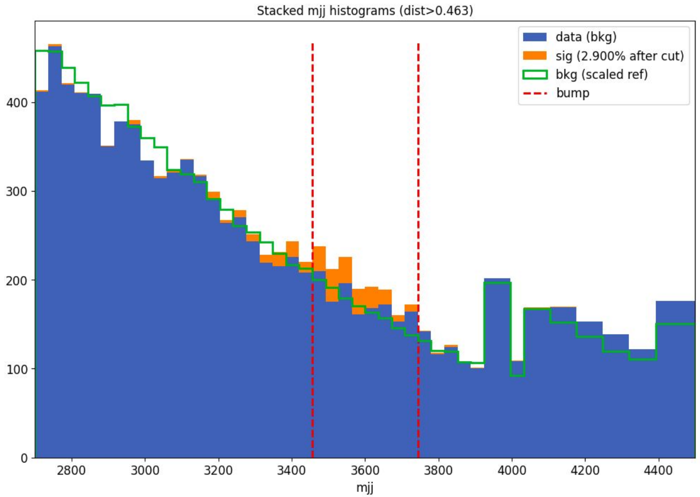
The Jensen-Shannon divergence is a distance metric for distributions which can be used to quantify mass sculpting.

The two model are comparable in terms of mass sculpting:

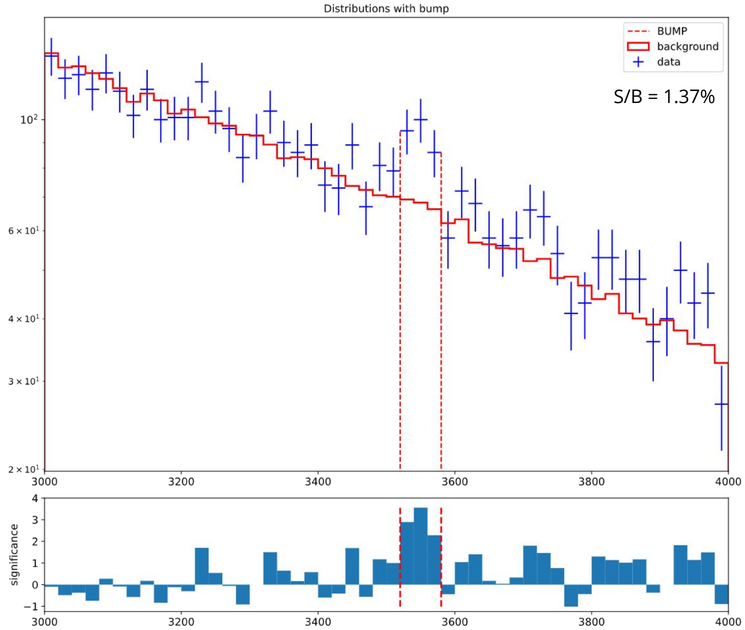


Test on RnD data - Signal Injection

GAN-AE



PAE

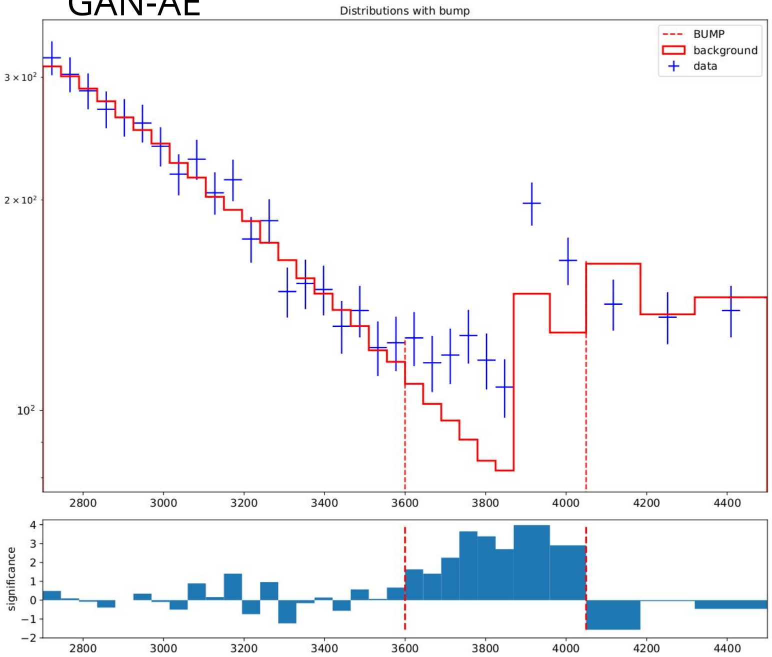


Results obtained with a cut on anomaly score at 95th percentile. Initial S/B ratio : 0.2%

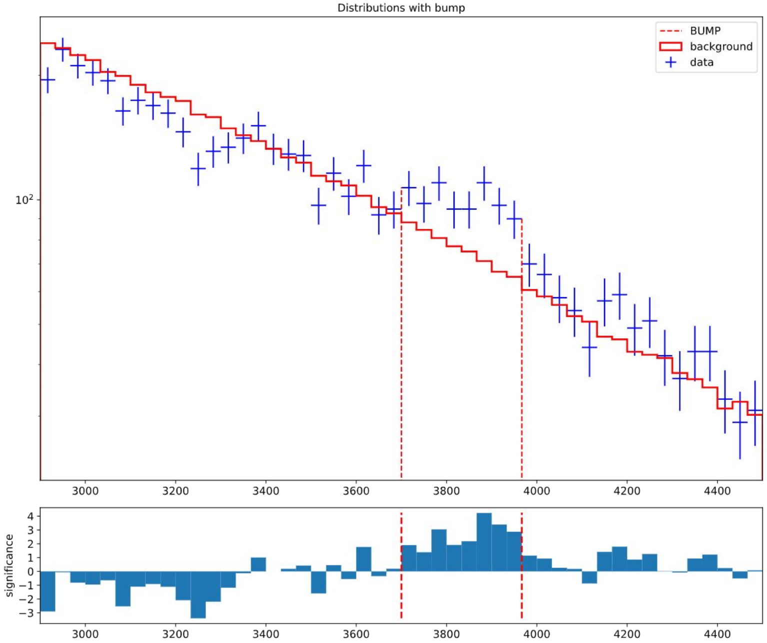
Both methods seem to enhance the signal and BumpHunter is able to find it with global significance > 3σ

Test on Black-box 1 data

GAN-AE



PAE

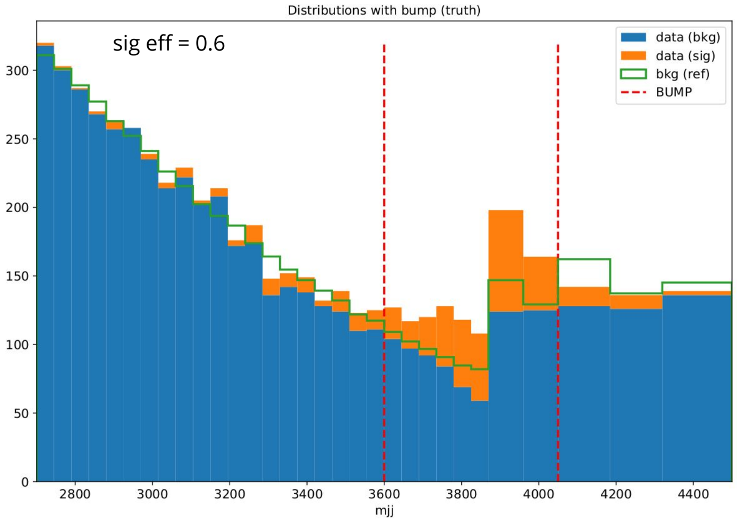


Cut threshold at the 99th percentile of the anomaly score

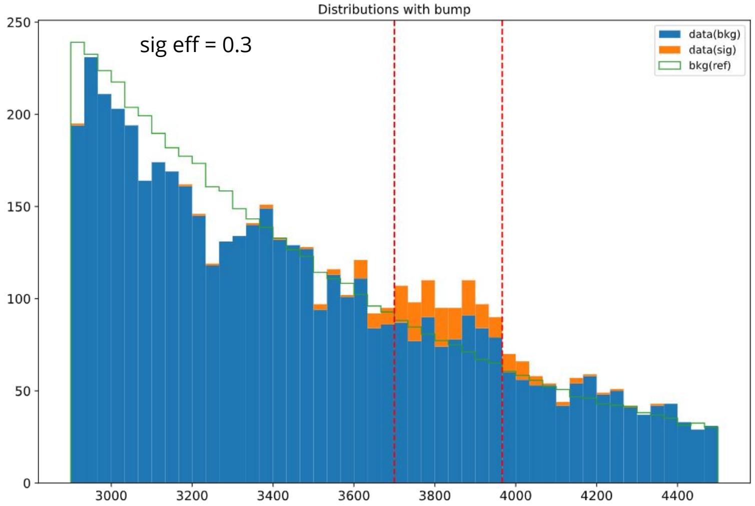
In both case a significant excess seems to appear in the same range

Black-Box 1 Unblinding

GAN-AE



PAE



Initial signal fraction : 834/1M

In both case, the bump found by BumpHunter seems to correspond to the signal with global significance > 5σ

However, the remaining mass sculpting seems to bias a little the significance.

Conclusion

- Results on LHC Olympics

  GAN-AE and PAE both are promising anomaly detection techniques

  Mass sculpting is a limiting factor, but mass decorrelation techniques keep it under control

  The bump hunting strategy is successful for the black-box 1 dataset

- Next steps

  Extend the techniques to trijet events and to the remaining black-boxes

  Adapt the method to work with jet images (Convolutional GAN-AE/PAE)

Thank you for your attention !

BACKUP

Training

- GAN-AE

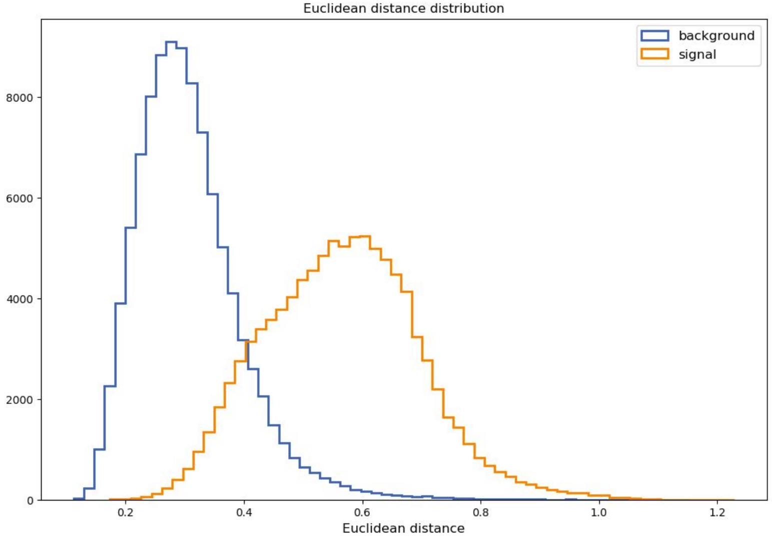
Decorrelation techniques : DisCo and sample reweighting based on dijet mass density
Hyperparameters : ε = 0.3 α = 10
Training on 100k events for 110 cycles (1 cycle = 5 D epochs + 7 AE epochs)

- PAE

Decorrelation techniques : Uniform distribution of features and sample reweighting
Training steps:

1. Train autoencoder on background events with MSE loss
2. Train normalizing flow on latent representation with NLL loss

Test on RnD data - Anomaly score



GAN-AE

GAN-AE use the Euclidean distance between the input and output of the AE.

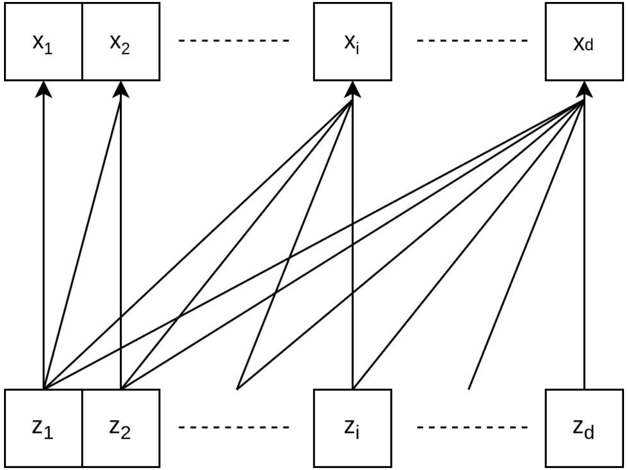
Normalizing Flows - Autoregressive Models

Learn a chain of triangular maps from a multivariate gaussian space to the data space

$$\mathbf{z} \xrightarrow{\mathbf{T}^{(1)}} \mathbf{z}_1 \xrightarrow{\mathbf{T}^{(2)}} \mathbf{z}_2 \dots \xrightarrow{\mathbf{T}^{(k)}} \mathbf{x}$$

Estimate density in the data space using the jacobian determinants of the maps (conservation of probability mass)

$$q(\mathbf{x}) = p(\mathbf{z}) \left|\nabla\mathbf{T}^{(1)}\right|^{-1} \left|\nabla\mathbf{T}^{(2)}\right|^{-1} \dots \left|\nabla\mathbf{T}^{(k)}\right|^{-1}$$



$$\mathbf{x} = \mathbf{T}(\mathbf{z})$$