clustering

June 30, 2021

Table of Contents

1 Clustering

Initially we looked at correlation between features and noticed that the largest ones (in terms of magnitude) appeared amongst the **opinion** question. So the question arose > Can we group participants into two classes based on their political opinion?

1.1 Metrics

Unsupervised learning is not so easy. How do we know that our clustering makes sense? These are some possible metrics to compare unsupervised learning methods.

- Silhouette score $[-1; 1]$, the closer to 1 - the better;
- Calinski-Harabasz score $[0; \infty]$, the larger - the better;
- Davies Bouldin score $[0, 1]$, the closer to 0 - the better.

1.2 Compare Methods

Lets compare standard methods according to the above scores. Spoiler alert: *Kmeans* performed the best.

1.2.1 Preparation

Check question First, we begin by removing the participants who answered the *check questions* incorrectly. For simplicity we exclude those people completely (meaning for all waves) even if they did answer correctly at a later wave.

The two *check questions* are `w2_q24x5` and `w1_q27x5`, meaning that such questions were only asked in wave 1 and 2.

```
The number of people who answered one of the check questions incorrectly is 953.
```

To be precise we are counting participants who answered both questions incorrectly twice, so the number might be a bit inflated.

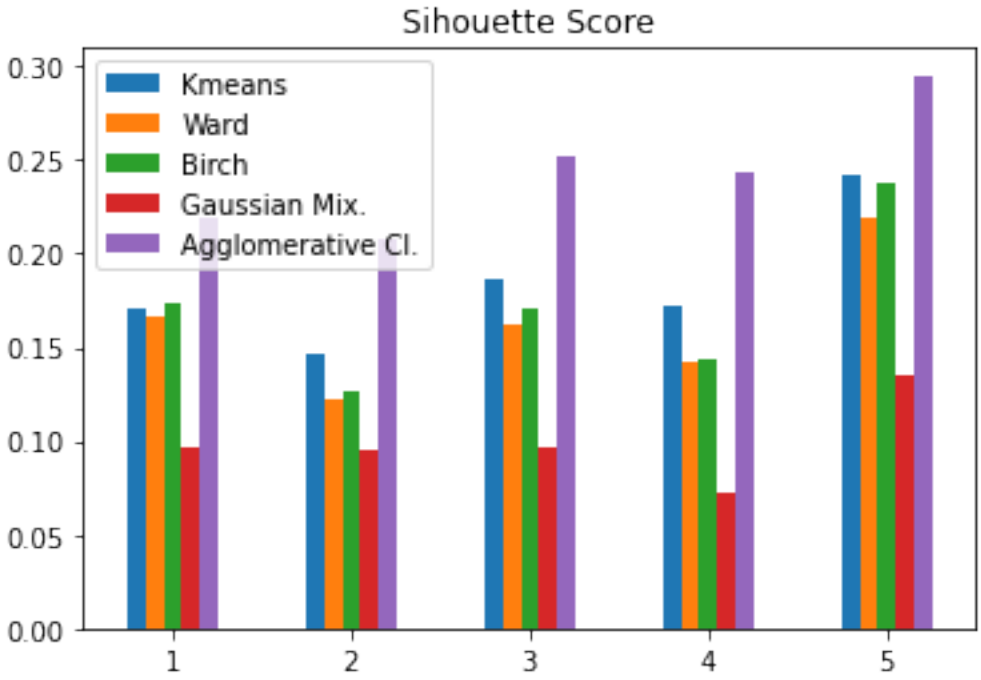
Remove no shows If people participated only in a single wave we cannot study their opinion over time so we remove them right away.

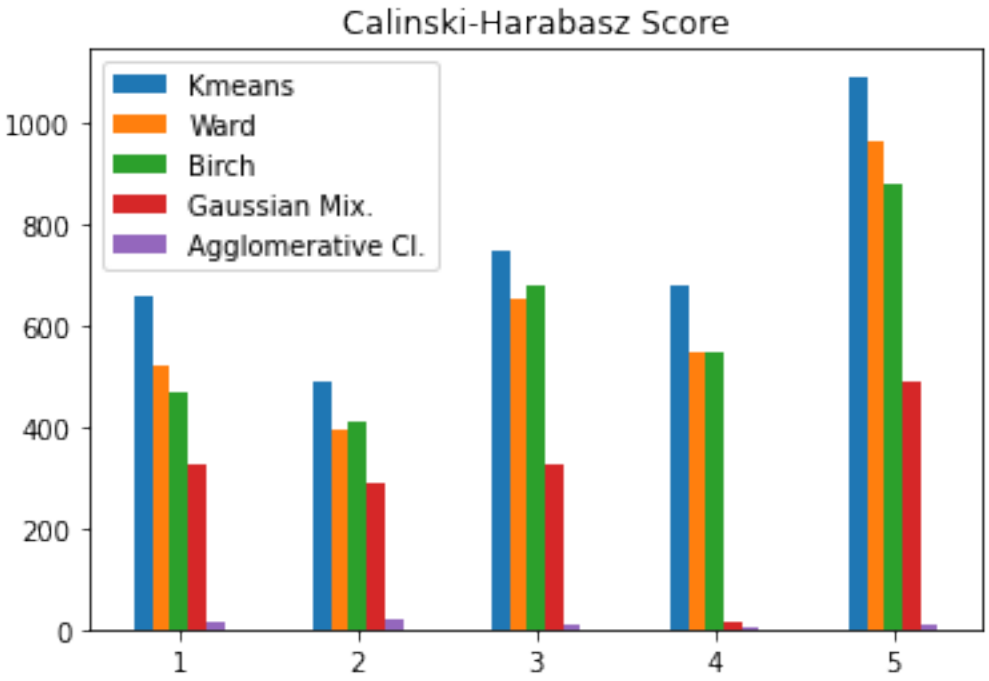
Opinion questions *only* We remove all non opinion related questions for the purpose of this analysis which can be conveniently done since we added the tag `OPINION` to all of those questions in the preprocessing notebook.

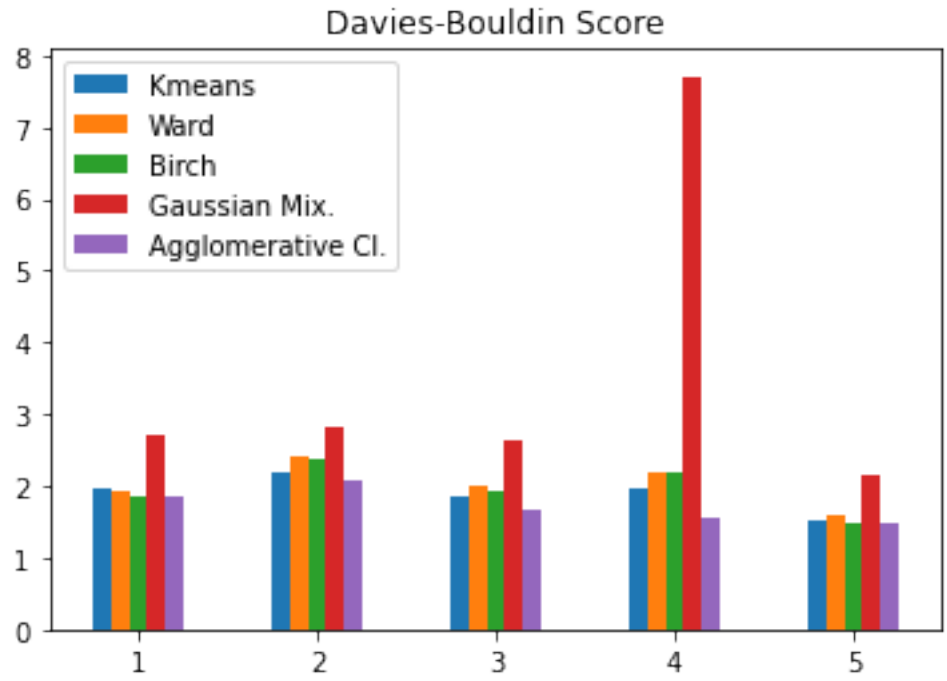
Scaling In the preparation notebook we already transformed the different answers to numerical values. Some questions use the wording `completely agree`, while others phrase it as `strongly agree` although both questions just give the same number of agreement options. So we converted everything to a numerical scale in order to make it comparable. If people picked `don't know` or refused to answer the question we assign the the value 2 which corresponds to the middle of the scale.

1.2.2 Investigate performance

Since we do not know what type of clustering method would be most appropriate for our data, we just run a couple of standard algorithms and see how the different metrics behave.







K-means performed either best or second best across the different metric. *Agglomerative clustering* is the best performing method with respect to *Silhouette score* and the *Davies-Bouldin score*, but the

worst with respect to the *Calinski-Harabasz score*, which might be worth investigating. (Spectral clustering took too long so it was not applied). In conclusion we use K-means for all of our next step.

1.3 Analyze clustering

We use the Kmeans algorithm since it performed the best in the previous evaluations. This method not only compute a clustering but also a *centroid* for every cluster and the distance of every point to those centers. Every point will simply be assigned to the centroid to which it has this smallest distance. If the distance of a point to its centroid is much smaller than its distance to the other centroid we tend to be more confident in it's assignment. So we compute quotient of the distances to the two centroids. Next we compute the median of all quotient of all points in a class and call the closest half *close* and the half that is further away than this median *far*.

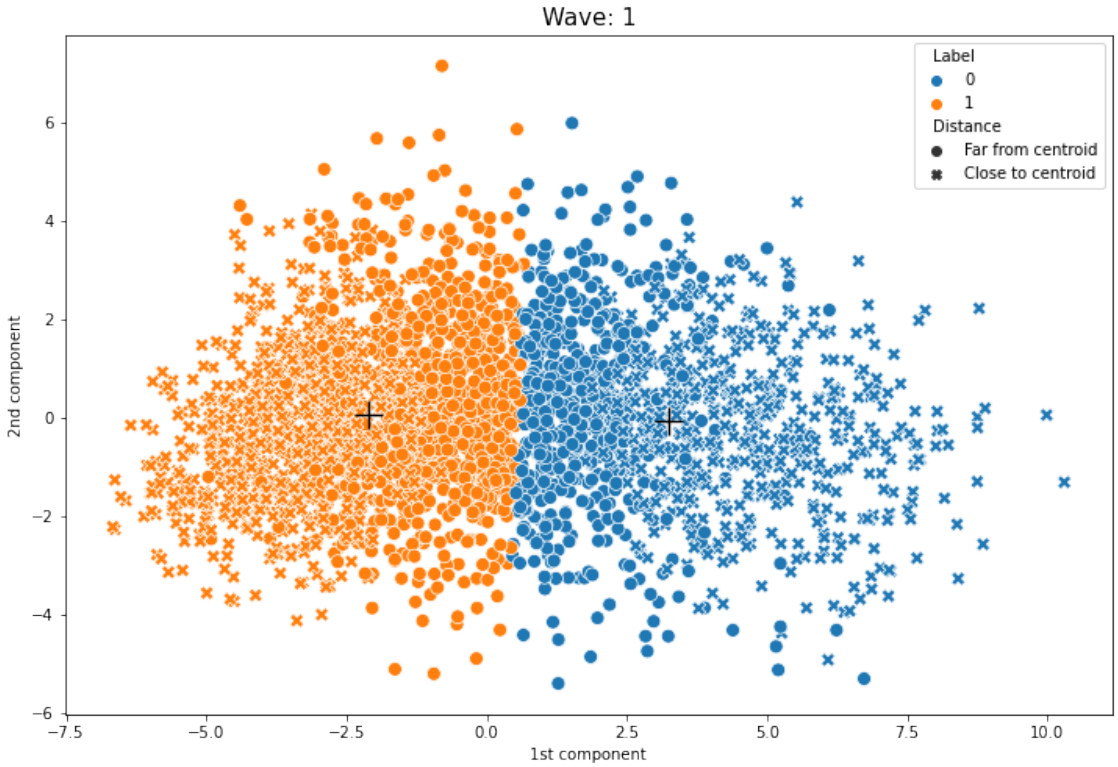
For illustration purposes we also compute a principal component analysis (PCA) of all the opinion question which allows us to visualize the obtained clustering and centroids in a lower dimensional space.

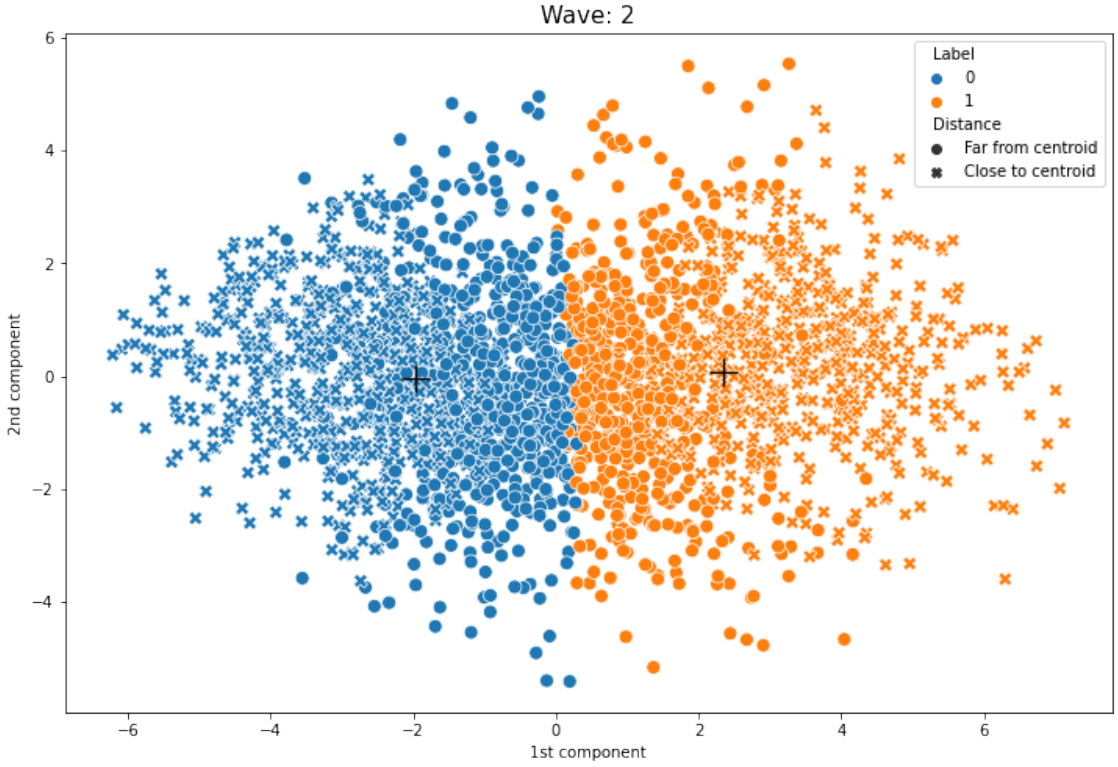
Additionally, for every question we compute the average strength of agreement for people of the same cluster and compare these values between the groups. We only list here the subset of question with the largest difference in agreement between the two groups.

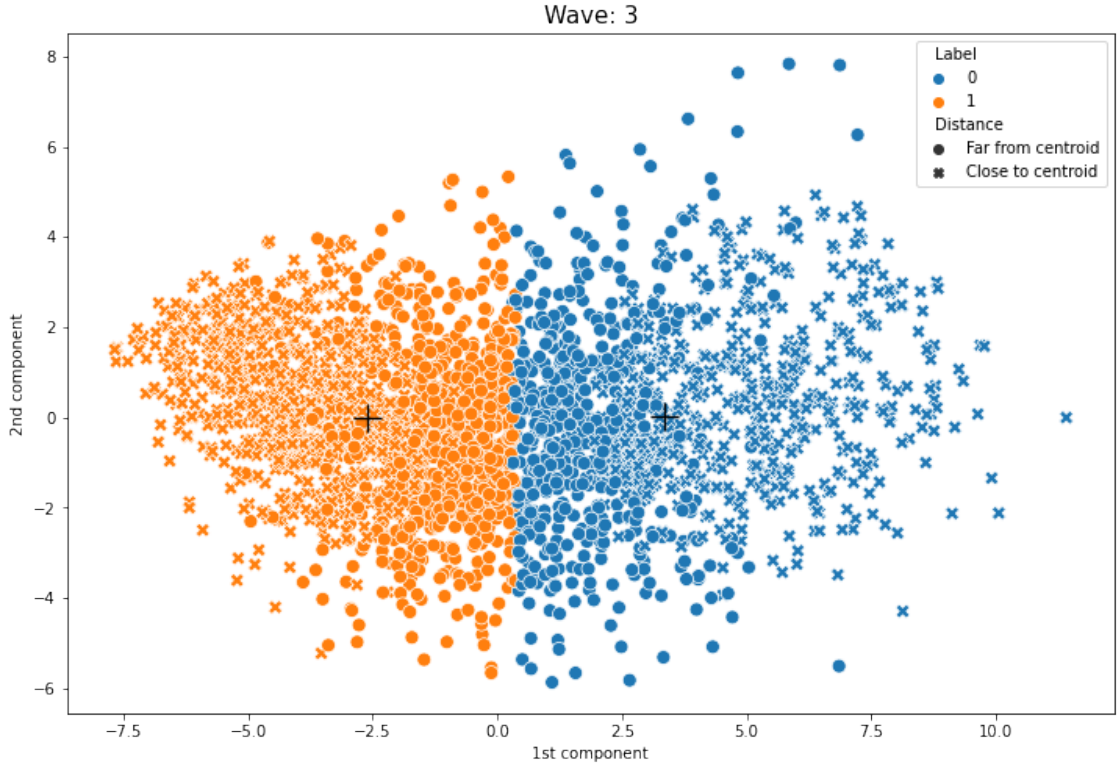
```
Wave:  1
             difference
OPINION: IMMIGRATION TO AUSTRIA ONLY IN EXCEPTIONAL CASES -w1_q44x7
1.800228
OPINION: CRIME RATES INCREASE IN AUSTRIA BECAUSE OF IMMIGRANTS -w1_q11x5
1.545444
OPINION: IMMIGRANTS ENRICH AUSTRIAN CULTURE -w1_q11x2
-1.477691
OPINION: SAME ACCESS TO SOCIAL BENEFITS: ASYLUM SEEKERS -w1_q37x2
-1.471789
OPINION: SAME ACCESS TO SOCIAL BENEFITS: NON-AUSTRIANS -w1_q37x1
-1.448840
OPINION: SAME ACCESS TO SOCIAL BENEFITS: EASTERN EUROPEANS (NEW EU MEMBER
STATES) -w1_q37x3   -1.402624
OPINION: IMMIGRANTS ARE GOOD FOR THE AUSTRIAN ECONOMY -w1_q11x3
-1.306389
OPINION: POLICE AUTHORITIES SHOULD BE EXTENDED -w1_q44x6
1.236557
OPINION: IMMIGRANTS PAY MORE INTO THE SOCIAL SECURITY SYSTEM -w1_q11x6
-1.228461
OPINION: IMMIGRANTS TAKE AWAY AUSTRIAN JOBS -w1_q11x4
1.207718

Wave:  2
       difference
OPINION: FEELING LIKE A STRANGER DUE TO THE MANY MUSLIMS -w2_q21x4
```

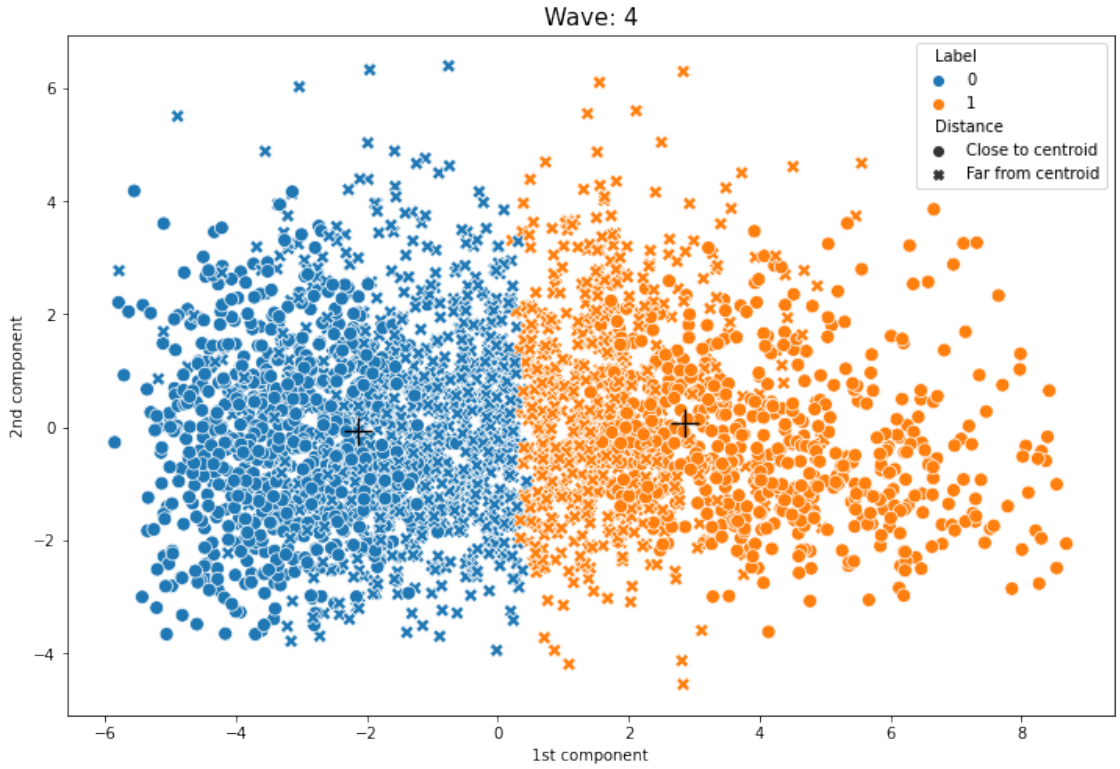
```
-2.061494
OPINION: HEINZ-CHRISTIAN STRACHE - DOES WHAT IS BEST FOR AUSTRIA -w2_q10x6
-1.506826
OPINION: ANGRY WHEN MUSLIMS ARE DISCRIMINATED AGAINST BECAUSE OF BELIEFS
-w2_q21x3       1.464736
OPINION: EUROPEAN AND MUSLIM LIFESTYLE ARE EASILY COMPATIBLE -w2_q21x5
1.178390
OPINION: THE PEOPLE SHOULD TAKE MOST IMPORTANT DECISIONS, NOT POLITICIANS
-w2_q24x7     -1.137606
OPINION: THE PARTIES ARE THE MAIN PROBLEM IN AUSTRIA -w2_q24x4
-0.982173
OPINION: CHRISTIAN KERN - DOES WHAT IS BEST FOR AUSTRIA -w2_q10x4
0.882206
OPINION: POLITICIANS DO NOT CARE ABOUT WHAT PEOPLE LIKE ME THINK -w2_q2x2
-0.871654
OPINION: MUSLIMS HAVE FEWER CHANCES IN AUSTRIA -w2_q21x1
0.845330
OPINION: POLITICIANS ONLY CARE ABOUT THE INTERESTS OF THE RICH AND POWERFUL
-w2_q24x2   -0.825724

Wave:  3
difference
OPINION: IMMIGRANTS HAVE RECEIVED MORE THAN THEY DESERVE -w3_q33x1
1.772062
OPINION: IMMIGRANTS GET MORE ATTENTION -w3_q33x2
1.709007
OPINION: DUTY TO ACCEPT ASYLUM SEEKERS -w3_q24x4
-1.667544
OPINION: CRIME RATES INCREASE IN AUSTRIA BECAUSE OF IMMIGRANTS -w3_q18x5
1.524966
OPINION: EQUAL MEANS-TESTED INCOME FOR ASYLUM SEEKERS -w3_q24x2
-1.482170
OPINION: IMMIGRANTS ENRICH AUSTRIAN CULTURE -w3_q18x2
-1.479515
OPINION: IMMIGRANTS ARE GOOD FOR THE AUSTRIAN ECONOMY -w3_q18x3
-1.301572
OPINION: IMMIGRANTS TAKE AWAY AUSTRIAN JOBS -w3_q18x4
1.228278
OPINION: ASYLUM SEEKERS TAKE AWAY JOBS -w3_q24x1
1.211813
OPINION: PEOPLE LIKE ME GET LESS ATTENTION THAN OTHERS -w3_q35x2
1.203807

Wave:  4
             difference
OPINION: IMMIGRATION TO AUSTRIA ONLY IN EXCEPTIONAL CASES -w4_q50x7
-1.797278
OPINION: CRIME RATES INCREASE IN AUSTRIA BECAUSE OF IMMIGRANTS -w4_q52x5
```

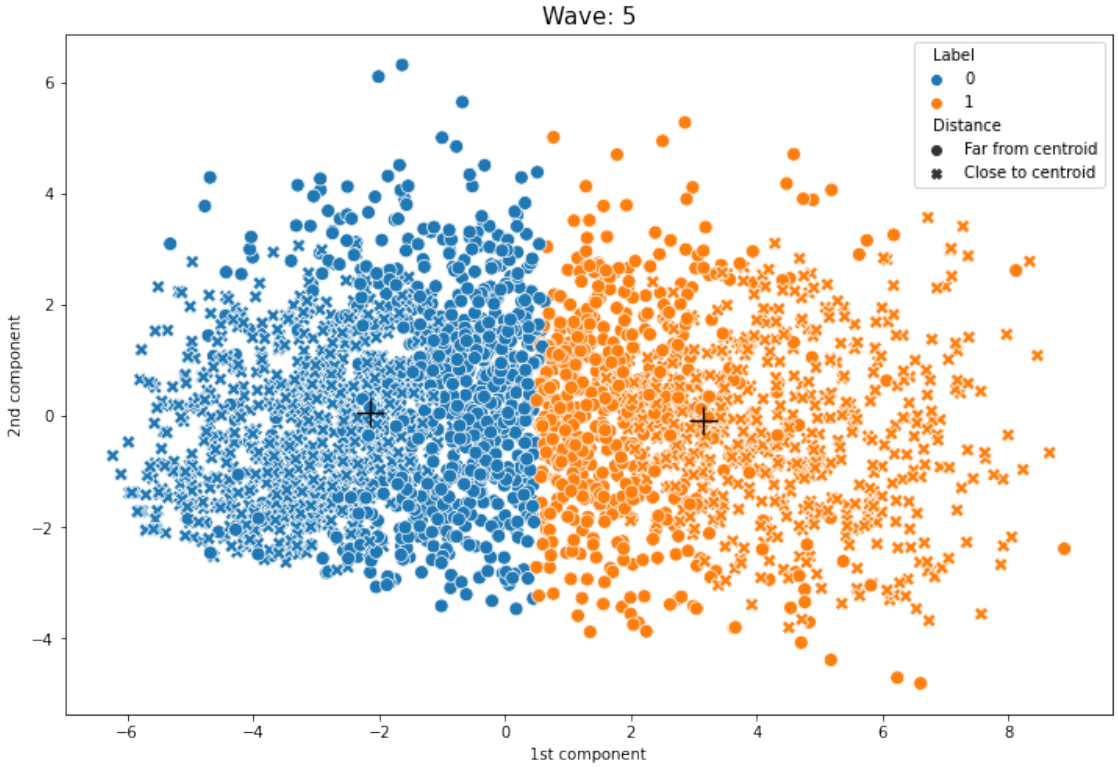
```
-1.649083
OPINION: SAME ACCESS TO SOCIAL BENEFITS: ASYLUM SEEKERS -w4_q65x2
1.593821
OPINION: IMMIGRANTS ENRICH AUSTRIAN CULTURE -w4_q52x2
1.544617
OPINION: SAME ACCESS TO SOCIAL BENEFITS: NON-AUSTRIANS -w4_q65x1
1.460928
OPINION: SAME ACCESS TO SOCIAL BENEFITS: EASTERN EUROPEANS (NEW EU MEMBER
STATES) -w4_q65x3    1.435847
OPINION: IMMIGRANTS ARE GOOD FOR THE AUSTRIAN ECONOMY -w4_q52x3
1.342664
OPINION: IMMIGRANTS PAY MORE INTO THE SOCIAL SECURITY SYSTEM -w4_q52x6
1.292485
OPINION: POLICE AUTHORITIES SHOULD BE EXTENDED -w4_q50x6
-1.225657
OPINION: IMMIGRANTS TAKE AWAY AUSTRIAN JOBS -w4_q52x4
-1.139211

Wave:  5
difference
OPINION: IMMIGRANTS GET MORE ATTENTION -w5_q30x2
-1.920959
OPINION: IMMIGRANTS HAVE RECEIVED MORE THAN THEY DESERVE -w5_q30x1
-1.886003
OPINION: CRIME RATES INCREASE IN AUSTRIA BECAUSE OF IMMIGRANTS -w5_q43x5
-1.701798
OPINION: IMMIGRANTS ENRICH AUSTRIAN CULTURE -w5_q43x2
1.523259
OPINION: FREEDOM OF MOVEMENT IN EU: THREATENS SECURITY -w5_q26x1
-1.513861
OPINION: IMMIGRANTS TAKE AWAY AUSTRIAN JOBS -w5_q43x4
-1.391518
OPINION: IMMIGRANTS ARE GOOD FOR THE AUSTRIAN ECONOMY -w5_q43x3
1.332105
OPINION: IMMIGRANTS PAY MORE INTO THE SOCIAL SECURITY SYSTEM -w5_q43x6
1.257536
OPINION: FREEDOM OF MOVEMENT IN EU: THREATENS EMPLOYMENT -w5_q26x3
-1.250120
OPINION: PEOPLE LIKE ME GET LESS ATTENTION THAN OTHERS -w5_q32x2
-1.243199
```


Wave: 1
Label
0
1
Distance
Far from centroid
Close to centroid
2nd component
1st component
6
4
2
0
−2
−4
−6
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


Wave: 2
Label
0
1
Distance
Far from centroid
Close to centroid
2nd component
1st component
6
4
2
0
−2
−4
−6
−4
−2
0
2
4
6


Wave: 3
Label
0
1
Distance
Far from centroid
Close to centroid
2nd component
1st component
8
6
4
2
0
−2
−4
−6
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


Wave: 4
Label
0
1
Distance
Close to centroid
Far from centroid
2nd component
1st component
6
4
2
0
−2
−4
−6
−4
−2
0
2
4
6
8



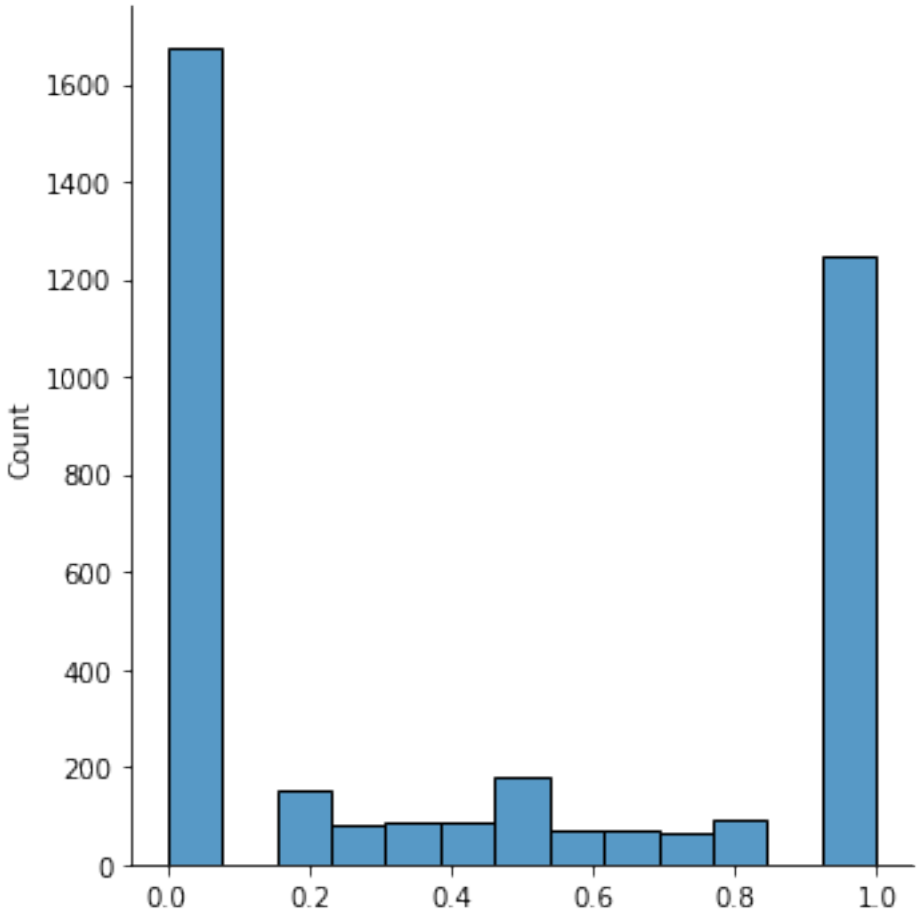
1.4 Is our clustering stable across waves?

We start with the following premise: > Political opinions of people in general don't vary much in short periods of time.

If this is indeed the case it makes sense to look at our clustering across different waves and check if participants are constantly assigned to the same cluster? It should be noted here that it's not so easy to define what we mean by "same" cluster as there are no external labels. However, by looking for example at the question with the biggest difference between the two clusters we can see a clear nationalist-liberal trend and can therefore manually give meaning to the two classes.

For each participant we then compute the `mean` of their respective labels across waves. If they have been assigned the label 1 in all waves in which they participated, then the mean is 1. Analogously for label 0. If the mean of the labels of a given participant is strictly between 0 and 1 this means that they have not been assigned the same group in all waves. The mean being close to 0 or 1 indicates that a person has been given the same label *most of the time*.

[16]: `<seaborn.axisgrid.FacetGrid at 0x7fab9c277100>`



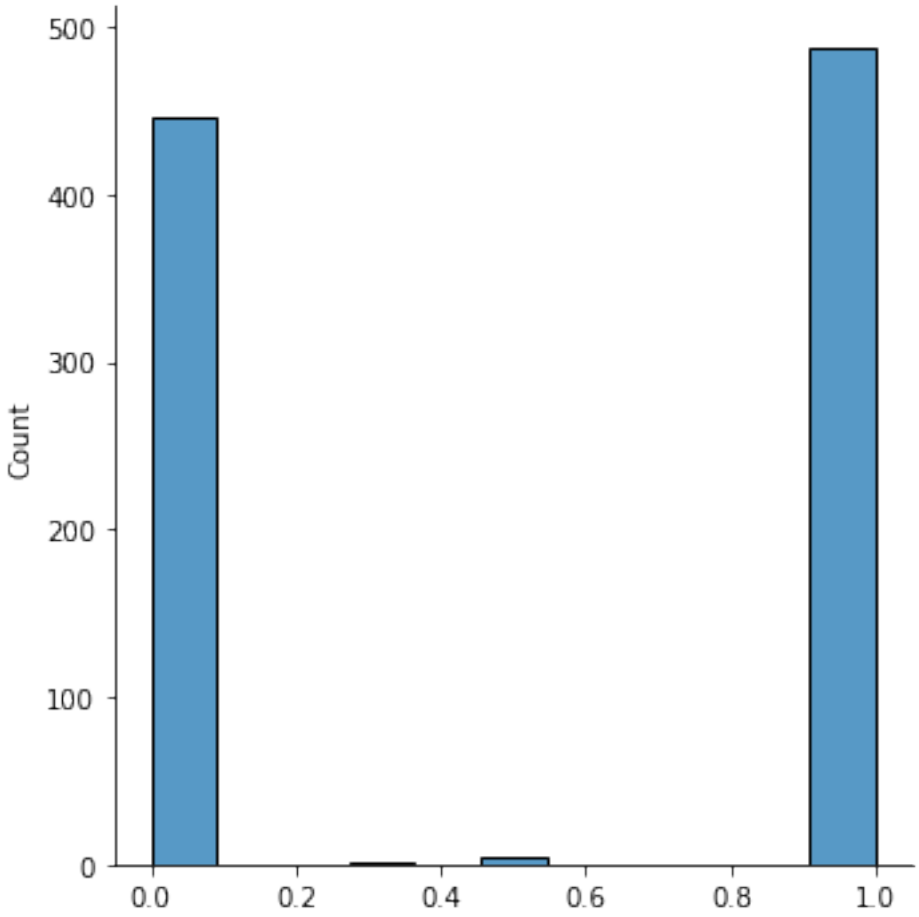
Evidently, most of the people are assigned to the same group all the time.

1.4.1 Respondents located close to centroids

Naturally, we assume that political opinions of people, no matter the topic, are not black or white, but located on a spectrum. People who are located somewhere in the middle might be assigned a different cluster in later waves even if their opinion did not change. For this reason we want to see if the assignment of clusters across waves becomes even more stable if we only consider people that are much closer to their centroid then they are to the other one. In some sense those people are clustered more confidently.

Repeating the same analysis of the *average label* assigned across waves yields the following plot.

[18]: `<seaborn.axisgrid.FacetGrid at 0x7fab9c3341c0>`



Evidently, for the selected subset of people we are extremely certain in the placement on the political spectrum.

```
[NbConvertApp] Converting notebook clustering.ipynb to pdf
[NbConvertApp] Writing 1190106 bytes to output_pdf/clustering.pdf
```