Unsupervised in-distribution anomaly detection of new physics through conditional density estimation

George Stein[1,2], Uroš Seljak[1,2], Biwei Dai[1]

[1] Berkeley Center for Cosmological Physics, University of California, Berkeley, [2]Lawrence Berkeley National Laboratory, Berkeley, CA, USA

LHC Olympics homepage
lhco2020.github.io/homepage/

Contact me at
gstein@berkeley.edu

Most commonly, **anomaly detection is synonymous with outlier detection** and the identification of Out-of-Distribution (OoD) examples. Applications typically involve the search for samples in low probability density areas of the data near the tails of various data distributions, in order to remove or flag them for further inspection.

Alternatively, a problem may instead motivate the search for *in-distribution anomalies,* loosely defined as a small set of samples that reside in areas of the data with high probability density, but have unique yet unknown properties when compared to their surroundings.

If the given data is expected to be smoothly distributed, **in-distribution anomalies may present themselves as local over-densities in a region of the parameter space**. In this regime it is no longer desirable to only determine if the probability density of a sample is high or low, we instead wish to compare the density of a sample to that of its neighbours along some conditional dimension of interest.

Such unsupervised **in-distribution anomaly searches are ideal for scientific applications with large amounts of data where the observable space, parameter space, or model space may be too large to perform directed searches for some set of already-known signatures**, and blind searches are required instead.

Method

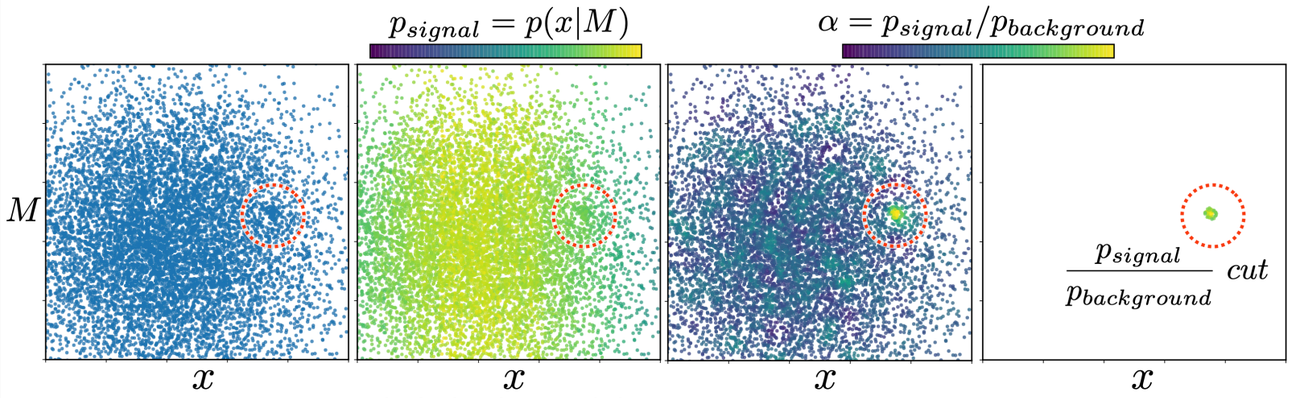
Our method relies on **conditional density estimation**, which aims to model the conditional distribution $p(x|x_c)$ of input data x with conditional parameter x_c by introducing a sequence of differentiable and invertible transformations to a Gaussian distribution.

We use an alternative approach to the current deep learning methodology - Gaussianizing Iterative Slicing (GIS), which iteratively matches the 1D marginalized distribution of the data to a Gaussian.

By determining the density at each data point in comparison to its neighbours along the conditional dimension, we construct an anomaly score.

In-distribution anomaly detection

Finding a needle in a haystack, without knowing you're looking for a needle



Consider samples of a 1D feature x and a conditional parameter of interest M drawn from a smooth Gaussian 'background' with a small number of anomalous 'signal' events added (inside red circle for clarity).

The density values at each data point do not allow the anomaly to be distinguished from the background, as they only identify the outliers in low density regions.

However, comparing the local density (p_{signal}) to neighbouring densities ($p_{background}$) along the conditional dimension - and constructing the anomaly ratio α - results in a peak at the anomalous data points.

Implementing a minimum cut on the anomaly ratio reveals the anomalous events, without knowing what we were looking for in the first place.

Dataset and competition

We apply our new in-distribution anomaly detection method towards the **detection of new physics in simulated Large Hadron Collider (LHC) particle collisions** as part of the *2020 LHC Olympics blind challenge.*

The data for each of the 1 Million collision events consists of the four momenta of the detected particles. By focusing on the jet summary statistics rather than the particle data from an event we vastly reduce the dimensionality of the data space. Each jet J is described by its:

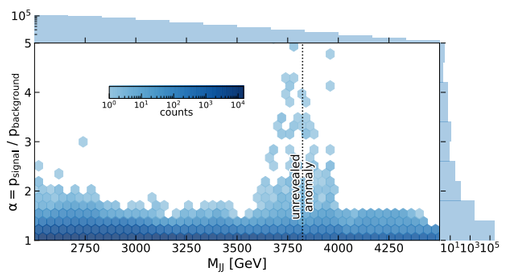
- mass m_J *and* linear momentum p=(p_T, η, ϕ).
- n-subjettiness ratios $\tau_{n,n-1}$, which describe the structure and number of sub-jets within each jet.

A pair of jets has an invariant mass M_{JJ}, which we choose at the conditional parameter of interest

Results

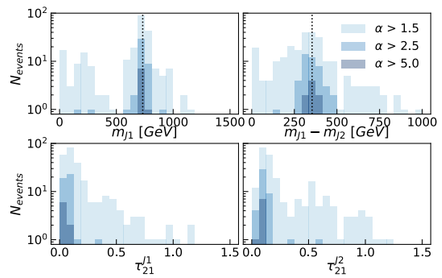
These results are from our original blind submission to the 2020 LHC Olympics, when neither the method or the authors knew if there was any anomaly, or what it might look like.

Step 1. Select desired physical variables and perform in-distribution anomaly search



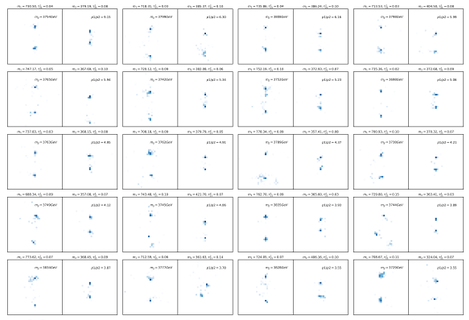
The anomaly score for each event as a function of the invariant mass of the leading two jets. A number of anomalous events are clearly seen near M_{JJ} ~ 3750 GeV.

Step 2. Determine statistics of remaining events after anomaly score cuts



Parameter distributions of the events that remain after imposing cuts on the anomaly score α. Strong peaks mean signal likely not from noise, and that there is a true overdensity at these parameter values

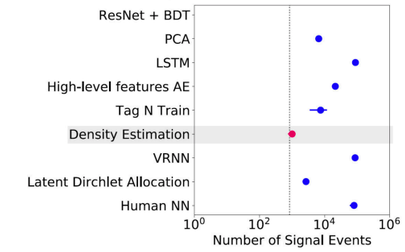
Step 3. Visualize most anomalous events



The 20 most anomalous events. Images show the particles belonging to two lead jets. They appear to be an anomalous particle decay

Step 4. report findings:

"a 3772.9±8.3 GeV particle decays into 2 particles, M_1=727.8±3.8 GeV and M_2=374.8±3.5 GeV. Each of these decayed into two-pronged jets."
Was this correct? yes:



Results of the competition. Our in-distribution anomaly score (density estimation) proved very powerful