Spectral clustering as a means of jet formation and IR safety

H.Day-Hall[1 2]
Supervised by; S.Dasmahapatra[1], S.Moretti[1], C.H.Shepherd-Themistocleous[2]

[1]University of Southampton, UK
[2]Rutherford Appleton Laboratory, UK

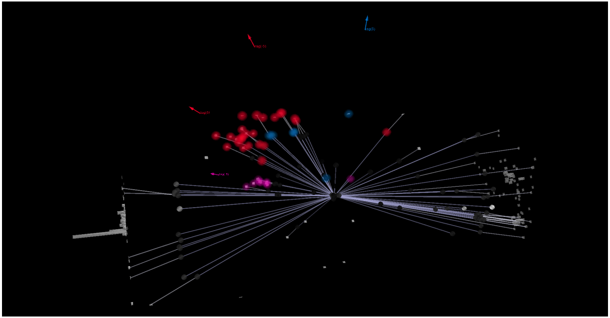
2nd April 2021

Outline

Potential algorithms

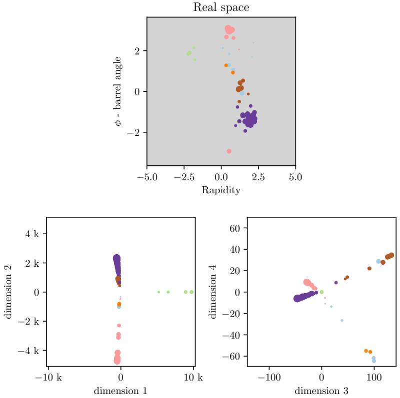
Jet classification or formation has many possibilities

- Recurrent Neural Networks
- Long-Short Term Memory networks
- Tree Recurrent Neural Networks
- Spectral clustering



Spectral Clustering

Spectral clustering is a machine learning technique with no black box.



It forms an embedding space to separate clusters.

Spectral Clustering

Theory

Pairs of particles are assigned similarity measures; $a_{i,j}$. From these, a cost function can be written

$$\text{NCut} = \sum_k \frac{W(G_k, \bar{G}_k)}{|G_k|}$$

Where $W(G_k, \bar{G}_k)$ is the similarity broken to separate the cluster G_k from the rest of the graph. The denominator seeks to balance cluster size.

Spectral Clustering

Theory

This can be minimised by finding the eigenvectors of a Laplacian.

Let

$$A = \begin{pmatrix} 0 & a_{0,1} & a_{0,2} & \dots \\ a_{0,1} & 0 & a_{1,2} & \\ a_{0,2} & a_{1,2} & 0 & \\ \vdots & & & \ddots \end{pmatrix}$$

and

$$D = \begin{pmatrix} \sum_k a_{0,k} & 0 & 0 & \dots \\ 0 & \sum_k a_{1,k} & 0 & \\ 0 & 0 & \sum_k a_{2,k} & \\ \vdots & & & \ddots \end{pmatrix}.$$

Then the symmetric Laplacian is written;

$L = D^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}}$.

Spectral Clustering
Adaptations

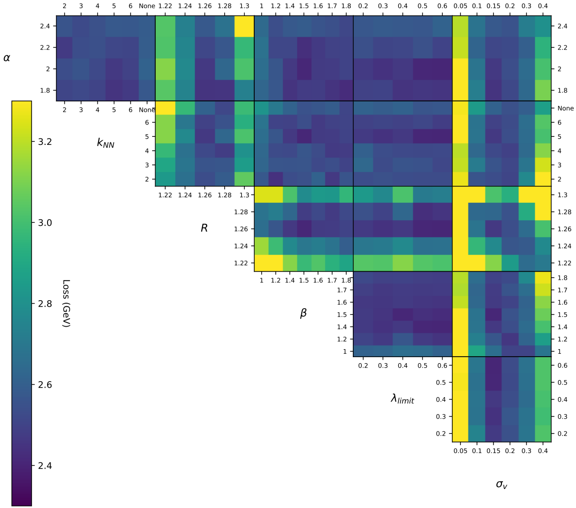
Classically this is used with a predetermined number of clusters in a euclidean space. Adaptions for use in jet formation;

- Use of information in eigenvalues.
- Recombination scheme in the embedding space.
- Stopping condition.

This results in 6 tunable parameters; σ_v, α, k_{NN}, λ_{limit}, β and R.

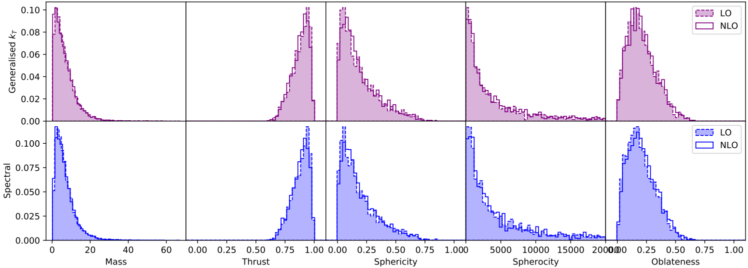
Spectral Clustering
Parameters

The parameters are not fine tuned;



Spectral Clustering
IR safety

Shape variables are sensitive to IR emissions.



Spectral Clustering

IR safety

The Jenson Shannon score can be used to compare two distributions. To begin with the Kullback-Leibler divergence;

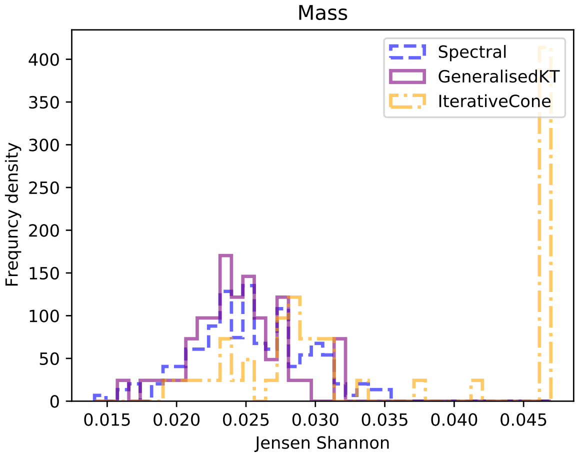
$$D_{\mathrm{KL}}(p|q) = \int_{-inf}^{\mathrm{inf}} p(x) log \left(\frac{p(x)}{q(x)} \right) dx.$$

From which the Jenson Shannon divergence can be written;

$$D_{\mathrm{JS}}(p, q) = \frac{1}{2} D \left(p | \frac{1}{2}(p+q) \right) + \frac{1}{2} D \left(q | \frac{1}{2}(p+q) \right)$$

Spectral Clustering

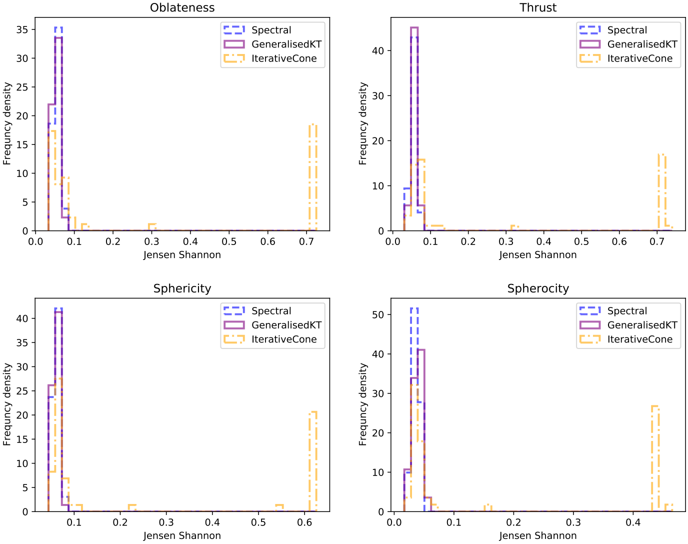
IR safety



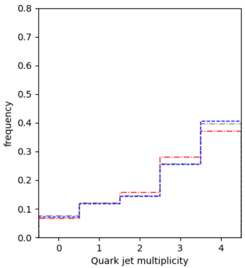
Each count is a clustering algorithm configuration. Low values indicate IR safety.

Spectral Clustering

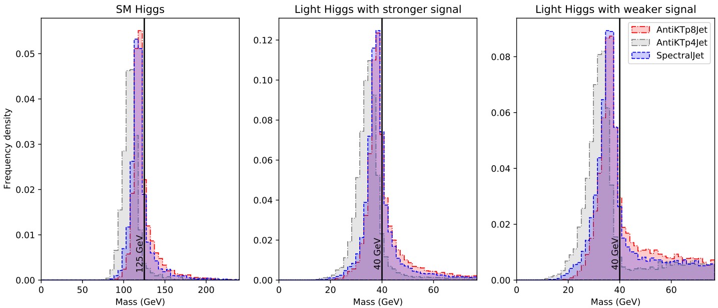
IR safety



Results
Light higgs cascade

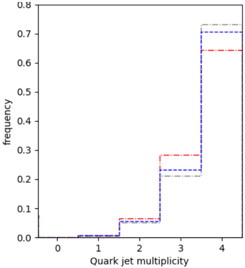


$$pp \rightarrow H_{\text{SM}} \rightarrow h_{40\text{GeV}} h_{40\text{GeV}} \rightarrow b\bar{b}b\bar{b}$$

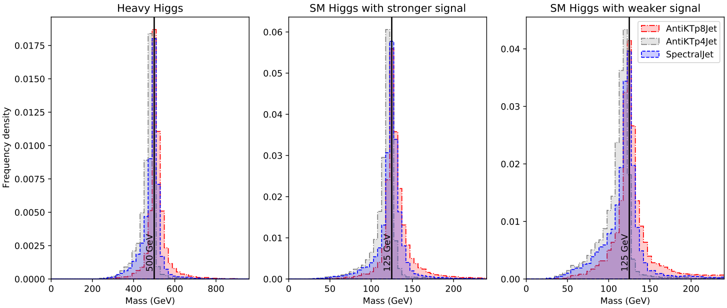


Results

Heavy higgs cascade

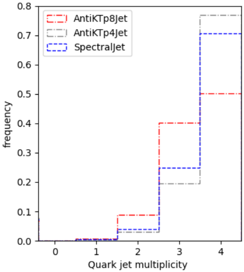


$$pp \to H_{500\mathrm{GeV}} \to h_{\mathrm{SM}} h_{\mathrm{SM}} \to b\bar{b}b\bar{b}$$

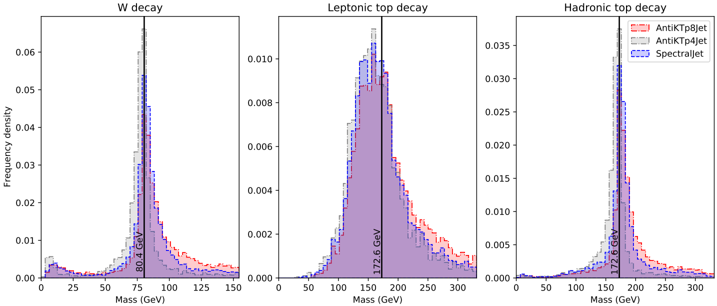


Results

Semileptonic top decay



$$pp \to t\bar{t} \to W^+W^-b\bar{b}$$



Conclusion

Spectral clustering makes an IR safe jet formation algorithm, without any use of learnt parameters or black box elements. It is remarkable for its ability to adapt to various data sets.