Benchmarking Initialization Algorithms With Three Model-Based Clustering Methods

Andy Qui Le, Yosha Mundhra, Xiang Yao

1. Research Design

This research compares initialization algorithms to determine which one offers the best initial clustering solution for mixture model-based clustering methods. Five initialization clustering techniques have been selected in advance. They are Fuzzy c-means (FKM), Partition around medoids (PAM), PD-clustering (PDC), K-means, and random methods. The three model-based clustering techniques in consideration are Gaussian Mixture Model (GMM), Multivariate t-Distribution Mixture Model (student-t), and Multivariate Contaminated Normal Mixture Model (MCNM). They are known as primary algorithms.

The output of the initialization algorithm serves as the starting input for primary algorithms. For each primary algorithm and various cluster counts, we will examine which initialization algorithm yields the best clustering results. To determine their performance, we applied these initialization methods to simulated data sets that have 2 particular features, overlaps and outliers in clusters. The above three model-based clustering techniques are selected to compare how well each of the five initialization methods handle outliers and overlaps. Simulated data sets are produced using overlaps across distributions, making it appear as though the clusters were not put together on purpose.

Visualization techniques like Box plots and Bar plots along with statistical metrics like BIC and Adjusted Rand Index (ARI) are used for comparison.

2. Simulation Design

Our designed data sets have two dimensions (p = 2) and a fixed number of clusters (G = 2). Each data set contains a total of 300 observations. Two components receive an equal number of observations ($\pi_1 = \pi_2 = 0.5$). Since we compare the performance of initialization algorithms using the proportion of data overlap between two components and the proportion of outliers, we have four scenarios to take into account: low overlap and low outlier, low overlap and high outlier, high overlap and low outlier, high overlap and high outlier. Twenty data sets are stimulated for each scenario. There are a total of 80 data sets.

Two functions, “MixSim” and “simdataset”, from the package MixSim are used to mimic the appropriate data sets. From a Gaussian normal distribution, MixSim is used to produce a mixing proportion, cluster means, and cluster covariance-variance matrix. The “simdataset” function is utilized to produce the necessary data sets using the MixSim results, providing the number of regular data points and the number of outliers depending on the four aforementioned situations. We use 5% and 15% for low and high percentages of outliers and 10% and 40% for low and high percentages of data overlap between two components.

The interval for outliers (int) is [-2,3][1], density level for simulating outliers (α) is 0.01. The “simdataset” function simultaneously generates 300 * (1 - Outlier Percentage) observations which are evenly assigned into two components, and 300 * (Outlier Percentage) observations as outliers which are labeled as 0. Since these outliers are randomly simulated, we manually assign

[1] Gaussian Normal Mixture and Multivariate t-Distributions uses [-1,2] as the interval for outliers to simulate data.

the first half outliers into group 1 and the other half into group 2.

Meanwhile, there is another function, “rCN”, which is a popular but different method to simulate Contaminated Normal Distribution data. The advantage of this simulation method is that it can specify Mixture Contaminated Normal parameters, such as the proportion of good points and the degree of contamination. Thus, we use “rCN” to generate a second list of 80 data sets. This simulation is similar to the one conducted by Tong and Tortora (2022). The proportions of good observations are $\alpha_1 = 0.95$ (low outlier) and $\alpha_2 = 0.85$ (high outlier); degrees of contamination are $\eta_1 = 20$, and $\eta_2 = 30$. Components’ centers are used to denote higher or lower overlap: low overlap: $\mu_1 = \begin{bmatrix} 0 \\ -3 \end{bmatrix}, \mu_2 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}$; High overlap: $\mu_1 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \mu_2 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}$. Variance-covariance matrices of the two components are $\Sigma_1 = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$. MCNM clustering will be performed on both data sets. A comparison will be conducted accordingly.

Scenario	**MixSim Data Sets**	**rCN Data Sets**
MixSim & rCN Common Parameters: n=300, G=2, p=2, $\pi_1 = \pi_2 = 0.5$		
rCN Common Parameters: $\Sigma_1 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$		
Specific Parameters		
1. Low Overlap & Low Outlier (20 Data Sets)	P(Overlap)=**10%**, P(Outlier)=**5%**	$\mu_1 = \begin{pmatrix} 0 \\ -3 \end{pmatrix}, \mu_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$ $\alpha_1 = 0.95,\ \eta_1 = 20$
2. Low Overlap & High Outlier (20 Data Sets)	P(Overlap)=**10%**, P(Outlier)=**15%**	$\mu_1 = \begin{pmatrix} 0 \\ -3 \end{pmatrix}, \mu_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$ $\alpha_2 = 0.85,\ \eta_2 = 30$
3. High Overlap & Low Outlier (20 Data Sets)	P(Overlap)=**40%**, P(Outlier)=**5%**	$\mu_1 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \mu_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$ $\alpha_1 = 0.95,\ \eta_1 = 20$
4. High Overlap & High Outlier (20 Data Sets)	P(Overlap)=**40%**, P(Outlier)=**15%**	$\mu_1 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \mu_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$ $\alpha_2 = 0.85,\ \eta_2 = 30$

Table 1 Parameters for Data Set Simulation

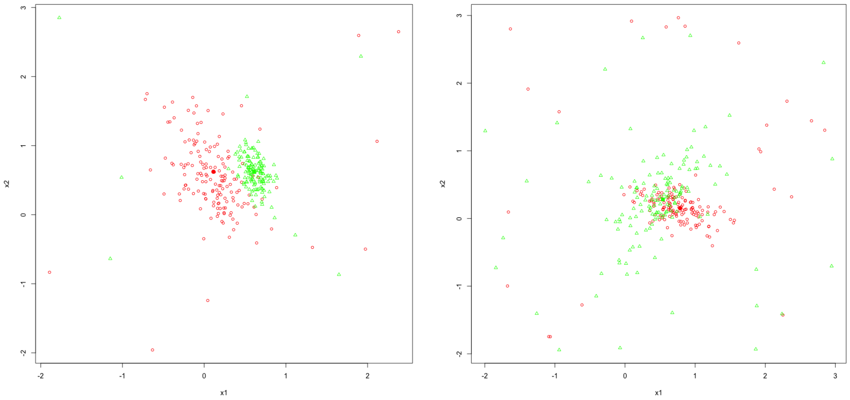


Figure 1 MixSim Data Set 1 (L Overlap & L Outlier) and 80 (H Overlap & H Outlier)

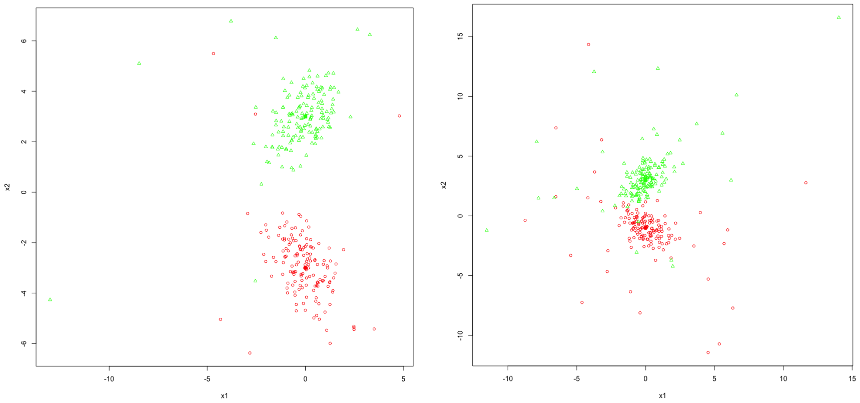


Figure 2 rCN Data Set 1 (L Overlap & L Outlier) and 80 (H Overlap & H Outlier)

3. Data Preparation and Analysis Criterion

The 80 simulated data sets are clustered using G=2:5 and five initialization algorithms, including K-means, PAM, Fuzzy c-means, PD Clustering, and Random method. There are 1,600 sets of clustering labels, or (80 data sets)*(4 G)*(5 Initialization). After that, each initialization algorithm's clustering labels from each data set are given to GMM, multivariate t-distribution and MCNM. Clustering function Teigen() in the Teigen package for GMM and student-t and CNmixt() in the ContaminatedMixt package for CNM. For MCNM, getBestmodel() with criterion = “BIC" is used to find which G gives the best model for each data set with each initialization algorithm. After checking the BIC and log likelihood, we find they have the same sign in CNmixt(). Thus, the largest BIC indicates the best model for MCNM.

Analysis Criterion. Firstly, the Adjusted Rand Index (**ARI**) value is then calculated using the labels of the best model of the data set for each initialization technique and the original label for that data set. An improved clustering outcome is indicated by a greater ARI, which denotes a better match between the clustering label and the original label. Currently, we have 400 best models and corresponding Gs, which equals (80 data sets)*(5 Initialization). Secondly, **Best G.** We also summarize, in each criteria and each initialization algorithm, how many best models of each model-based clustering is G=2, which is the original G we use to simulate the data. Larger frequency of G=2 indicates a better clustering result based on that initialization algorithm. Thirdly, for MCNM only, **Outlier Detection.** We look at the percentage of outliers that the MCNM model detected, based on each initialization algorithm in each scenario, and compare it to the original outlier proportion. A percentage of detected outliers of a MCNM that is closer to the original proportion of a data set indicates a better initialization performance.

4. Simulation Results

Gaussian Mixture Model

Following are the results of the comparison of 5 different initialization algorithms on GMM.

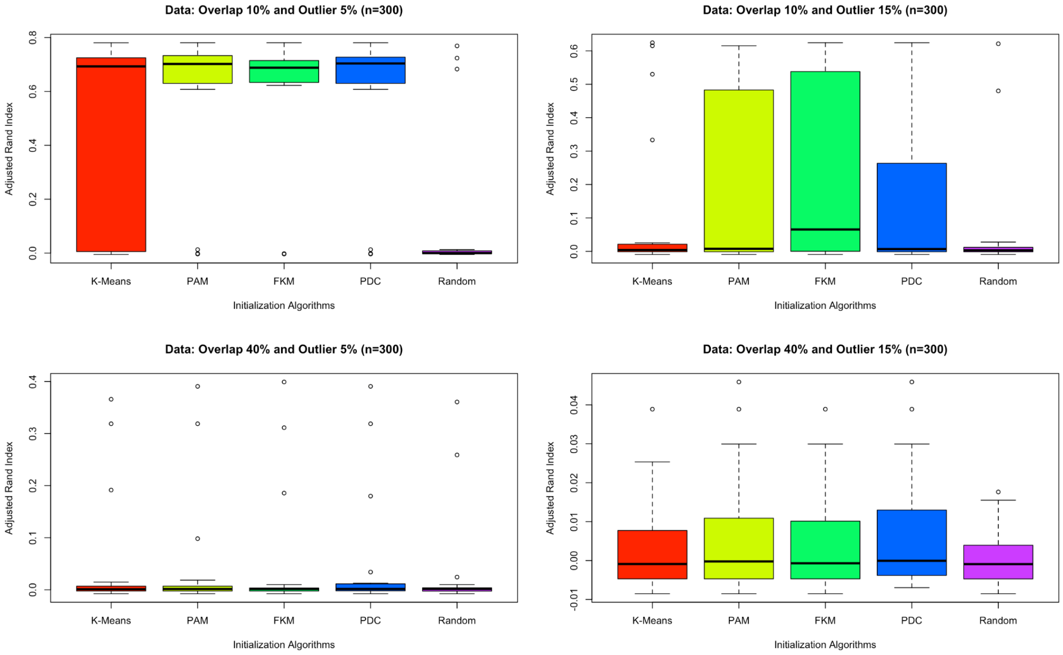


Figure 3 ARI per scenario with percentage of overlap & outliers using GMM on MixSim data sets based on five initialization algorithms

10% Overlap, 5% Outliers - This is the best across the four scenarios. Intuitively, it is reasonable that this scenario has the lowest effects from outliers and overlap because it would have been easier to distinguish components with less overlap and outliers. The median of ARI for K-means, PAM, FKM and PDC are very similar to each other. Higher median and lower interquartile range for PAM, FKM and PDC suggests more stability and robustness in the clustering results. However, a wider range in K-means indicates it is more sensitive to change in the dataset. Random initialization performs the worst as its ARI indicates no better than a random assignment of data points into clusters.

10% Overlap, 15% Outliers – The percentage of outliers increases from 5% to 15% here. The median of ARI for K-means, PAM, PDC, and Random initialization are very similar to each other and FKM has the highest median. Both K-means and Random initialization perform worst as seen in the plot with ARI indicating no better than a random assignment of data points into clusters and their failure to detect outliers. FKM performs best among all with highest median indicating higher agreement or similarity in clustering results, but wider interquartile range indicates that results vary with the change in dataset.

40% Overlap, 5% Outliers - the percentage of overlap increases from 10% to 40% here. All the algorithms have ARI approximately equal to 0 indicating no better than a random assignment of data points into clusters.

40% Overlap, 5% Outliers - This scenario has a comparatively higher percentage of both outliers and overlap. K-means, FKM and Random initialization have negative ARI indicating performance even worse than a random assignment of data points into clusters.

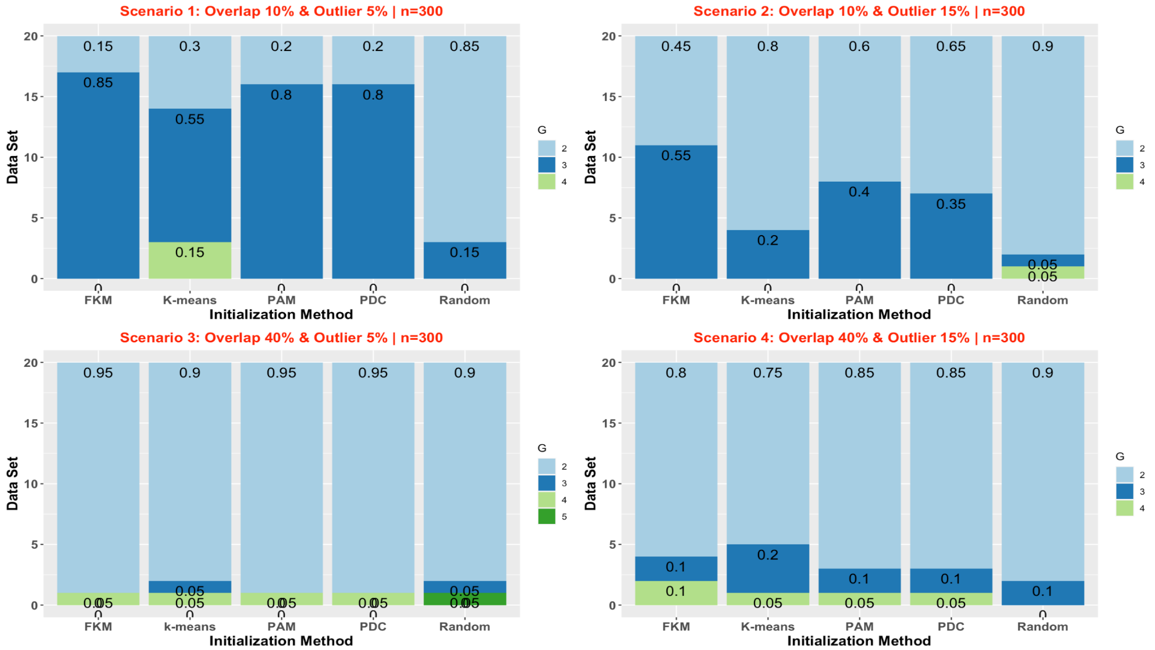


Figure 4 Frequency of best G per scenario with percentage of overlap & outliers using GMM on MixSim data sets based on five initialization algorithms

10% Overlap, 5% Outliers - 80% models of GMM based on PAM and PDC have G=3 and 20% of the models pick G=2. 85% models of GMM based on FKM have G=3 and only 15% of the models pick the best G=2. K-means selects 15% of models with G=4, 55% with G=3 and only 30% with G=2. Randoms selects 85% with G=2 and rest 15% as G=3

10% Overlap, 15% Outliers - 60% models of GMM based on PAM have G=2 and 40% of the models pick G=3. 65% models of GMM based on PDC have G=2 and 35% of the models pick G=3. 45% models of GMM based on FKM have G=3 and only 55% of the models pick the best G=2. K-means selects 80% of models with G=2 and only 20% with G=3. Random selects 90% with G=2 and rest 5% each as G=3,4.

40% Overlap, 5% Outliers - at least 90% models of GMM based on all five algorithms select G=2

40% Overlap, 5% Outliers - 85% models of GMM based on PAM have G=2 and rest 10% each as G=3,4. 85% of models of GMM based on PDC have G=2. 80% of models of GMM based on FKM have G=2. K-means selects 75% of models with G=2. Random selects 90% with G=2.

Multivariate t-Distribution Mixture Model

In this section, we will provide analysis along with two figures in which there are four plots. For clarity, we analyze the performance of the initialization algorithms in the order presented in each plot from left to right.

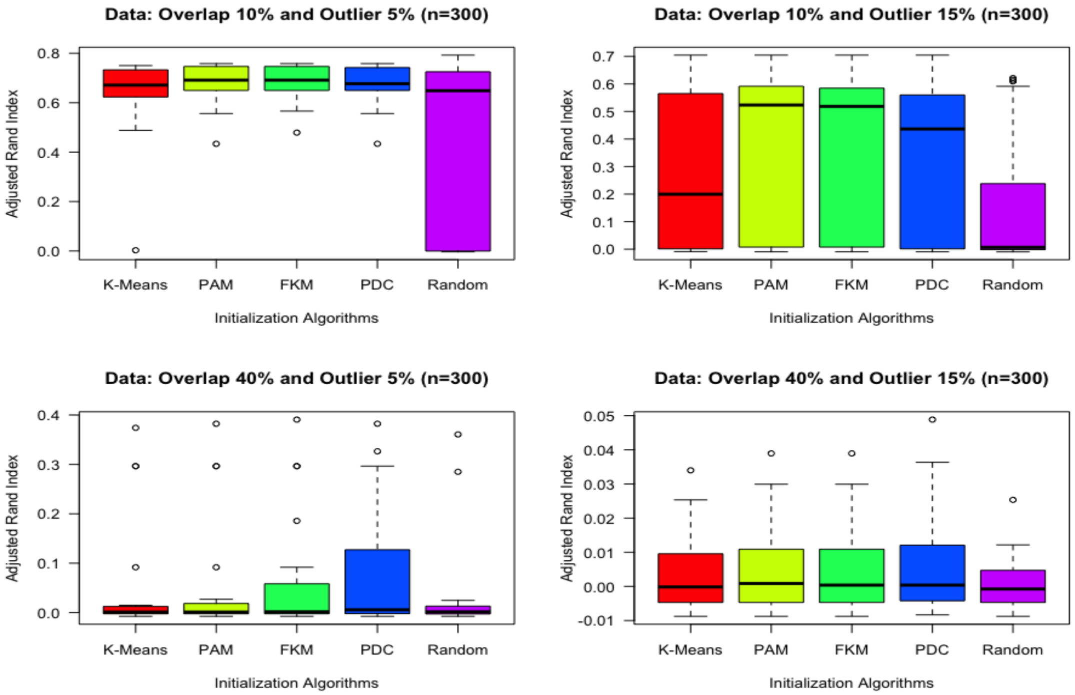


Figure 5 Adjusted Rand Index (ARI) per different mix of data in simulation

Scenario 1 (10% Overlap, 5% Outlier): The average ARI values are found to be 0.639, 0.672, 0.678, 0.671, and 0.423, respectively. From the boxplot, it is clear that the interquartile range is roughly the same for methods 2 through 4, with method 1 showing weaker performance in some datasets and method 5 underperforming in many datasets indicated by a larger range of ARI values. Moreover, methods 1 through 4 show that they agree on their clustering solutions for many of the 20 datasets while method 5 produces different outputs, which is explainable through the random assignment of the data points to clusters. Among the five different methods, fuzzy c-means method outperforms the rest; however, PAM and PDC show relatively similar results as that of fuzzy c-means.

Scenario 2 (10% Overlap, 15% Outlier): The average ARI values are 0.279, 0.388, 0.362, 0.302, and 0.147, respectively. Although the interquartile range for methods 1 through 4 look roughly the same, it is a larger range, which means less agreement on ARI for the 20 datasets. In this scenario, the performance of all five methods drops significantly. Except for the random method, the other four methods have an ARI of 0.56 to 0.59 at the third quartile mark.

Scenario 3 (40% Overlap, 5% Outlier): The average ARI values are 0.0535, 0.0541, 0.0640, 0.0754, and 0.0348, respectively. In this scenario, method 1, 2 and 5 have more agreement over ARI due to a tight interquartile quartile, but this ARI value is near zero which is not appealing at all. K-means, PAM, and Random methods consistently perform badly.

Scenario 4 (40% Overlap, 15% Outlier): The average ARI values are 0.00375, 0.00456, 0.00463, 0.00699, and 0.00130, respectively. The clustering method does not perform well in this scenario either; however, there are a few instances where ARI reaches 0.3 or more. Similar to scenario 3, the median ARI hovers around zero.

Overall, looking at the four scenarios, the five initialization methods perform the best in the first scenario.

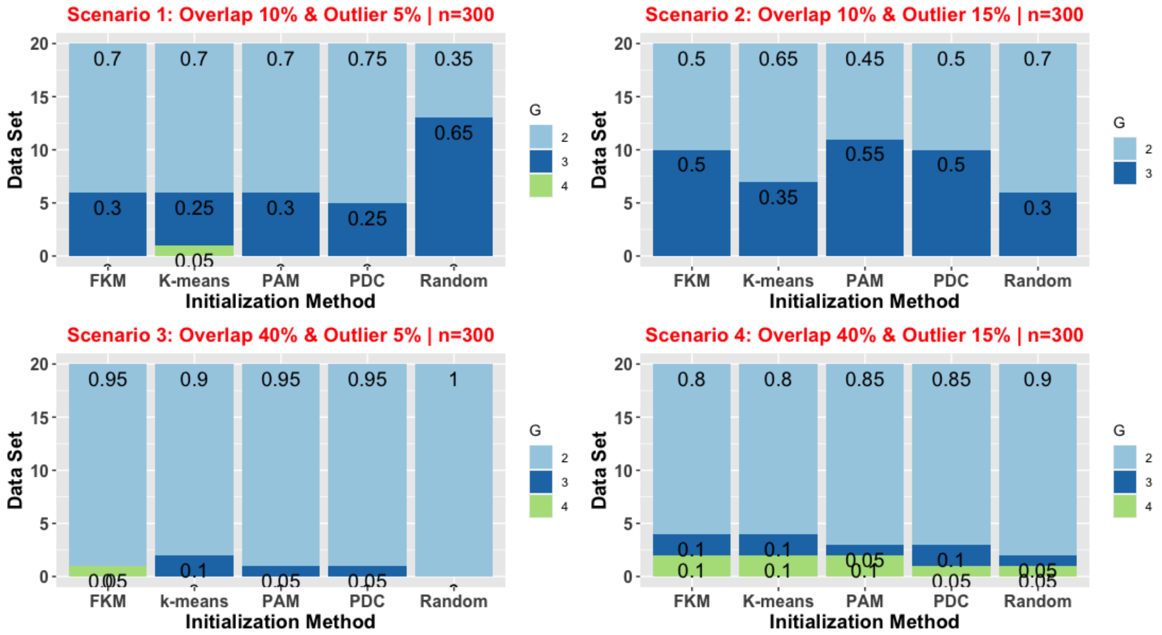


Figure 6 Frequency of selected G per different mix of data in simulation

Scenario 1 (10% Overlap, 5% Outlier): Among the five initialization algorithms, four methods, except for the Random method, picks 2 as the number of clusters for most of the 20 clusters. The next best G is 3. Only the K-means initialization method picks G to be 4 in one dataset (5%). The worst method is random initialization which picks G to be 3 in 60% of the cases.

Scenario 2 (10% Overlap, 15% Outlier): Here, the best G is 2 in K-means, PDC, and Random initialization methods. Both FKM and PAM were not able to give a 2-cluster solution definitively.

Scenario 3 (40% Overlap, 5% Outlier): All five initialization methods select G to be 2 in the majority of the 20 datasets, with FKM, PAM, and PDC leading the way by 95% of the time.

Scenario 4 (40% Overlap, 15% Outlier): Similar to scenario 4, all five initialization methods choose G to be 2. The next best G is 3 for all of them. However, the Random initialization method is the only one that does not choose G to be 4 in any instances.

Even though we see that all five initialization methods choose the best G to be 2 in most of the cases in scenario 3 and 4, they may not cluster the data points correctly, compared to the original labels. This can be explained by the low ARI values in scenario 3 and 4 in the previous section.

Multivariate Contaminated Normal Mixed Model (MCNM)

In the following, the initialization algorithm performance will be analyzed in each scenario using three criteria, ARI, Best G and Outlier Detection, on MixSim data sets and rCN data sets separately. After that, a comparison of the MCNM clustering results between the two data sets will be discussed.

I. MixSim Data Sets. From the perspective of ARI, regardless of the percentage of outlier and initialization algorithm, the data sets with lower overlap (10%, Scenario 1 & 2) between two components generally have much higher ARI than those with higher overlap (40%, Scenario 3 & 4). Meanwhile, it is also interesting to find that, regardless of outlier and initialization algorithm, data sets with higher overlap have a higher percentage to get best G=2. For lower overlap data sets (10%, Scenario 1 & 2), across the five initialization algorithms, approximately 5%-20% get best G=2. For higher overlap data sets (40%, Scenario 3 & 4), about 45%-60% get best G=2. Besides, the overall percentage of outliers that are detected by the best MCNM model in each scenario and based on each initialization algorithm is much lower than its original percentage. **(See Figure 7, 8 & 9)**

Scenario 1: Low overlap (10%) & Low outlier (5%). ARI. The average ARI of best MCNM clustering models for the 20 data sets based on k-means, PAM, FKM, PDC and Random initialization algorithms are 0.68, 0.71, 0.72, 0.69 and 0.70 respectively. All the five median ARI are about 0.75 and distribution ranges are between 0.68-0.75. This is the best across the four scenarios. Intuitively, it is reasonable that this scenario has the lowest effects from outliers. The overlap between two components is also lower. It should be easier to distinguish components' borders than data with higher overlap and more outliers. Generally, there is no big difference in terms of ARI for the MCNM clustering based on the five initialization algorithms . But k-means and PDC models have data sets with 0 ARI.

Best G. For all five initialization algorithms, the most frequent best G is 3 rather than 2. There are 60%, 55%, 70%, 70% and 55% best MCNM models based on FKM, k-means, PAM, PDC, and Random initialization have best G=3. There are 20% best models of MCNM based on PDC and Random initialization have G=2. For the other three initialization algorithms, there are 10% best models that have G=2. **Outlier Detection.** MCNM based on PDC detected the highest percent of outliers, 1.15%. Thus, based on ARI, best G and outlier detection, we pick **PDC** as the best initialization algorithm for MCNM with lower overlap & lower outlier. MCNM based on PDC has an average ARI of 0.69 and 10% of the best models get G=2.

Scenario 2: Low overlap (10%) & High outlier (15%). ARI. The average ARI for the 20 data sets based on k-means, PAM, FKM, PDC and At Random initialization algorithms are 0.31, 0.54, 0.56, 0.56 and 0.55 respectively. Generally, they are lower than that of Scenario 1, but higher than that of Scenario 3 & 4. A 10% increase in outliers probably leads to the decrease in correctly distinguishing the observation labels. The median and distribution range of ARI for FKM, PAM, PDC and Random initialization algorithms are very similar. FKM and PAM have a little bit larger median and 1st quartile ARI. Comparably, clustering based on k-means initialization performs the worst and varies much more than other initialization algorithms. It has more than one data set that has an almost 0 ARI value.

Best G. Most frequent best Gs for MCNM clustering based KFM, k-means, PAM, PDC and Random initializations are 4 (45%), 2 (45%), 4 (40%), 3 (75%) and 3 (70%). K-means has the most frequent best G=2. However, its ARI varies the most across the 20 data sets. Next highest frequent of best G=2 is MCNM based on PAM initialization (10%). **Outlier Detection.** MCNM based on k-means detected the highest percent of outliers, 0.22%. MCNM best models based on other four initialization algorithms averagely detected 0% outliers. Generally, the average percent of detected outliers are very low. Thus, based on ARI, best G and outlier detection, we pick **PAM** as the most stable initialization algorithm for MCNM with lower overlap & higher outlier. MCNM based on PAM has an average ARI of 0.54 and 10% of the best models get G=2.

Scenario 3: High overlap (40%) & Low outlier (5%). **ARI.** The average ARI for the 20 data sets based on k-means, PAM, FKM, PDC and At Random initialization algorithms are 0.15, 0.15, 0.15, 0.16 and 0.16 respectively. The median and distribution range of ARI of the five initialization algorithms are similar to each other as well. Generally, it is much worse than Scenario 1 & 2 and is similar to Scenario 4. PDC has the highest 3rd quartile and median. But FKM, k-means, PAM and PDC have at least one dataset's ARI close to 0.

Best G. For all five initialization algorithms, the most frequent best G is 2. There are 50%, 50%, 50%, 60% and 50% best MCNM models based on FKM, k-means, PAM, PDC, and Random initialization have best G=2. **Outlier Detection.** MCNM based on PDC detected the highest average percent of outliers, 0.75%. Thus, based on ARI, best G and outlier detection, we pick **PDC** as the best initialization algorithm for MCNM with higher overlap & lower outlier. MCNM based on PDC has an average ARI of 0.16 and 60% of the best models get G=2.

Scenario 4: High overlap (40%) & High outlier (15%). **ARI.** The average ARI for the 20 data sets based on k-means, PAM, FKM, PDC and At Random initialization algorithms are 0.06, 0.09, 0.11, 0.10 and 0.12 respectively. All median and 3rd quartiles of ARIs are lower than 0.2. Generally, this Scenario has the lowest ARIs. All five 1st quartiles of ARI are around 0. The FKM and Random initialization algorithm has the largest 3rd quartile and median ARI. K-means has the smallest median ARI.

Best G. For all five initialization algorithms, the most frequent best G is 2. There are 45%, 65%, 55%, 50% and 55% best MCNM models based on FKM, k-means, PAM, PDC, and Random initialization have best G=2. **Outlier Detection.** MCNM based on the Random algorithm detected the highest percent of outliers, 1.95%. Thus, based on ARI, best G and outlier detection, we pick **Random** as the best initialization algorithm for MCNM with higher overlap & higher outlier. As the mixture component overlap and percentage of outliers increase, MCNM clustering based on FKM, k-means, PAM and PDC doesn't perform better than Random initialization. MCNM based on Random has an average ARI of 0.12 and 55% of the best models get G=2.

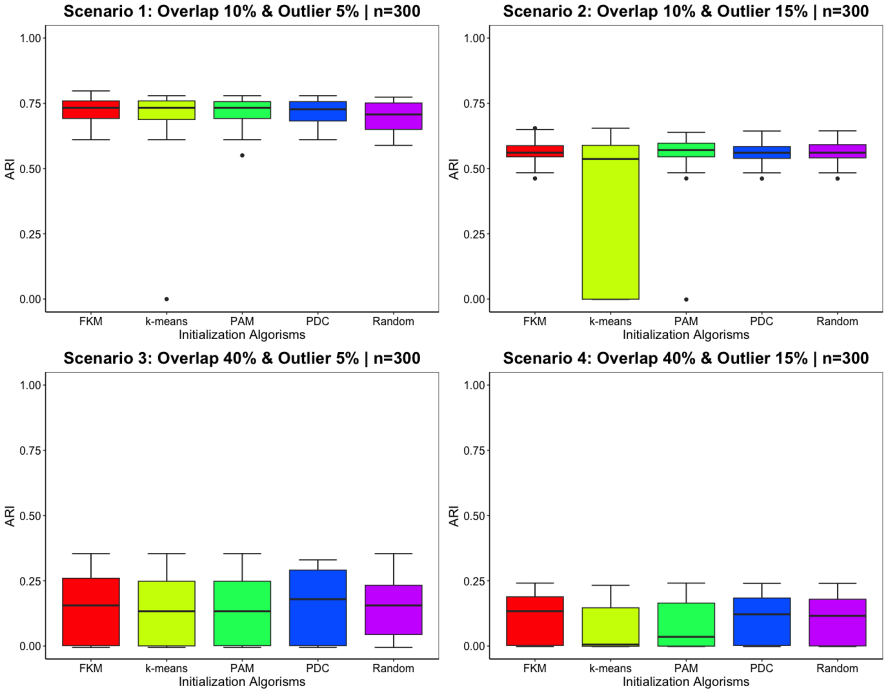


Figure 7 ARI per scenario and percentage of overlap & outliers using MCNM on MixSim data sets based on five initialization algorithms

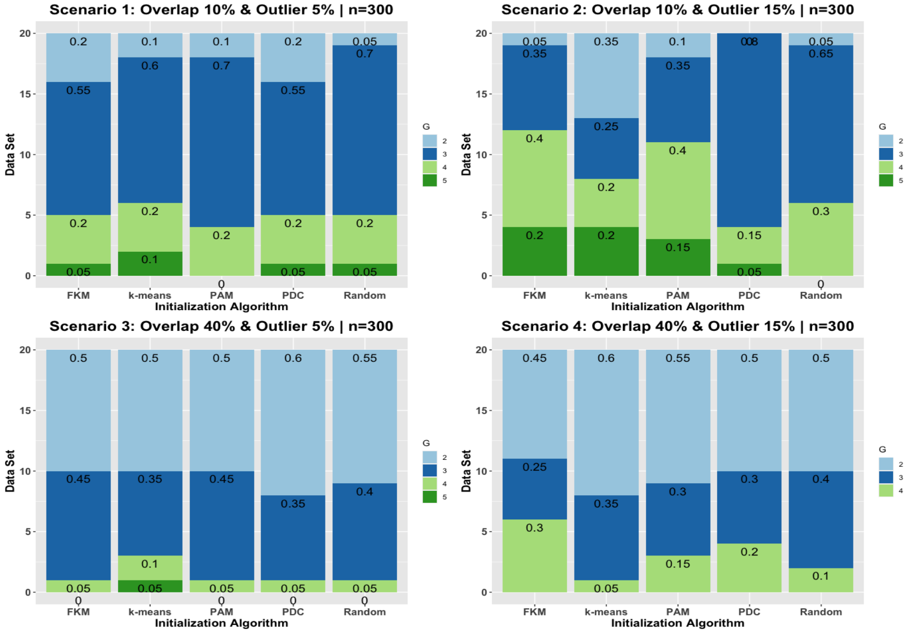


Figure 8 Frequency of best G per scenario and percentage of overlap & outliers using MCNM on MixSim data sets based on five initialization algorithms

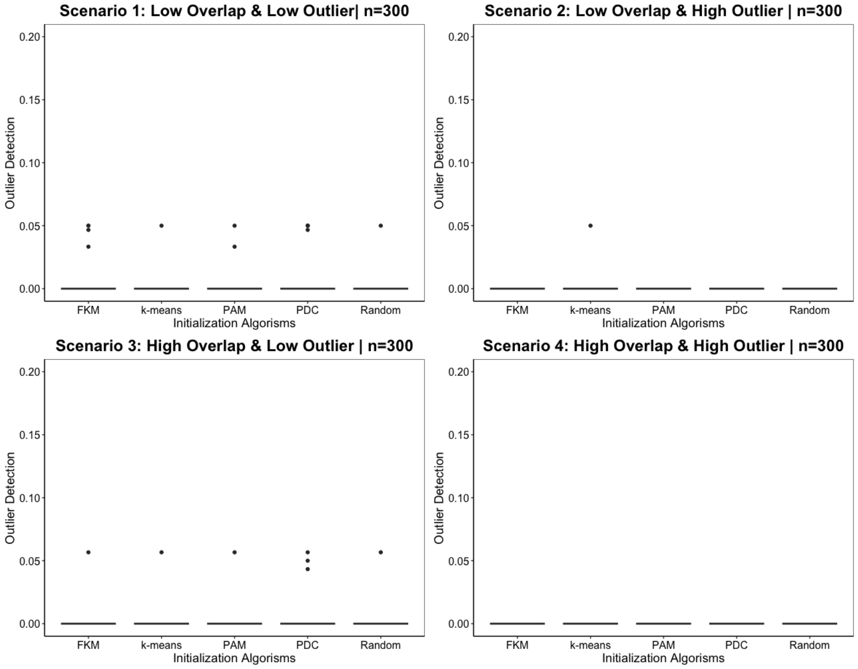


Figure 9 Outlier Detection of MCNM on MixSim data sets based on five initialization algorithms

II. rCN Data Sets. From the perspective of ARI, lower overlap & lower outlier has the best performance and higher overlap & higher outlier scenario performs the worst, regardless of initialization algorithms. Overall, the ARI of MCNM clustering in each Scenario with each initialization algorithm is higher than that of MixSim data sets. For the best G, except Random initialization, in all four Scenarios, the most frequent best G of MCNM clustering based on FKM, k-means, PAM and PDC are G=2. This frequency is very high in Scenario 1, Scenario 2 and Scenario 3. They are all larger than 0.8. It is lower in Scenario 4, though all are larger than 0.5. **(See Figure 10, 11,& 12)**

Scenario 1: Low overlap & Low outliers ($\alpha = 0.95, \eta = 20$)**. ARI.** The average ARI of MCNM clustering for the 20 data sets based on k-means, PAM, FKM, PDC and Random initialization algorithms are 0.94, 0.94, 0.94, 0.94 and 0.92 respectively. All the five initializations have median ARI across 20 data sets of above 0.90 and 1st quantile larger than 0.85. This is also the best across the four scenarios. Generally, there is no big difference in terms of ARI for the MCNM clustering based on the five initialization algorithms. Random initialization performs a little bit worse than the other four initialization algorithms in terms of ARI.

Best G. For all five initialization algorithms, the most frequent best G is 2. There are 85%, 85%, 90%, 90% and 50% best MCNM models based on FKM, k-means, PAM, PDC, and Random initialization have best G=2. **Outlier Detection.** MCNM based on PAM & PDC detected the highest percent of outliers, 3.40%, which is 1.6 percentage lower than the original percentage of outliers. Thus, based on ARI, best G and outlier detection, we pick **PAM & PDC** as the best initialization algorithm for MCNM with lower overlap & lower outlier. MCNM based on both PDC PAM have average ARI of 0.94 and 90% best models get G=2. Both indicators are the highest among the five initialization algorithms.

Scenario 2: Low overlap & High outliers ($\alpha = 0.85$, $\eta = 30$)**. ARI.** The average ARI of MCNM clustering for the 20 data sets based on k-means, PAM, FKM, PDC and Random initialization algorithms are 0.86, 0.86, 0.86, 0.86 and 0.75 respectively. Similar to the median of ARI. Overall, both the average and median ARIs for the MCNM clustering based on the five initialization algorithms are a little bit lower than that of Scenario 1, but the decrease is not that much. There is no big difference in terms of ARI across five initialization algorithms. Random initialization performs a little bit worse than the other four initialization algorithms.

Best G. Except for the Random initialization, MCNM clustering based FKM, k-means, PAM and PDC pick G=2 as the best model in G=2:5 across all 20 data sets. However, 65% MCNM models based on Random initialization algorithms pick G=3 as the best one. **Outlier Detection.** MCNM based on k-means, PAM, FKM & PDC detected almost the same percent of outliers, about 11.50%. It is much higher than that of Random initialization and is about 3.5 percent lower than its original percentage of outliers. Thus, based on ARI, best G and outlier detection, **k-means, PAM, FKM & PDC** perform equally well as initialization algorithms for MCNM clustering. Their average ARI for 20 data sets are 0.86 and 100% pick G=2 as the best model.

Scenario 3: High overlap & Low outliers ($\alpha = 0.95$, $\eta = 20$)**. ARI.** The average ARI of MCNM clustering for the 20 data sets based on k-means, PAM, FKM, PDC and Random initialization algorithms are 0.83, 0.85, 0.85, 0.85 and 0.82 respectively. Similar to the median of ARI. Overall, both the average and median ARIs for the MCNM clustering based on the five initialization algorithms are a little bit lower than that of scenario 1, but the decrease is not that much. There is no big difference in terms of ARI across five initialization algorithms. Random initialization performs a little bit worse than the other four initialization algorithms.

Best G. For all five initialization algorithms, the most frequent best G is 2. There are 95%, 80%, 100%, 100% and 50% best MCNM models based on FKM, k-means, PAM, PDC, and Random initialization have best G=2 across G=2:5. **Outlier Detection.** MCNM based on PAM, FKM & PDC detected the highest percent of outliers, about 3.7%. It is about 1.3 percent lower than its original percentage of outliers. Thus, based on ARI, best G and outlier detection, we pick **PAM & PDC** as the best initialization algorithm for MCNM in this Scenario. Their average ARI for 20 data sets are 0.85 and 100% pick G=2 as the best model.

Scenario 4: High overlap & High outliers ($\alpha = 0.85$, $\eta = 30$)**. ARI.** The average ARI of MCNM clustering for the 20 data sets based on k-means, PAM, FKM, PDC and Random initialization algorithms are 0.71, 0.70, 0.70, 0.70 and 0.67 respectively. k-means has the largest median ARI, while median ARI for FKM, k-means, PAM and PDC initialization are very similar to each other. Overall, both the average and median ARIs for the MCNM clustering based on the five initialization algorithms are a little bit lower than that of Scenario 2 & 3. There is no big difference in terms of ARI across five initialization algorithms. Random initialization performs a little bit worse than the other four initialization algorithms.

Best G. Except for the Random initialization, MCNM clustering based FKM, k-means, PAM and PDC pick G=2 as the best model in G=2:5 across all 20 data sets. There are 50%, 65%, 55%, 55% and 25% best MCNM models based on FKM, k-means, PAM, PDC, and Random initialization pick G=2 across G=2:5. **Outlier Detection.** MCNM based on k-means detected the highest percent of outliers, 7.10%. Thus, based on ARI, best G and outlier detection, we pick **k-means** as the best initialization algorithm for MCNM in this Scenario. Its average ARI for 20 data sets are 0.71 and 65% pick G=2 as the best model.

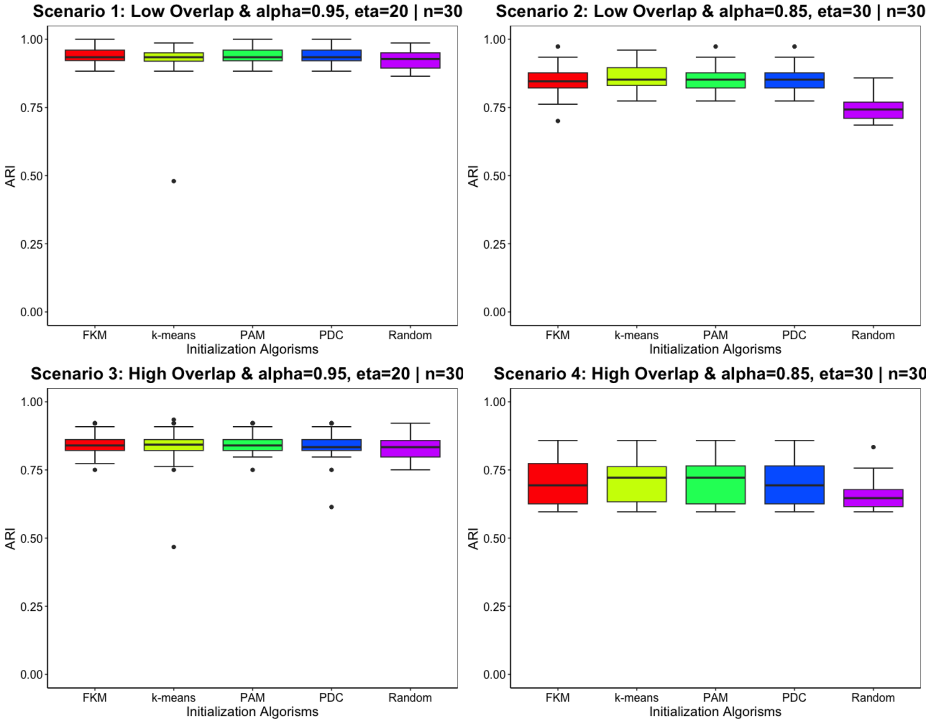


Figure 10 ARI per scenario and percentage of overlap & outliers using MCNM on rCN data sets based on five initialization algorithms

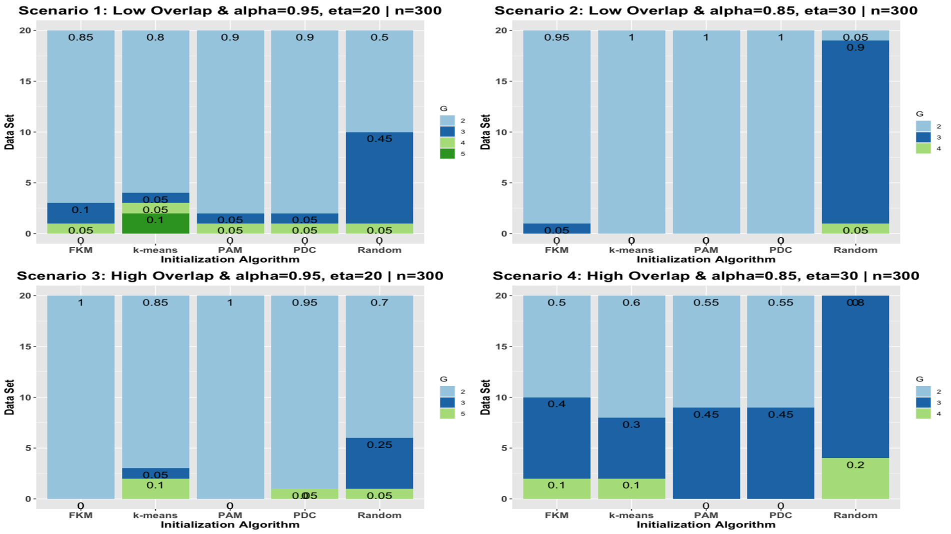


Figure 11 Frequency of best G per scenario and percentage of overlap & outliers using MCNM on rCN data sets based on five initialization algorithms

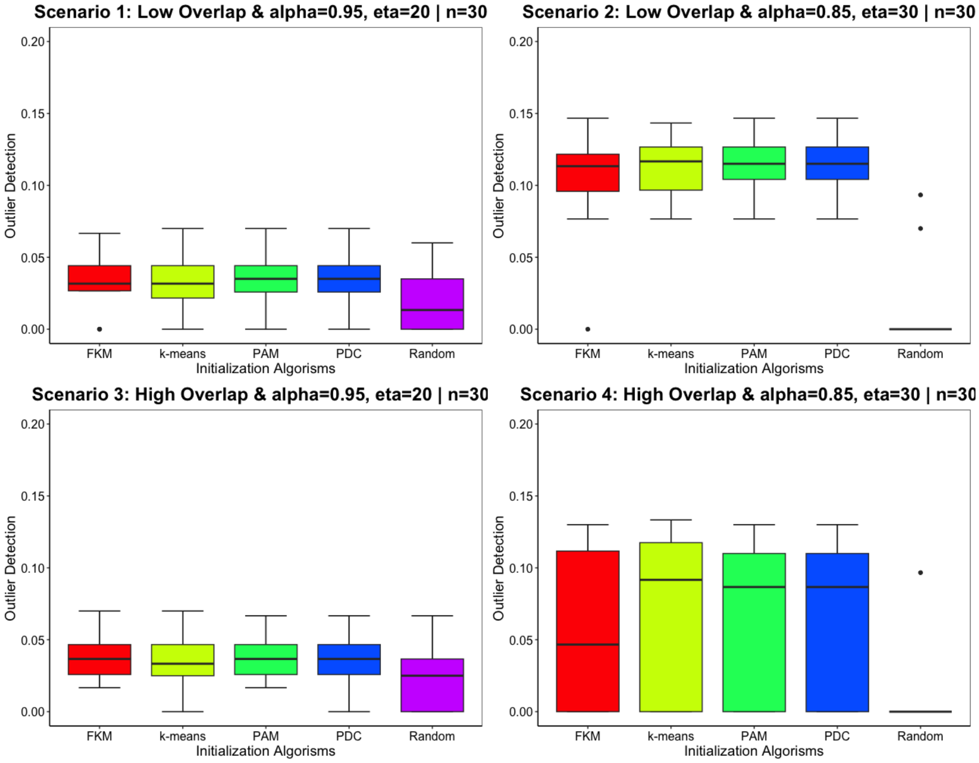
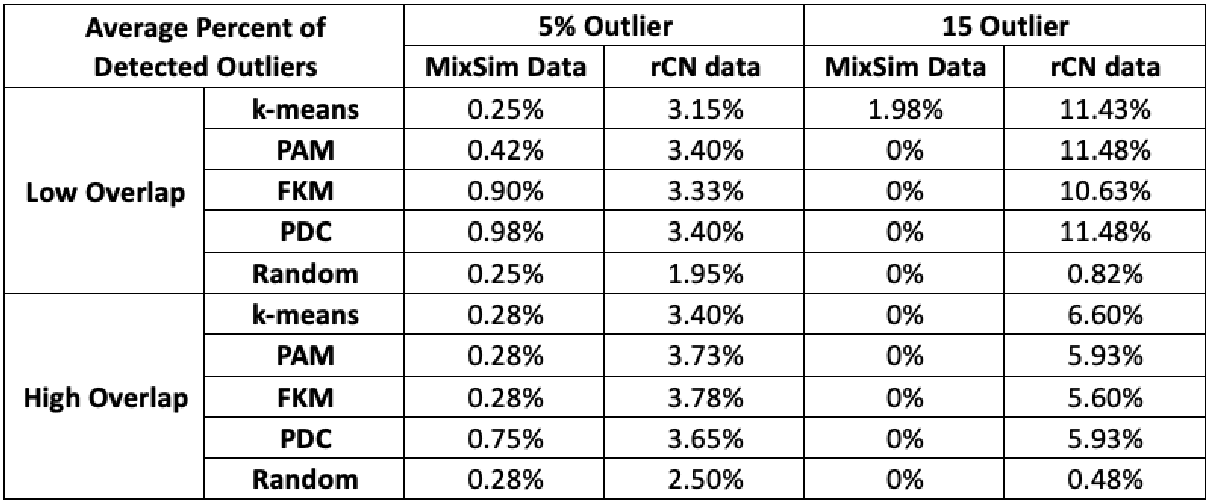


Figure 12 Outlier Detection of MCNM on rCN data sets based on five initialization algorithms

III. MixSim Data Sets vs. rCN Data Sets. From the perspective of average ARI, MCNM clustering using rCN data sets performs much better than that on MixSim data sets, regardless of initialization algorithm and scenarios. Particularly, as the overlap percentage and outlier percentage increase, this performance difference is enlarged. From the perspective of best G, regardless of initialization algorithm and scenario, the percentage of MCNM clustering models that get G=2 as the best model is significantly higher if using rCN data sets than MixSim data sets. Except for the Random initialization algorithm, all MCNM based on the other four initializations in all scenarios have G=2 as the most frequent best Gs. From the perspective of outlier detection, the average percent of outliers detected by MCNM clustering using rCN data sets, regardless of initialization and scenarios, perform much better than that using Mixsim data sets. The percentage of outliers detected by the former is closer to the original percentage of outliers.

The possible reason is that the rCN() function could specify the proportion of good points (α) & the degree of contamination (η), which makes the distribution of simulated data more matches the contaminated normal distribution. Thus, the MCNM model-based clustering could have better performance in general, regardless of initialization algorithm and scenario. Besides, rCN() data sets have a larger frequency to detect G=2 as the best model. It is noticeable that if G=2, the MCNM clustering is more likely to detect the outliers. Otherwise, the probability is greatly decreased. **(See Table 2)**



Average Percent of Detected Outliers		5% Outlier		15 Outlier	
		MixSim Data	rCN data	MixSim Data	rCN data
Low Overlap	k-means	0.25%	3.15%	1.98%	11.43%
	PAM	0.42%	3.40%	0%	11.48%
	FKM	0.90%	3.33%	0%	10.63%
	PDC	0.98%	3.40%	0%	11.48%
	Random	0.25%	1.95%	0%	0.82%
High Overlap	k-means	0.28%	3.40%	0%	6.60%
	PAM	0.28%	3.73%	0%	5.93%
	FKM	0.28%	3.78%	0%	5.60%
	PDC	0.75%	3.65%	0%	5.93%
	Random	0.28%	2.50%	0%	0.48%

Table 2 Average Percent of Detected Outliers

Reference

[1] H. Tong and C. Tortora. Model-based clustering and outlier detection with missing data. Advances in data analysis and classification 1-26, 2022.