KMedian Ballot Clustering with Integer Programs

Overview

The notebook (OptimalBallotClustering.ipynb) contains three types of models for kMedians (using L1 distance metric) clustering of the ballots each representing different restrictions on possible centroids: continuous (any point), possible ballots (continuous_rest), and true ballots in the election (discrete). For each of these model types, the ballots can be represented with three possible embeddings: head-to-head (hh), pessimistic borda (bordaP), and average borda (bordaA).

Models

Continuous (All Points)

All of these models are based on the `continuous_model` function, which abstracts away the particular embedding. To use, you pass the embedded ballot matrix, weights representing the count, the number of clusters desired, and the value set. The value set represents the possible values for each dimension of the ballots and is dependent on the embedding.

The `continuous_hh`, `continuous_bordaP`, and `continuous_bordaA` functions generate the models with their respective embeddings. All that is passed to these functions is the raw election data dictionary matching ballots to counts (generated from `parse` in the utility function section).

The resulting centroids and assignments of this model are represented by the following named variables in Gurobi:
`z[i,r,v]` : 1 if the ith dimension of the median of the rth cluster is value v
`x[j,r]` : 1 if ballot j is assigned to cluster r.

The constraints for the model are simply the constraints specified by the overleaf writeup. The variables in the writeup and in the code should (roughly) correspond.

Restricted to Possible Ballots

Since the constraints for each of these models vary based on the embedding, they do not share a common abstracted function.

The `continuous_rest_hh`, `continuous_rest_bordaP`, and `continuous_rest_bordaA` functions generate the model with their respective embeddings with identical interfaces as the `continuous` functions.

`continuous_rest_hh` is not yet implemented.

The resulting centroids are represented in the model the same as in the `continuous` functions.

The constraints are also explained in the overleaf writeup.

Discrete (Restricted to Real Ballots)

These models rely on the `formulate_new` function which takes in a dictionary of possible pairs of ballots (`possible_pairs`), a dictionary matching each ballot to its count (`multiplicities`), a dictionary matching each pair of ballots to their distance (`distances_dict`), and `NUM_CLUSTERS`. There is also functionality for outliers (`NUM_OUTLIERS`), but this is not currently used. The function returns a variety of information on the model, but we must make use of the returned model in the dependent functions.

The `discrete_HH`, `discrete_bordaP`, `discrete_bordaA` functions generate models for their respective embeddings. They take in the same arguments as the `continuous` interface and do the work of calculating distances between points using functions in the utility section.

The resulting centroids are embedded in the `isCenter[i]` variables where `i` is the original index of the centroid ballot, and `isCenter[i] == 1` when `i` is a centroid.

Note: These models are directly from a previous edition of the code, so Professor Shmoys has a better idea of how they were implemented.

Brief Note on Saving/Loading Models

The running/saving models section of the notebook optimizes and saves models. They can often take hours, even on a M2 Max chip with 12 cores. The saved files follow the same naming conventions as the functions, particularly: `{model_type}_{embedding}_{num_clusters}`. Where `model_type` in `['continuous', 'continuous_rest', 'hh']` and `embedding` in `['bordaA', 'bordaP', 'hh']`.

The models are stored as `.mps` (model) and `.sol` (solution) files.

Results

We currently have results on 8 of the 9 possible models on 2 clusters (so no `continuous_rest_hh`).

There are various functions to generate results, but these are abstracted away to the `perform_analysis` function which allows one to specify `model_type`, `embedding_type`, and `num_clusters`. This uses the same naming conventions as specified above. The function loads in the model using the `.mps` and `.sol` files. It uses global variables `VALUE_SETS`, `BALLOTS`, and `DIMS` to get the proper setup for each `embedding_type`.

It then extracts the centroids. This is super hardcoded to the Gurobi variable names. There is also some weirdness in that the results are not properly loaded to the model as solutions, so we must access them with `Var.Start` rather than `Var.X`.

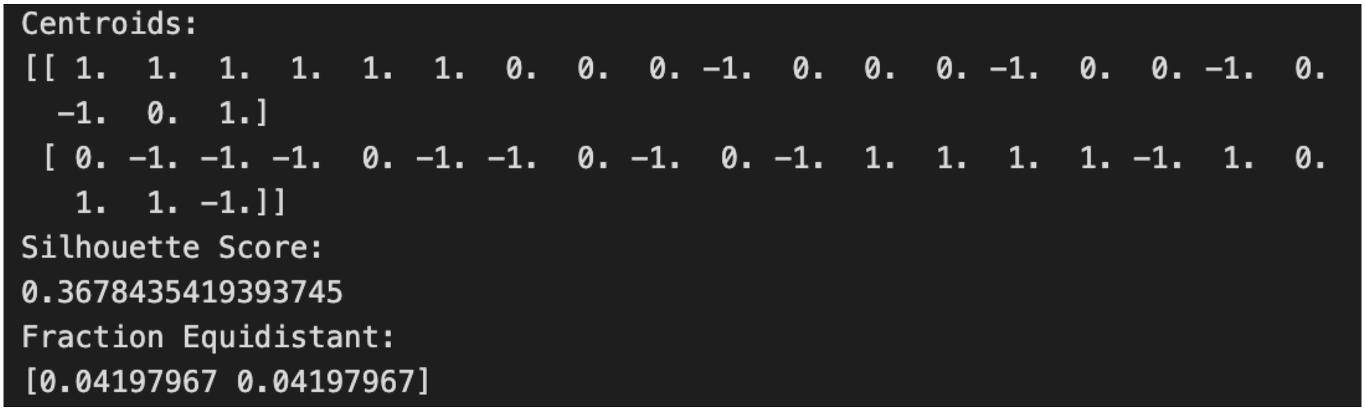
It prints the following:
`Centroids` representing the centroids in their embedded form.
`Silhouette Score` representing the average silhouette score over all ballots.
`Fraction Equidistant` representing the fraction of equidistant ballots for each cluster.
`MDS` graph with color coded clusters and centroids displayed. This is done with the proper L1 distance metric now. Note that this does not currently factor in counts of ballots so some points may represent multiple ballots.
`Histogram` of distances of centroids to points in its cluster. The y-axis is density and factors in counts.
`PCA` graph showing the graph of transformed points and centroids
`PCA Components` showing the PCA components of the ballots (reduced to 2 dimensions).
`PCA Explained Variance Ratio` Ratio of explained variance for each component.

The results for each of the 8 models are below. They are grouped by embedding.

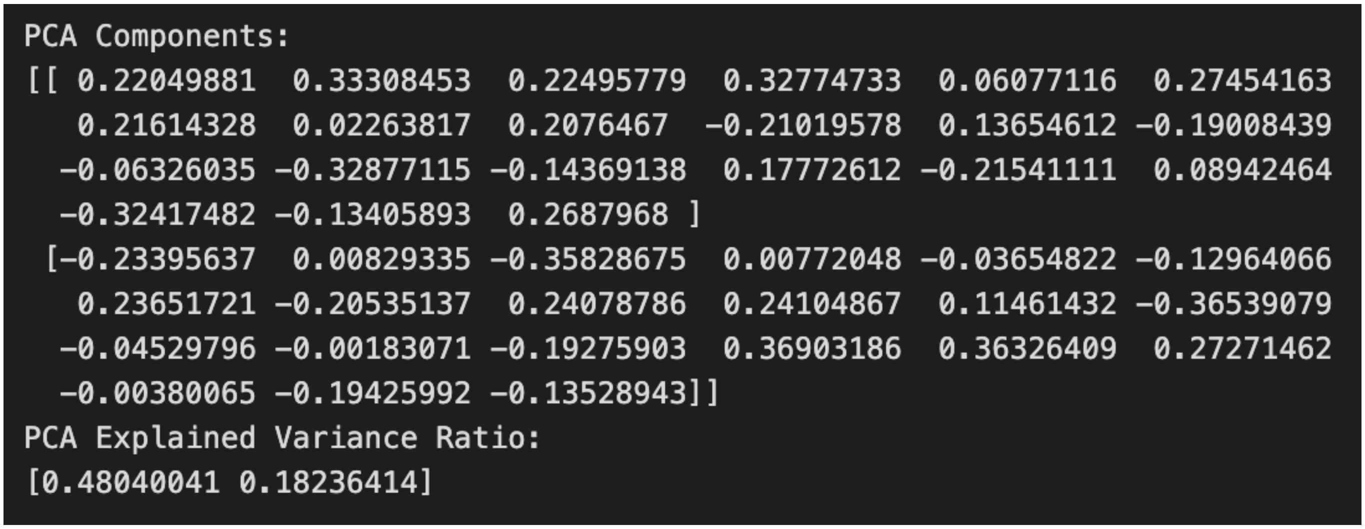
One interesting thing of note is that in the `continuous_rest_bordaA` model, the centroids produced are bullet ballots (only ranking one candidate).

Head-to-Head

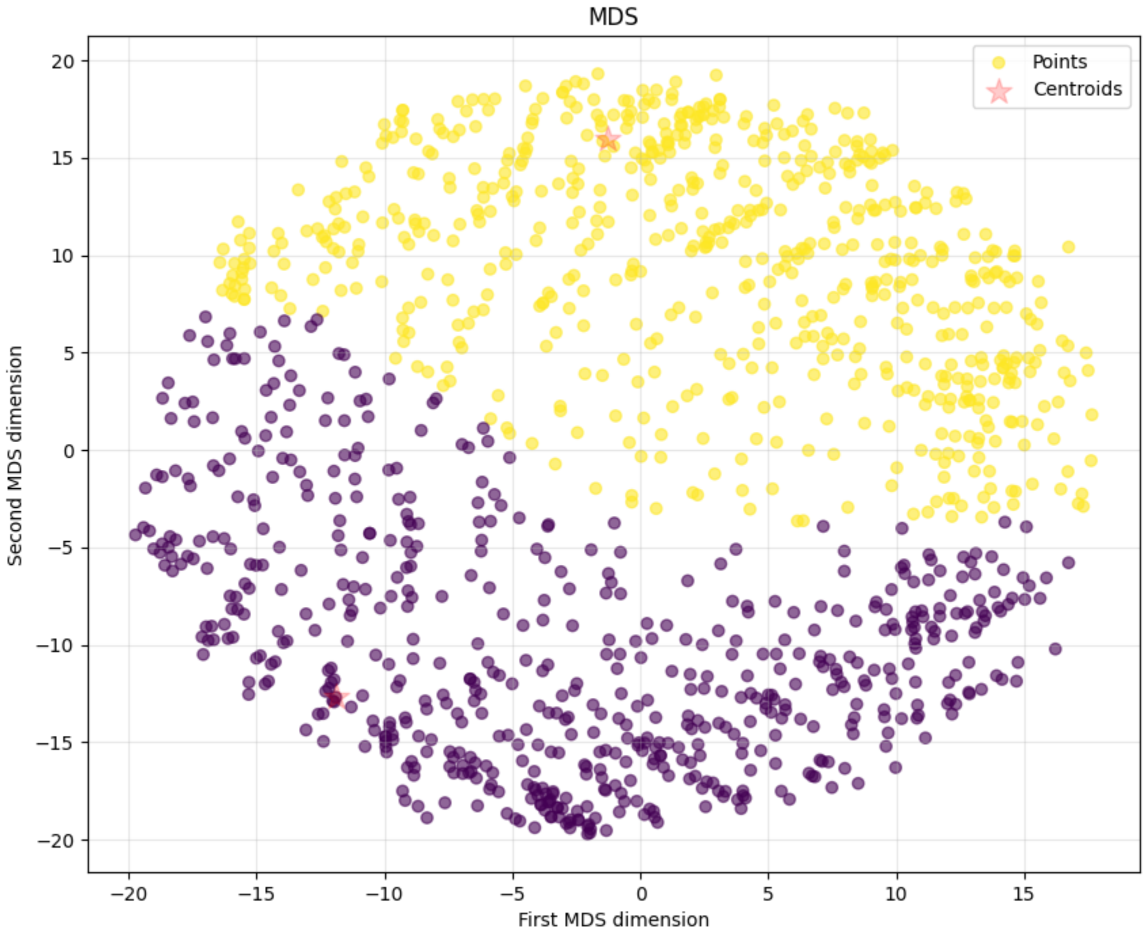
Continuous

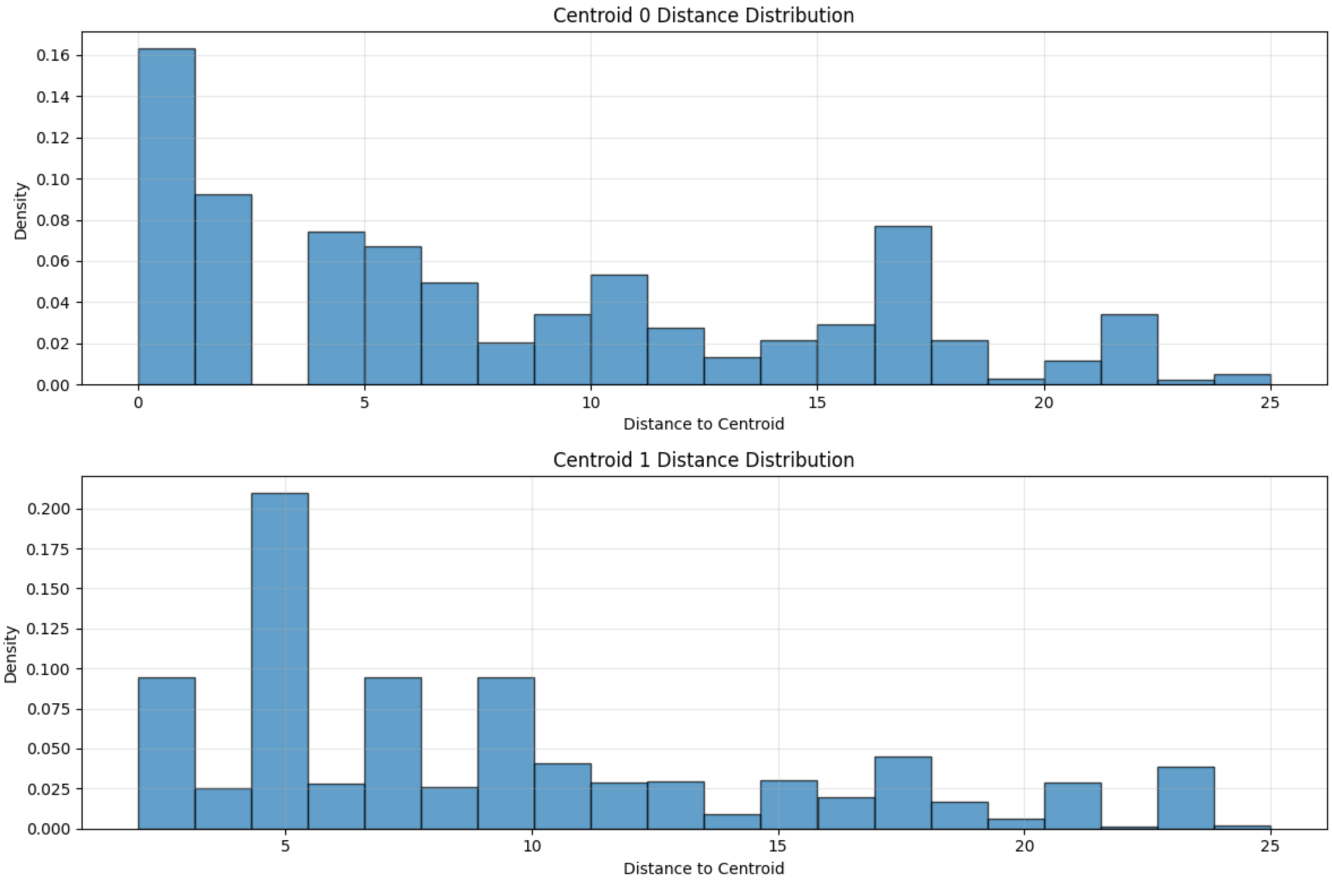


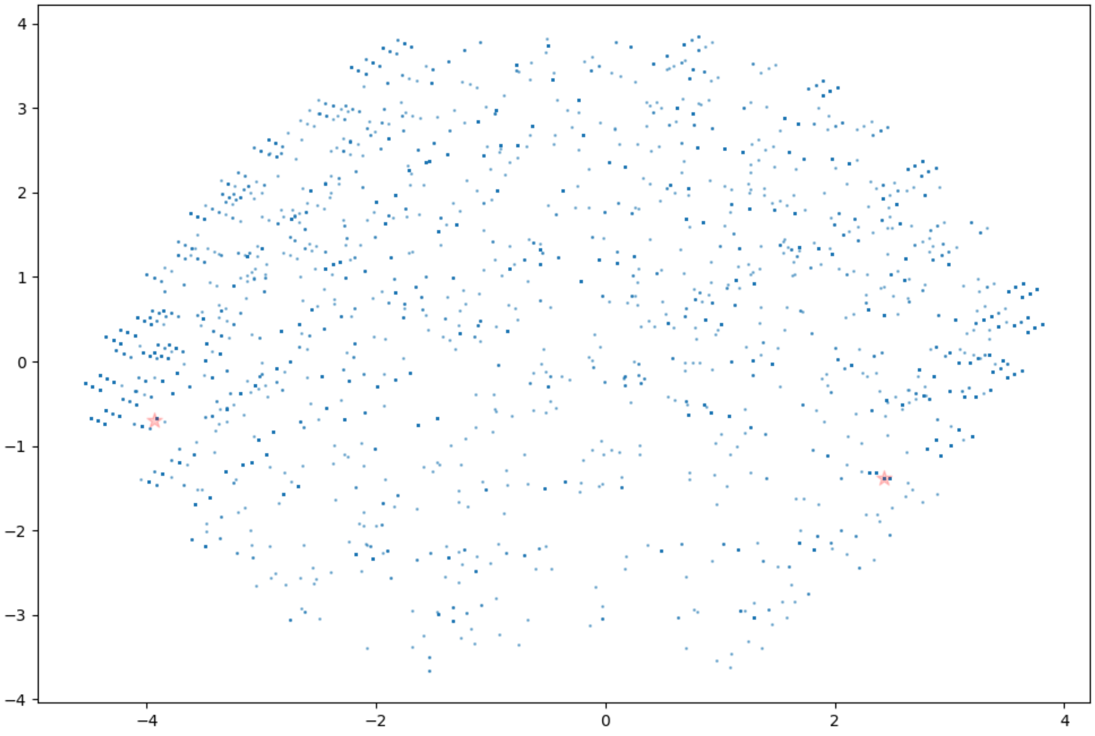
```
Centroids:
[[ 1.  1.  1.  1.  1.  1.  0.  0.  0. -1.  0.  0.  0. -1.  0.  0. -1.  0.
  -1.  0.  1.]
 [ 0. -1. -1. -1.  0. -1. -1.  0. -1.  0. -1.  1.  1.  1.  1. -1.  1.  0.
   1.  1. -1.]]
Silhouette Score:
0.3678435419393745
Fraction Equidistant:
[0.04197967 0.04197967]
```



```
PCA Components:
[[ 0.22049881  0.33308453  0.22495779  0.32774733  0.06077116  0.27454163
   0.21614328  0.02263817  0.2076467  -0.21019578  0.13654612 -0.19008439
  -0.06326035 -0.32877115 -0.14369138  0.17772612 -0.21541111  0.08942464
  -0.32417482 -0.13405893  0.2687968 ]
 [-0.23395637  0.00829335 -0.35828675  0.00772048 -0.03654822 -0.12964066
   0.23651721 -0.20535137  0.24078786  0.24104867  0.11461432 -0.36539079
  -0.04529796 -0.00183071 -0.19275903  0.36903186  0.36326409  0.27271462
  -0.00380065 -0.19425992 -0.13528943]]
PCA Explained Variance Ratio:
[0.48040041 0.18236414]
```

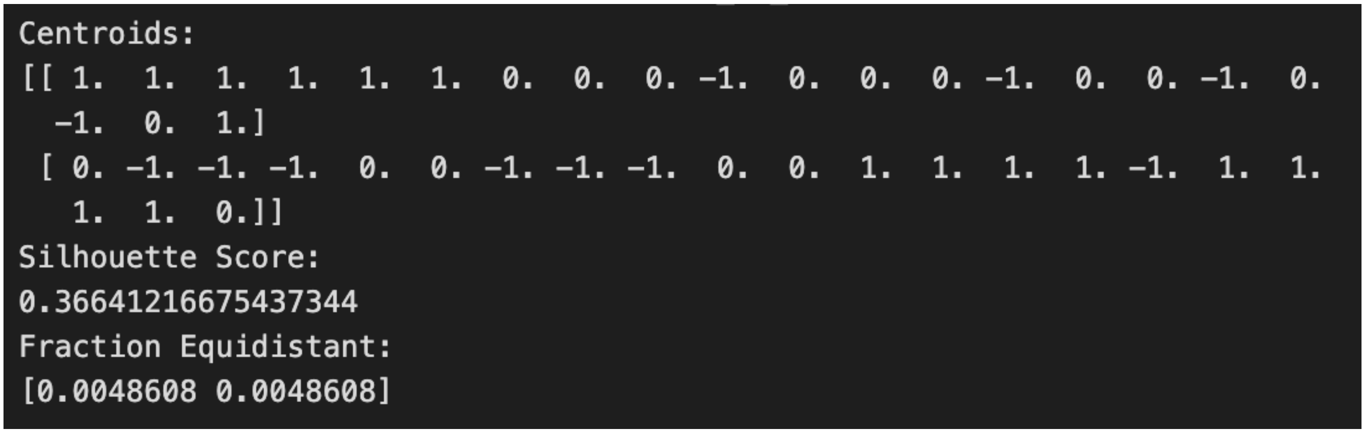

MDS
Points
Centroids
20
15
10
5
0
−5
−10
−15
−20
Second MDS dimension
−20
−15
−10
−5
0
5
10
15
First MDS dimension



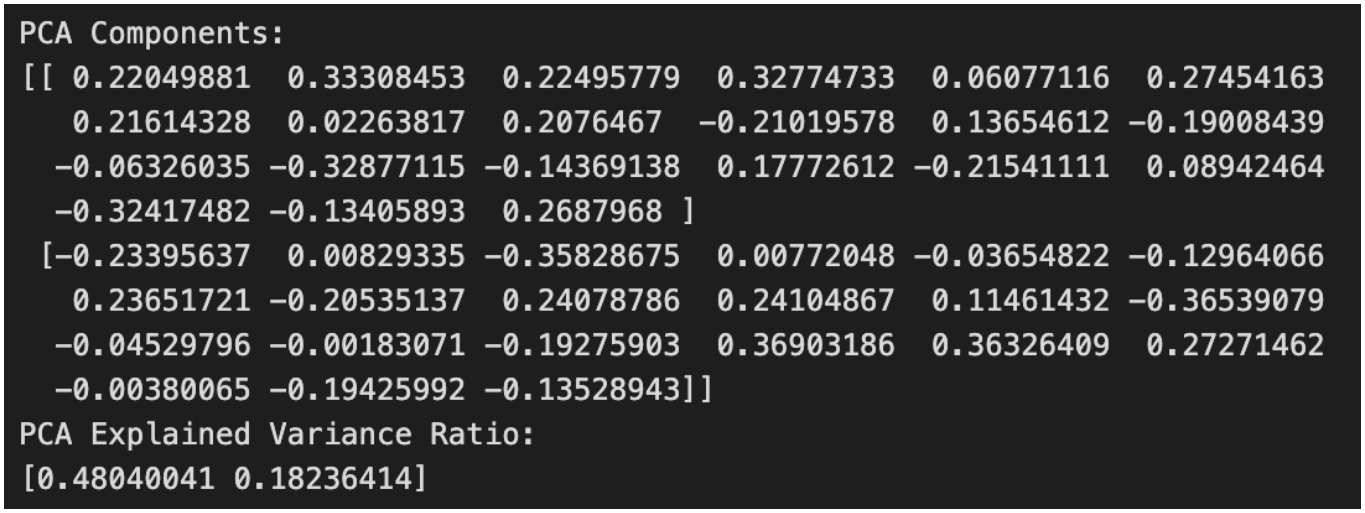


Restricted to Possible Ballots

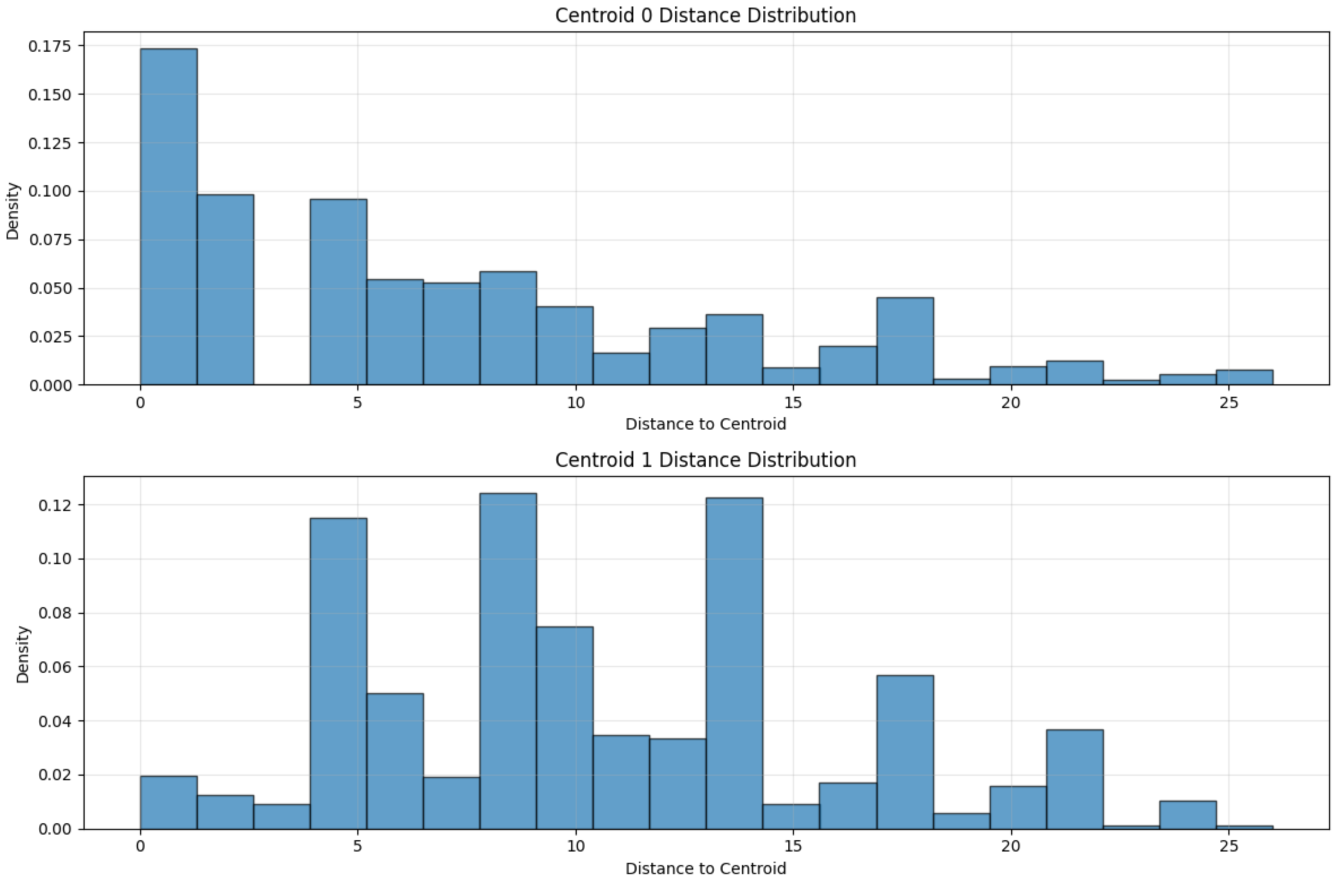
Restricted to Real Ballots



```
Centroids:
[[ 1.  1.  1.  1.  1.  1.  0.  0.  0. -1.  0.  0.  0. -1.  0.  0. -1.  0.
  -1.  0.  1.]
 [ 0. -1. -1. -1.  0.  0. -1. -1. -1.  0.  0.  1.  1.  1.  1. -1.  1.  1.
   1.  1.  0.]]
Silhouette Score:
0.36641216675437344
Fraction Equidistant:
[0.0048608 0.0048608]
```

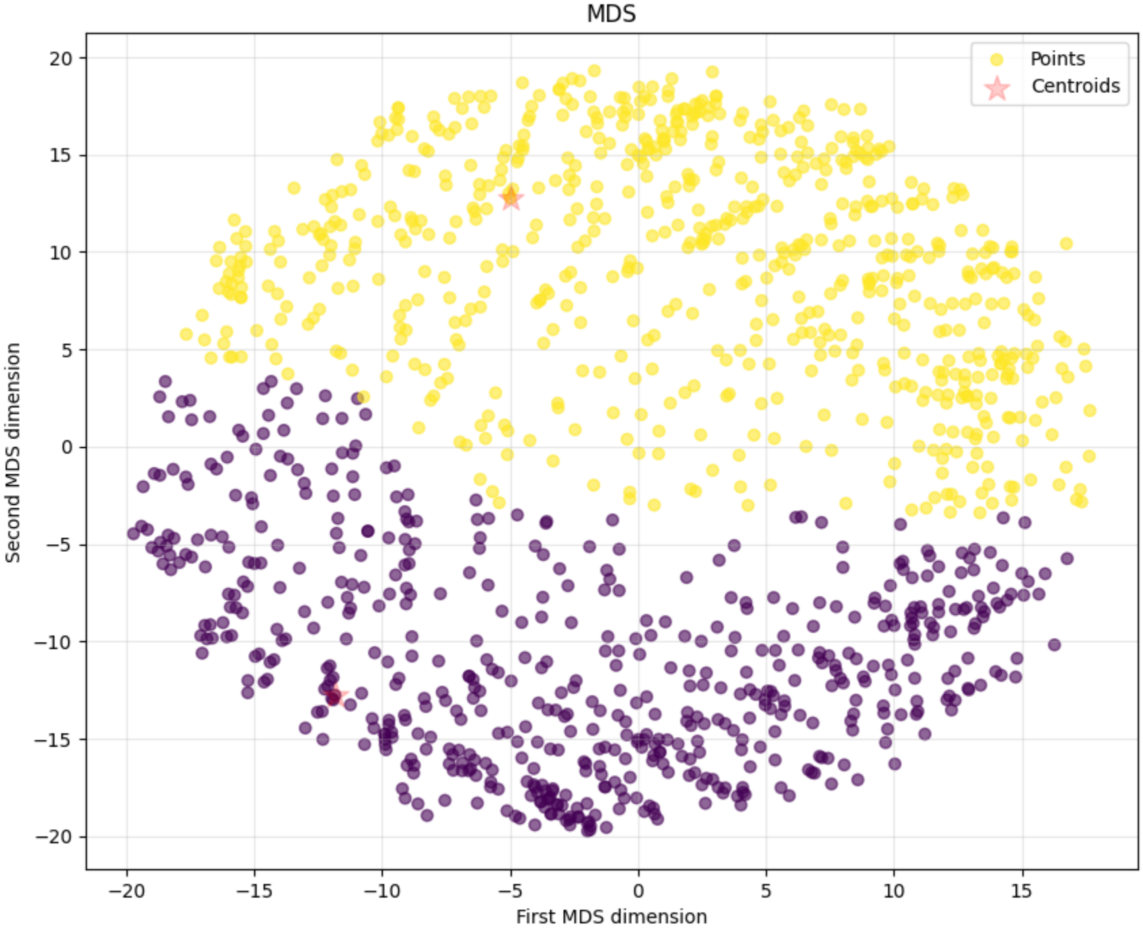


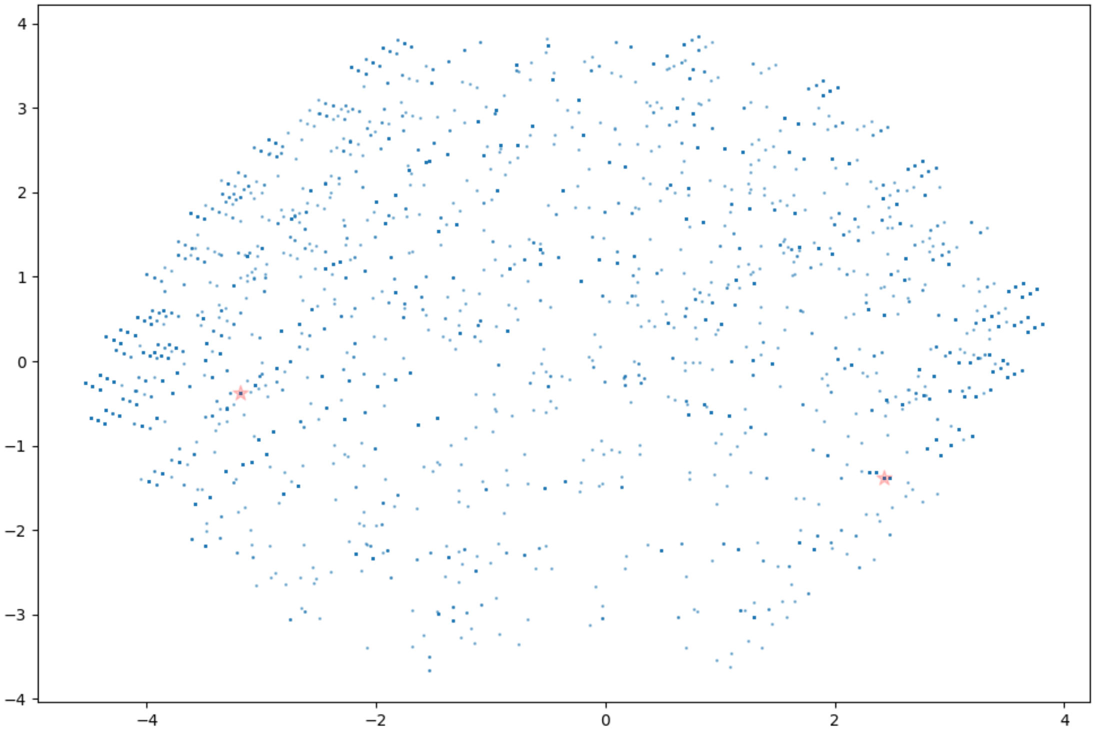
```
PCA Components:
[[ 0.22049881  0.33308453  0.22495779  0.32774733  0.06077116  0.27454163
   0.21614328  0.02263817  0.2076467  -0.21019578  0.13654612 -0.19008439
  -0.06326035 -0.32877115 -0.14369138  0.17772612 -0.21541111  0.08942464
  -0.32417482 -0.13405893  0.2687968 ]
 [-0.23395637  0.00829335 -0.35828675  0.00772048 -0.03654822 -0.12964066
   0.23651721 -0.20535137  0.24078786  0.24104867  0.11461432 -0.36539079
  -0.04529796 -0.00183071 -0.19275903  0.36903186  0.36326409  0.27271462
  -0.00380065 -0.19425992 -0.13528943]]
PCA Explained Variance Ratio:
[0.48040041 0.18236414]
```



Centroid 0 Distance Distribution

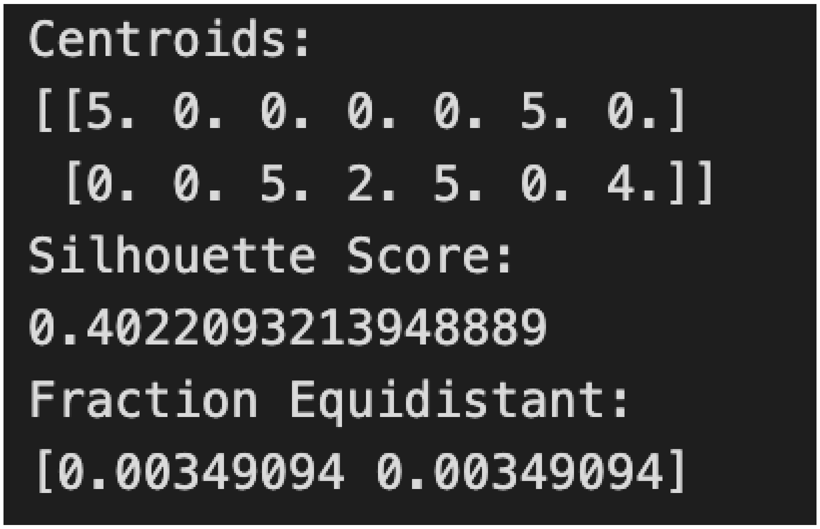
Centroid 1 Distance Distribution


MDS
Points
Centroids
Second MDS dimension
20
15
10
5
0
−5
−10
−15
−20
First MDS dimension
−20
−15
−10
−5
0
5
10
15

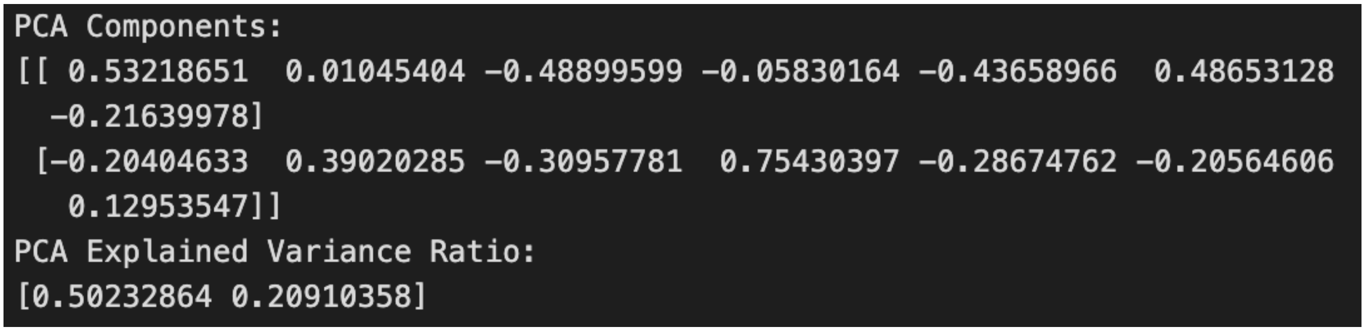


Pessimistic Borda

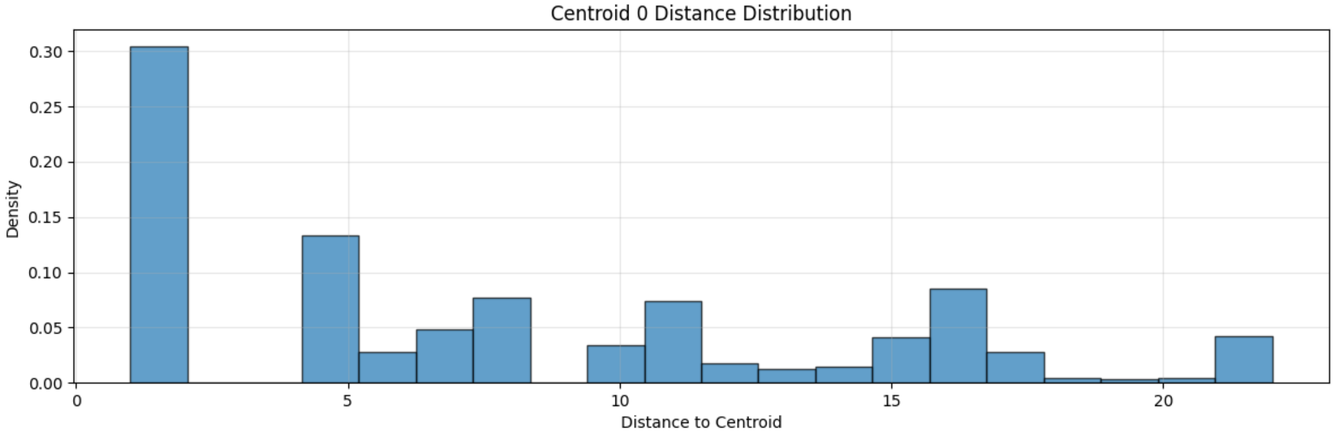
Continuous

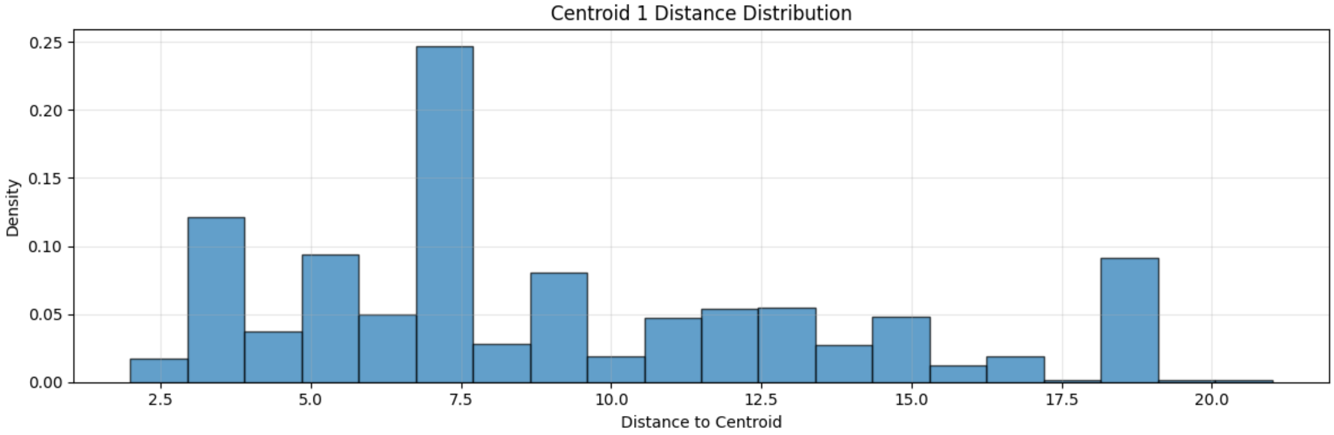


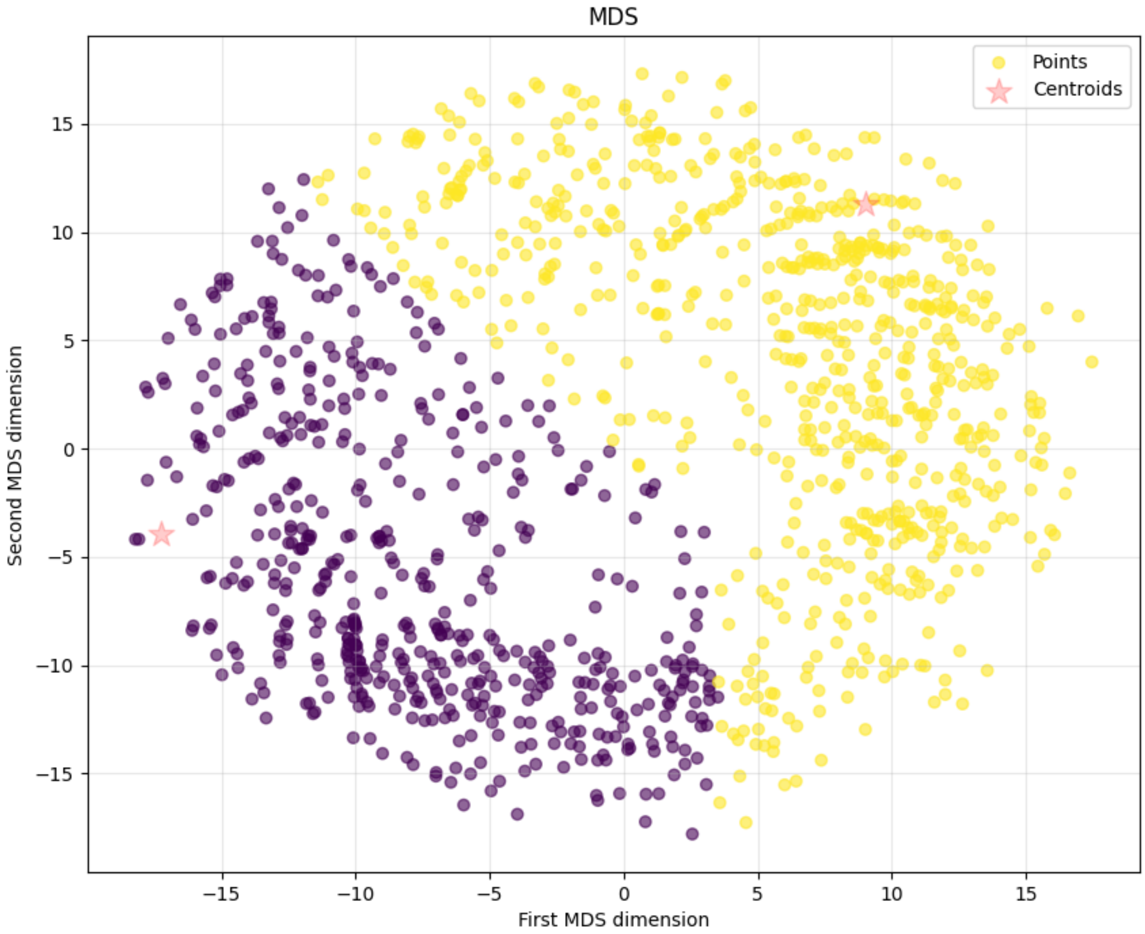
```
Centroids:
[[5. 0. 0. 0. 0. 5. 0.]
 [0. 0. 5. 2. 5. 0. 4.]]
Silhouette Score:
0.4022093213948889
Fraction Equidistant:
[0.00349094 0.00349094]
```

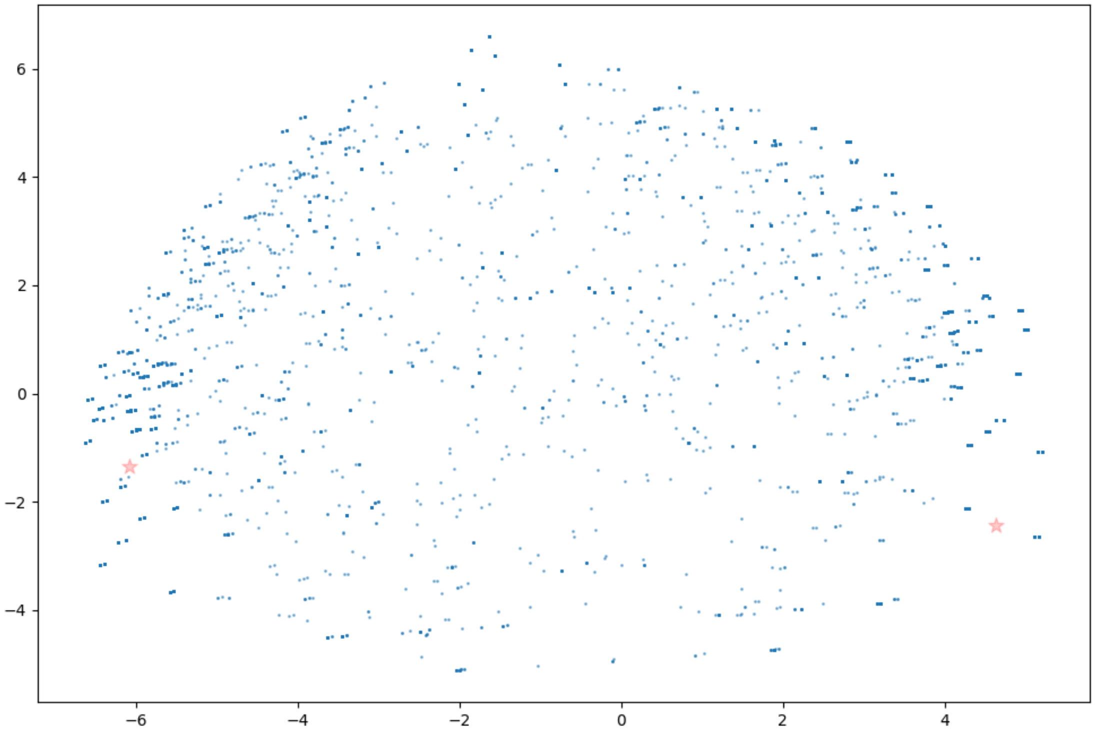


```
PCA Components:
[[ 0.53218651  0.01045404 -0.48899599 -0.05830164 -0.43658966  0.48653128
  -0.21639978]
 [-0.20404633  0.39020285 -0.30957781  0.75430397 -0.28674762 -0.20564606
   0.12953547]]
PCA Explained Variance Ratio:
[0.50232864 0.20910358]
```

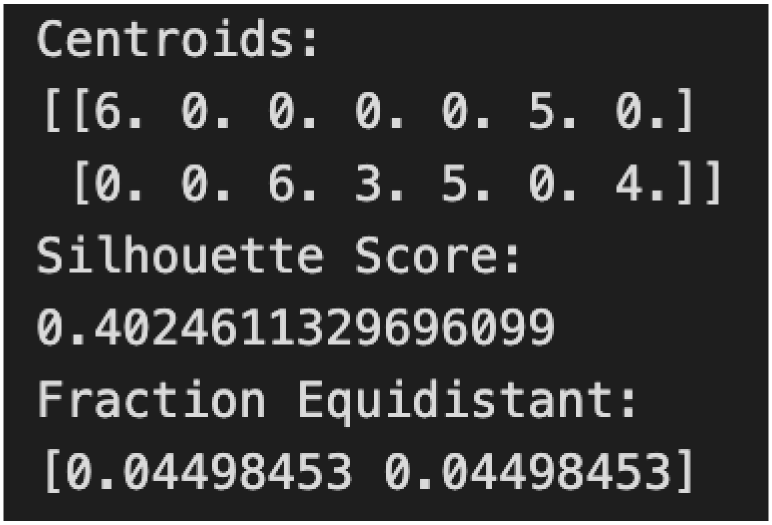




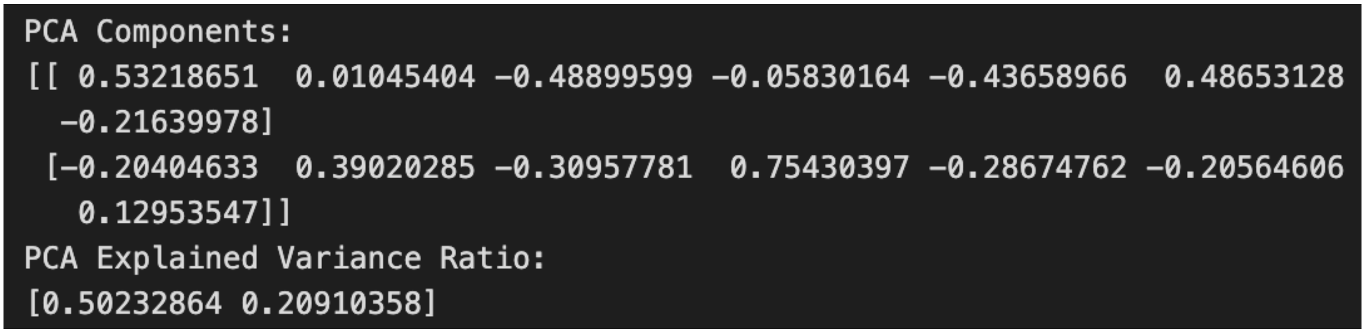

MDS
Points
Centroids
Second MDS dimension
First MDS dimension
15
10
5
0
−5
−10
−15
−15
−10
−5
0
5
10
15



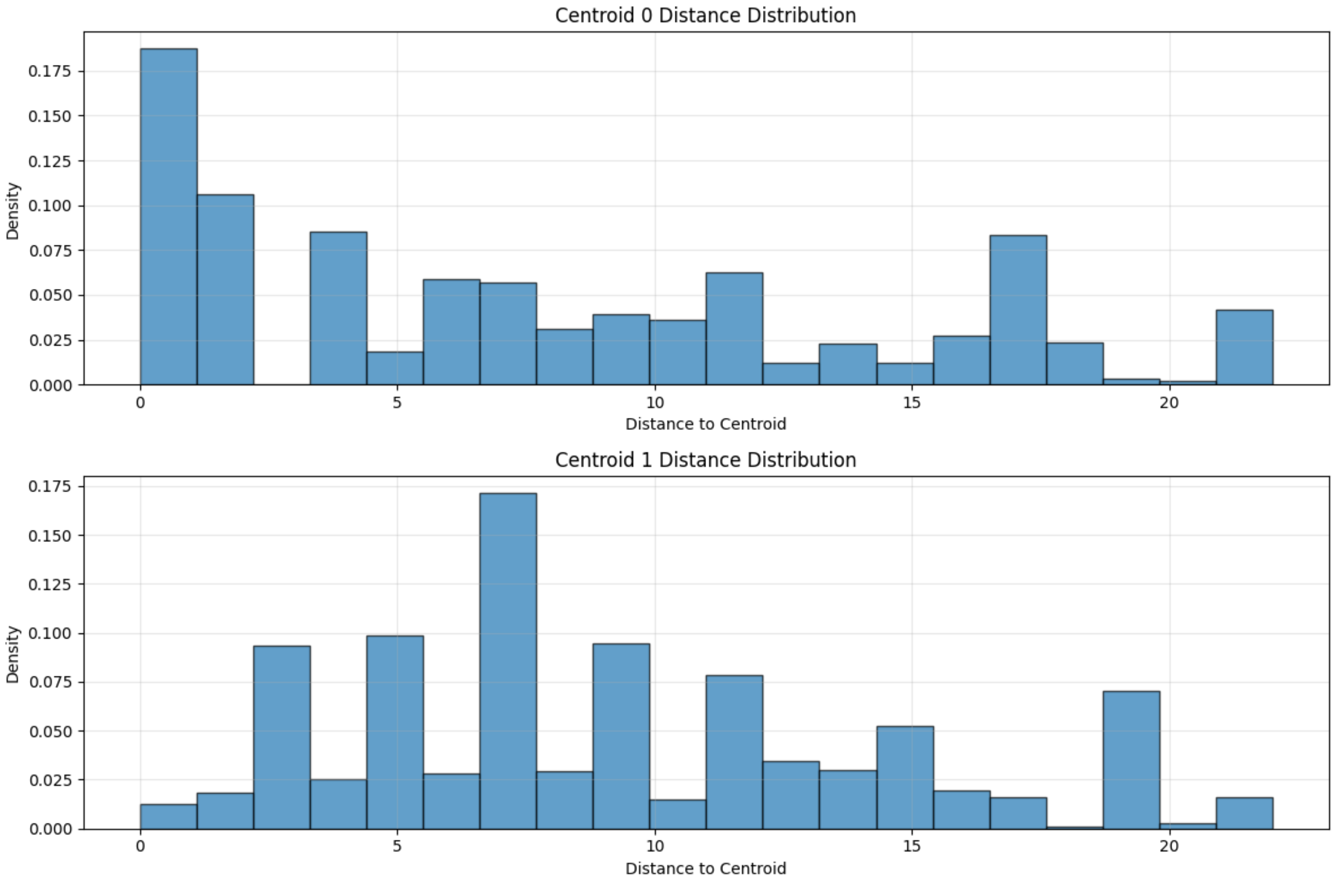
Restricted to Possible Ballots



```
Centroids:
[[6. 0. 0. 0. 0. 5. 0.]
 [0. 0. 6. 3. 5. 0. 4.]]
Silhouette Score:
0.4024611329696099
Fraction Equidistant:
[0.04498453 0.04498453]
```

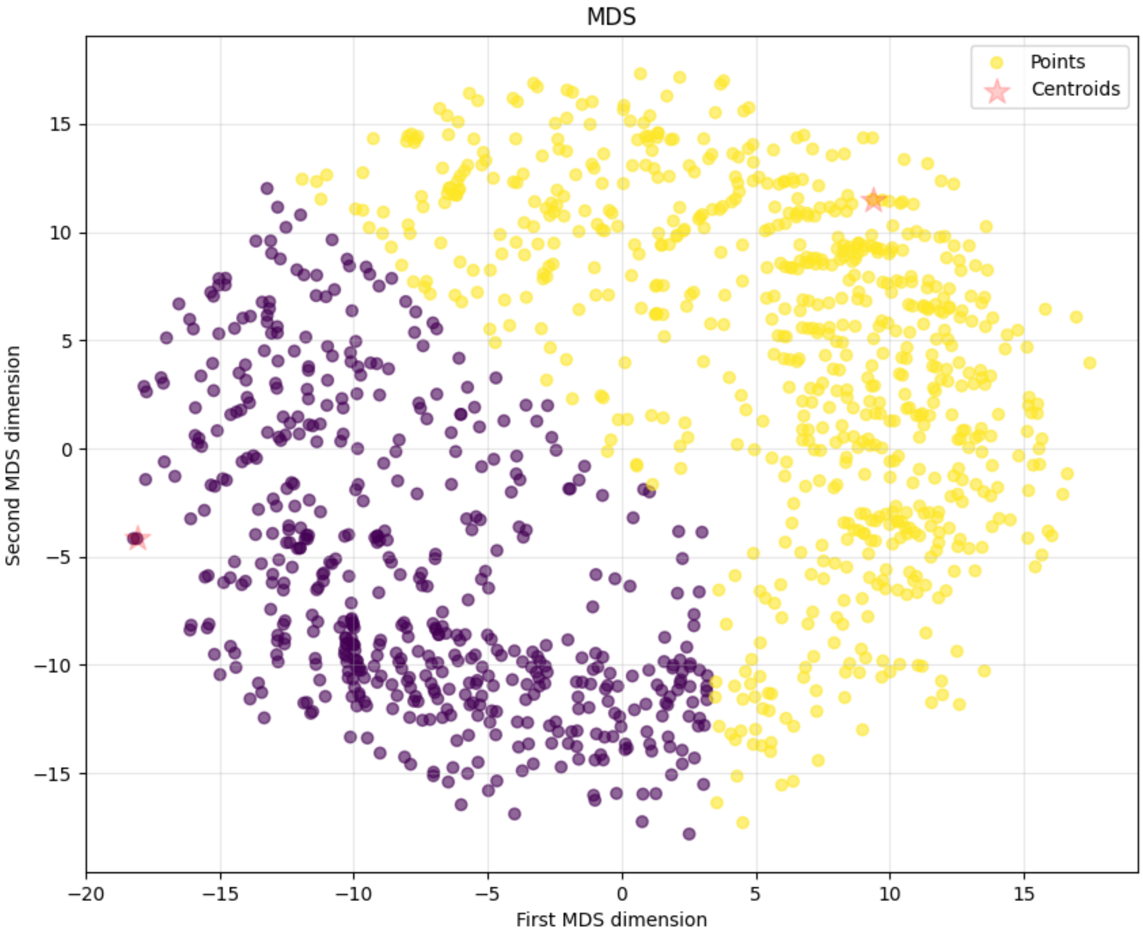


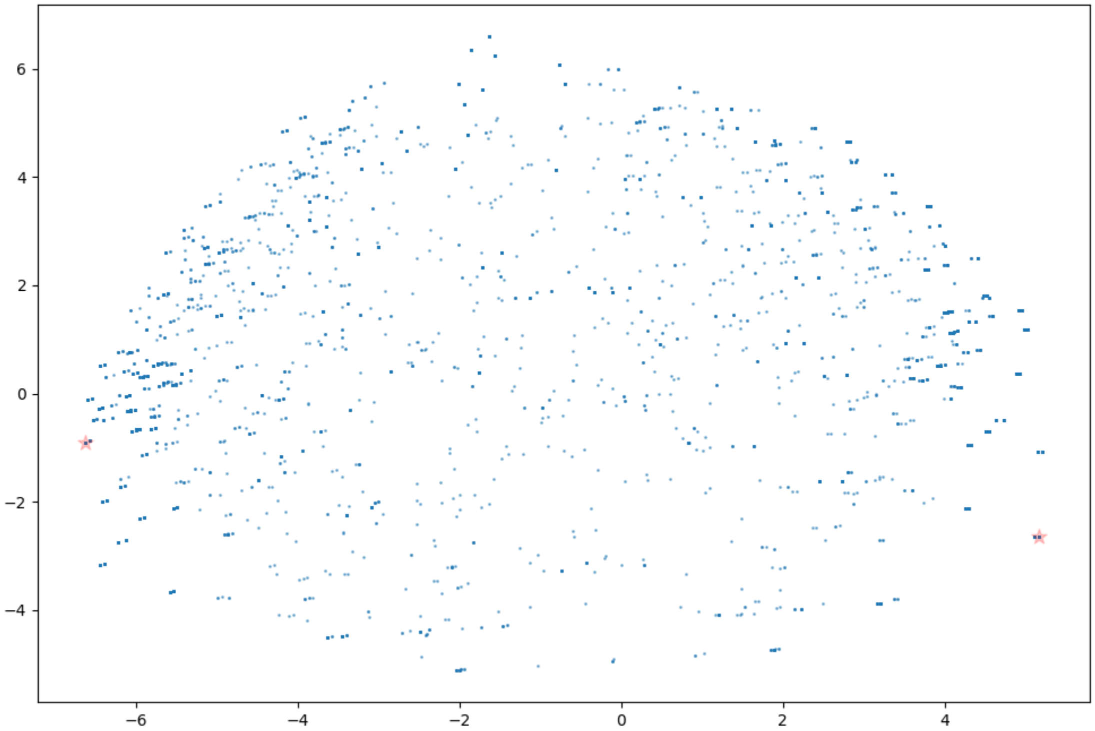
```
PCA Components:
[[ 0.53218651  0.01045404 -0.48899599 -0.05830164 -0.43658966  0.48653128
  -0.21639978]
 [-0.20404633  0.39020285 -0.30957781  0.75430397 -0.28674762 -0.20564606
   0.12953547]]
PCA Explained Variance Ratio:
[0.50232864 0.20910358]
```



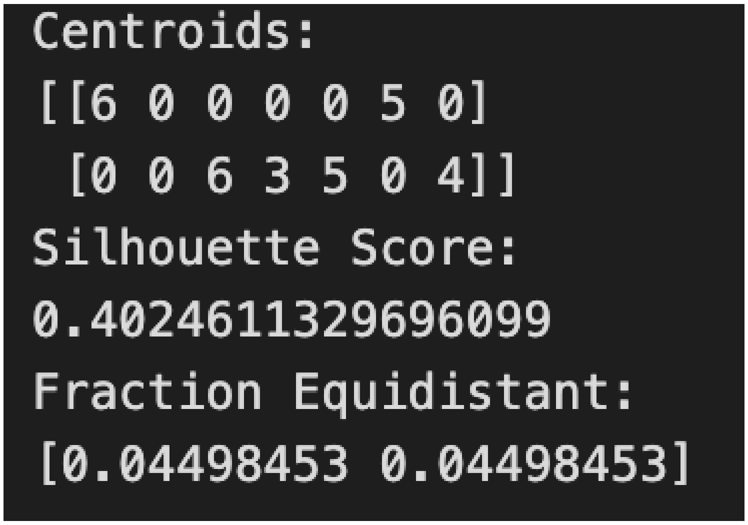
Centroid 0 Distance Distribution

Centroid 1 Distance Distribution

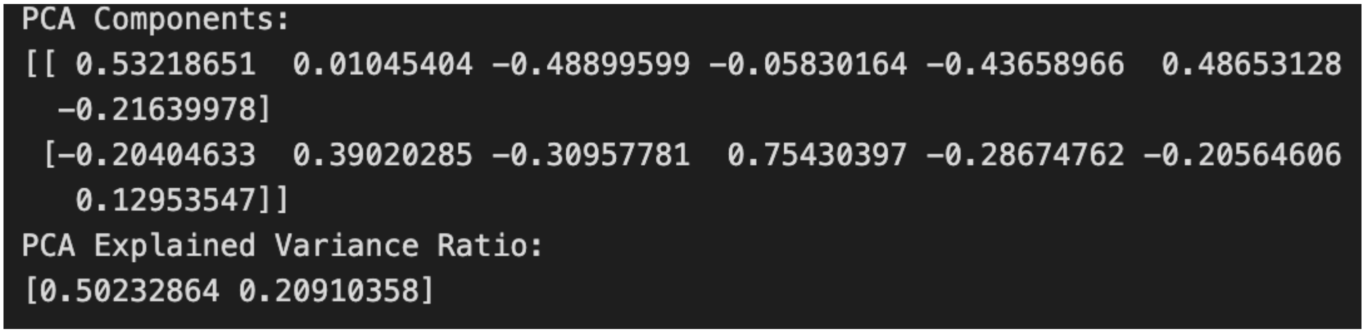

MDS
Points
Centroids
Second MDS dimension
First MDS dimension
15
10
5
0
−5
−10
−15
−20
−15
−10
−5
0
5
10
15



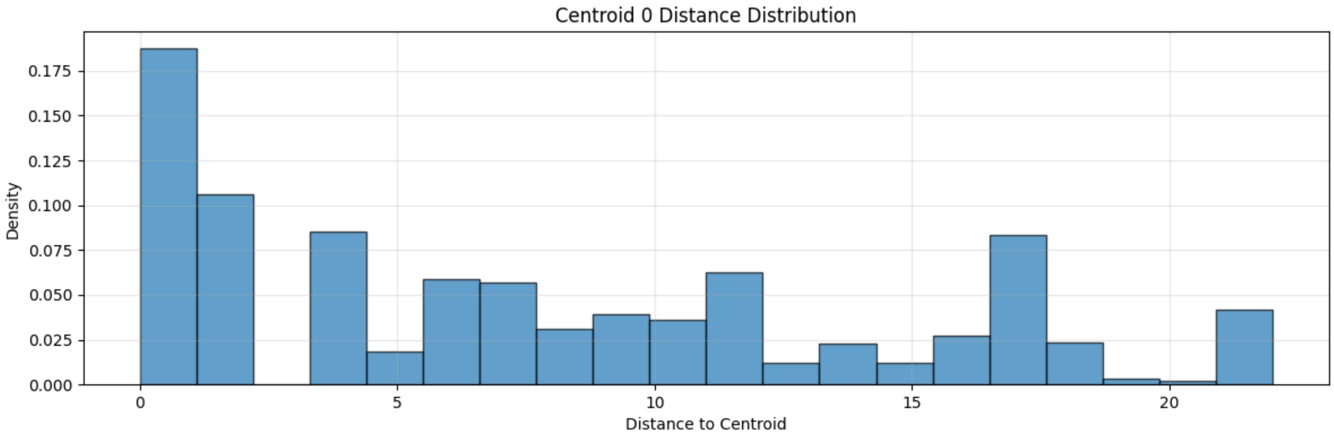
Restricted to Real Ballots

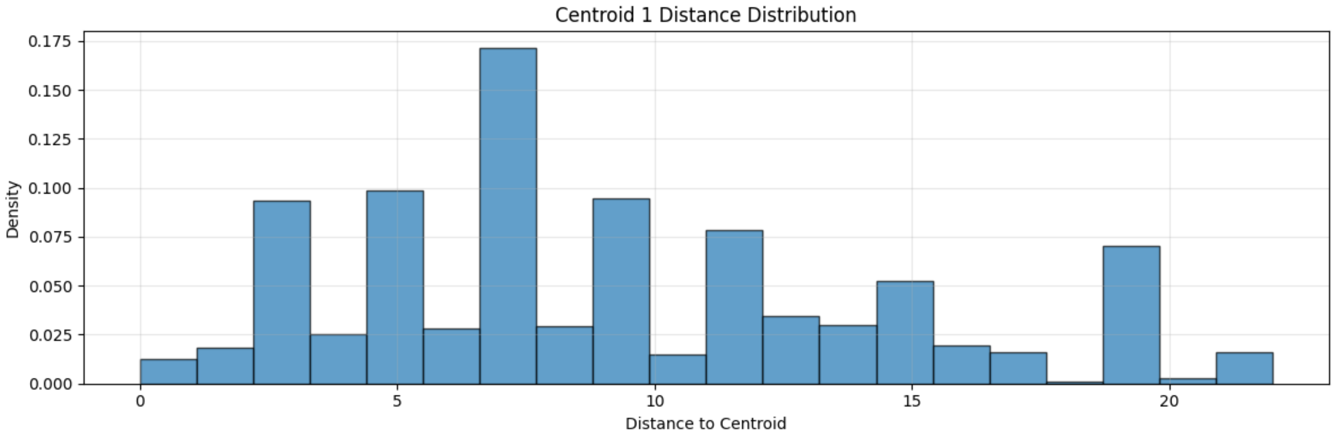


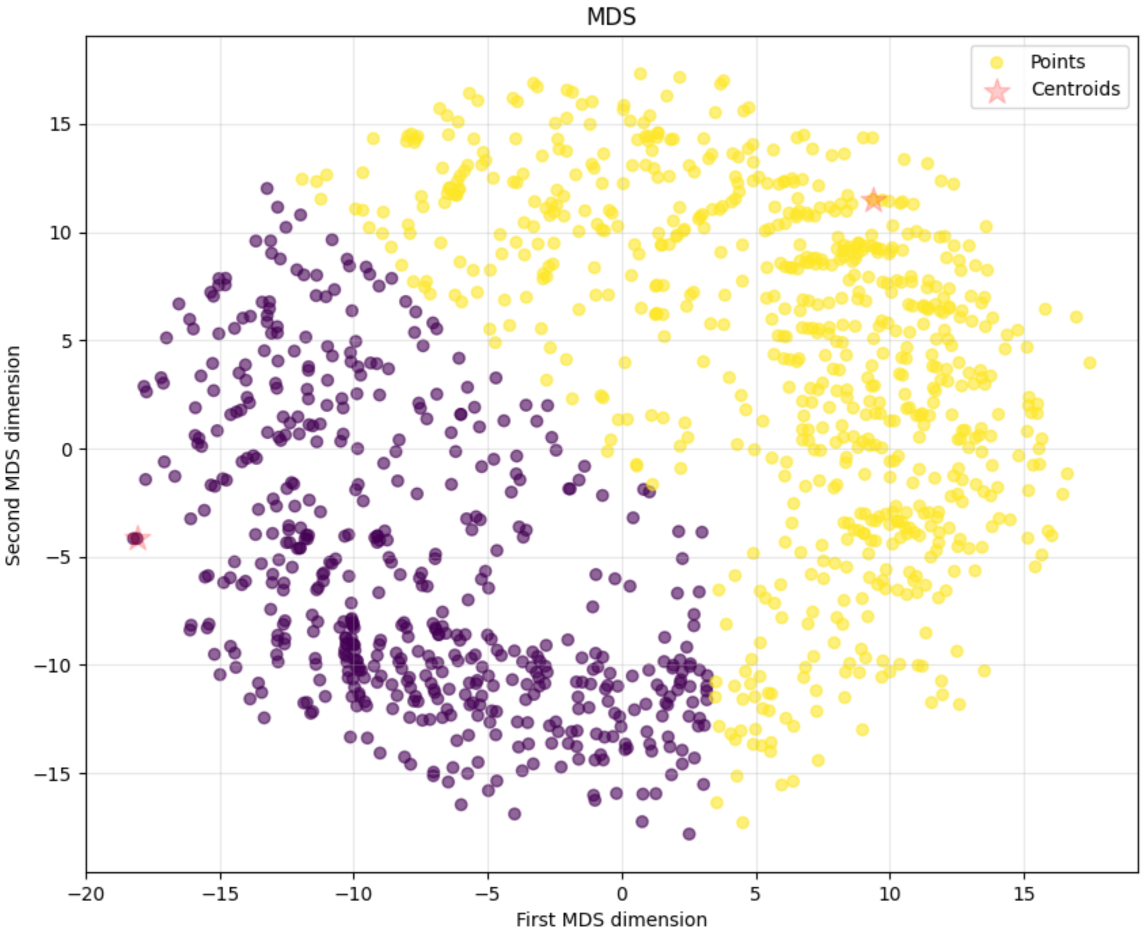
```
Centroids:
[[6 0 0 0 0 5 0]
 [0 0 6 3 5 0 4]]
Silhouette Score:
0.4024611329696099
Fraction Equidistant:
[0.04498453 0.04498453]
```

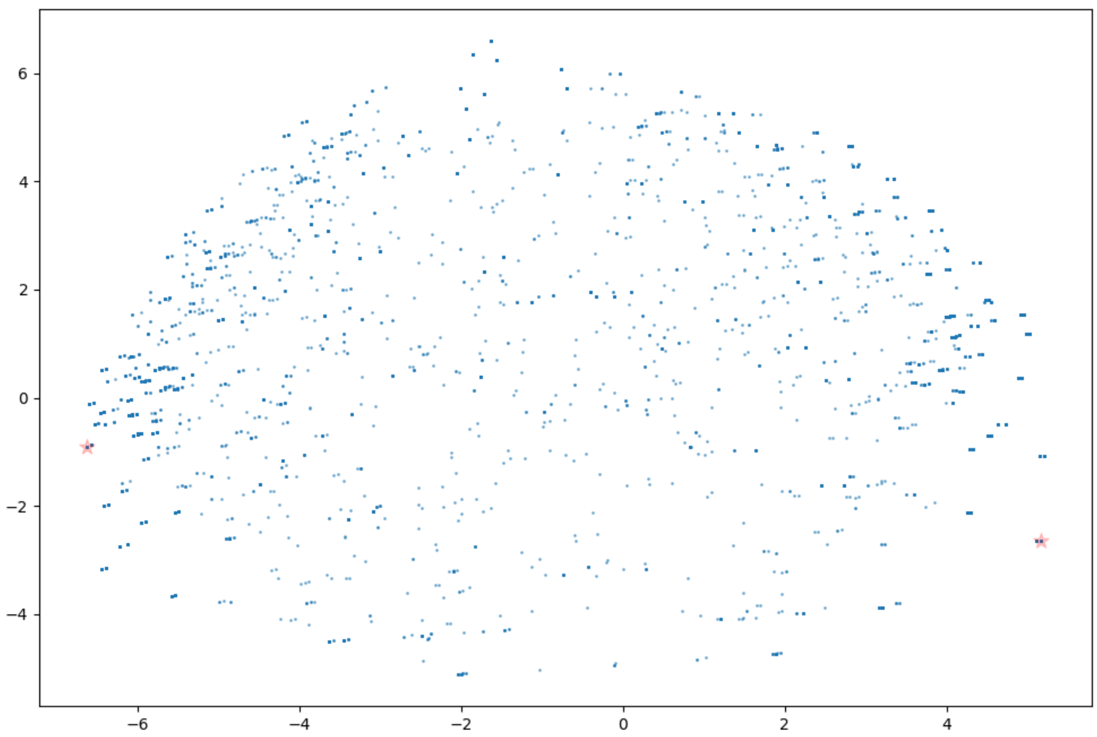


```
PCA Components:
[[ 0.53218651  0.01045404 -0.48899599 -0.05830164 -0.43658966  0.48653128
  -0.21639978]
 [-0.20404633  0.39020285 -0.30957781  0.75430397 -0.28674762 -0.20564606
   0.12953547]]
PCA Explained Variance Ratio:
[0.50232864 0.20910358]
```


Centroid 0 Distance Distribution
Density
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0
5
10
15
20
Distance to Centroid

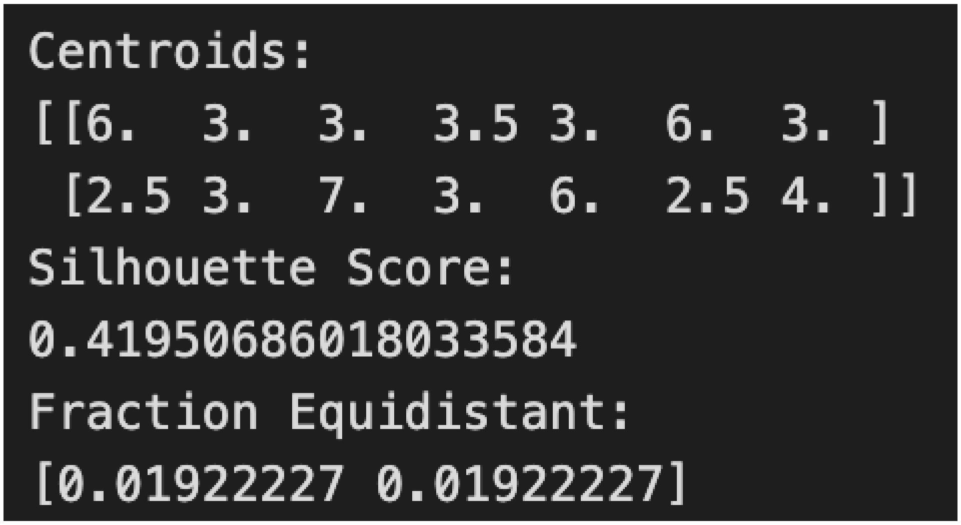

Centroid 1 Distance Distribution
Density
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0
5
10
15
20
Distance to Centroid


MDS
Points
Centroids
15
10
5
0
−5
−10
−15
−20
−15
−10
−5
0
5
10
15
Second MDS dimension
First MDS dimension

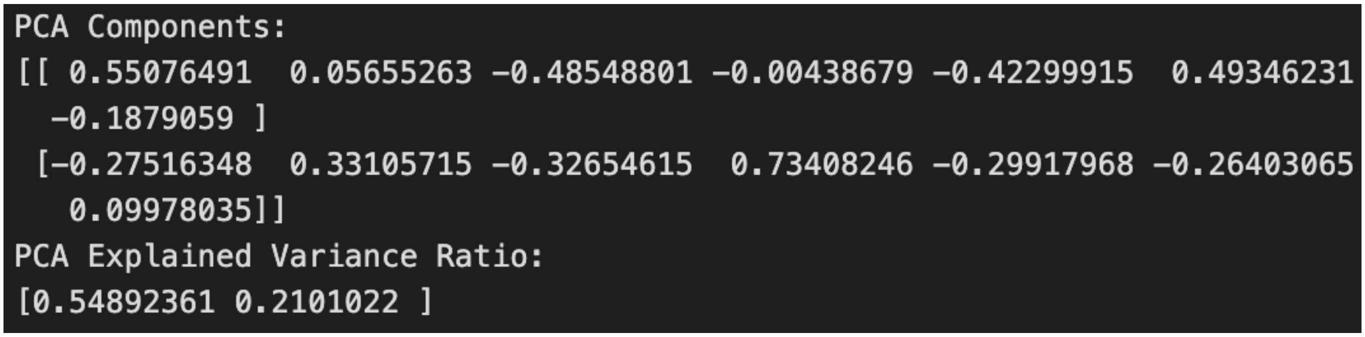


Average Borda

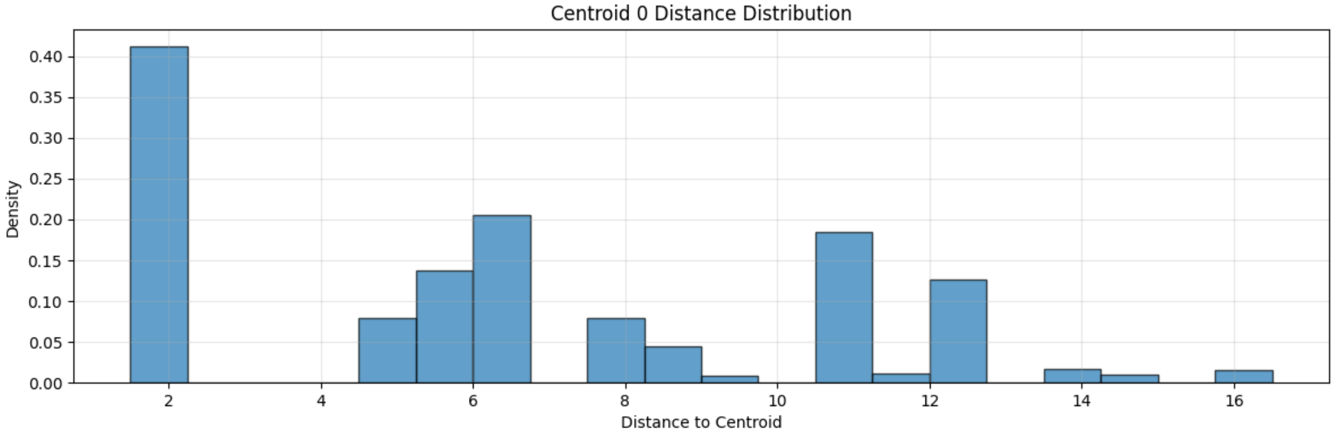
Continuous



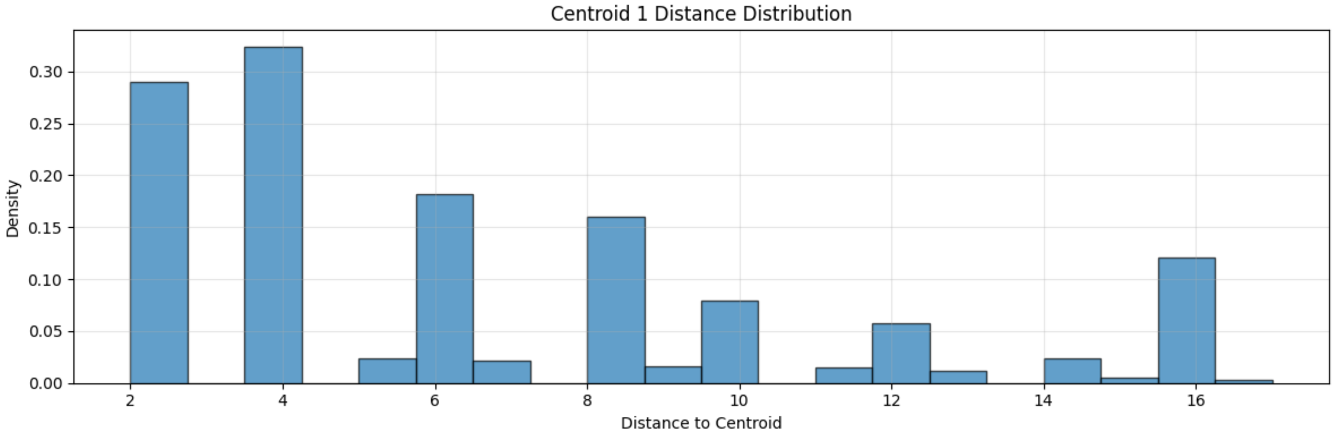
```
Centroids:
[[6.  3.  3.  3.5 3.  6.  3. ]
 [2.5 3.  7.  3.  6.  2.5 4. ]]
Silhouette Score:
0.41950686018033584
Fraction Equidistant:
[0.01922227 0.01922227]
```



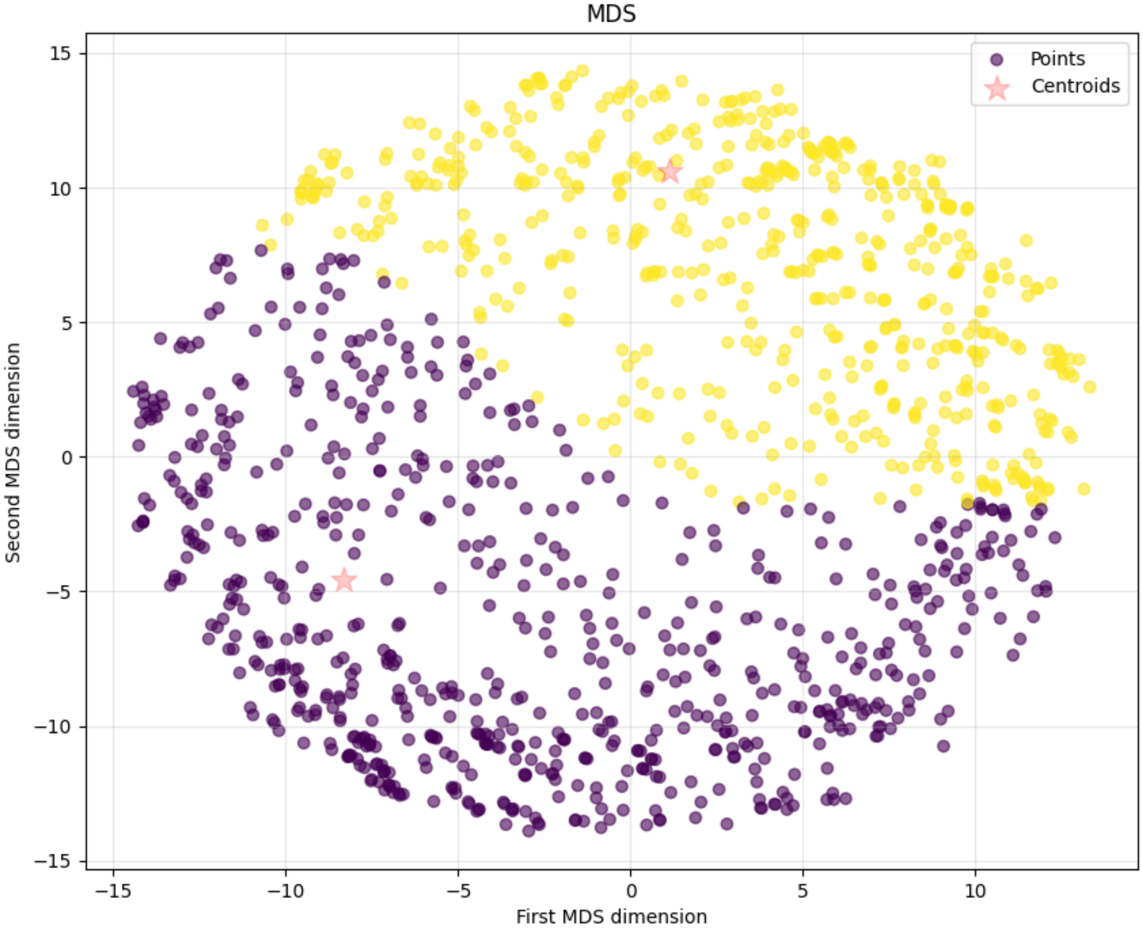
```
PCA Components:
[[ 0.55076491  0.05655263 -0.48548801 -0.00438679 -0.42299915  0.49346231
  -0.1879059 ]
 [-0.27516348  0.33105715 -0.32654615  0.73408246 -0.29917968 -0.26403065
   0.09978035]]
PCA Explained Variance Ratio:
[0.54892361 0.2101022 ]
```

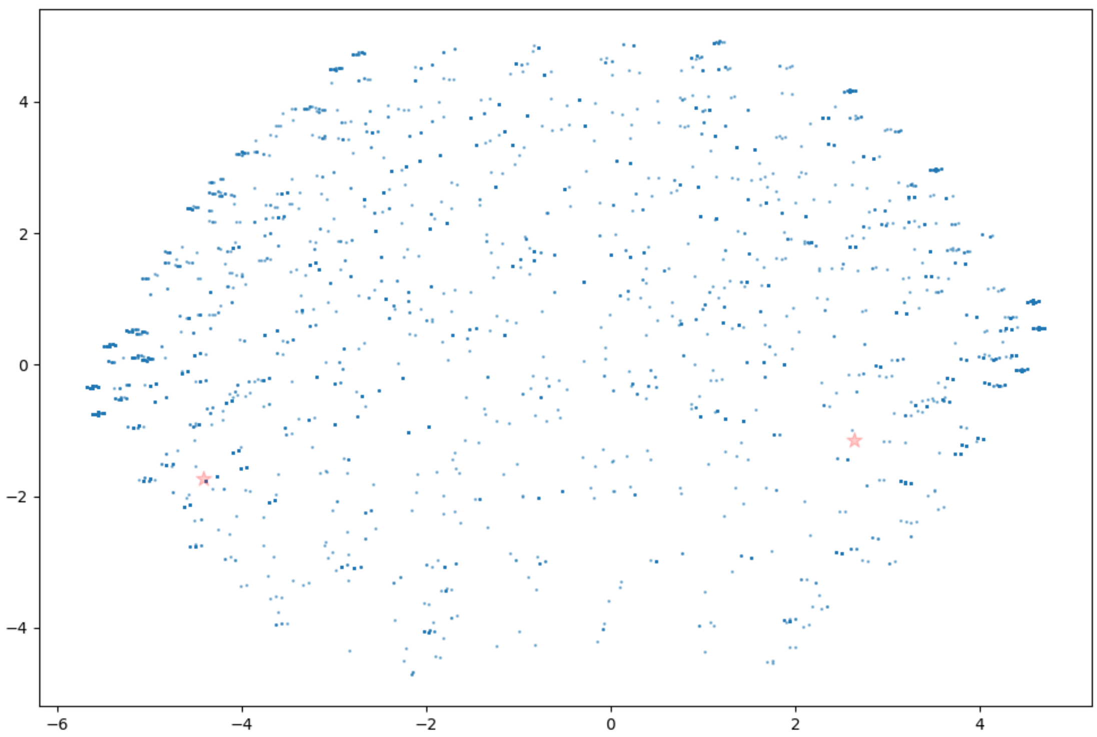


Centroid 0 Distance Distribution

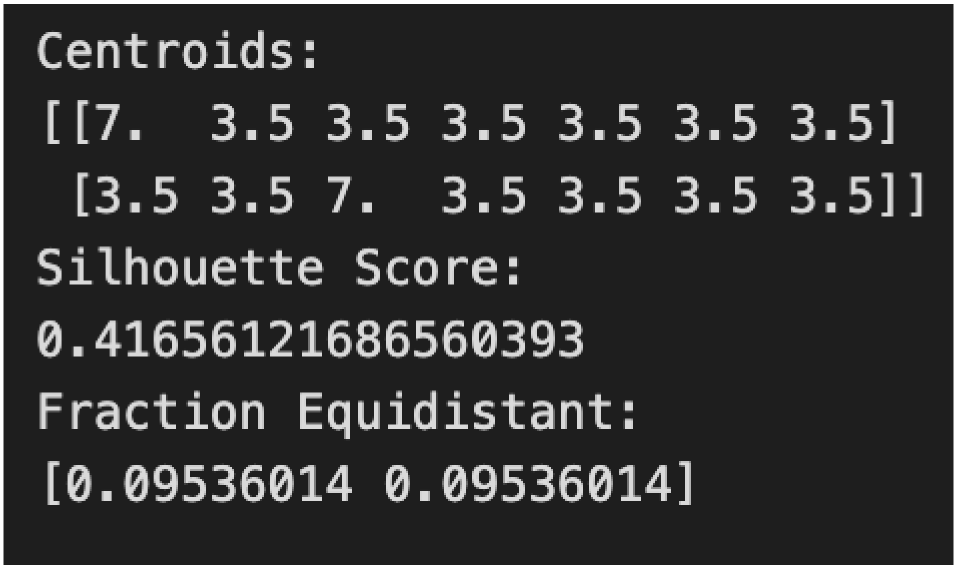


Centroid 1 Distance Distribution

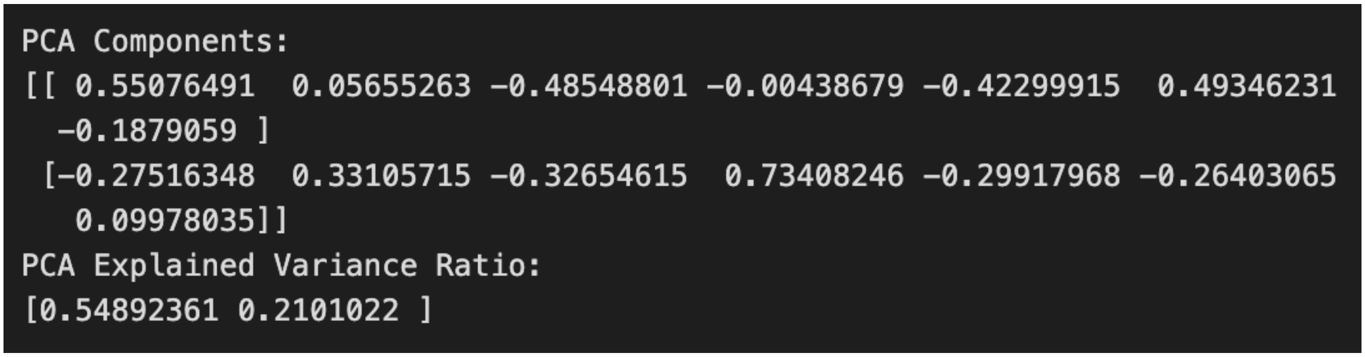

MDS
Points
Centroids
First MDS dimension
Second MDS dimension
−15
−10
−5
0
5
10
15
−15
−10
−5
0
5
10



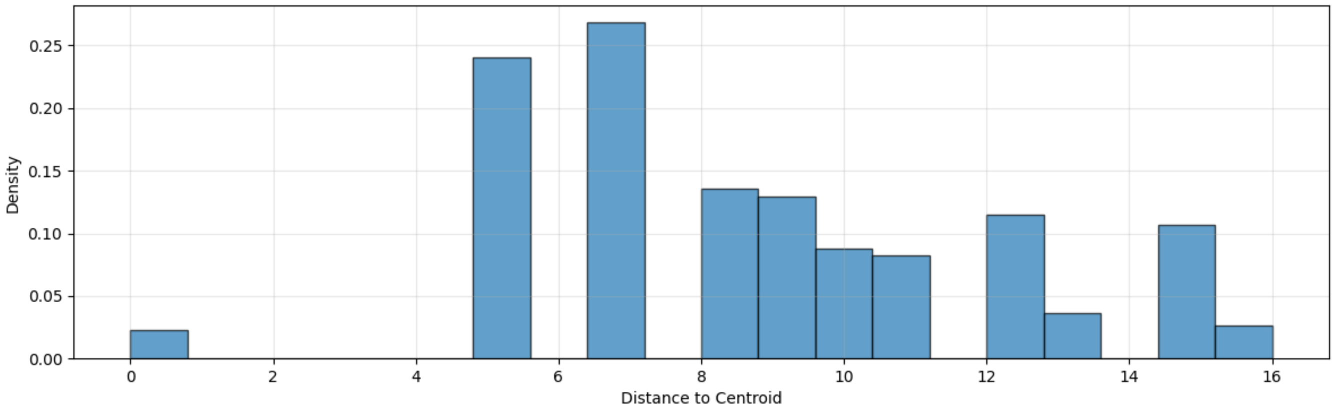
Restricted to Possible Ballots



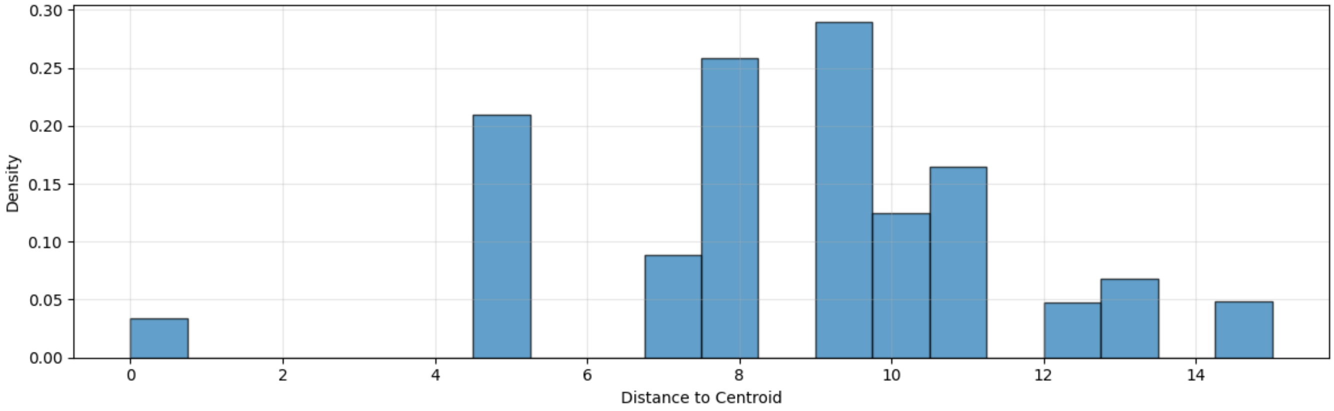
```
Centroids:
[[7.  3.5 3.5 3.5 3.5 3.5 3.5]
 [3.5 3.5 7.  3.5 3.5 3.5 3.5]]
Silhouette Score:
0.41656121686560393
Fraction Equidistant:
[0.09536014 0.09536014]
```



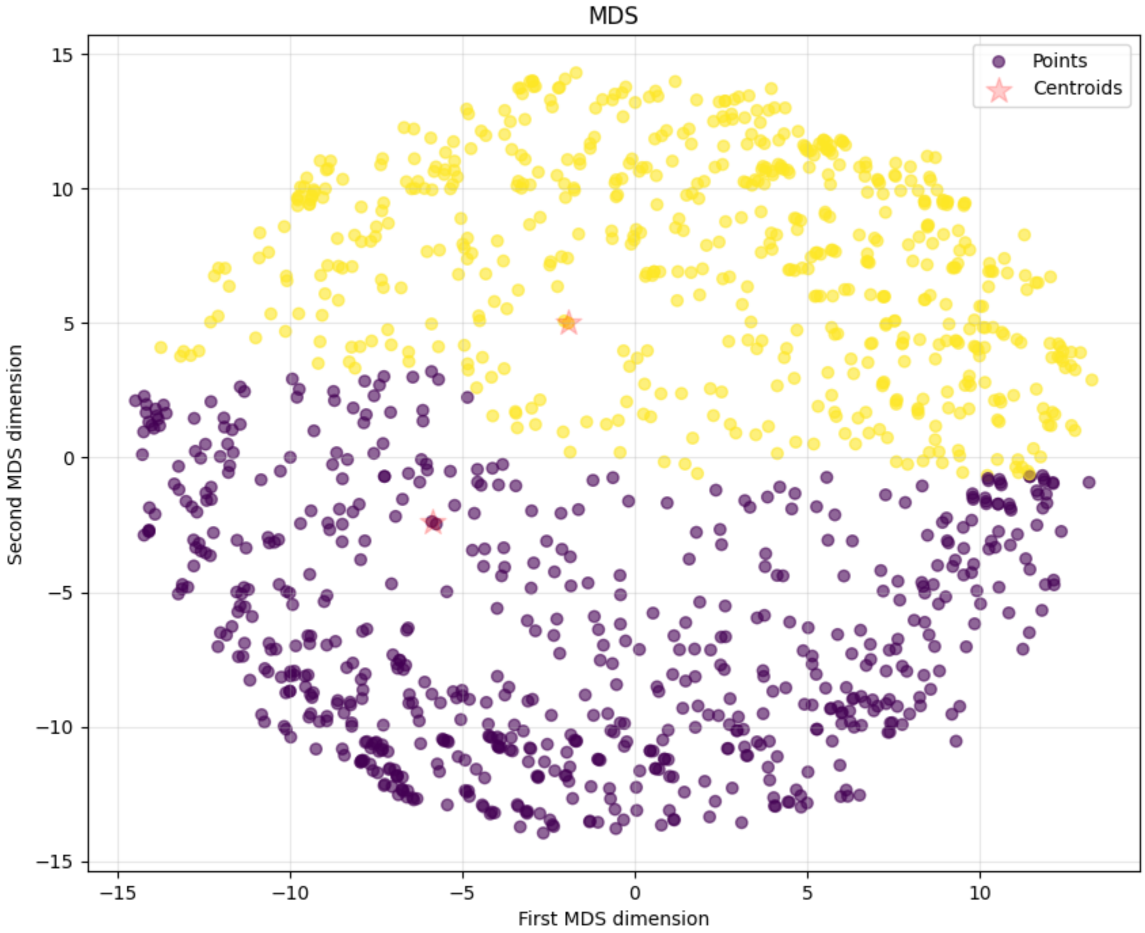
```
PCA Components:
[[ 0.55076491  0.05655263 -0.48548801 -0.00438679 -0.42299915  0.49346231
  -0.1879059 ]
 [-0.27516348  0.33105715 -0.32654615  0.73408246 -0.29917968 -0.26403065
   0.09978035]]
PCA Explained Variance Ratio:
[0.54892361 0.2101022 ]
```

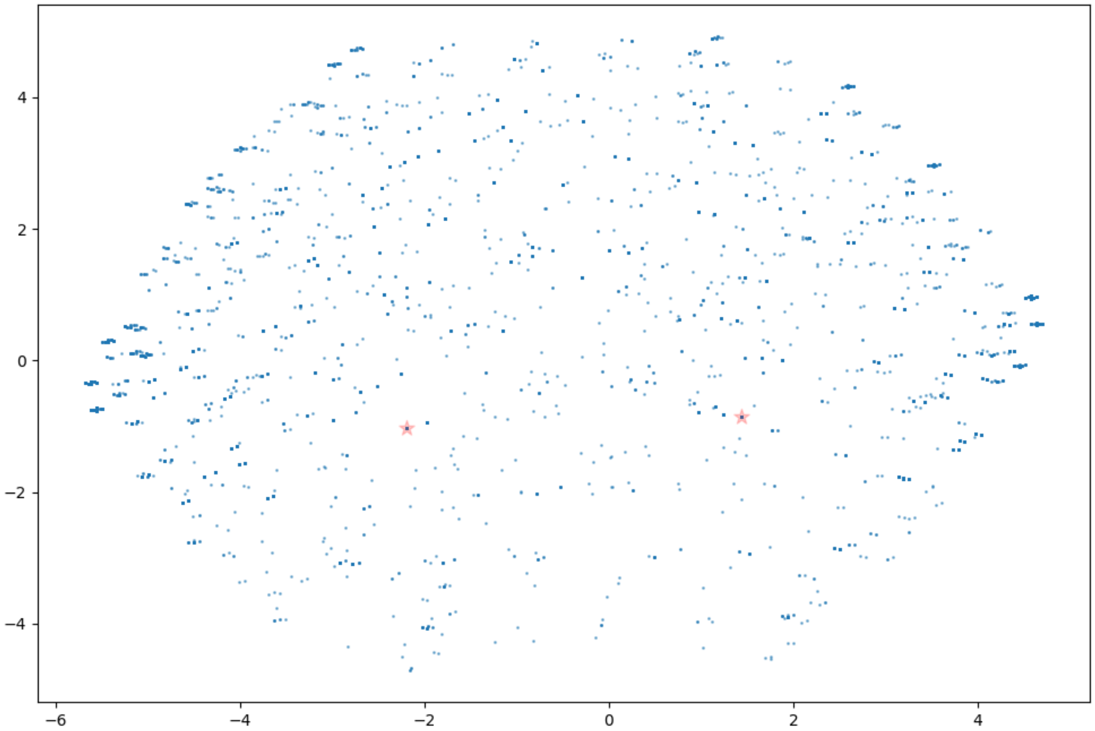


Centroid 0 Distance Distribution

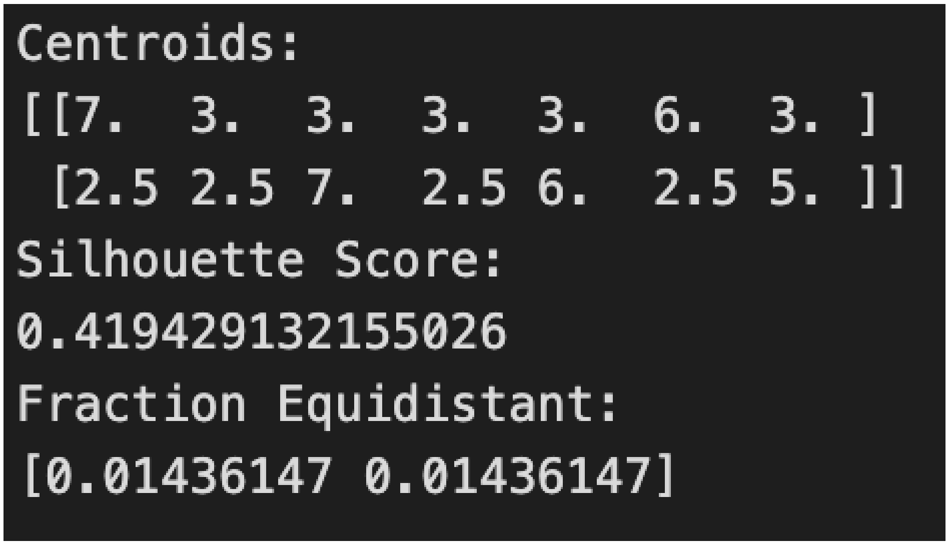


Centroid 1 Distance Distribution

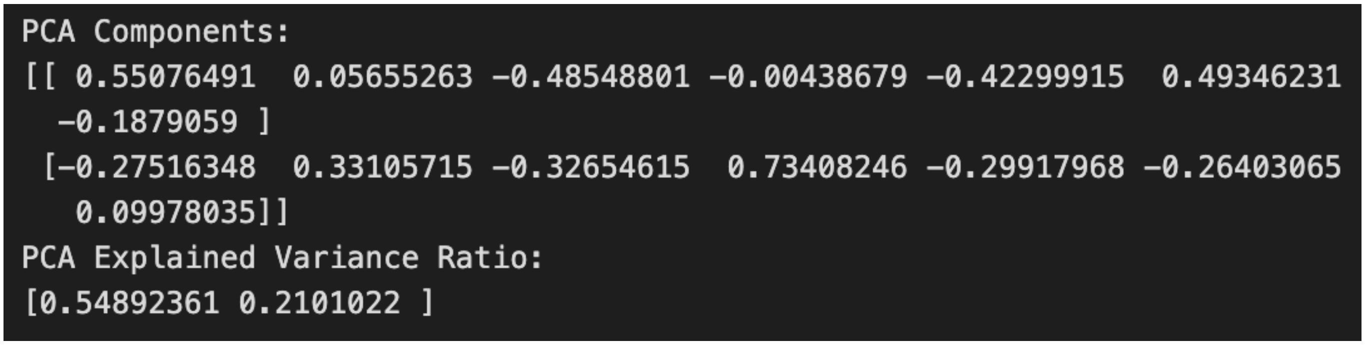

MDS
Points
Centroids
Second MDS dimension
First MDS dimension
15
10
5
0
−5
−10
−15
−15
−10
−5
0
5
10



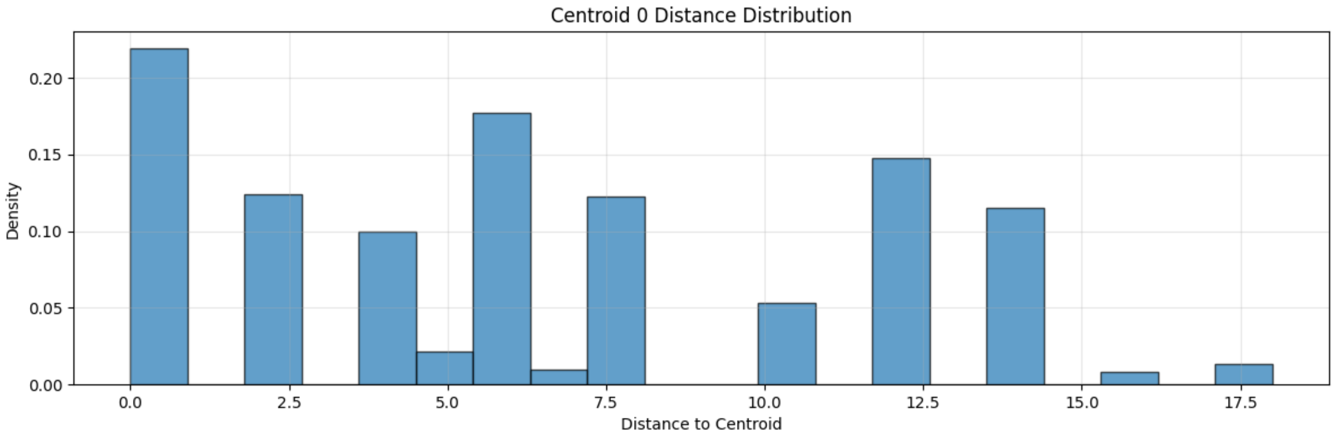
Restricted to Real Ballots



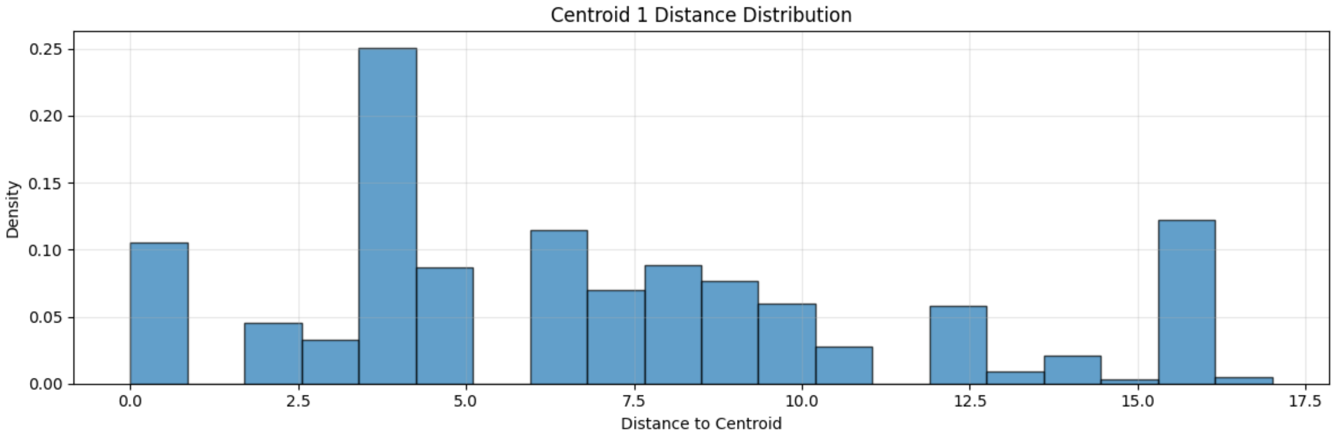
```
Centroids:
[[7.  3.  3.  3.  3.  6.  3. ]
 [2.5 2.5 7.  2.5 6.  2.5 5. ]]
Silhouette Score:
0.419429132155026
Fraction Equidistant:
[0.01436147 0.01436147]
```



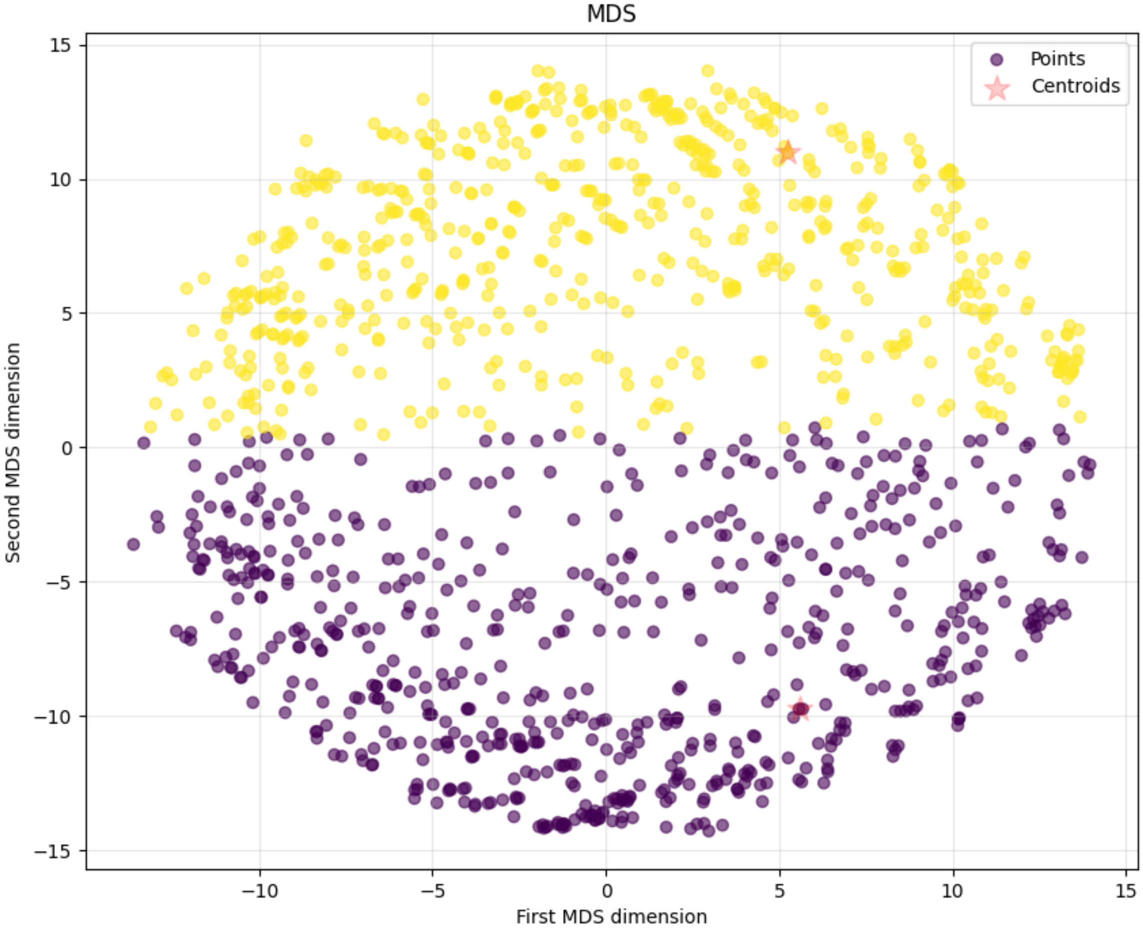
```
PCA Components:
[[ 0.55076491  0.05655263 -0.48548801 -0.00438679 -0.42299915  0.49346231
  -0.1879059 ]
 [-0.27516348  0.33105715 -0.32654615  0.73408246 -0.29917968 -0.26403065
   0.09978035]]
PCA Explained Variance Ratio:
[0.54892361 0.2101022 ]
```

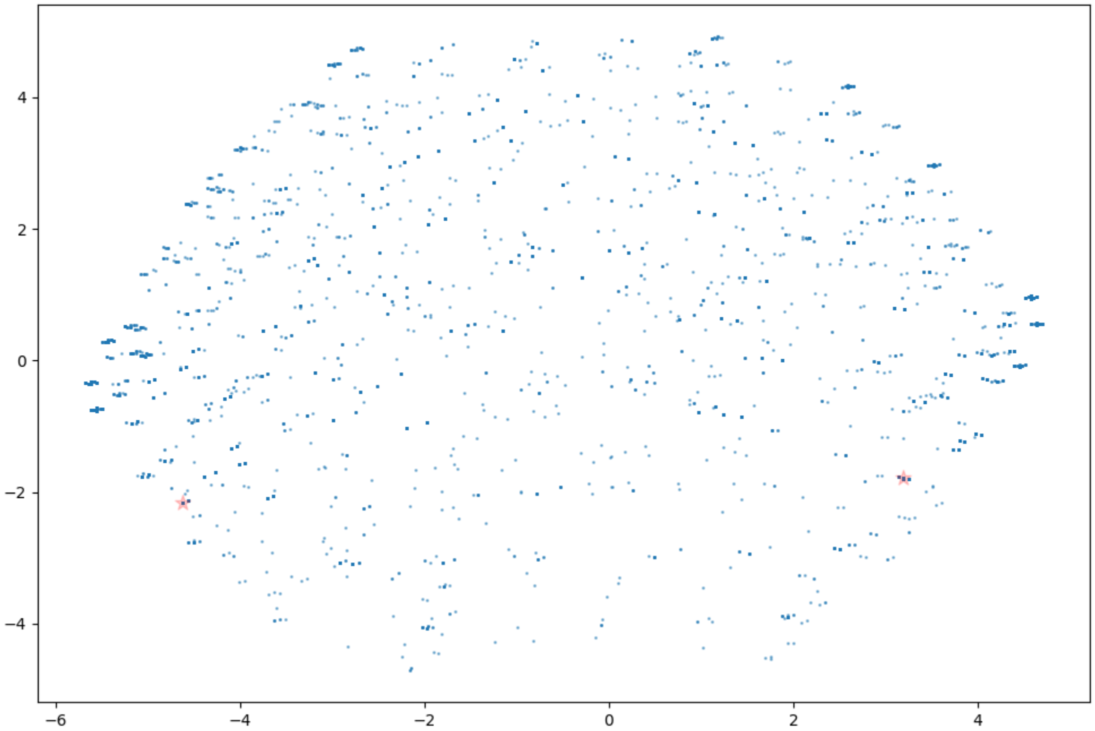


Centroid 0 Distance Distribution



Centroid 1 Distance Distribution


MDS
Points
Centroids
15
10
5
0
−5
−10
−15
−10
−5
0
5
10
15
First MDS dimension
Second MDS dimension


4
2
0
−2
−4
−6
−4
−2
0
2
4