TSEClustering = Threshold Smoothing Ensemble Clustering – is an unsupervised clustering algorithm that allows smoothing over noisey data with low observational size in high dimensional space.

Smooth Ensemble Workflow, Steps:  
 1. Input data of *n x p* (observations x predictors)

2. Intra-observational similarity matrices are generated through Gap statistic in various subsets of predictor space

3. Subset predictor space (*i.e.* subsets of the features) is calculated manually by selecting groups of features by use / type

4. Analysis without each predictor subset is performed (*i.e. leave-one-out*)

5. Resulting similarity matrices are used to independently cluster observations through the optimal number of clustersin Kmeans

6. Resulting cluster assignments are used to obtain co-occurrence of observations in relation to total number of analyses

7. A threshold approximation is applied to perform observational similarity dropout

8. Smooth Ensemble of observational correlations for Kmeans analysis   
Smooth Ensemble Workflow, Part 1:

Smooth Ensemble Workflow, Part 2:

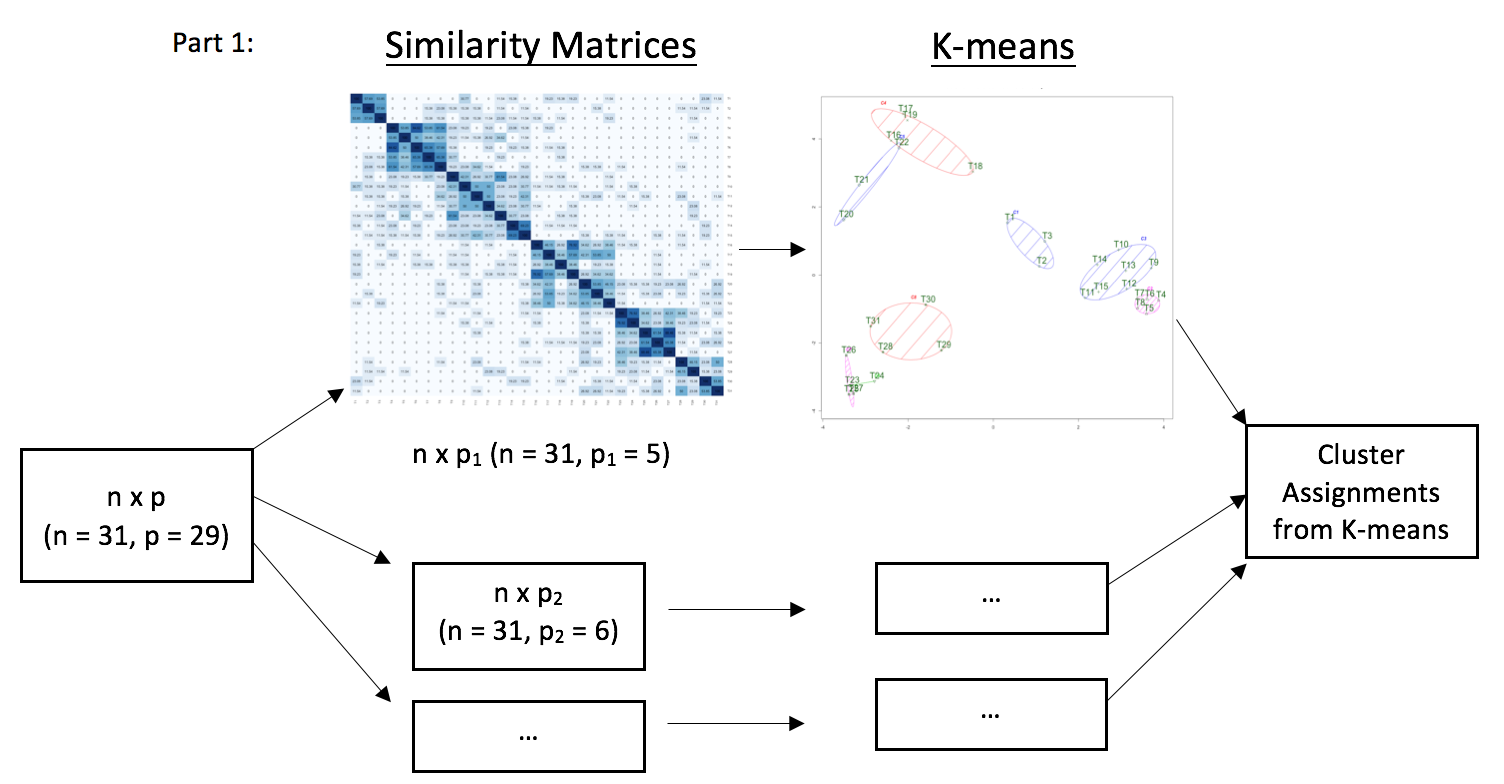
Kmeans -> pairwise co-occurrence = raw co-occurrence matrix -> smoothing function (math function) -> smooth ensemble (corr matrix) = probabilistic co-occurrence matrix  
  
  
Steps for feature importance:

1. Each subgroup of features has
   1. within- vs. between-cluster variance
      1. measure of cluster goodness
   2. actual cluster assignment & prob. Of assignment
      1. measure of park goodness / sureness
2. Compare:
   1. 1.a.i across subgroups to get influence of features on cluster assignments
   2. 1.a.ii across subgroups to get influence of features on park assignment
3. Take high probs of assignments / importance from 1. and 2. to find what factors determined cluster assignment
   1. To get equation (e.g. linear combination) to get to cluster assignment
   2. E.g. only use probs > 0.9

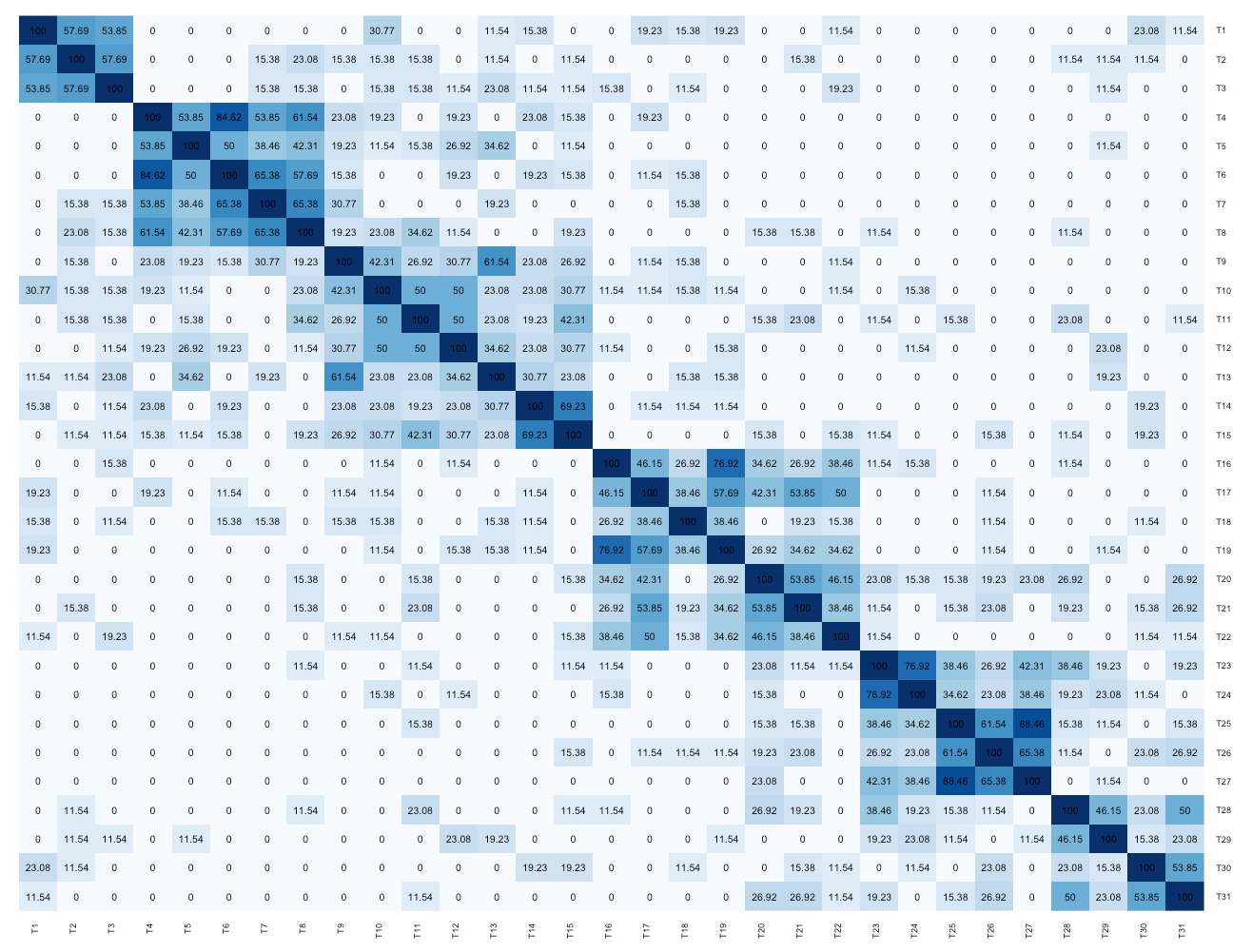
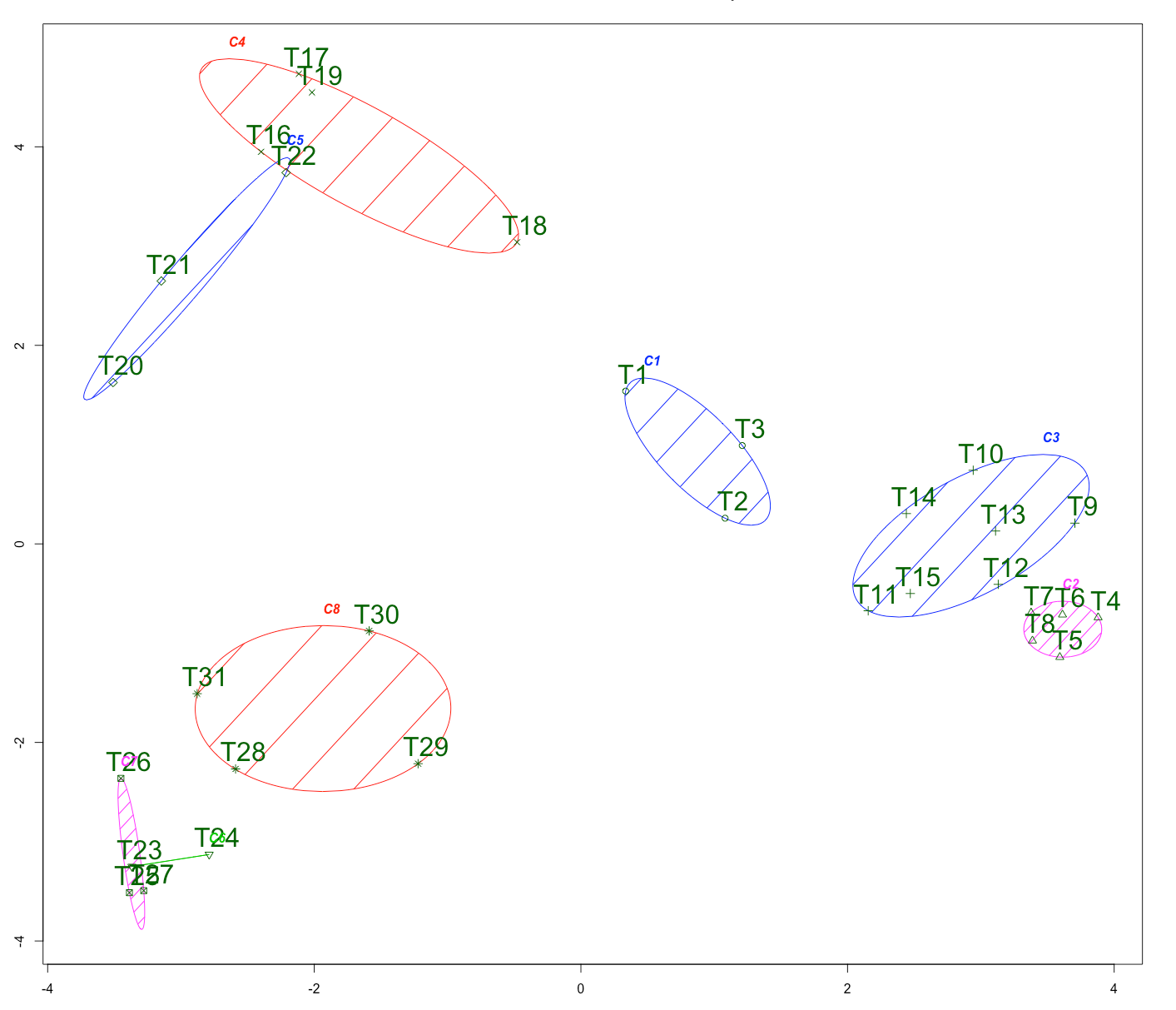
References:

Tibshirani, R., Walther, G. and Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), pp.411-423.

Ciss, S., 2015. Random Uniform Forests for Classification, Regression and Unsupervised Learning.  
  
  
Alternative Figure 2:



Part 1:



n x p  
n = 31 (Tools)

p = 29

(Allfeatures)

G = {g1, g2, g3, g4, g5, g6}

g1 = p1 to p6

…

g6 = p26 to p29

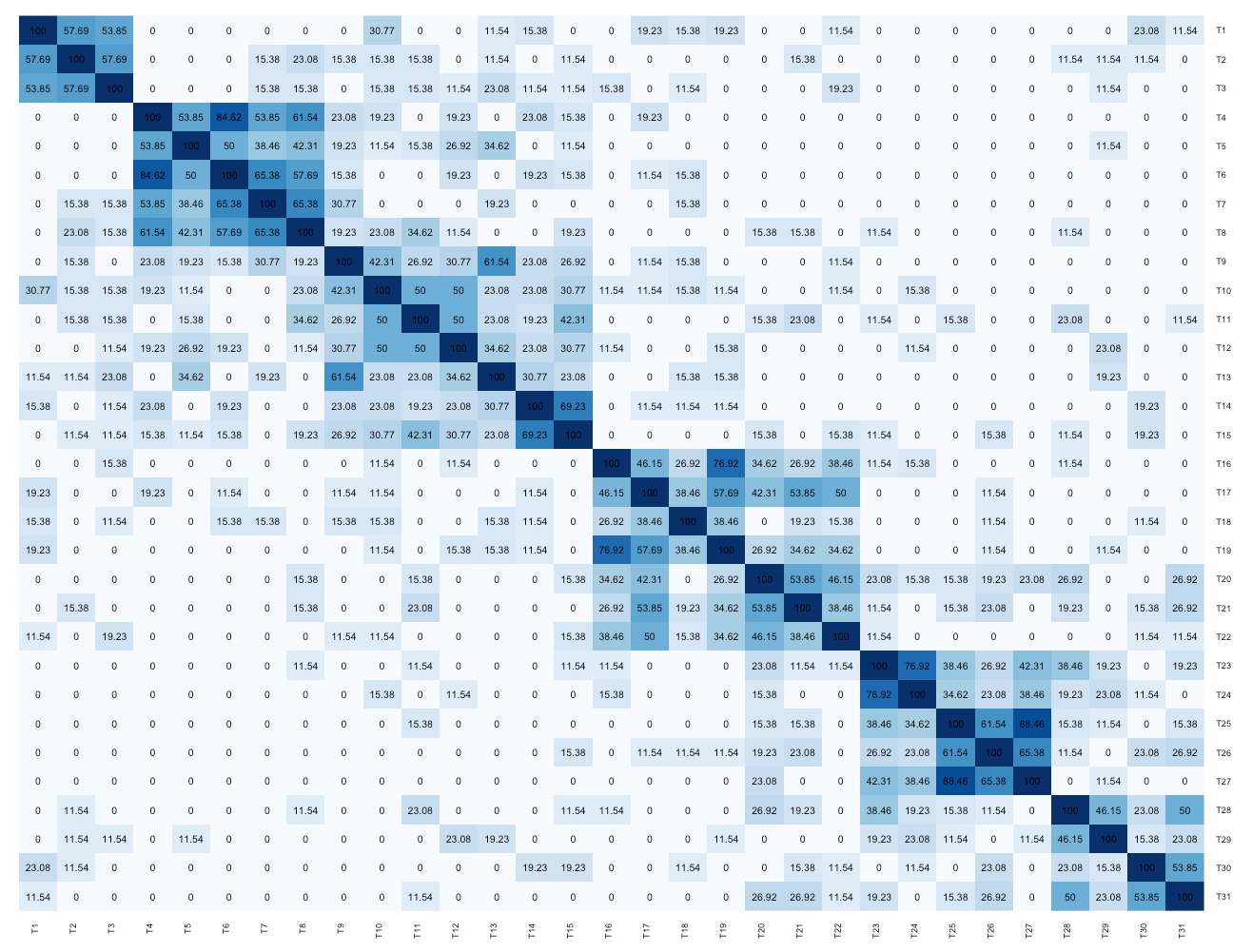
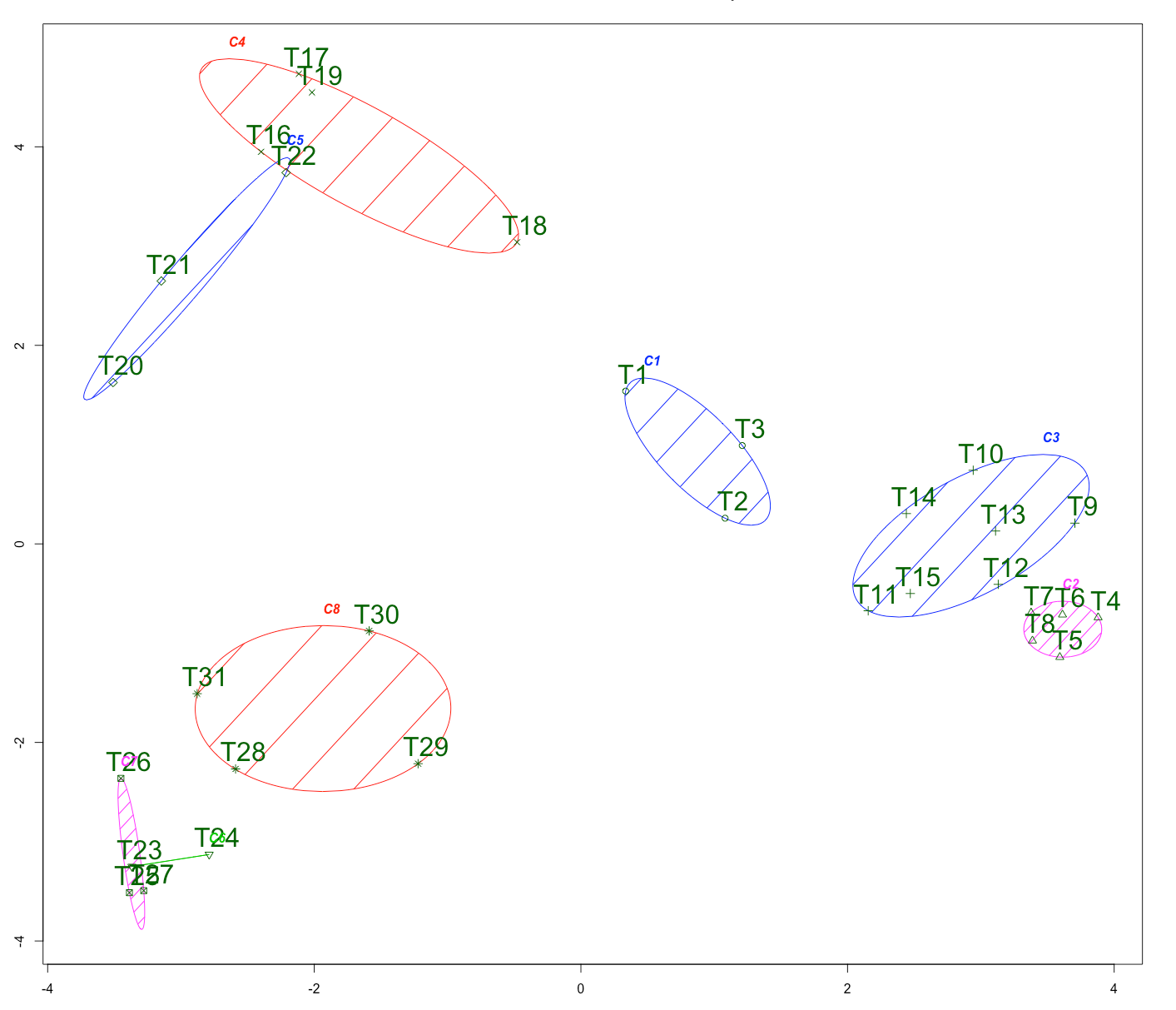
for each gi in G:

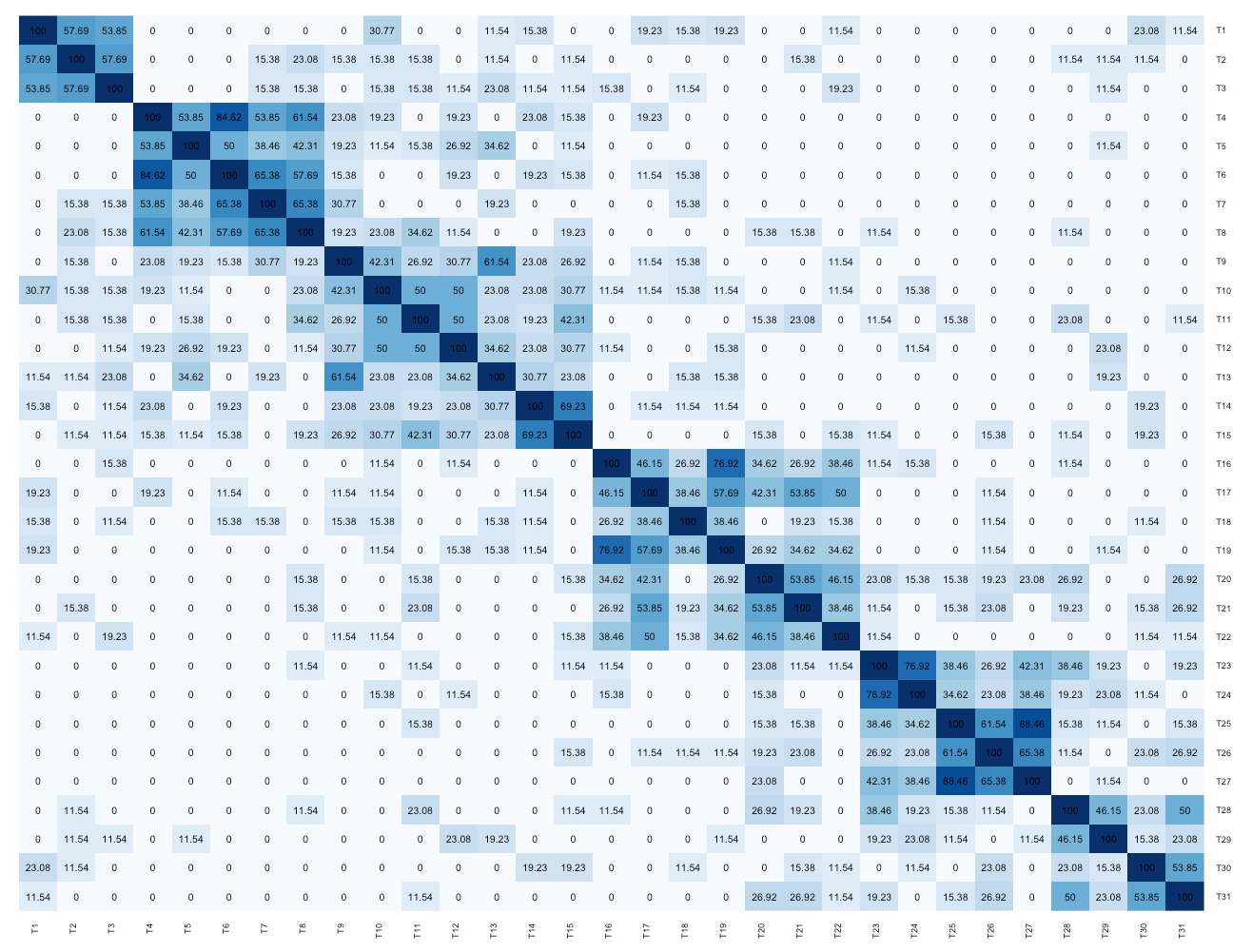
i) generate similarity matrix for n x gi

ii) generate similarity matrix for n x (for all gx not gi)

Similarity Matrices

K-means





Aggregate cluster assignments from K-means for each gi in G

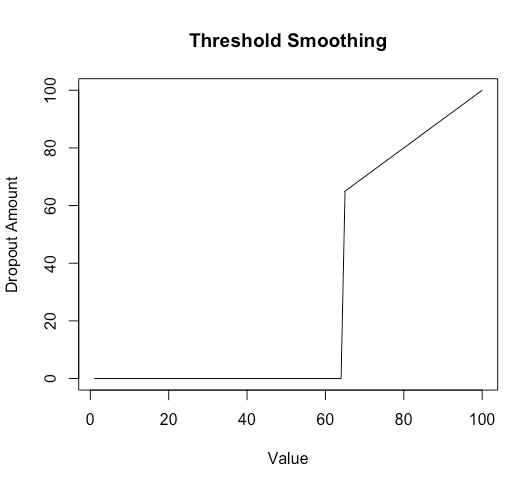
(i.e. from each group of features)

Pairwise Co-Occurrences

K-means

Pairwise Occurrences

Cluster Assignments from K-means



Threshold Smoothing