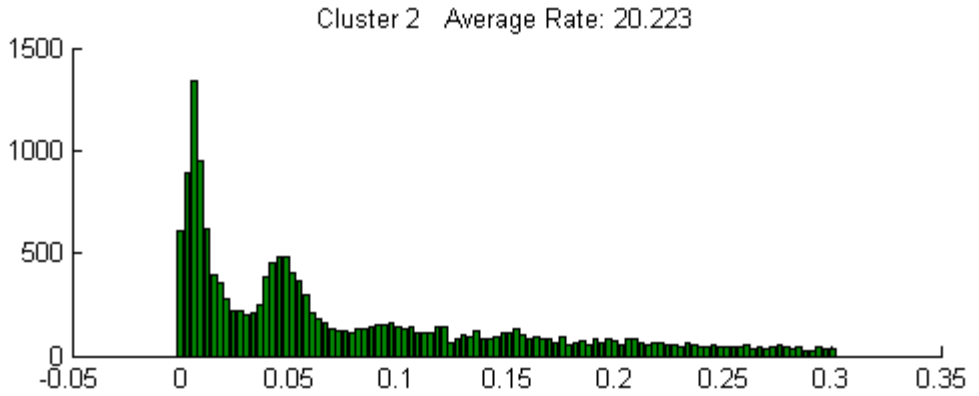
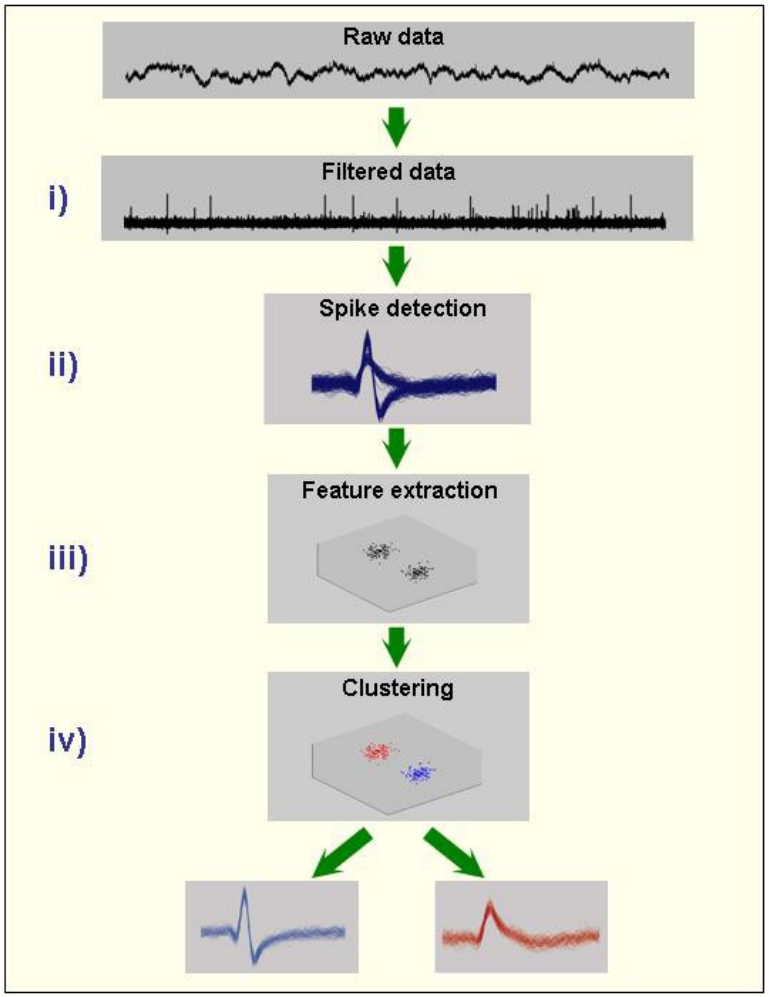


Unsupervised Batch Clustering of Neural Spikes

Scott Grimes - HIWI

Max-Planck-Institut
für biologische Kybernetik

Manual Spike Sorting is Tedious



General Sorting Steps:

i) The raw data is filtered

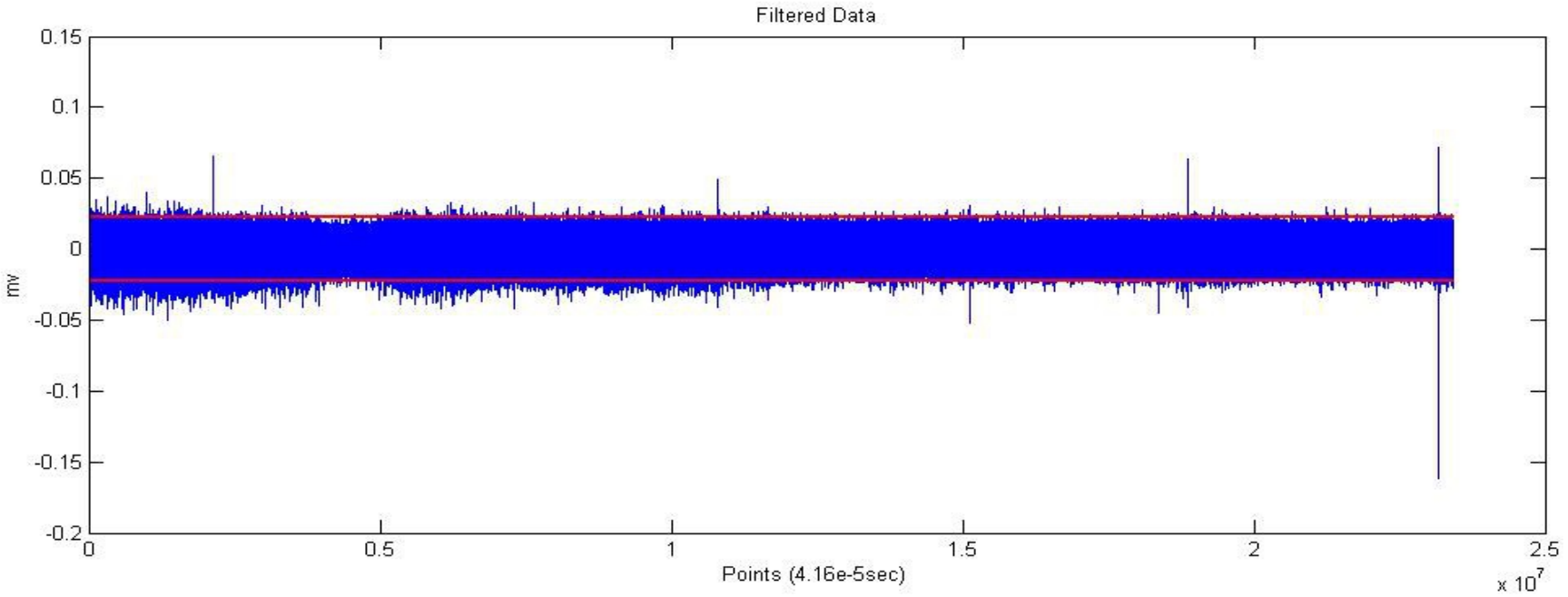
ii) Spikes are extracted from the data using an amplitude threshold

iii) Spikes features are determined (giving a dimensionality reduction)

iv) Spikes are clustered

Adapted from Quian Quiroga et al. (2004)

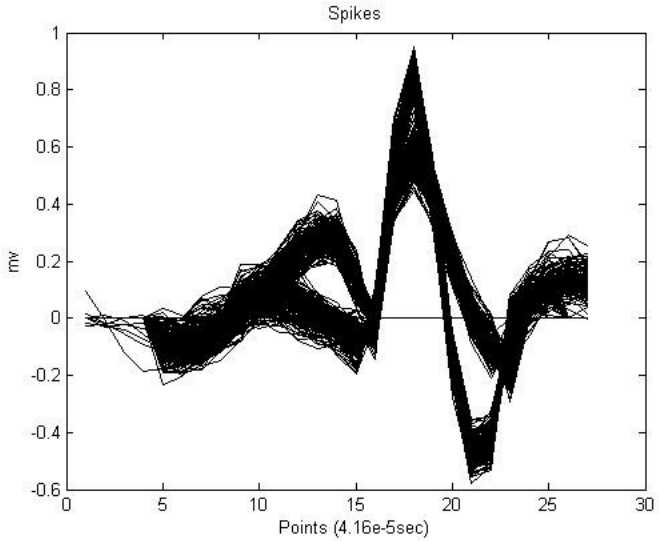
Automated Procedure

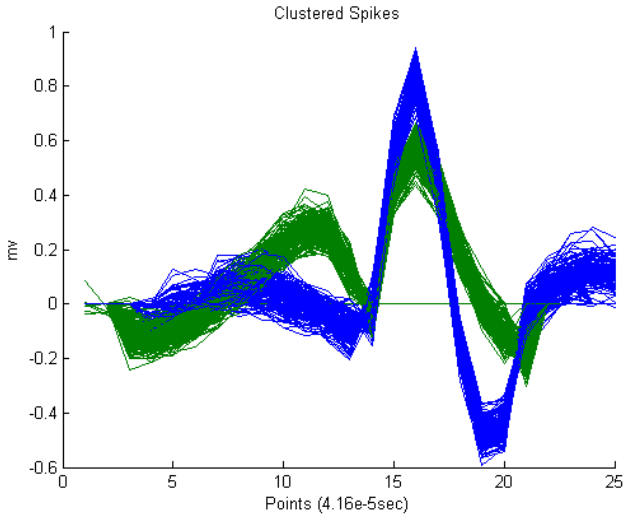


Spikes are extracted from filtered data using amplitude detection (positive, negative, or both) in standard deviations

Automated Procedure

Spikes are aligned to each other using their extreme amplitudes for sorting. There are two primary clustering methods available: Principal Component Analysis and Superparamagetic Clustering. Artifacts may be removed at this point using an amplitude threshold if they were not already removed during the initial importation.





Pre-Clustered

Post-Clustered

Superparamagnetic Clustering

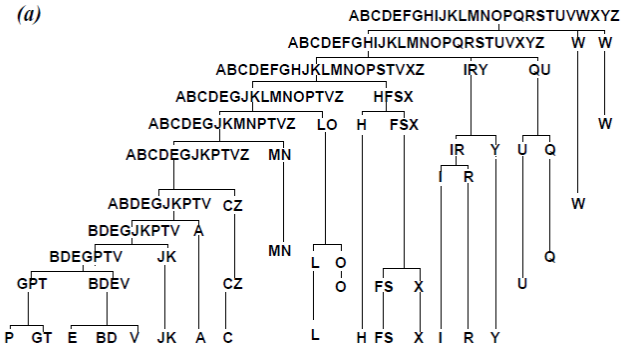
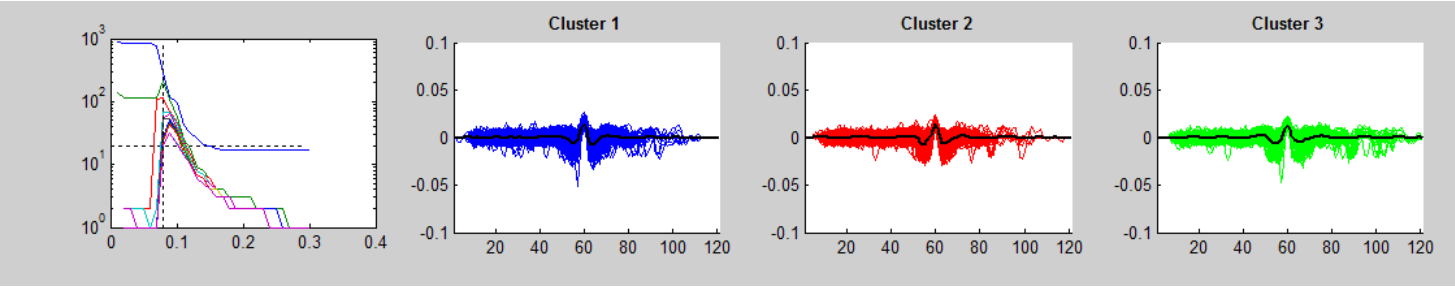


Fig. 3. Isolated–letter speech–recognition hierarchy obtained by (*a*) the Super–Paramagnetic method.

Low T
Not Clustered

High T
Over-Clustered

As T increases, the number of clusters increases and the size of the clusters decreases. The cost function which determines which spikes to include in a given cluster is fixed, which can leave many spikes un-clustered or ill-clustered.



For sorting the Principal Component Loadings without knowing the final number of clusters initially we can work backwards, over-clustering and then merging similar clusters until convergence

Superparamagnetic clustering of data--The definitive solution of an ill-posed proble
E Domany - Physica A: Statistical Mechanics and its Applications, 1999 - Elsevier

Principal Component Analysis for N-Unknown Clusters

Compute principal components for the `ingredients` data in the Hald data set, and the variance accounted for by each component.

```
load hald;
[pc,score,latent,tsquare] = princomp(ingredients);
pc,latent

pc =
   0.0678 -0.6460  0.5673 -0.5062
   0.6785 -0.0200 -0.5440 -0.4933
  -0.0290  0.7553  0.4036 -0.5156
  -0.7309 -0.1085 -0.4684 -0.4844

latent =
 517.7969
  67.4964
  12.4054
   0.2372
```

The following command and plot show that two components account for 98% of the variance:

```
cumsum(latent)./sum(latent)
ans =
      0.86597
      0.97886
       0.9996
            1
```

The loadings for n spikes are computed. In this example The vector *latent* is the eigenvalues of the covariance matrix of *ingredients*

Here the first two components account for ~98% of the variance in the set. We can reduce each spike vector to loadings with a small dimension which reduces the computation time required for clustering.

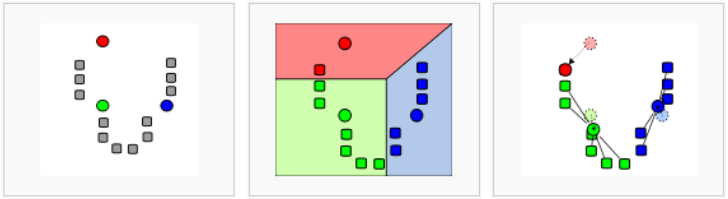
Example Adapted from MATLAB User Guide

Principal Component Analysis for N-Unknown Clusters

Once the number of required components are know we can begin clustering.

K-means clustering attempts to minimize the sum of squares within each cluster for an initial cluster guess M.

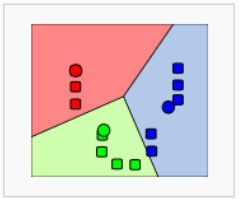
Demonstration of the standard algorithm



1) *k* initial "means" (in this case *k*=3) are randomly selected from the data set (shown in color).

2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3) The centroid of each of the *k* clusters becomes the new means.



4) Steps 2 and 3 are repeated until convergence has been reached.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

However we do not know the final number of clusters, so k-means is initialized with an arbitrary value M, known to be greater than N, the final number of clusters.

kmeans infographic

Principal Component Analysis for N-Unknown Clusters

A similarity matrix is computed for M clusters where M_{ij} is the similarity of the clusters i and j.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

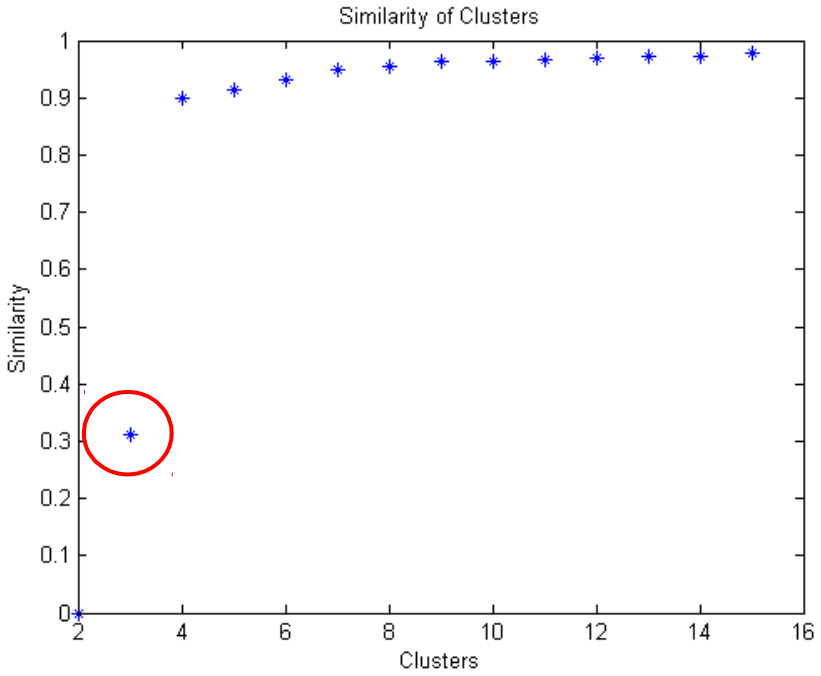
For M-1 iterations the clusters with the greatest similarity are combined, and a new similarity matrix is computed…

0	-0.9559	-0.8860	0.9543	-0.8947	-0.3402	0.9205	-0.9574	-0.8932	-0.9875	-0.9765	-0.9236	0.8941	0.9616	-0.9674
0	0	0.9283	-0.9472	0.9736	0.3104	-0.9254	0.9559	0.8906	0.9513	0.9610	0.9625	-0.9354	-0.9768	0.9340
0	0	0	-0.9220	0.9508	0.0659	-0.7938	0.8197	0.9662	0.8855	0.9473	0.9175	-0.9656	-0.9367	0.9176
0	0	0	0	-0.9268	-0.3153	0.8281	-0.9045	-0.9338	-0.9561	-0.9576	-0.9457	0.8857	0.9782	-0.9552
0	0	0	0	0	0.2342	-0.8673	0.9050	0.9055	0.9118	0.9318	0.9480	-0.9551	-0.9580	0.8873
0	0	0	0	0	0	-0.5165	0.3740	0.1719	0.3801	0.2164	0.4215	-0.2183	-0.2736	0.3272
0	0	0	0	0	0	0	-0.9243	-0.7925	-0.9261	-0.8825	-0.9117	0.8905	0.8765	-0.8890
0	0	0	0	0	0	0	0	0.7888	0.9551	0.9368	0.8766	-0.8448	-0.9398	0.8767
0	0	0	0	0	0	0	0	0	0.9127	0.9321	0.9286	-0.9438	-0.9349	0.9576
0	0	0	0	0	0	0	0	0	0	0.9712	0.9345	-0.9069	-0.9720	0.9673
0	0	0	0	0	0	0	0	0	0	0	0.9156	-0.9320	-0.9764	0.9575
0	0	0	0	0	0	0	0	0	0	0	0	-0.9420	-0.9499	0.9544
0	0	0	0	0	0	0	0	0	0	0	0	0	0.9159	-0.9158
0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.9595
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

In this example cluster 14 is merged into cluster 4

Principal Component Analysis for N-Unknown Clusters

When the number of clusters has been reduced from M to 2, the maximum similarity between clusters for every iteration of M is taken:

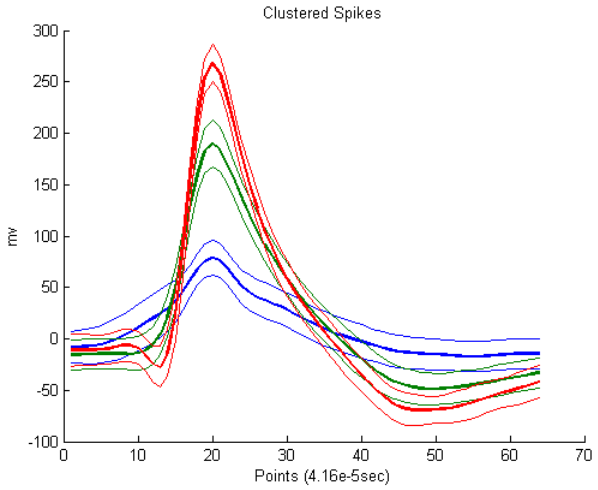


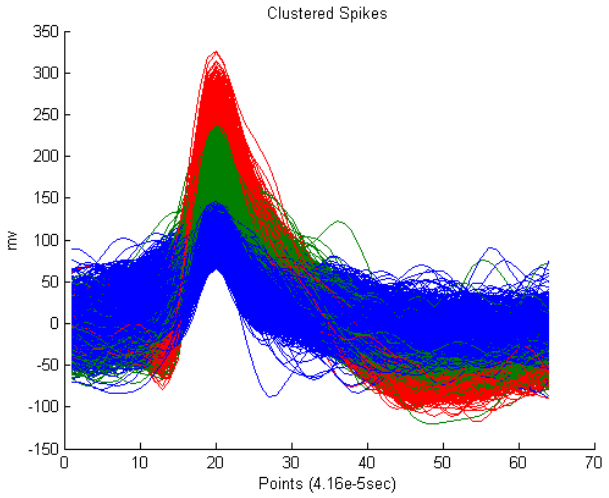
The maximum cluster similarity in iteration 3 was far greater than previous iterations, all similar clusters have been collected together at this point. We can use the cluster indexes of this iteration (3 final number of clusters).

Post-Processing Review

Several tools are included to review the clusters after they have been sorted:

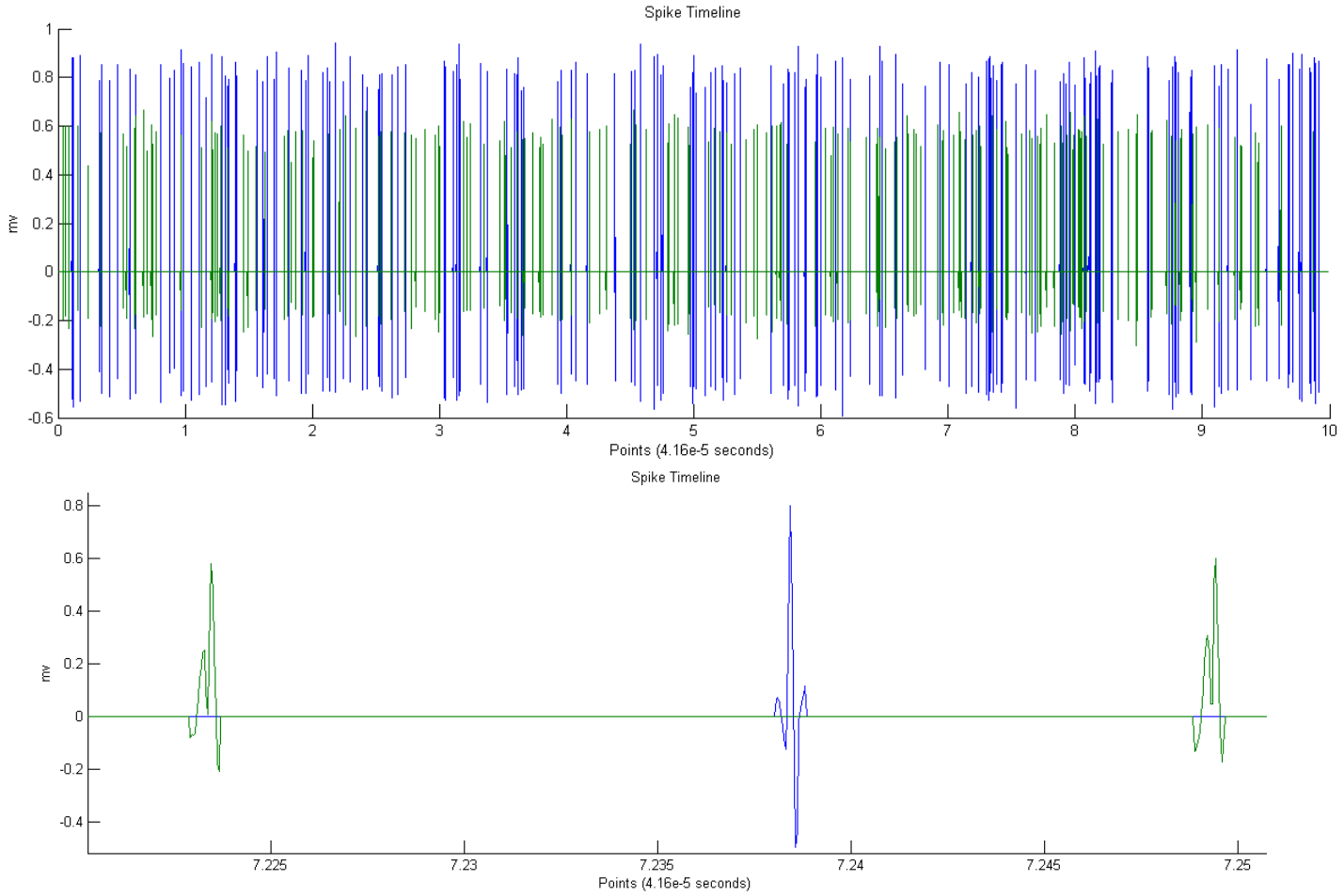
Clusters can be combined or discarded. Spikes shapes and ISI histograms can be plotted. Clusters may be overlaid for comparison.



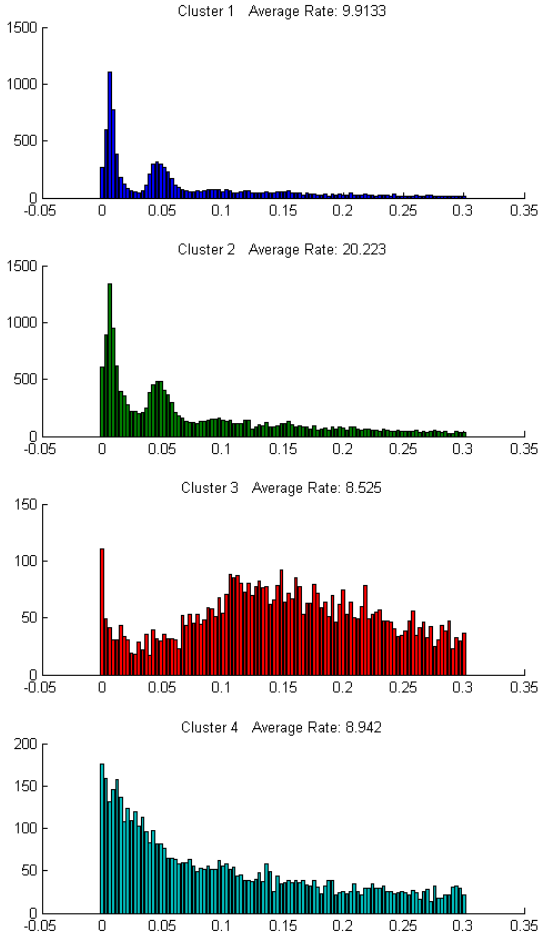


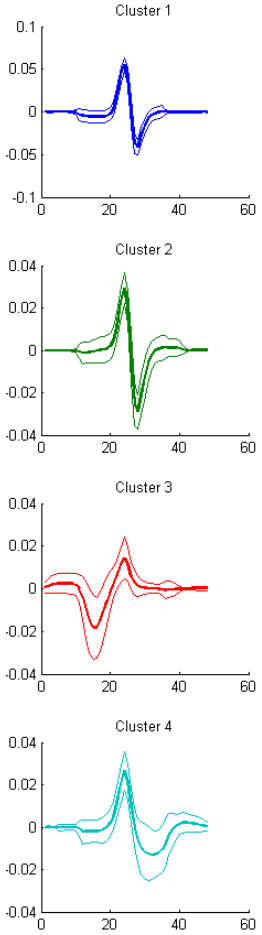
Post-Processing Review

Timelines for each cluster can be created using the spike shapes and timestamps for each event (plotted together or separately for comparison)



Post-Processing Review

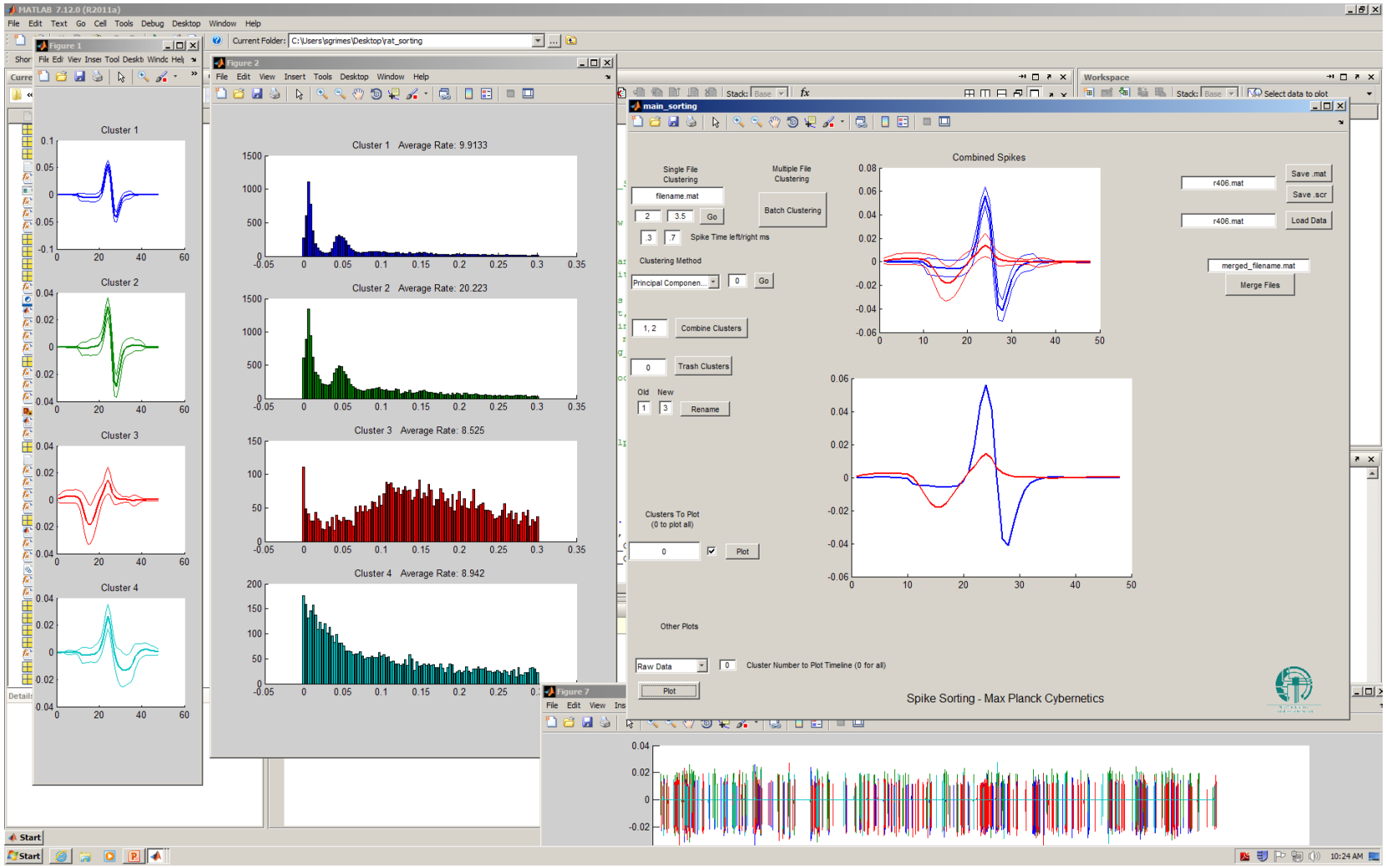




The inter-spike interval histogram and average rate of fire for each cluster is shown on the left.

Cluster shapes with standard deviation shown on the right

Interface


Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 1 Average Rate: 9.9133
Cluster 2 Average Rate: 20.223
Cluster 3 Average Rate: 8.525
Cluster 4 Average Rate: 8.942
main_sorting
Single File Clustering
Multiple File Clustering
filename.mat
Batch Clustering
Go
Spike Time left/right ms
Clustering Method
Principal Componen...
Combine Clusters
Trash Clusters
Old New
Rename
Combined Spikes
r406.mat
Save .mat
Save .scr
Load Data
merged_filename.mat
Merge Files
Clusters To Plot (0 to plot all)
Plot
Other Plots
Raw Data
Cluster Number to Plot Timeline (0 for all)
Spike Sorting - Max Planck Cybernetics
Workspace