Benchmarking in Clustering

K Medoids, Student T Model and DBSCAN

Math 252 Project II
By

Qian Meng

November, 2020

Outlines

- ❖ Purpose & Background
- ❖ Clustering Methods (DBSCAN)
- ❖ Optimal Number of Clusters

- ❖ 4 experiments:
  Ⅰ) High Dim II) Outliers III) Clusters IV) Correlation

- ❖ Summary

Purpose & Background

Purpose:

Benchmark clustering methods by performing experiments

Using silhouette index(SI) and adjusted random index(ARI).

Clustering methods :

K-medoids, Student-t model,

DBSCAN (density-based spatial clustering of applications with noise)

Datasets:

Real data (USPS MNIST handwritten digits) (-> test high dimension)

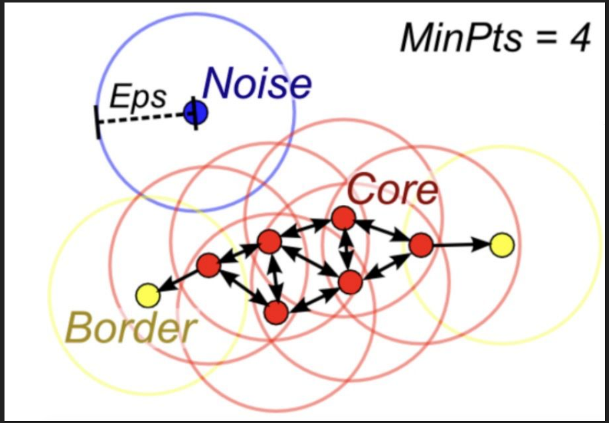
Simulated Contaminated data (-> test outliers, number of clusters and correlation).

Clustering Methods

- ❖ K-medoids: creates clusters by using a data point in the data set as the centroid.
- ❖ Student t model : is an EM algorithm which maximizes the likelihood of t distribution to model data.
- ❖ **DBSCAN**(density-based spatial clustering of applications with noise)

  : groups points that are close to each other based on a distance measurement (Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions.

** All three methods are supposed to be resilient to **outliers**.

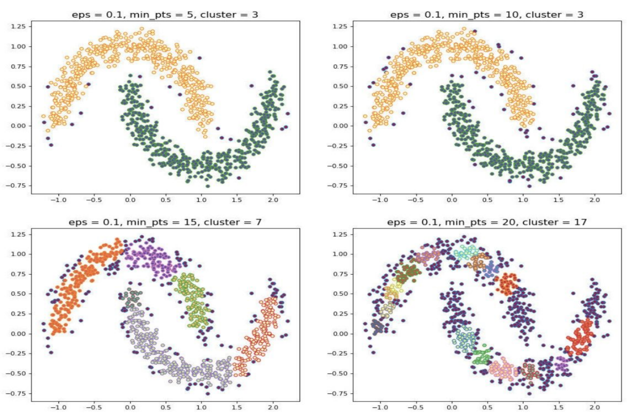


DBSCAN (Density-based Spatial clustering of applications with noise)

2 parameters: **epsilon** (radius around the point) and **k (**minimum neighbor points)

2 properties: needn’t initiate the number of clusters

determine outliers (using the distance of k nearest neighbors.)



In R

(fpc package)

```
dbscan(data, eps, MinPts = k, ...)
```

Use knn to pick epsilon

```
kNN(x, k, query = NULL, sort = TRUE, search =
"kdtree", …)
```

How we choose Optimal Number of Clusters

- K-medoids:
  - based on SI or ARI
- Student t model:
  - based on the BIC.
- DBSCAN:
  - computed by DBSCAN algorithm automated.

Ⅰ) High Dimension

Raw Data : USPS MNIST handwritten digits (70,000x784 matrix, 10 clusters)

Limited by R, a smaller subset was used to do the experient.

PCA

$$\sum_{i=1}^{k} \lambda_i / \sum_{i=i}^{p} \lambda_i$$

Fig.1 Percent scatter preserved formula

λ_i ordered eigenvalues of the covariance matrix of the dataset.

p the number of total dimensions.

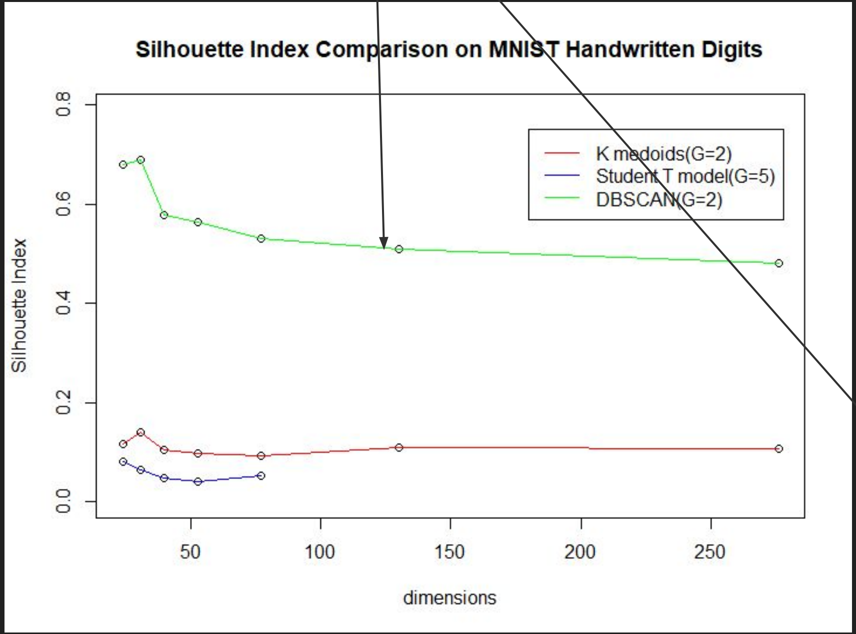
k is the number of dimensions in a lower dimensional plane.

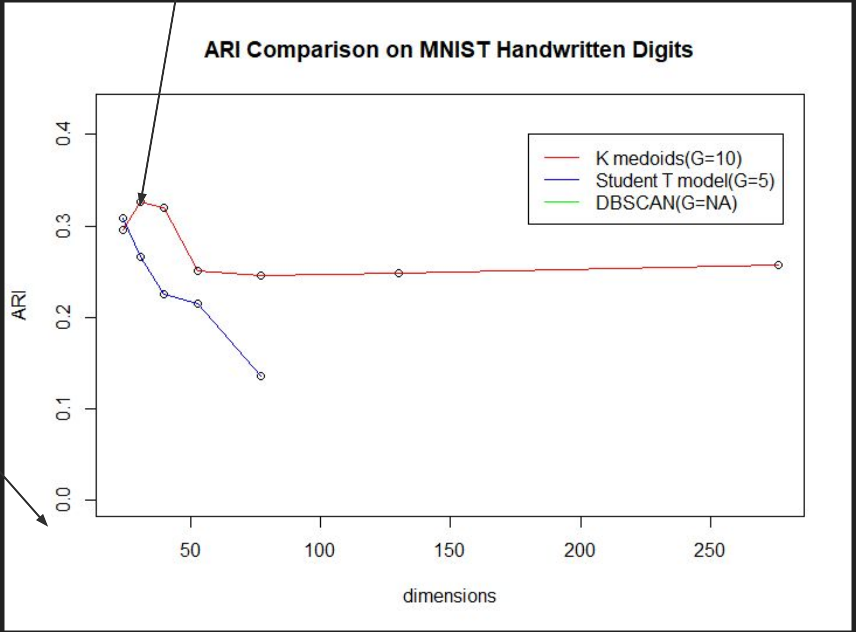
Data set preserving 70% to 99% -> subspace dims = c(24,31,40,53,77,130,276)

Models’ Performance(High Dim)

K-medoids was also able to pick out 10 clusters (the true number of labels).

DBSCAN considered most data points to be an outlier





**** All three methods performed poorly in high dimensions.**

(G are number of clusters found in the best model per method)

II) Outliers

Simulated Dataset: contaminated normal distribution

$$f(\boldsymbol{x};\vartheta) = \alpha\phi(\boldsymbol{x};\boldsymbol{\mu},\Sigma) + (1-\alpha)\,\phi(\boldsymbol{x};\boldsymbol{\mu},\eta\Sigma),$$

Cluster Number =5

Means: different between clusters

Variance = 100, 1000 (low sig = 100, high sig = 1000)

Alpha = 0.05, 0.1,... , 0.3 (the percent of ‘bad’ data points, high sig)

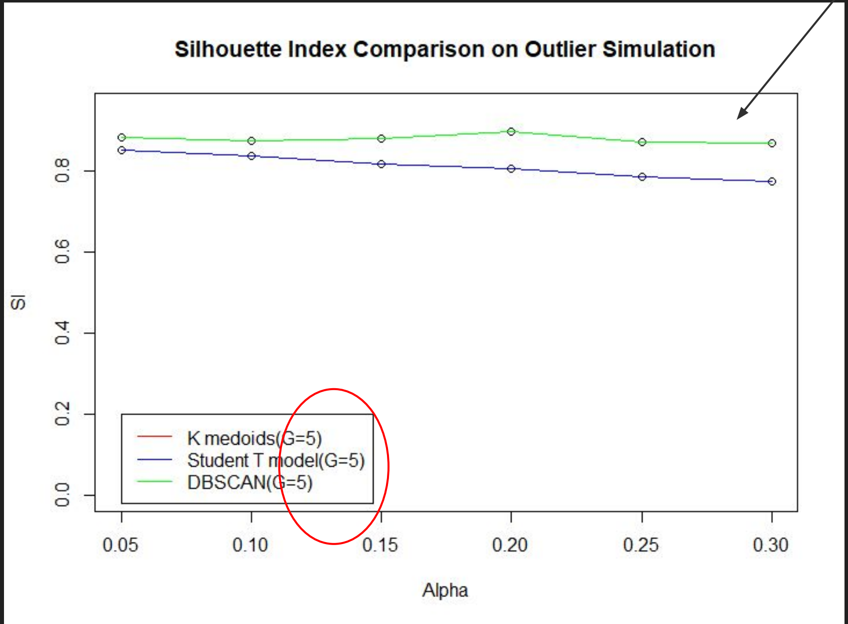
Outliers : ten times greater than the original variance.

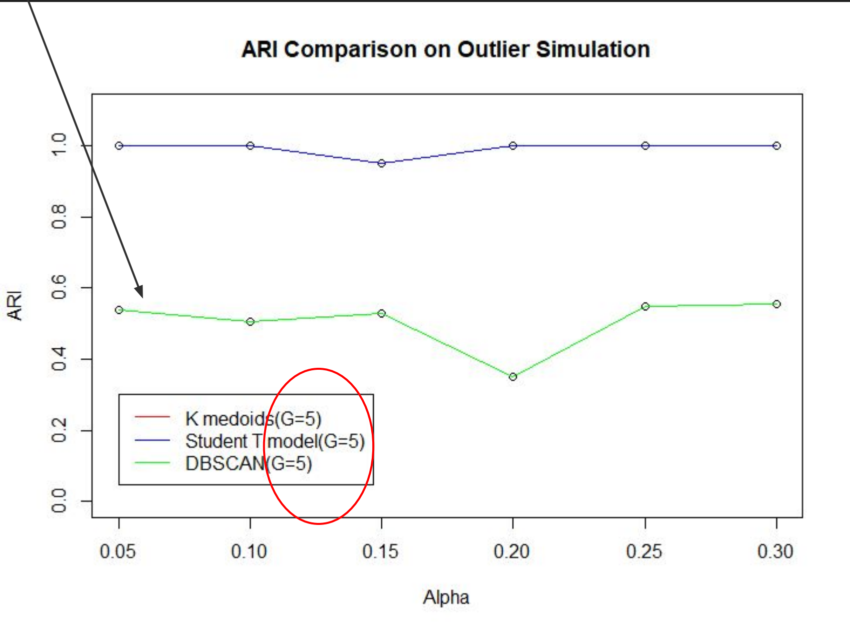
Models' Performance(Outliers)

All three methods performed well with less or more outliers.

K-medoids line is overlap by Student T.

DBSCAN finds outliers and omits them from clusters





III) Number of Clusters

Simulated Dataset: contaminated normal distribution

$$f(\boldsymbol{x}; \vartheta) = \alpha\phi(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) + (1 - \alpha)\,\phi(\boldsymbol{x}; \boldsymbol{\mu}, \eta\Sigma),$$

Cluster Number =5,10,...,30

Variables Number =10

Means: different between clusters

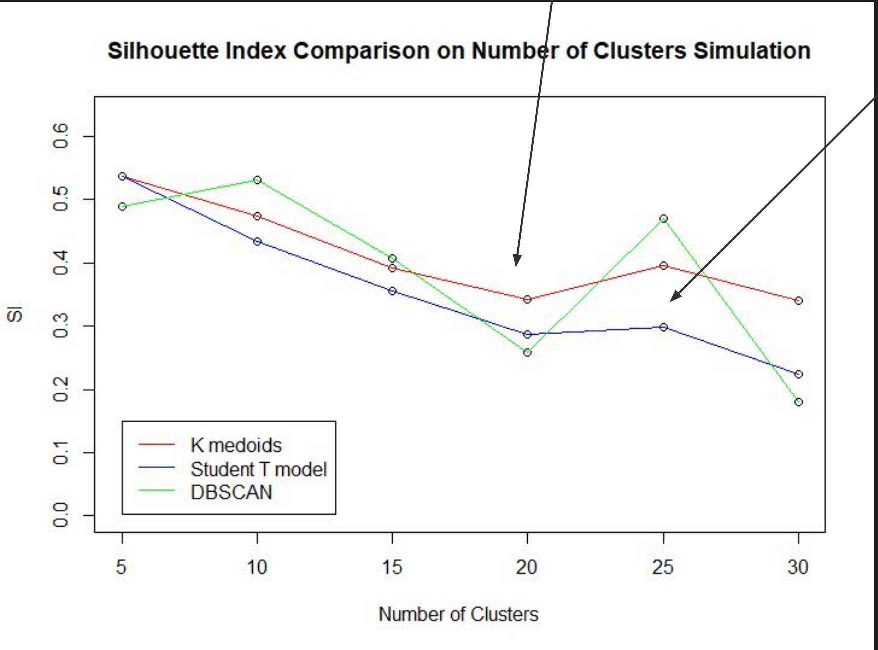
Variance = 100 (for most points)

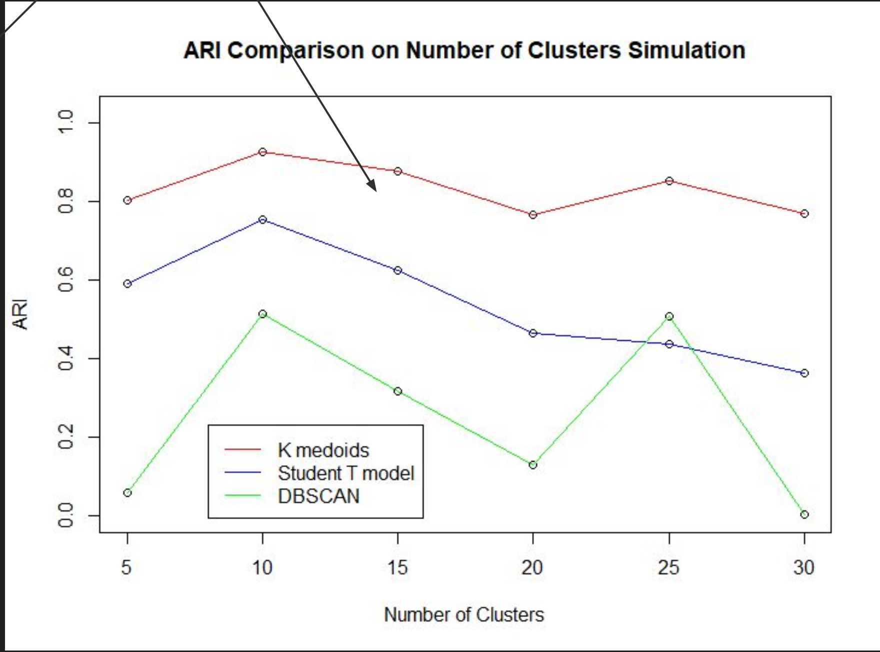
Alpha = 0.05 (the percent of ‘bad’ data points, sig = 1000)

Models' Performance (Cluster Number)

K medoids performed better than the student t. (no outliers, var stable)

Downward sloping pattern — the more classes, the worse the performance is.





IV) Correlation

Simulated Dataset: contaminated normal distribution

$$f(\boldsymbol{x};\vartheta) = \alpha\phi(\boldsymbol{x};\boldsymbol{\mu},\Sigma) + (1-\alpha)\,\phi(\boldsymbol{x};\boldsymbol{\mu},\eta\Sigma),$$

Cluster Number =5

Correlations = 0, 0.2,..., 1

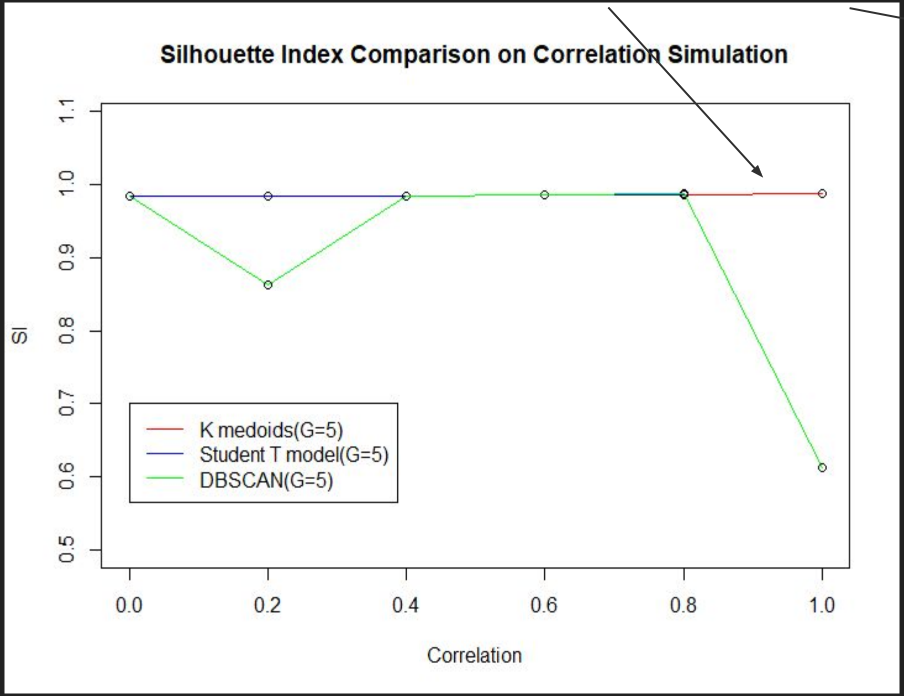
Variance = 1

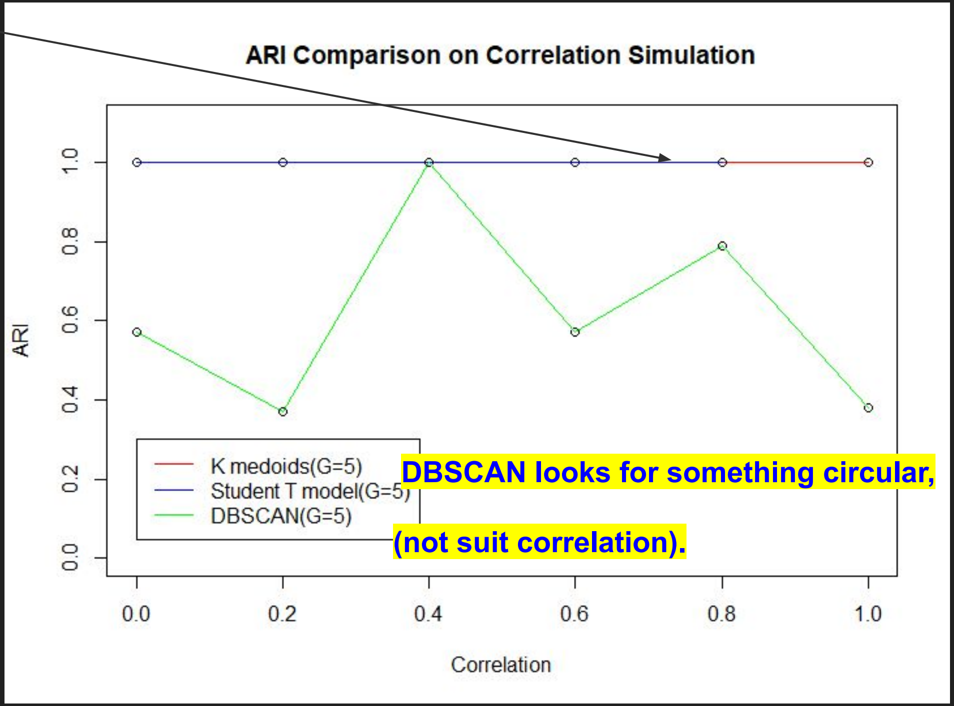
Means: different between clusters

Models' Performance(Correlation Far Clusters)

The SI and ARI values for k medoids and student t are close to 1 (not actually 1).

Student t model not converge when the correlation was 1.

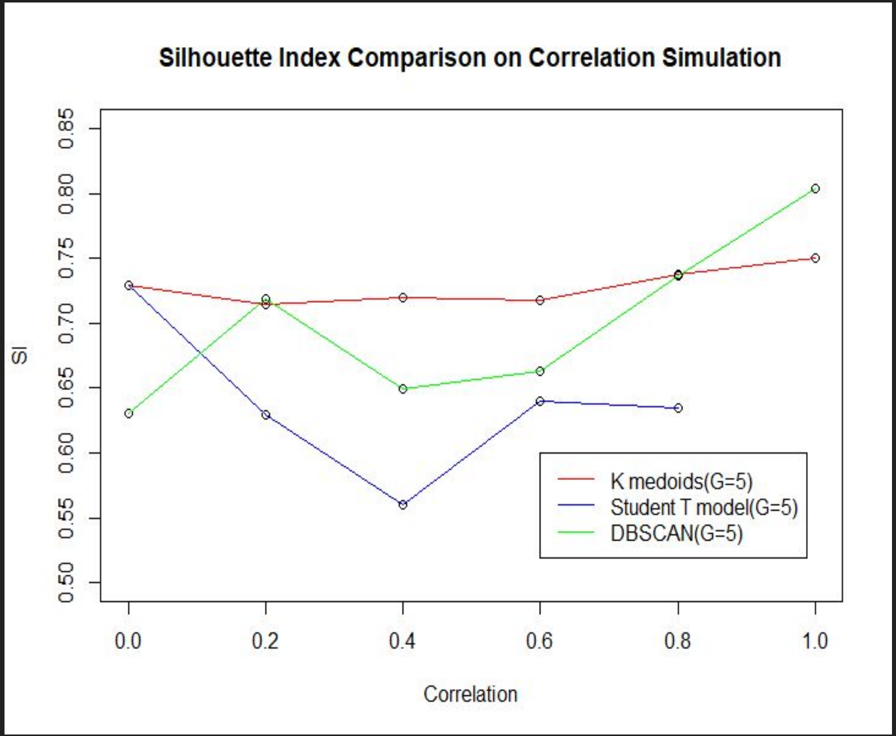


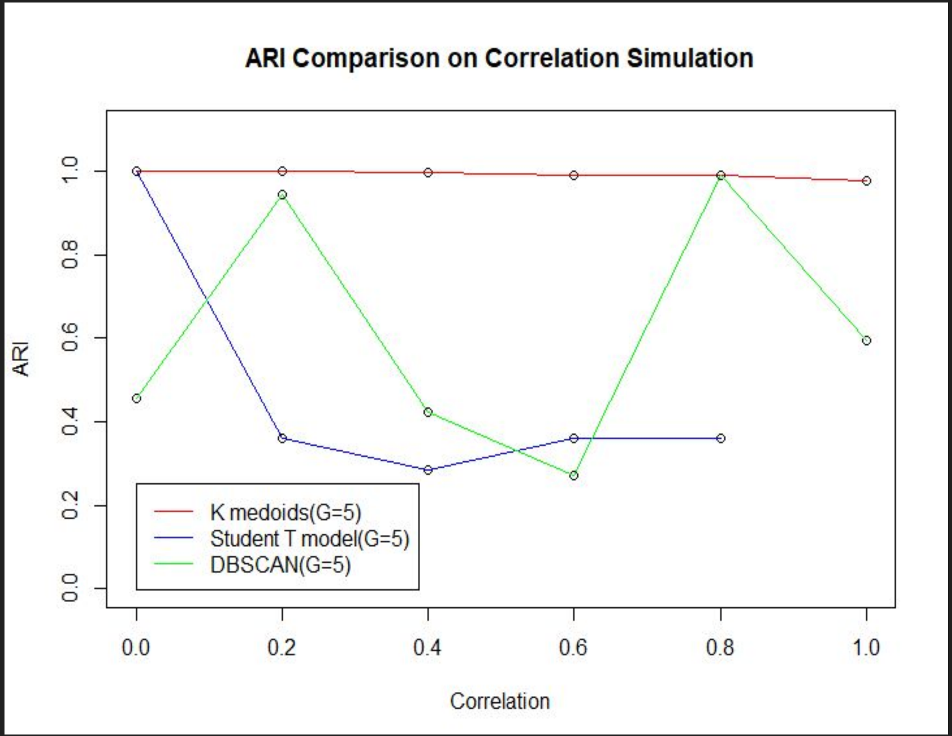


DBSCAN looks for something circular,

(not suit correlation).

Models' Performance(Correlation Close Clusters)





Summary

- K medoids and student t model seems to be more similar to each other than DBSCAN.
- K medoids and student t performed well with outliers and correlation but poorly with high dimensions.
- DBSCAN performed well with high dimensions and outliers but poorly with correlation.

	Dimensions	Outliers	Number of Clusters (more clusters)	Correlation
K Medoids	X	✓	X	✓
Student T Model	X	✓	X	△
DBSCAN	△	✓	X	X

Bibliography

[1] Math 252 Course slides (by Dr. Tortora)

[2] https://rdrr.io/cran/dbscan/man/kNN.html

[3] https://medium.com/@agarwalvibhor84/lets-cluster-data-points-using-dbscan