## Question 5 & 6

Here we reduce dimension with SVD and NMF respectively. Let r denote the dimension that we want to reduce of the data. We try r = 1,2,3,5,10,20,50,100,300 and plot the 5 measure scores discussed before. Figure x shows the result of K-means after SVD. Figure x shows the result of K-means after NMF.

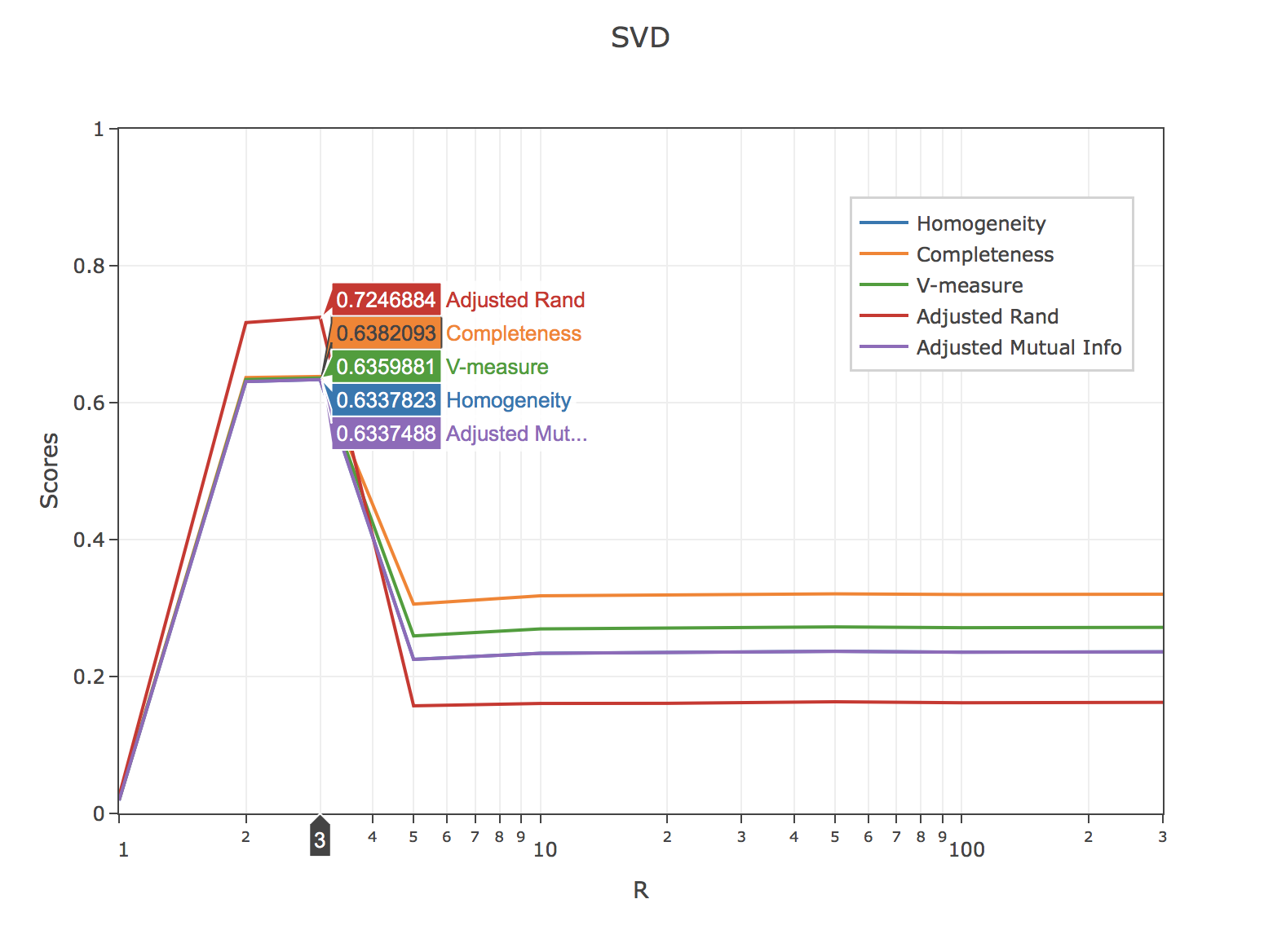


Figure Results for different r of SVD

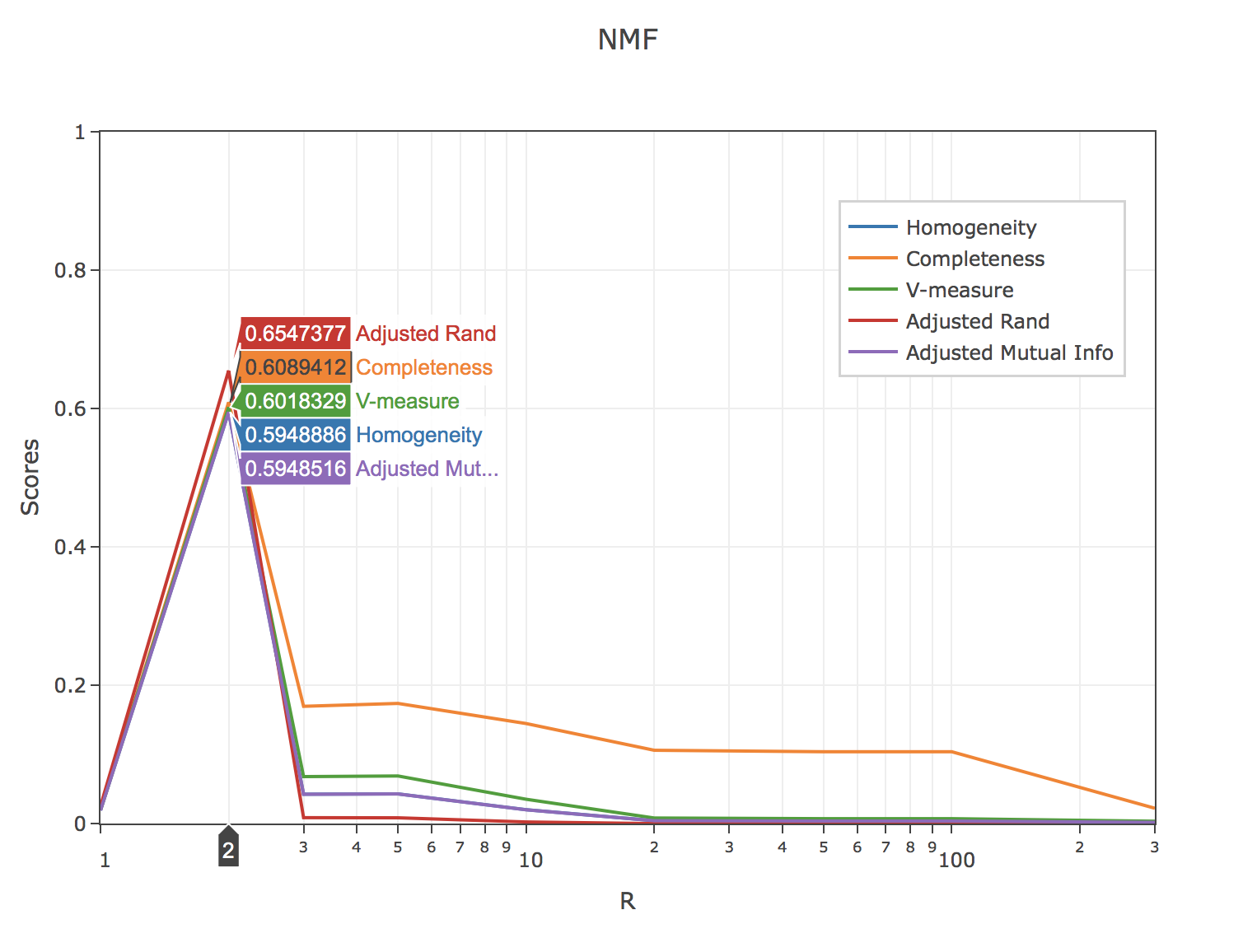


Figure Results for Different r of NMF

From these two figures, we can see that the best r for SVD is 3 and the best r for NMF is 2. Also, we notice that all metrics including homogeneity, completeness, V-measure, adjustable rand and adjusted mutual information, reach best value when r equals to 3 for SVD and 2 for NMF.

Moreover, we can conclude from these two figures that the relationship between dimension r and K-means results of both two-dimensional reduction methods is not monotonic. When r is small, limited information is stored after reducing dimension that the clustering process of K-means relies only on the information far from accurate enough. Therefore, the result will not be good. However, when r is large, the vector space is in extreme high dimension that Euclidean distance is no longer a good metric anymore while K-means we used in this project is based on Euclidean distance. In this way, the clustering results with such high dimension is not reliable.

## Question 7

The ground truth of clustering is shown in figure x below.

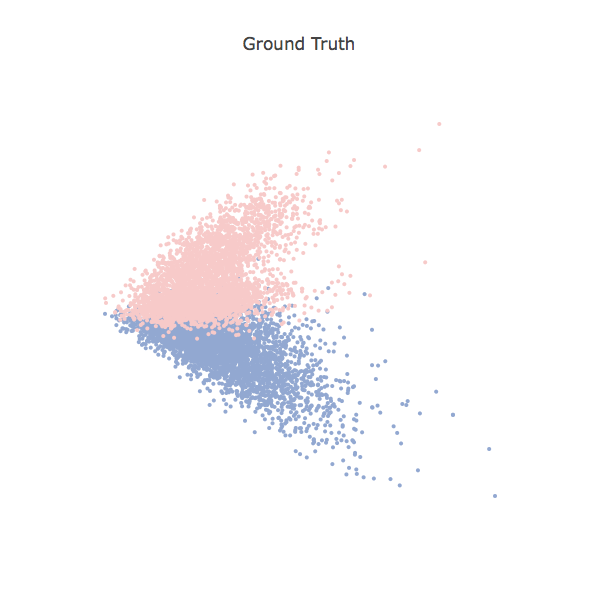


Figure Ground Truth

The result r or SVD is 3. The performance of K-means after applying SVD for dimension reduction with r equals to 3 is

Homogeneity: 0.634

Completeness: 0.639

V-measure: 0.637

Adjusted Rand-Index: 0.725

Adjusted Mutual Information Score: 0.634

Contingency Table:

[[3435 468]

[ 117 3862]]

The clustering result is shown in figure x.

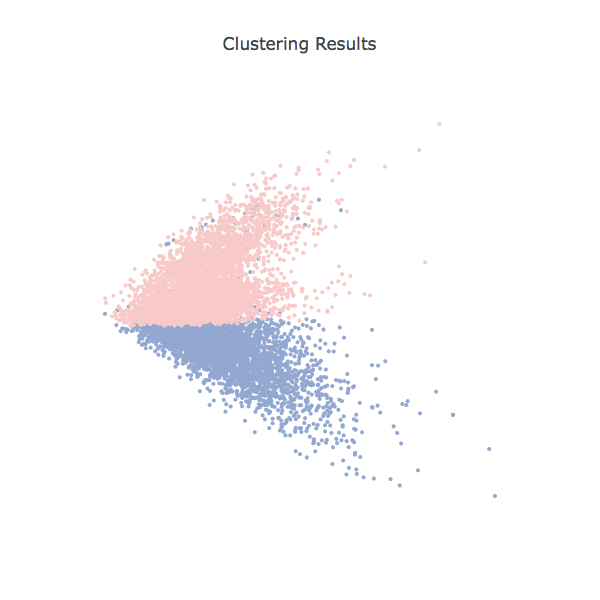


Figure Result of SVD with r=3

The result r or NMF is 2. The performance of K-means after applying NMF for dimension reduction with r equals to 2 is

Homogeneity: 0.595

Completeness: 0.609

V-measure: 0.602

Adjusted Rand-Index: 0.655

Adjusted Mutual Information Score: 0.595

Contingency Table:

[[3194 709]

[ 43 3936]]

The clustering result is shown in figure x.

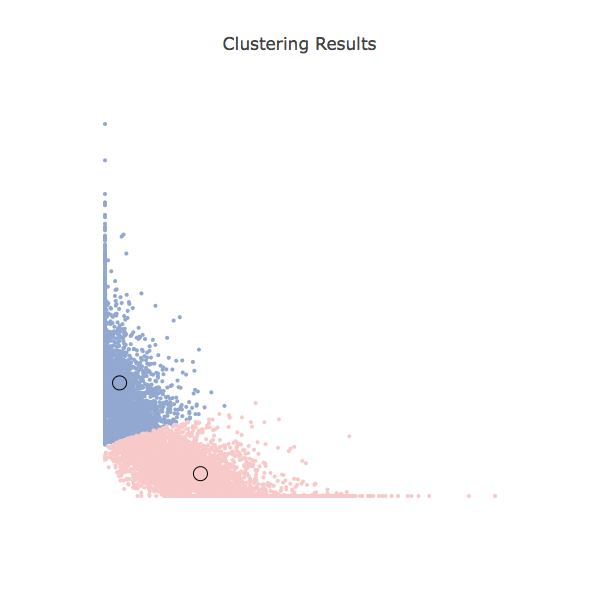


Figure Result of NMF with r=2