Detection of Clustering Instabilities for Sequential Recombination Algorithms

C. S. Cowden and I. Volobouev
Texas Tech University, Physics Department, Box 41051, Lubbock TX 79409-1051
E-mail: christopher.cowden@ttu.edu

Abstract. We explore clustering stability of sequential recombination jet reconstruction algorithms. Events are reconstructed many times, using random variations of kinematic properties of the jet fragmentation process. Sensitivity of different algorithms to initial conditions are quantified by introducing probabilistic assignment of initial particles to jets (fuzzy clustering). A criterion detecting unstable configurations (bifurcation points) is proposed, based on the overall fuzziness of the event.

1. Introduction
Studies of jet substructure can provide insights into a number of interesting phenomena at high energy colliders by providing such tools as identification of boosted resonances or differentiation of quark and gluon jets. In a typical current jet substructure analysis (see [1, 2, 3] for a few examples and [4] for a recent comparison of different methods), one finds large jets by some sequential recombination algorithm then un-clusters the target jet by another sequential recombination algorithm (Cambridge/Aachen is the usual choice for the second stage). In this work we consider a different approach in which the sequential recombination is performed in a manner appropriate for substructure studies without the necessity of unclustering the results of the initial cluster finding. We call a sequential recombination algorithm “multiscale” (or multiresolution) if it has the following features:

• Stopping condition is implemented as a cutoff on the recombination distance function.
• The distance function is dimensionless.
• The minimum pairwise distance is almost always increasing in the recombination process\(^1\).

An algorithm with such properties will result in identical clusterings no matter whether we cluster to a distance cutoff \(D_1\) or whether we interrupt the clustering process at at some smaller cutoff \(D_2\) and then resume and continue to \(D_1\). Also, identical clusterings will be obtained by a class of distances which differ from each other only by a monotonous transformation (with the same transformation applied to the stopping criterion). Thus we associate the distance cutoff with the concept of resolution scale (or jet size). The use of dimensionless distances allows us to compare events with different partonic center of mass energy on equal footing which is essential in hadronic collisions.

Multiscale jet clustering algorithms provide a wealth of new information regarding the hadronic portion of collider events by adding the scale dimension. In other words, instead of returning just a set of four-momenta for the reconstructed jets, multiscale algorithms return a clustering tree (a.k.a. hierarchical

\(^1\) Ideally the minimum distance should monotonically increase, but no known algorithms employing four-vector recombination have such a property. Monotonous clustering methods without pairwise recombinations do exist: see for example [5].
clustering dendrogram). A good choice of distance function ensures that the recombination scale is almost always increasing in the process of tree construction. In a true multiscale jet clustering algorithm, one could stop at any distance by merely navigating the history of a more inclusive clustering. Adopting a multiscale jet clustering algorithm, one obviates the need of the two phase analysis in jet substructure studies. In the context of global jet finding algorithms, the ideas of multiscale jet reconstruction were pioneered by FFTJet [6].

Not all commonly used sequential recombination algorithms have the necessary properties to adapt to a multiscale application nicely. Most notably, algorithms akin to $k_T$ [7] and anti-$k_T$ [8] do not adapt well because these algorithms employ dimensionful distance functions which do not compare well event to event in hadron collider events. For the following analysis, the jet clustering algorithms employed do not perform any beam removal. The algorithms simply work by comparing inter-particle distances defined by a given distance function and selecting the particles with the smallest distance to recombine. This repeats until the algorithm encounters a predefined stopping criterion. The Cambridge/Aachen [9] algorithm used in this study defines a distance function based on separation of a pair of particles $i$ and $j$ in the $\eta$-$\phi$ space given by

$$d_{ij} = \Delta R_{ij}^2$$

where $\Delta R_{ij}^2 = \Delta \eta_{ij}^2 + \Delta \phi_{ij}^2$.

In addition to the algorithm which employs the distance defined by Equation 1, we consider more general distances represented by

$$d_{ij} = \Delta R_{ij} f \left( \frac{\min(p_{Ti},p_{Tj})}{\max(p_{Ti},p_{Tj})} \right),$$

where $f$ is some continuous function defined on the $[0, 1]$ interval. In particular, we can choose the function $f$ in Equation 2 so that $d_{ij}$ becomes the minimum width of a Gaussian filter in the $\eta$-$\phi$ space for which only a single peak remains when two initial energy deposits are smeared by the filter$^2$. One can also think of this distance as being proportional to the square root of the time for the two energy deposits to “diffuse” into a single blob, so we call the resulting distance function the “diffusion distance”. This distance function is inspired by the scale-space theory [10], and we expect it to perform well in multiscale applications. Figure 1 shows the function $f$ versus the input argument $x = \frac{\min(p_{Ti},p_{Tj})}{\max(p_{Ti},p_{Tj})}$.

![Figure 1. The function $f$ defined by the diffusion distance versus the input argument $x = \frac{\min(p_{Ti},p_{Tj})}{\max(p_{Ti},p_{Tj})}$.

A number of studies have suggested using clustering stability as a tool for model selection (see [11] for a review). In what follows we study the clustering stability as a function of scale, i.e., jet size or distance cutoff. We intend to associate stable clustering configurations with the physical structures

\[\text{C++ implementation of this distance function is available upon request.}\]
present in high energy events (such as inter-jet distances in the $\eta$-$\phi$ space). Intuitively, one expects the stability, defined in some sensible way, to decrease when the distance cutoff exists in a band where two or more jets merge together. Conversely, one expects a relatively high degree of stability when jets are well separated.

Figure 2. An illustration showing two possible clusterings in which objects may be assigned to two clusters, J0 and J1. Clustering $C_1$ is on the left while clustering $C_2$ is on the right.

One can effectively measure the stability of a particular jet configuration by perturbing the system a number of times and comparing the clustering outcomes at the same distance cutoff after the perturbations. To quantify the clustering stability, we use modifications of the Rand [12] and Jaccard [13] indices. These indices are defined by considering what happens to each pair of input particles in two different clusterings $C_1$ and $C_2$ (Figure 2 shows an example of such a comparison). Each pair of particles falls into one of the following four categories:

I Both particles belong to same cluster in both clusterings (for example, particles 0 and 1 in Figure 2).
II Particles belong to different clusters in both clusterings (particles 0 and 4).
III Particles belong in the same cluster in clustering $C_1$, but belong to different clusters in clustering $C_2$ (particles 3 and 4).
IV Particles belong to different clusters in clustering $C_1$, but belong to the same cluster in clustering $C_2$ (particles 1 and 2).

In the statistical literature, the Rand and Jaccard indices are defined via the number of pairs in these categories; we instead weight each pair by $w_{ij} = p_T^i p_T^j$. The $p_T$-weighted Rand index is defined by:

$$ R = \frac{\Sigma I + \Sigma II}{\Sigma} $$

where $\Sigma$ is the sum of all weights, $\Sigma I$ is the sum of weights $w_{ij}$ in category I, and so on. The $p_T$ weighted-Jaccard index is defined by:

$$ J = \frac{\Sigma I}{\Sigma I + \Sigma III + \Sigma IV} $$

The subsequent analysis also makes use of another stability measure referred to as the “$p_T$-weighted minimal matching distance”, or MMD. This measure is defined by forming an $M \times N$ matrix of the $M$ clusters in $C_1$ and $N$ clusters in $C_2$. The matrix element in row $i$ and column $j$ is set to the scalar sum of the $p_T$ of the particles which belong simultaneously to the cluster $i$ in clustering $C_1$ and to the cluster $j$ in clustering $C_2$. Then, in case $M < N$ we find a set of $M$ matrix elements, all from different rows and columns, with maximum sum of the elements ($N$ such elements are summed in case $N \leq M$). The MMD is defined as one minus this maximized sum divided by the scalar sum of $p_T$ values of all initial particles\(^3\). Since all stability measures so far considered ($R$, $J$, and MMD) range between zero and one, we compute the corresponding instability as one minus the stability measure.

\(^3\) An approximate greedy implementation of this algorithm is used in practice as the computational complexity of the exact algorithm scales factorially with the number of clusters.

...
Fuzziness, on the other hand, quantifies how the clustering changes as a result of reasonable jet energy flow fluctuations. We employ the following fuzziness definition \[6\]:

\[
F_j = \sqrt{\sum_i w_{ji}(1-w_{ji})(p_T^i)^2} / \sum_i w_{ji}p_T^i
\]  

(5)

where the weight \(w_{ij}\) represents the probability of particle \(i\) to belong to jet \(j\) in the ensemble of perturbed clustering outcomes. Subsequently, we define the event fuzziness as

\[
F = \frac{\sum_j F_j \sum_i w_{ji}p_T^i}{\sum_i p_T^i}
\]  

(6)

which can be understood as the average \(p_T\)-weighted fuzziness of the individual jets. One also has a choice regarding the outer summation in the numerator of Equation 6 to be a regular sum or in quadrature. The results shown below were found using the regular summation.

2. Event Generation and Simulation

We study clustering stability and fuzziness in the context of a controlled dijet toy MC environment as well as hadronic \(t\bar{t}\) events in high energy \(p\bar{p}\) collisions. We generate events using the Pythia event generator. For the toy dijet MC sample, we generate the events by randomly selecting a point in the \(\eta-\phi\) space and shooting a parton for which Pythia performs the showering and hadronization. Then we select a point at a given \(\Delta R\) of that first parton to shoot the second parton for which Pythia also simulates the showering and hadronization. We also separately generate a sample of \(t\bar{t}\) events using standard Pythia cards.

The stability and fuzziness measures employed in this analysis are defined with respect to the statistical ensemble formed by multiple perturbations of the initial event. We perturb the event according to one of two schemes. In the first scheme (referred to as “Scheme 1”) we perturb the jet showering process and detector response in a semi-realistic manner. We treat each particle as if it were a gluon and calculate the probability for each particle pair that the pair originated from the \(g \rightarrow gg\) splitting:

\[
p \sim C_A 4\pi \alpha_s (p_{soft}^2) \frac{p_{mother} \cdot p_{hard}}{p_{mother} \cdot p_{soft} p_{hard} \cdot p_{soft}}
\]  

(7)

where \(p_{mother} = p_{hard} + p_{soft}\) and \(p_{soft}\) is the 4-momentum of the particle with the smaller \(p_T\) in the pair\(^4\). Then we re-decay the mothers of the most probable pairs. The scheme then excludes the particle in the pair with lower \(p_T\) from further consideration in the fluctuation process. More specifically, this scheme of QCD like fluctuation works in the following manner.

- For each pair of particles, estimate the probability that these particles have the same mother according to Equation 7.
- Sort all pairs of particles in the decreasing order of probability.
- For the highest probability pair, re-decay the mother particle according to a flat phase space.
- Mark the softer particle in the pair as “used” to exclude it from further consideration.
- Repeat the previous two steps until the 4-momentum of half of the event particles has been fluctuated.

We also apply simple efficiency and resolution simulation with effective parameters similar to those of the CMS detector. Typically we execute the perturbation sequence and re-cluster 100 times per event.

The second scheme “Scheme 2” is simpler: in this scheme we perturb the event by adding noise to each particle’s \(\eta, \phi, \) and energy according to a set of Gaussian distributions. Some results below are shown for the case where \(\sigma_\eta = \sigma_\phi = 0.087\) and \(\sigma_E = 0.15\).

\(^4\) The authors thank Steve Mrenna for guidance regarding Equation 7.
3. Results

3.1. Toy Dijet MC Stability

We first tested this method on a sample of toy MC dijet events in which we controlled the initial partons’ \( \Delta R \) and \( p_T \) ratio (this ratio is denoted \( r \) in the figures). Figure 3 illustrates these results by plotting instability averaged over many events vs. the distance cutoff. One can see from Figure 3(a) that the three stability measures employed (Rand, Jaccard, and \( p_T \)-weighted minimal matching distance) all share a peak of instability at the same location but have varying degrees of sensitivity to the instability. Also as one may expect, applying Scheme 2 to perturb the system results in a magnified instability peak found at the unstable points as shown by Figure 3(c). One can also compare the behavior of these stability measures in the C/A algorithm from Figures 3(b) and Figure 3(d). One can observe from these Figures that instability peak magnitude remains roughly the same for both clustering algorithms studied if events are perturbed by the same perturbation Scheme. One may also note that the location of instability peaks for events clustered using the diffusion distance depends upon the \( p_T \) ratio of the jets, unlike the C/A algorithm.

**Figure 3.** These plots illustrate the instability in a few toy MC dijet configurations and the two clustering algorithms previous described. One can see excesses in the instability near cutoff values where the two jets merge. One may also notice differences in stability sensitivity by choice of smearing scheme but not quite such a dependence upon choice of algorithm.

Figure 4 illustrates the fuzziness in these dijet events for the smearing and clustering algorithms discussed. Upon comparing Figure 4(a) to Figure 4(c), one observes a similar structure of peaks but
events perturbed by Scheme 1 show instability peaks with a suppressed significance compared to events perturbed by Scheme 2. The C/A algorithm exhibits a similar, but more drastic, suppression of average fuzziness structure for events perturbed by Scheme 1 as opposed to Scheme 2.

![Diffusion Fuzziness](image1)

![C/A Fuzziness](image2)

(a) Fuzziness of dijet configurations clustered with the diffusion distance averaged over many events perturbed by Scheme 1.

(b) Fuzziness of dijet configurations clustered with the C/A distance averaged over many events perturbed by Scheme 1.

![Diffusion Fuzziness](image3)

![C/A Fuzziness](image4)

(c) Fuzziness of dijet configurations clustered with the diffusion distance averaged over many events perturbed by Scheme 2.

(d) Fuzziness of dijet configurations clustered with the C/A distance averaged over many events perturbed by Scheme 2.

Figure 4. These plots illustrate the fuzziness in a few toy MC dijet configurations and the two clustering algorithms previously described. The fuzziness shows peaks around distance cutoffs similar to the instability.

3.2. $t\bar{t}$ MC Stability

Since jets produced in $t\bar{t}$ events have a wide range of $\Delta R$ separations and $p_T$ ratios, it makes sense to consider the stability and fuzziness of single events rather than averages over many events. Figure 5(a) shows the instability of a $t\bar{t}$ event using the diffusion distance for jet clustering and perturbed by Scheme 1. In this figure, peaks emerge indicating points where jets merge together; the peak occurring at $d\text{Cut} \sim 3.75$ shows the point at which the entire event has been reconstructed into a single jet. Figure 5(b) depicts the fuzziness in the same $t\bar{t}$ event; this plot shows a similar pattern of peaks as Figure 5(a). The peak in instability at $d\text{Cut} \sim 1$ corresponds to a transition from a 6 jet configuration to a 3 jet configuration. Likewise, over the range $0 \leq d\text{Cut} \lesssim 0.5$ there is a transition from some initial unstable high multiplicity clustering to a more stable 6 jet configuration one very well expects in a fully hadronic $t\bar{t}$ event.
4. Conclusions

In this note we have developed the concept of multiscale sequential recombination algorithms appropriate for jet reconstruction at hadron colliders. The jet substructure information is produced by these algorithms in the most natural manner. We have also introduced the diffusion distance clustering algorithm based on ideas from the scale-space theory. The use of clustering stability as a function of scale was explored as a tool for identifying physically meaningful jet configurations. Such configurations are characterized by a substantial range of scales free of bifurcation points (clustering instabilities). We construct the clustering stability measures in such a way as not to be affected by infrared or collinear divergences. The concepts presented here can potentially become powerful tools in the analysis of event energy flow in hadronic collisions and deserve further study.

References

[1] Jonathan M. Butterworth, Adam R. Davison, Mathieu Rubin, and Gavin P. Salam. Jet substructure as a new higgs-search channel at the large hadron collider. *Phys. Rev. Lett.*, 100:242001, Jun 2008.

[2] David Krohn, Jesse Thaler, and Lian-Tao Wang. Jet Trimming. *JHEP*, 02:084, 2010.

[3] Stephen D. Ellis, Christopher K. Vermilion, and Jonathan R. Walsh. Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches. *Phys. Rev.*, D81:094023, 2010.

[4] Paloma Quiroga-Arias and Sebastian Sapeta. A comparative study of jet substructure taggers. 2012, 1209.2858.

[5] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[6] I. Volobouev. FFTJet: A Package for Multiresolution Particle Jet Reconstruction in the Fourier Domain. 2009, 0907.0270.

[7] S. Catani, Yuri L. Dokshitzer, M.H. Seymour, and B.R. Webber. Longitudinally invariant $K_t$ clustering algorithms for hadron hadron collisions. *Nucl.Phys.*, B406:187–224, 1993.

[8] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The Anti-$k(t)$ jet clustering algorithm. *JHEP*, 0804:063, 2008, 0802.1189.

[9] Thomas H. Cormen, Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition, 2001.

[10] Tony Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, pages 224–270, 1994.

[11] U. von Luxburg. Clustering Stability: An Overview. *ArXiv e-prints*, July 2010, 1007.1075.

[12] William M. Rand. Objective criteria for the evaluation of clustering methods. *JASA*, 66:846, 1971.

[13] P. Jaccard. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44:223–270, 1908.