Unsupervised feature learning for waveform classification

Ken

June 22, 2018

~~CNN for waveform classification~~

Unsupervised feature learning for waveform classification

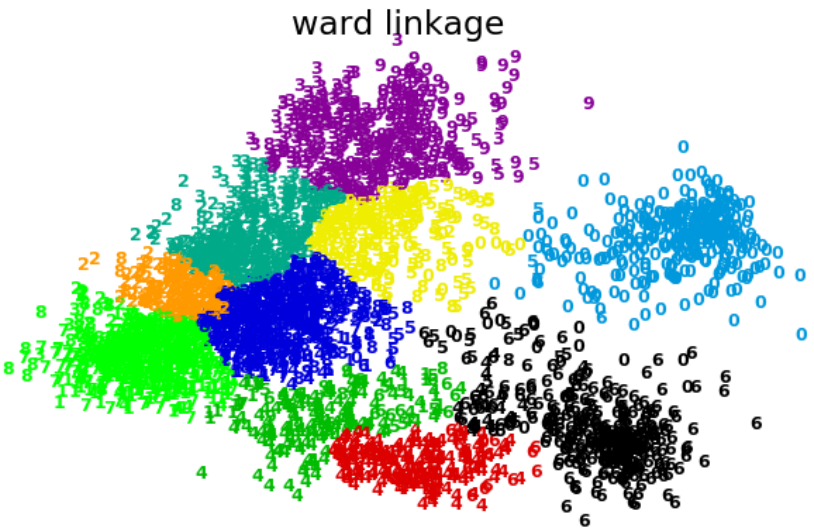
2 main concept:-

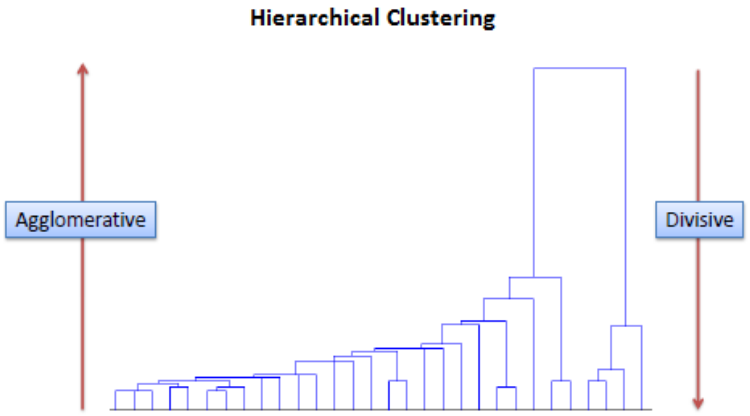
- Hierarchical Clustering
- Auto-encoder

These 2 methods falls under a category of unsupervised learning within machine learning.

Specific labels are not given, the algorithm itself have to deduce the label itself.

Hierarchical Clustering





The "distance" between each data is calculated. This "distance" indicates how closely the data are related to one and another. This "distance" between all data points are calculated. Hierarchical clustering can be best visualise through a dendrogram where the data are arranged and connected in a tree like method. The height of the tree represents the "distance" between data points. Data that have similar features will tend to have shorter "distance".

"Distance" for Hierarchical clustering - Ward's Method

Ward's minimum variance method

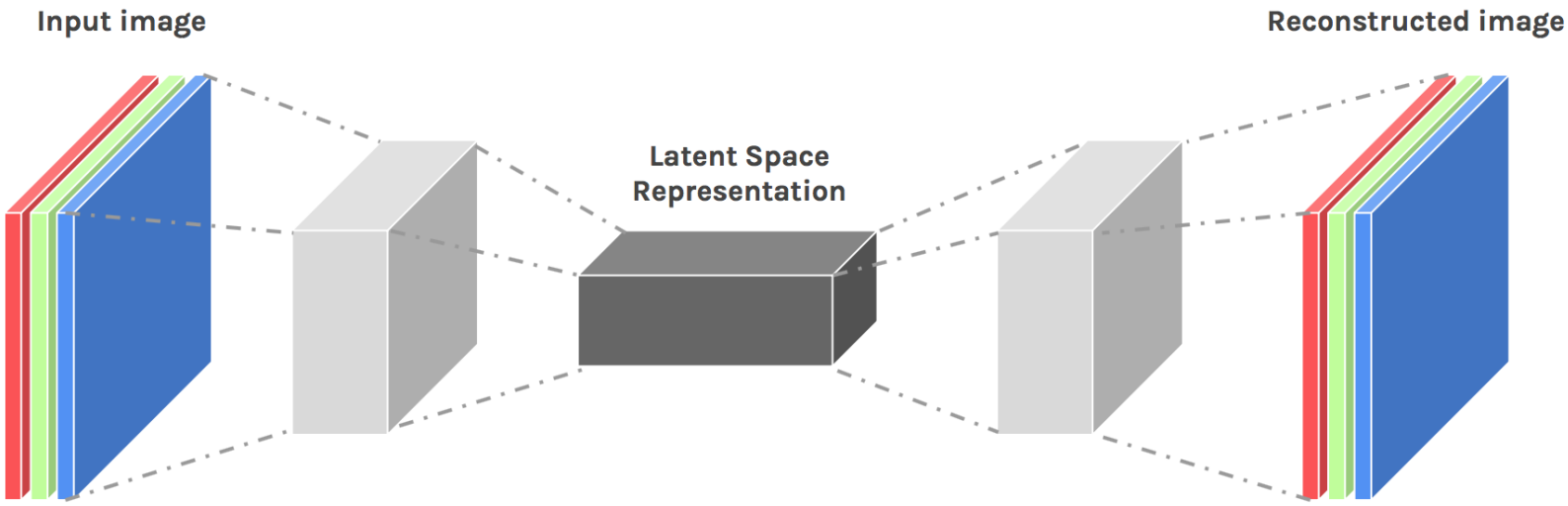
$$\Delta(A,B) = \sum_{i \in A \cup B} ||\vec{x_i} - \vec{m_{A \cup B}}||^2 - \sum_{i \in A} ||\vec{x_i} - \vec{m_A}||^2 - \sum_{i \in B} ||\vec{x_i} - \vec{m_B}||^2$$
$$= \frac{n_A n_B}{n_A + n_B} ||\vec{m_A} - \vec{m_B}||^2$$

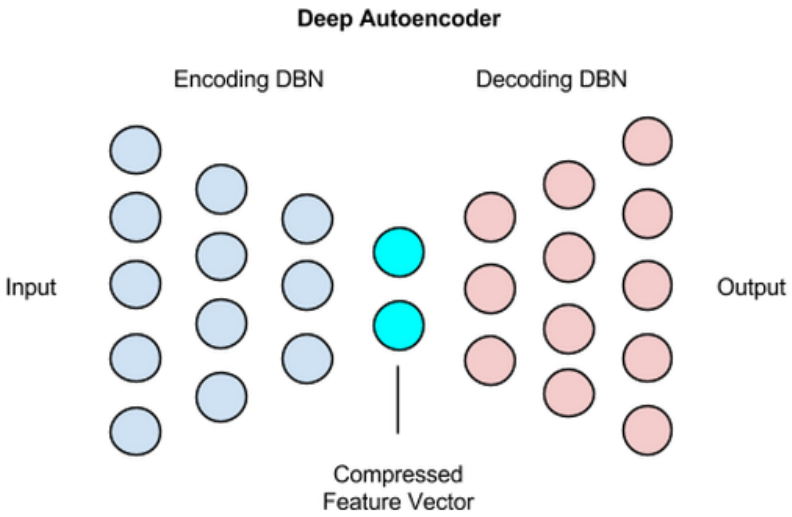
- The distance between two clusters, A and B
- The formula shows the merging cost in Ward's method.
- It is the increase of the sum of squares when merging A and B

How to determine the number of clusters?

This method does not tell us directly how many clusters is there. In any clustering methods, the number of clusters is heuristic (ie:- not optimal but sufficient). *my homework to do!

Auto-encoders



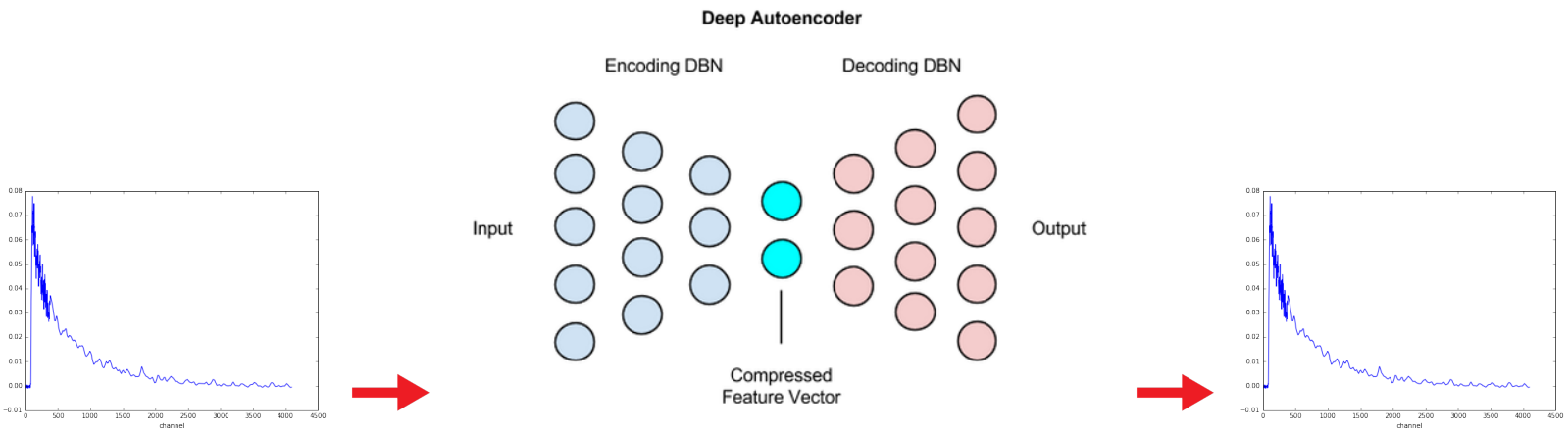


The information of the data(waveform) is reduced through the encoder (4096→512). The size of the data is reduced and the job of the decoder is to reconstruct the input of the encoder from the reduced encoder output (512→4096).

Training the auto-encoder

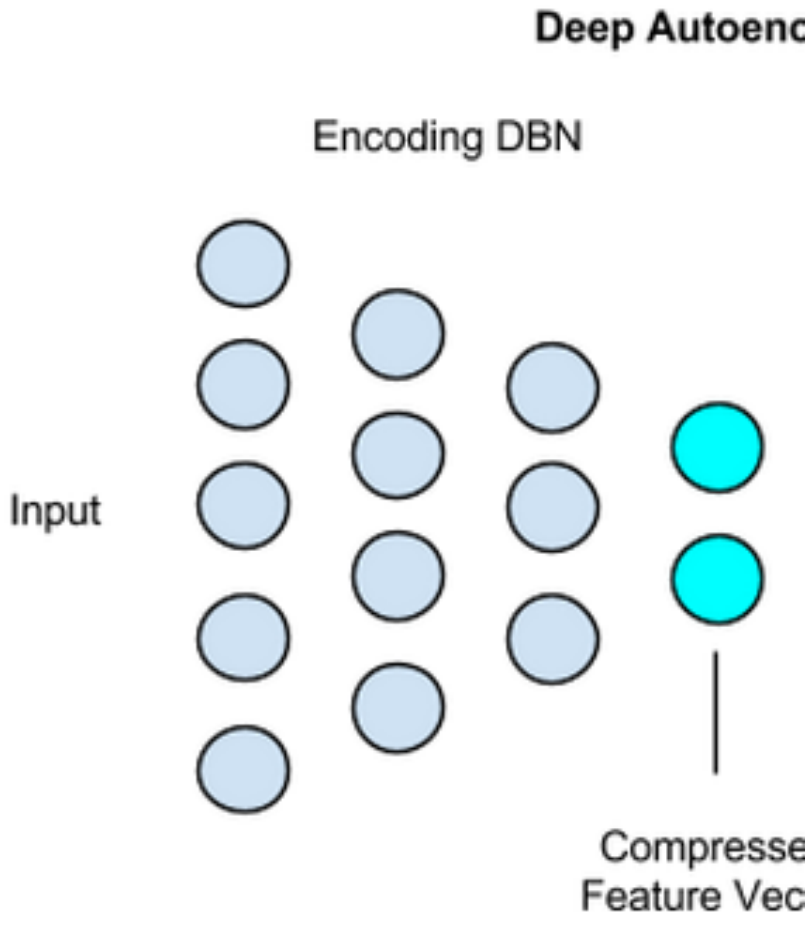
Purpose: train the auto-encoder to reproduce the input and the output to be as close as possible to the input.

input Auto-encoder output



* The waveform are normalized before being fed into the auto-encoder.

Taking the **encoder** of the auto encoder and do hierarchical clustering on it

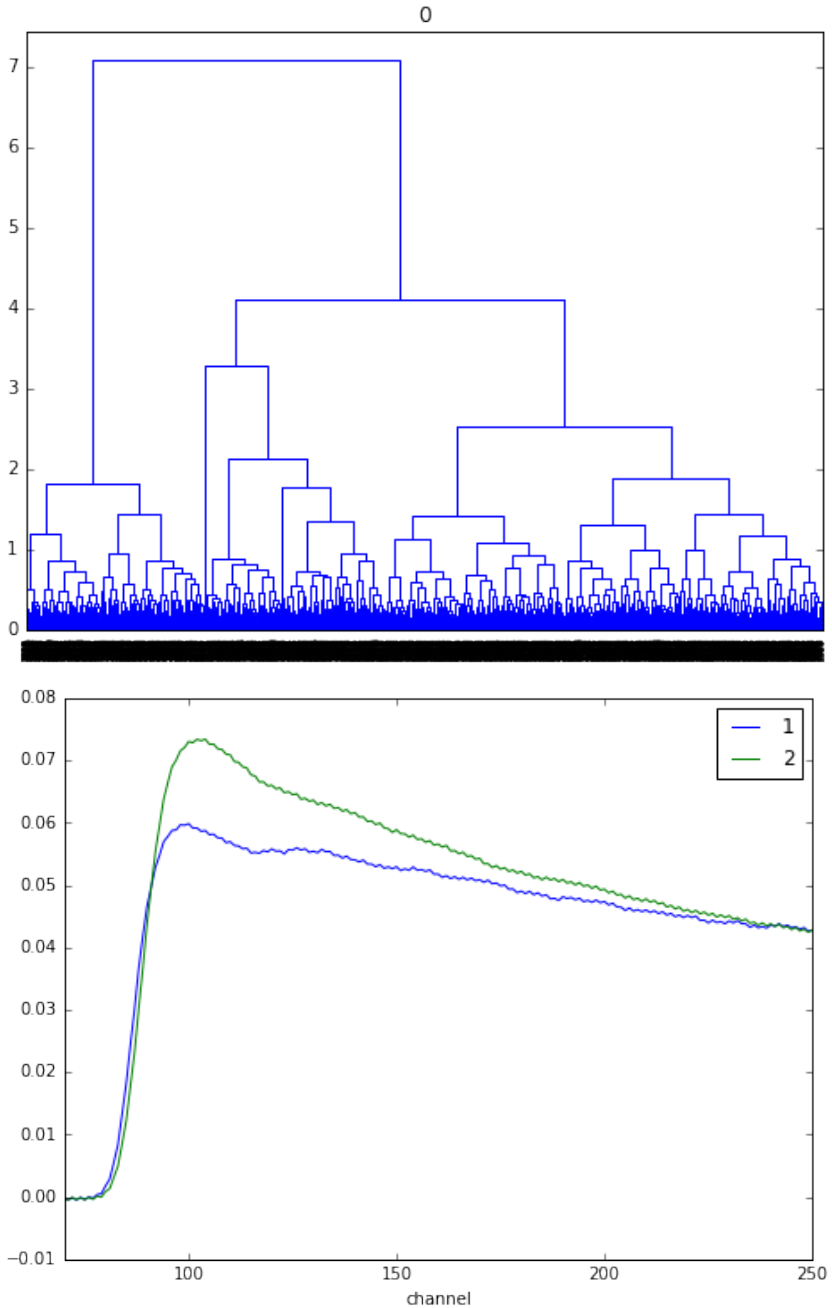


The features of waveform (low level parameter) is reduced to few parameters (high level parameters/engineered features).

These high level features are similar to the reconstructed parameters that we obtain through χ^2/gradient/etc with the difference that the algorithm learn to create their own unique parameters.

Hierarchical clustering on these high level features where waveform with similar features can be grouped together (visualise through dendrogram.)

Testing Algorithm to differentiate α and β/γ



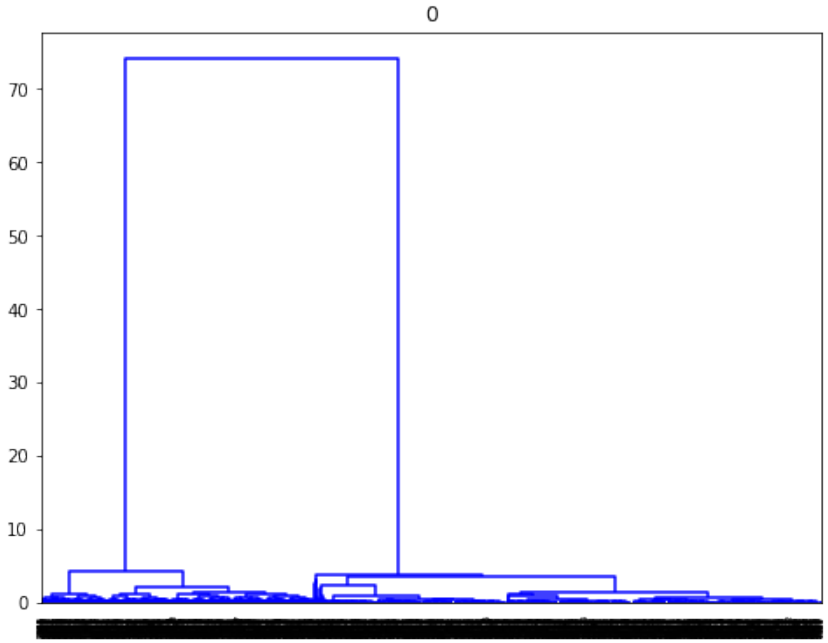
The tested data is made up by combining the dataset used to make the α list and β/γ reference pulse.

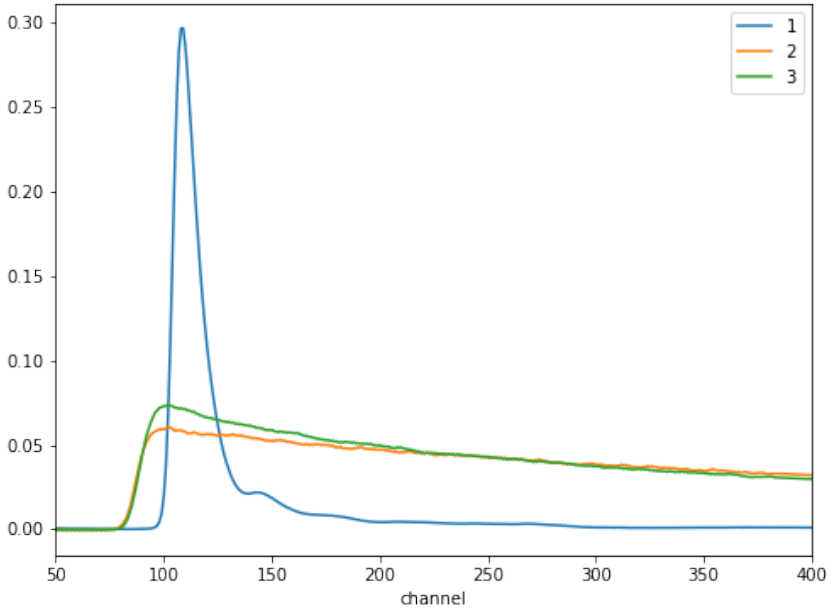
Only 4096 channels of waveform is fed into the trained encoder.

The algorithm managed to identified the 2 type of waveform at 2 clusters level.

Overall, 96.99% accurately identified
β/γ, 95.63% correctly identified
α, 97.38% correctly identified

Adding LS reference pulse list into the mix





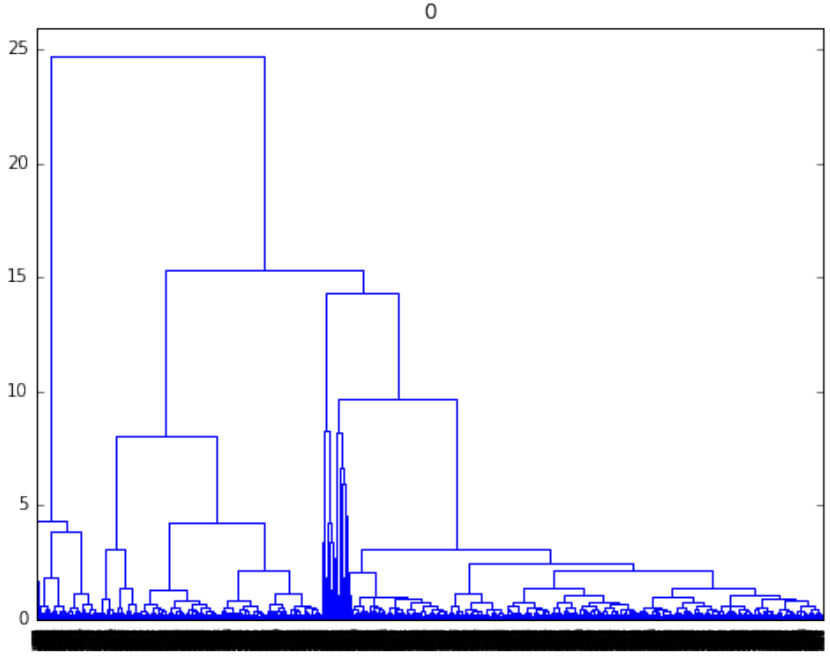
tested again to identify 3 different type of waveform. (3type of single pulse, ls, β/γ and α)

Overall, 98.92% accurately identified
LS, 100% correctly identify
β/γ, 96.84% correctly identified
α, 96.57% correctly identified

Testing on dataset from Run009

Dataset:- obtained via cut condition, PSDPara[1]<1.5, dual gate trigger, good data quality, Energy>3000keV

Total number of events in the dataset: 5145 events

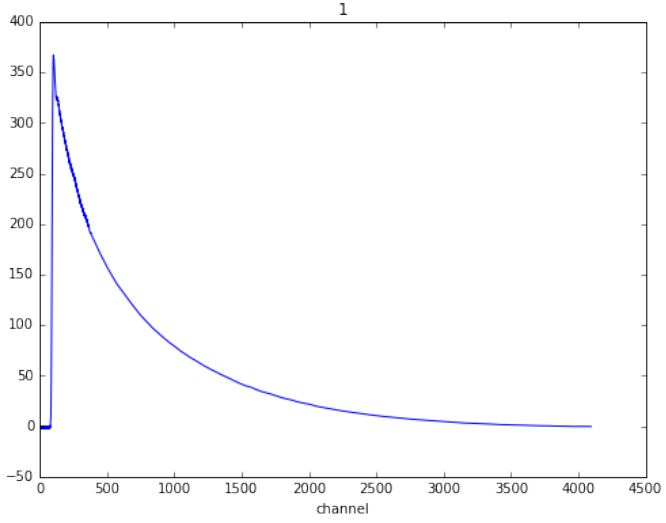


The dataset can be grouped into 9 clusters(heuristic).

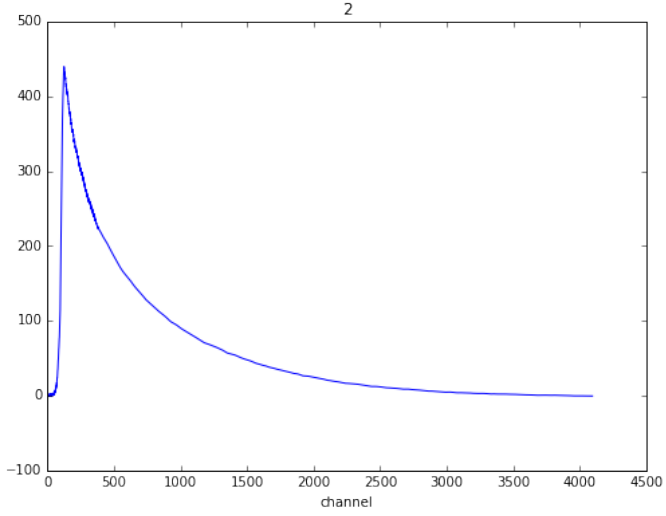
Cluster#	no. of events
1	2552
2	640
3	71
4	58
5	48
6	420
7	617
8	300
9	439

Dataset from Run009 - Examine the clusters - 1,2

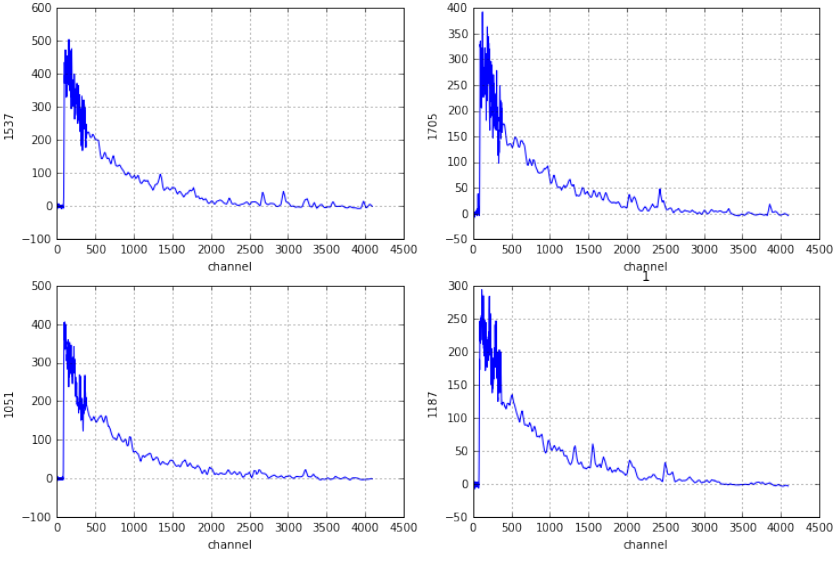
Averaged waveform of cluster#1



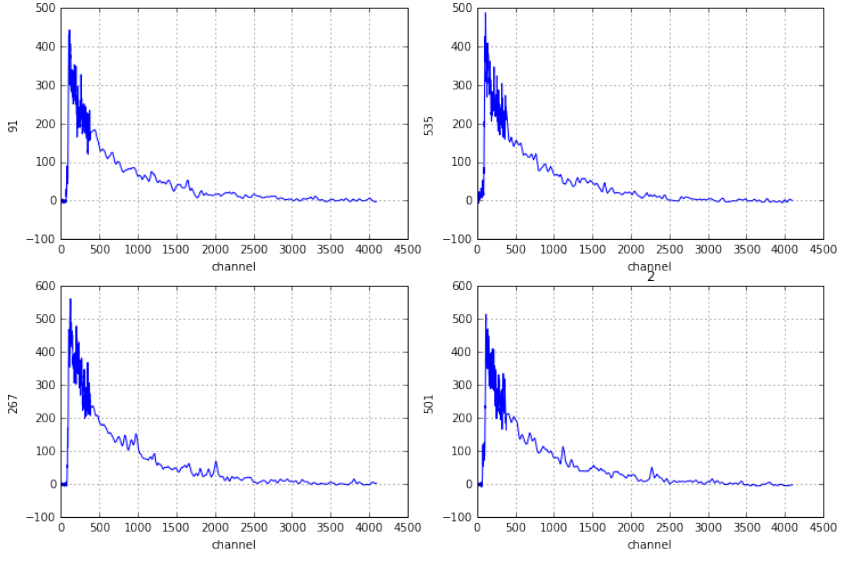
Averaged waveform of cluster#2



randomly sampled from cluster#1

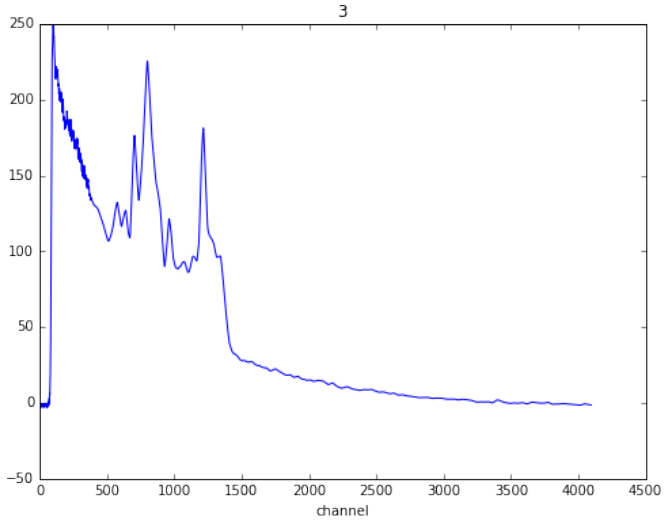


randomly sampled from cluster#2

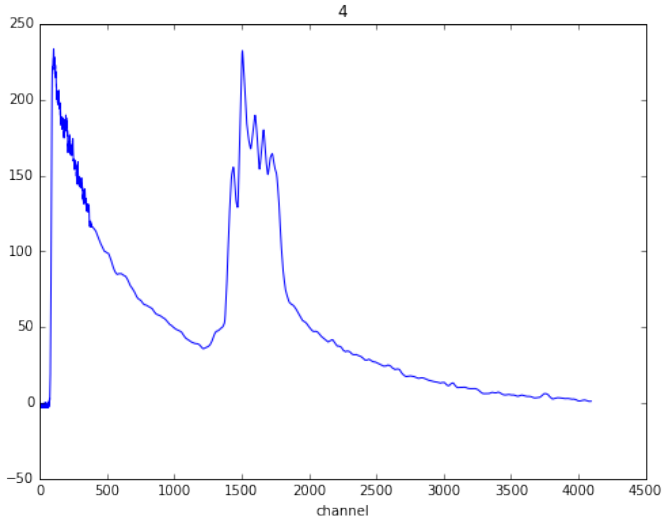


Dataset from Run009 - Examine the clusters - 3,4

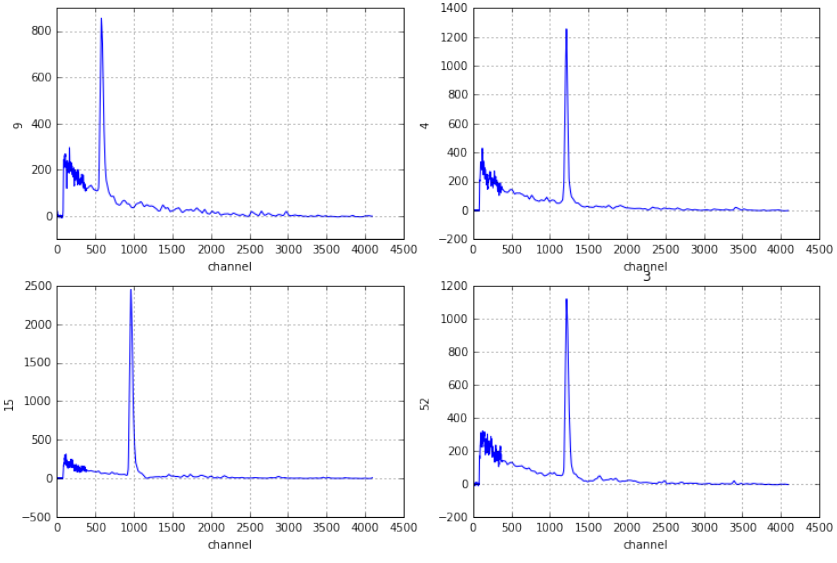
Averaged waveform of cluster#3



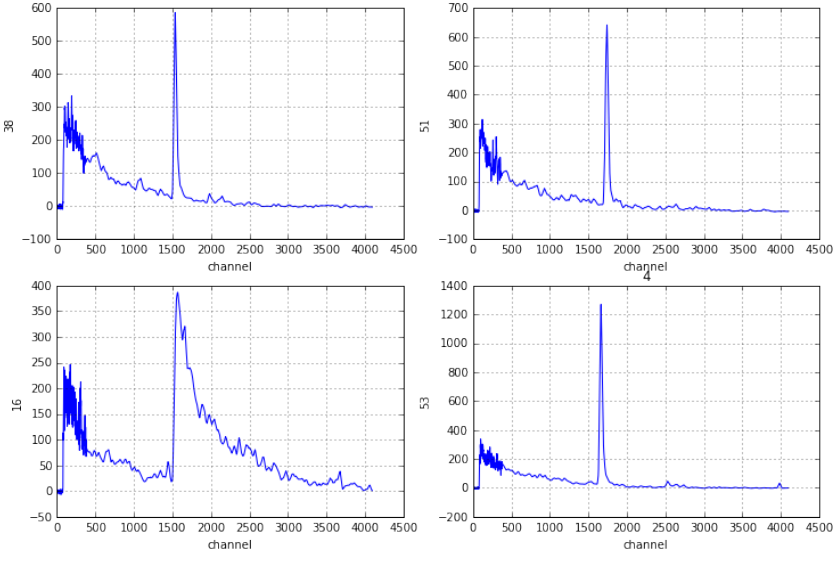
Averaged waveform of cluster#4



randomly sampled from cluster#3

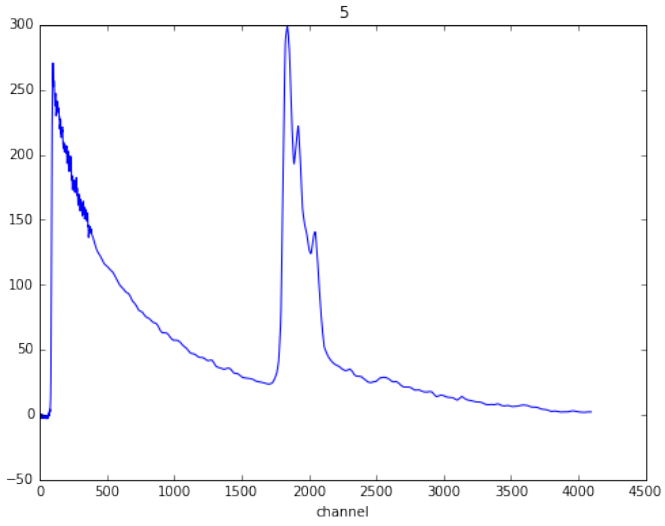


randomly sampled from cluster#4

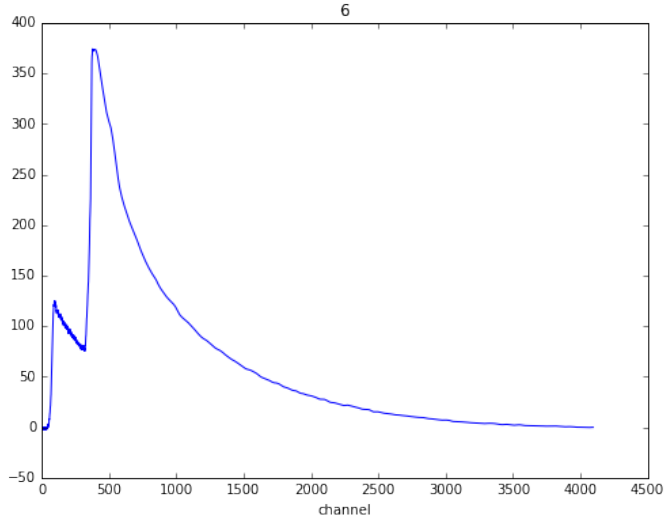


Dataset from Run009 - Examine the clusters - 5,6

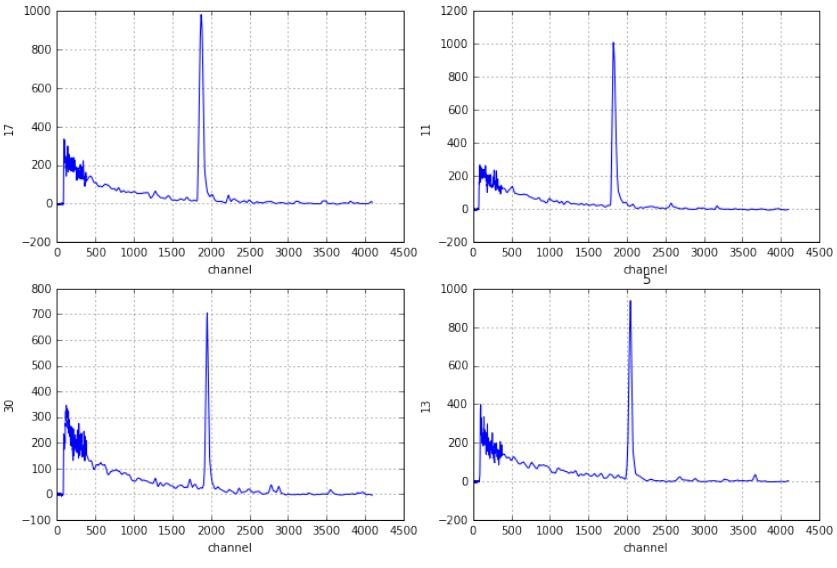
Averaged waveform of cluster#5



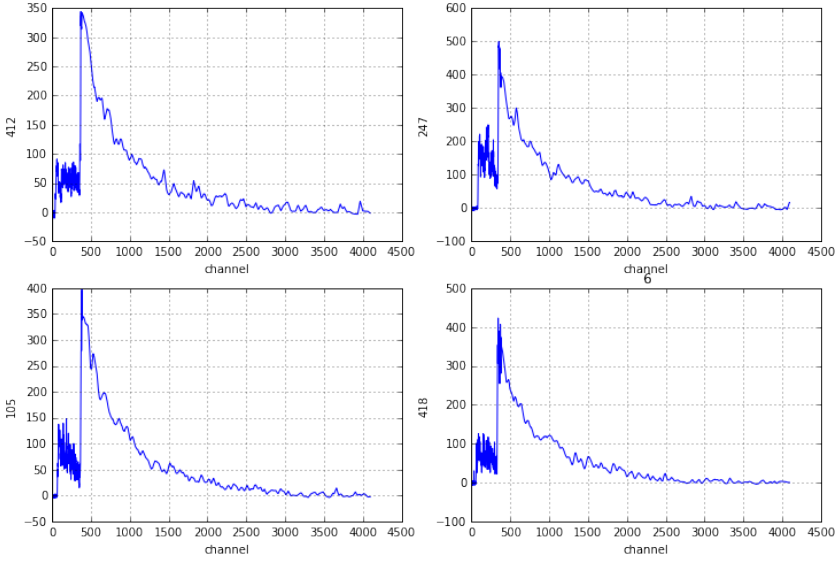
Averaged waveform of cluster#6



randomly sampled from cluster#5

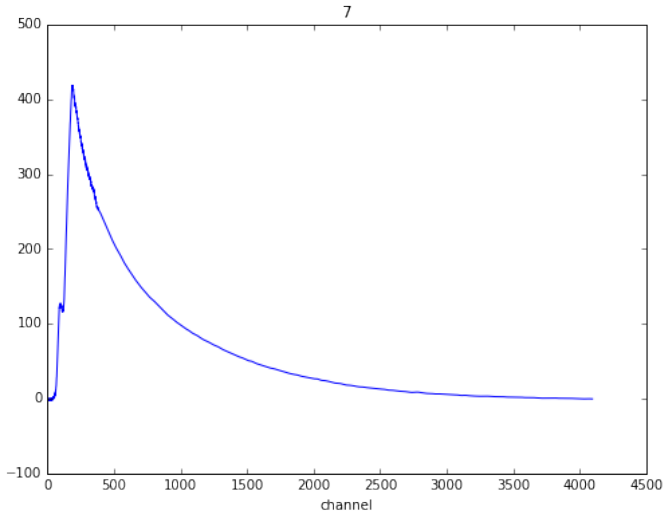


randomly sampled from cluster#6

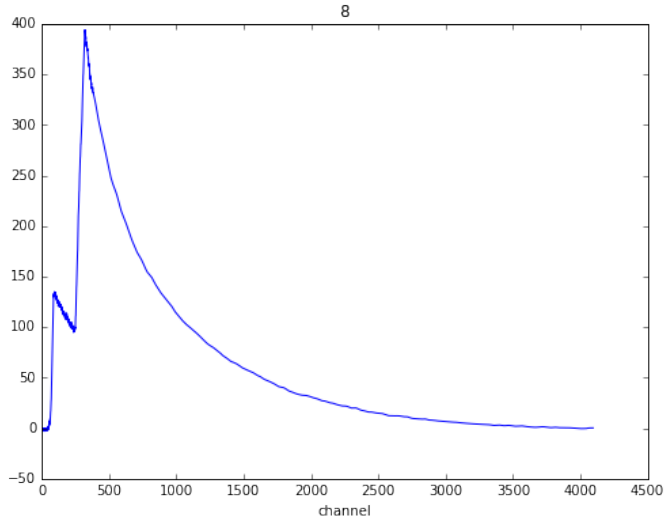


Dataset from Run009 - Examine the clusters - 7,8

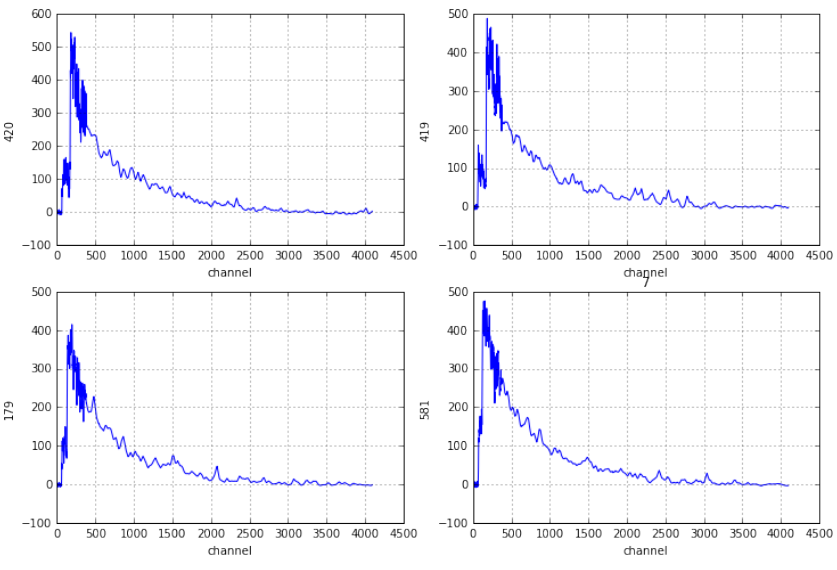
Averaged waveform of cluster#7



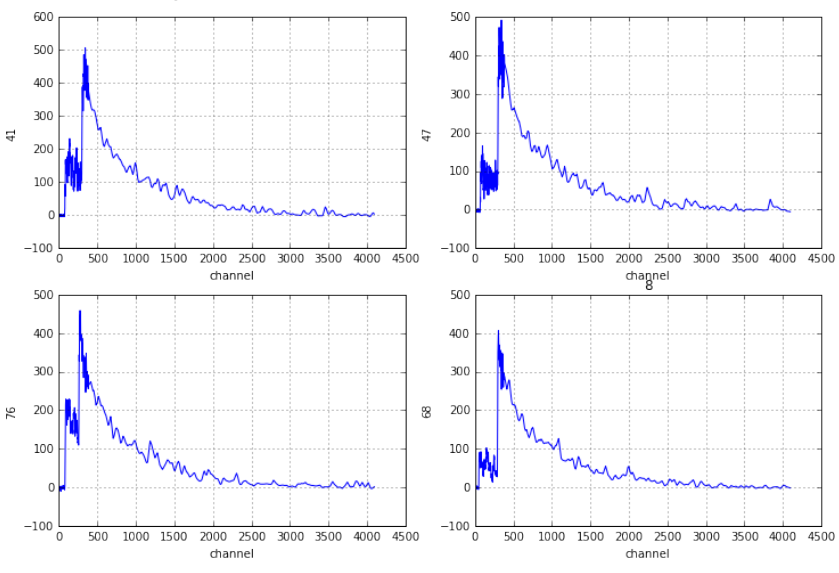
Averaged waveform of cluster#8



randomly sampled from cluster#7

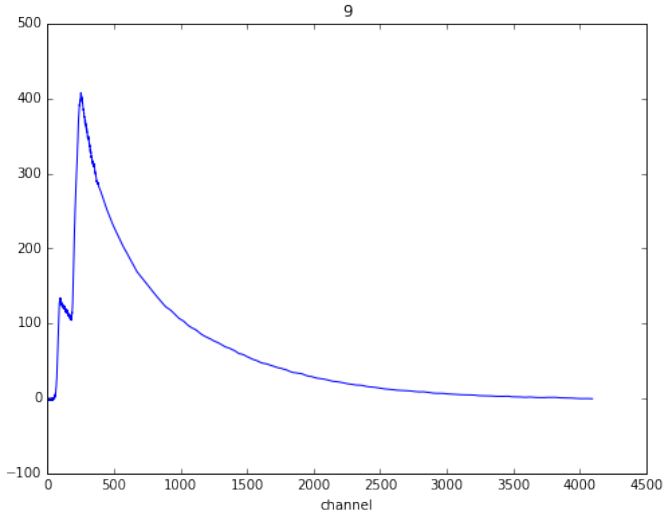


randomly sampled from cluster#8

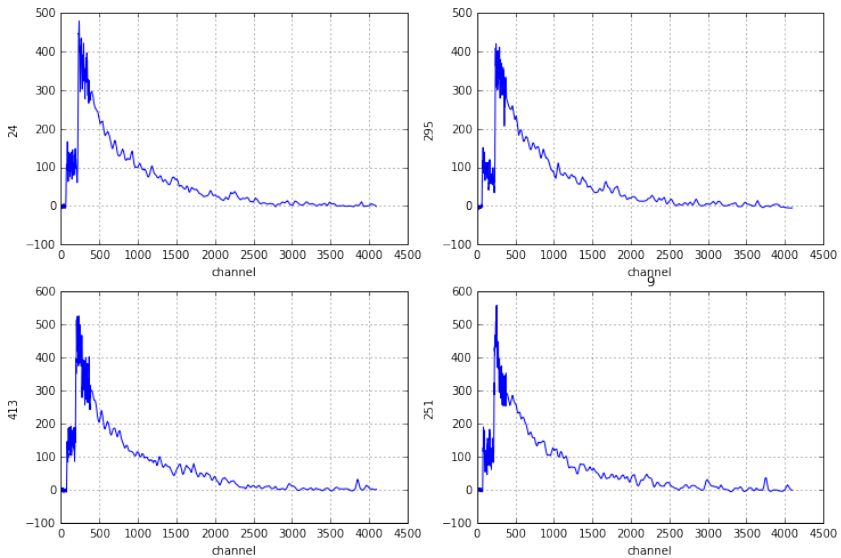


Dataset from Run009 - Examine the clusters - 9

Averaged waveform of cluster#9

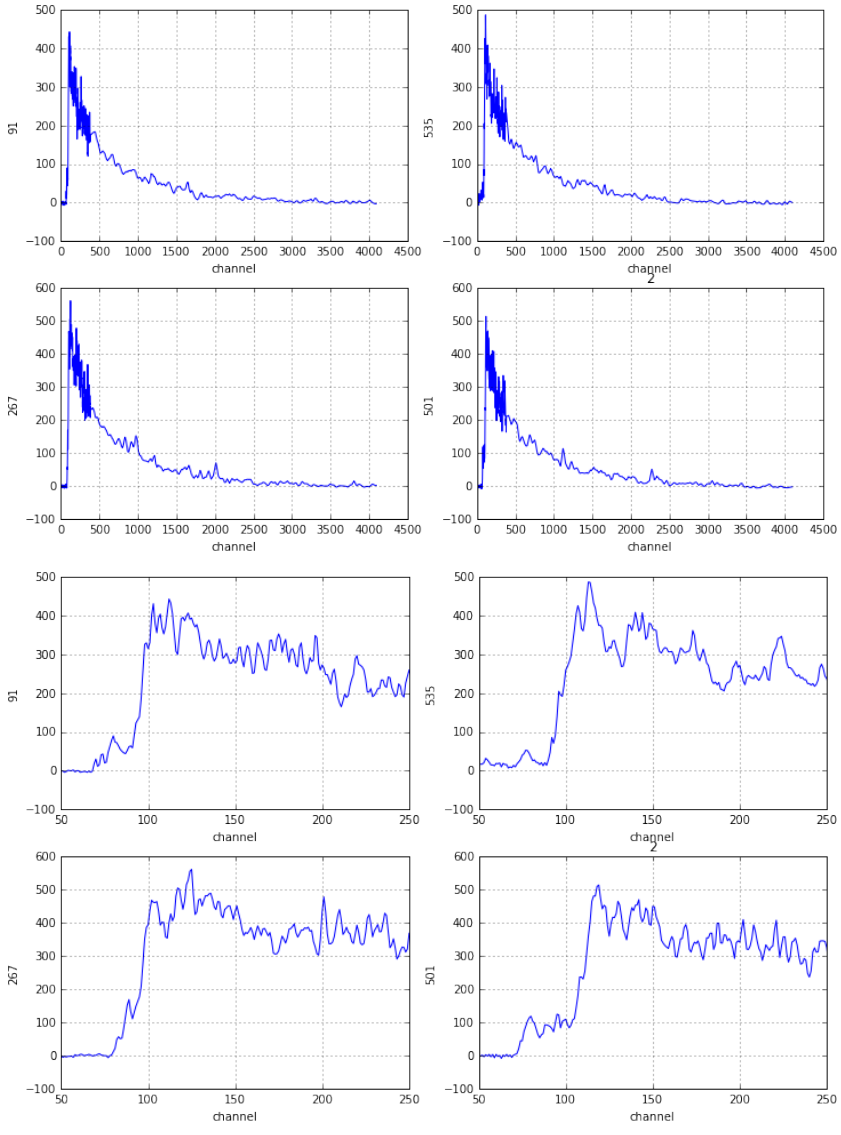


randomly sampled from cluster#9

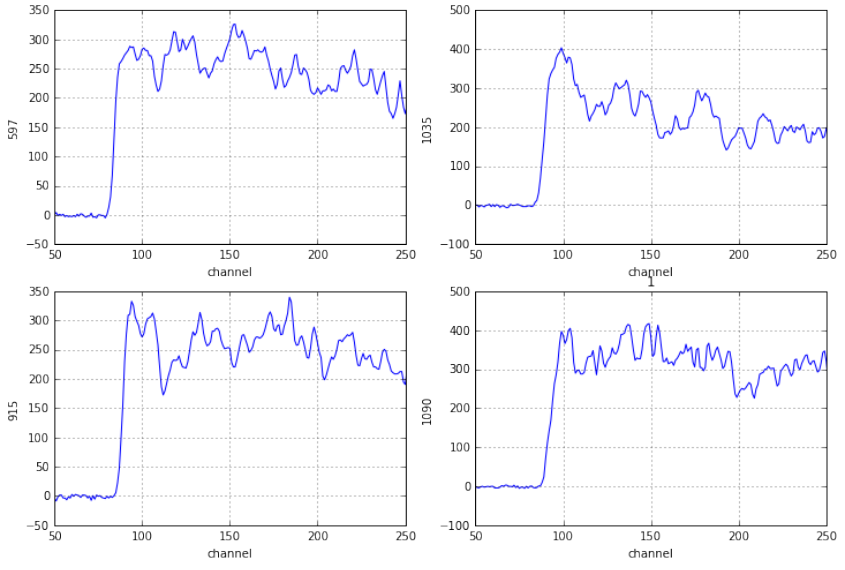


Cluster#2 in detail - 640 events

Cluster#2, double pulse within $\sim 0.1\mu s$?



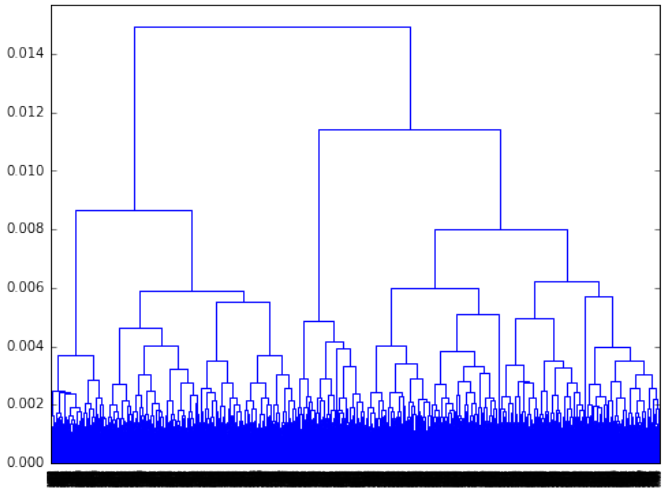
Comparing with cluster#1,

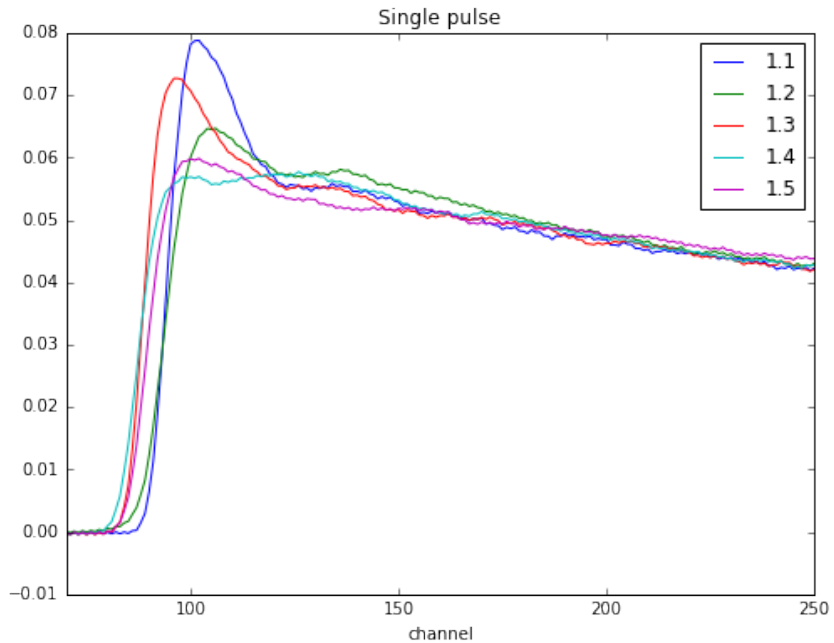


The minimum gap to classify double pulse? It was hard to determine this due to the low statistic available, but theoretically how far can we go?

Single Pulse(Cluster #1) in depth - 2552 events

My attempt to identify β/γ directly.



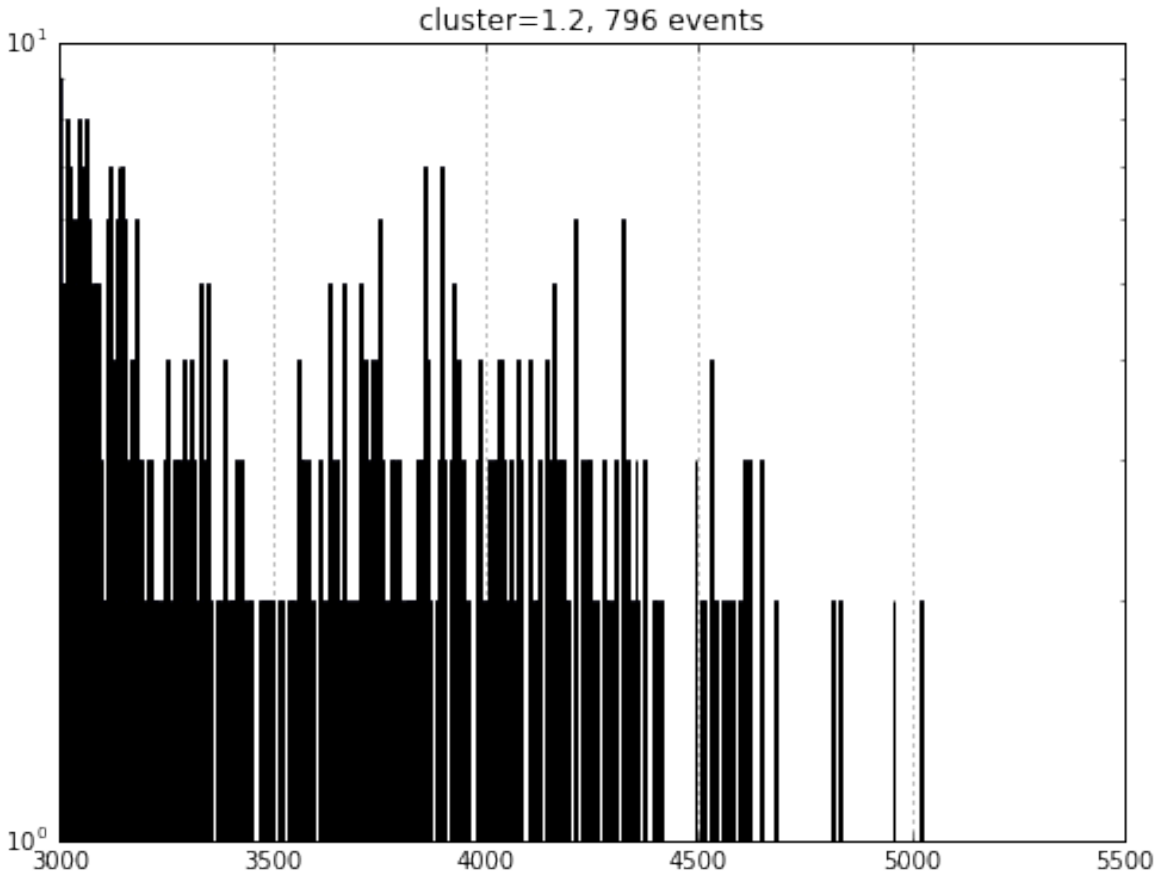


At 5 clusters,

Cluster#	no. of events
1.1	234
1.2	796
1.3	290
1.4	557
1.5	675

Based on Alpha list and β/γ list, Clusters#1.2 has the closest averaged waveform to reference pulse?

Single Pulse(Cluster #1) in depth - most β/γ like?



no. of events 4000 keV $<$ Energy[1] $<$ 4500 keV, without Tl208 cut $=$ 179 events

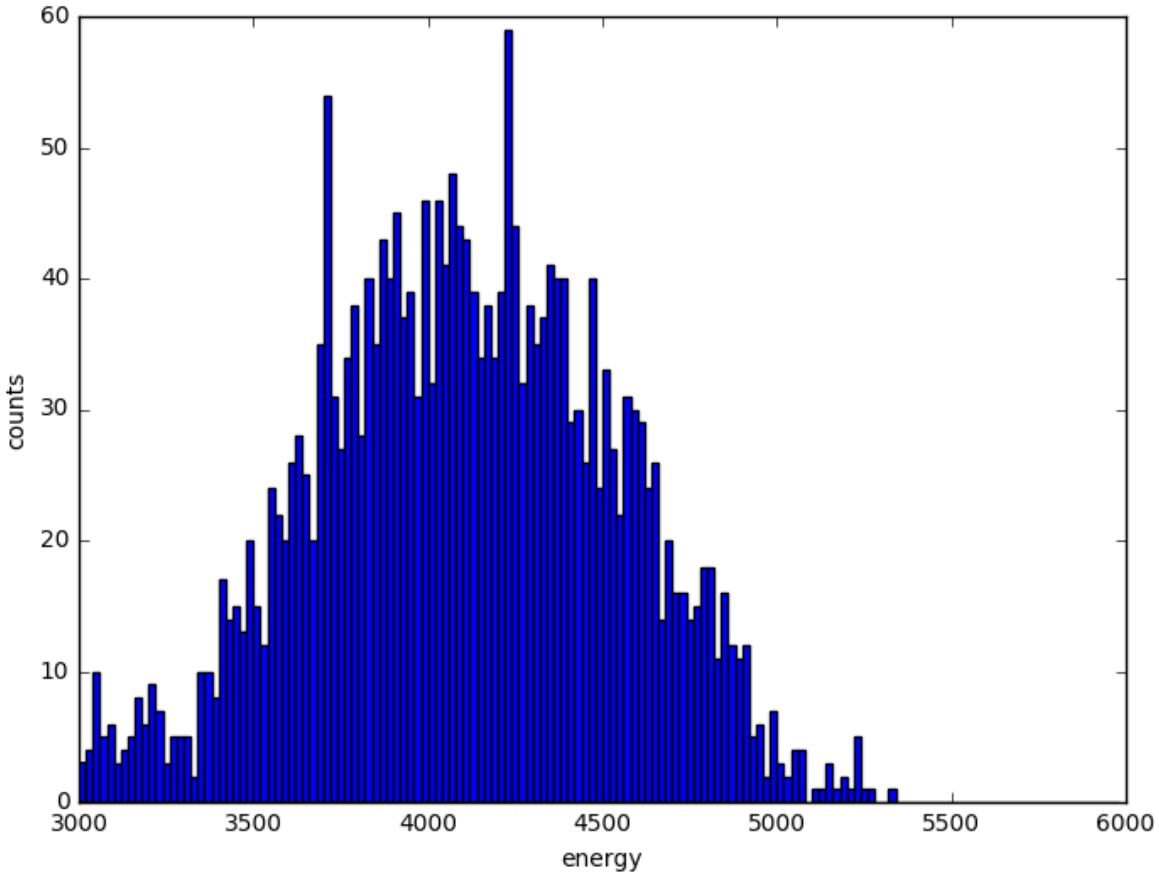
A work in progress

- An algorithm that is capable to clusters waveform into categories based on the similarity of their shape.
- Investigating how small the condensed latent space of encoder can go (This study, the encoder reduce 4096 to 512).
- A better way to determine the number of cluster? (Cross-validation/etc.)
- ....

End

Extra

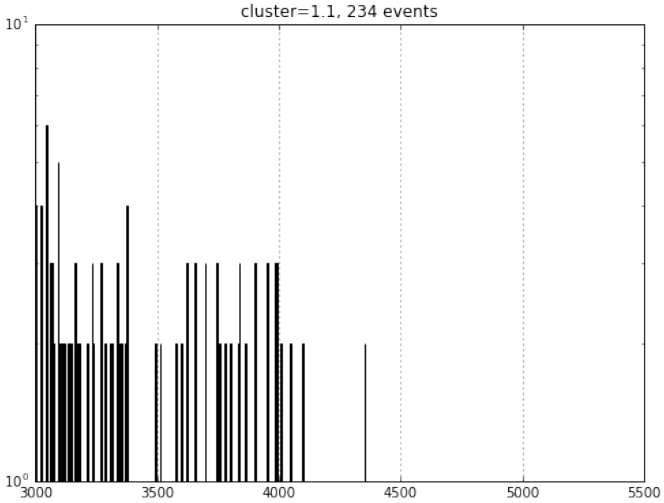
Double pulse spectrum after PSDPara[1]$<$ 1.5

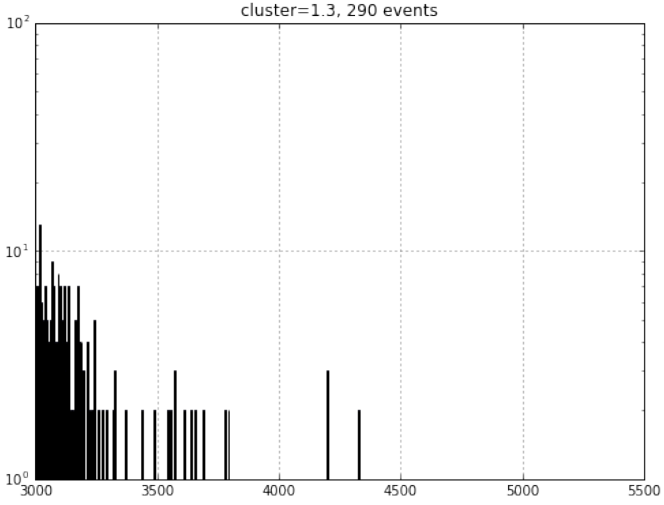


2416 events

Energy spectrums of Single Pulses, cluster#1.1, #1.3, #1.4, #1.5

Alpha?





??