# Technique Report

Cheng Qiao, Kenneth N.Brown, Fan Zhang and Zhihong Tian

In this report, we evaluate the performance of proposed method to describe the clusters with nested bounding boxes. The performance is measured by comparing the summary description of generated data points with the ground truth summary description. To measure how close the new summary description is to the ground truth summary description, we calculate the difference of three measurements before and after regeneration:

1) $E_d$: Euclidean Distance between the original centroids and new centroids,
2) $E_s$: shift distance of bounding box (measures the shift distance of four coordinates (leftmost, rightmost, bottom, top) of the box), and
3) $E_r$: rotation degree from old PCs to new PCs.

Actually, the arrows plotted in Figure 1 should be two-way arrows, which means we could rotate the PCs from both direction. We define the rotation degree as the minimum rotate degree of the four directions of two PCs to remap new PCs from old PCs. Since the underlying data distribution, the number of data points and percentiles play crucial roles in the performance of proposed method, we measure the performance over these parameters.

## I. Measurement with different underlying distributions

In this section, we evaluate the performance with different underlying distributions. First, we consider the scenario that data points are generated from binormal distributions, where the standard deviations are varied. Table I shows that, the shift distance of centroids and bounding boxes have slightly changed. The centroids move by only a small distance (below 0.07 with a small standard deviation around 0.04). In term of rotation degree, the worst average rotation is less than 3.5 degree, which indicates that the new PCs are quite faithful to the original PCs. The shift distance of boxes increase when the range of variance decrease. Meanwhile, the rotation degree required to remap new PCs increase when the range of variance decrease.

| S.D | Measures | Mean | S.D |
|---|---|---|---|
| 5*$Sig_1$= 0.6,$Sig_2$= 0.1 | $E_d$ | 0.063 | 0.043 |
| | $E_r$ | 0.175 | 0.87 |
| | $E_s$ (100% box) | [0.117, 0.099,0.076, 0.074] | 0.062 |
| | $E_s$ (80% box) | [0.087,0.085,0.047,0.043 ] | 0.046 |
| | $E_s$ (40% box) | [[0.086, 0.063, 0.075, 0.072] | 0.052 |
| 5*$Sig_1$= 0.5,$Sig_2$= 0.2 | $E_d$ | 0.044 | 0.031 |
| | $E_r$ | 0.76 | 2.85 |
| | $E_s$ (100% box) | [0.101, 0.137,0.081, 0.071] | 0.065 |
| | $E_s$ (80% box) | [0.114,0.14,0.081,0.079 ] | 0.07 |
| | $E_s$ (40% box) | [[0.077, 0.071, 0.068, 0.071] | 0.041 |
| 5*$Sig_1$= 0.4,$Sig_2$= 0.3 | $E_d$ | 0.050 | 0.031 |
| | $E_r$ | 3.51 | 11.71 |
| | $E_s$ (100% box) | [0.191, 0.327,0.17, 0.258] | 0.21 |
| | $E_s$ (80% box) | [0.177, 0.251, 0.165, 0.22 ] | 0.18 |
| | $E_s$ (40% box) | [[0.187, 0.205, 0.179, 0.205] | 0.14 |

TABLE I: Measurement for different normal distributions with standard basis

Then data points generated from uniform distributions are considered. To ensure that the comparison is fair, data points generated from normal or uniform distributions should be located in same range. Each time a set of data is generated from normal distributions, the 100% bounding box are saved, and it is used to generate another set of data uniformly. Figure 1 shows that data created from normal distributions and its 100% bounding box. Based on the green bounding box, another set of data was generated uniformly(refer to Figure 4), so both generated data are ready to compare.

Following the steps we have done previously, the next step is generating new set of data based on the summary description of created original data. Here we consider two different methods to generate data points here:

1) Randomly generate an x coordinate and find its right place by comparing with the percentiles, then generate a random number in the corresponding place of the percentile value. Data points are uniformly generated in 5 * 5 cells. This method is denoted by M1.
2) The second method is quite similar to the M1 but data points are randomly generated in only one cell (the 100% bounding box), and it was denoted by M2.

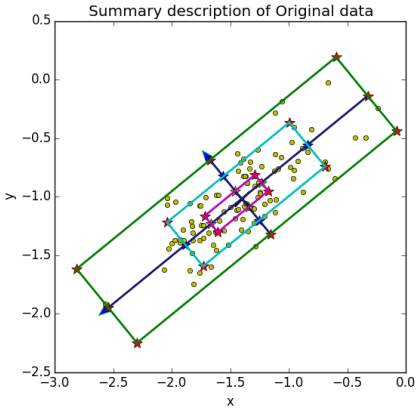


Fig. 1: Original data (Normal)

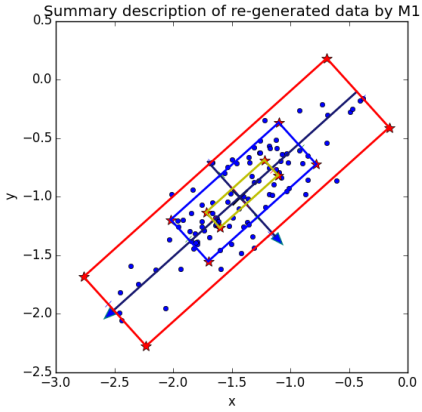


Fig. 2: Original data (Uniform)

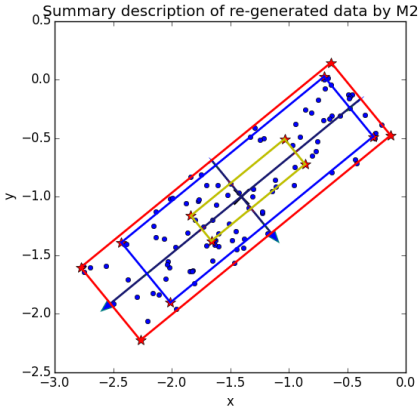


Fig. 3: re-generated data by M2 (Normal)

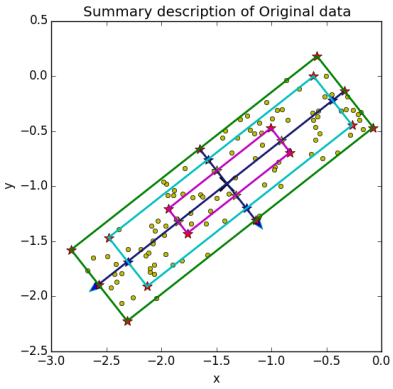


Fig. 4: Original data (Uniform)

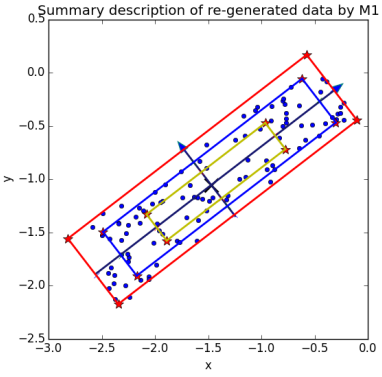


Fig. 5: re-generated data by M1 (uniform)

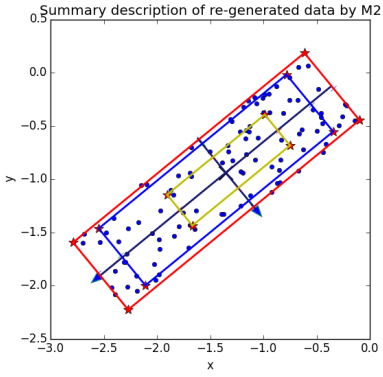


Fig. 6: re-generated data by M2 (uniform)

Figure 2 and 3 show the re-generated data by method M1 and M2 separately according to the summary description of data points created from normal distributions in Figure 1. It shows that the re-generated data by M1 is similar to the created data. Figure 5 and 6 show the re-generated data by method M1 and M2 separately according to the summary description of data points created from uniform distributions in Figure 4. Results shows that there is a little difference between the data generated by M1 and M2.

Table II shows that shift distance of bounding boxes and centroids of data points generated by normal distributions is larger than that by uniform distributions. However, the standard deviation of rotation degree is much smaller. Compared normal (M1) with normal (M2), we can see that the M1 performs better than M2. We expect this result, since more information are provided. However, there is a little difference between uniform (M2) and uniform (M1). The reason behind it is that there is a little difference between randomly generate data points in a big box and that in multiple sub-boxes when the underlying dataset is uniform.

To conclude, the summary descriptions do appear to be stable after data regeneration, changing the coordinates, shape and rotation by only small amounts. Descriptions for clusters that are elliptical are more stable than those for circular clusters.

## II. Measurement with different number of data points

In this section, we use different number of data points to measure the performance. Data points are generated from different distributions but with a fixed standard deviation. Table III shows that the shift distance between centroids increase when the number of data points decrease from 200 to 50. But for N= 50, the shift distance is still low. The rotation is low and the boxes have moved, though, so 50 is close to the lower limit. Note that although the mean of rotation is low for N=50, the standard deviation is higher, and increases as N decreases.

Similarly, the shift distance between bounding boxes increase as well. However, the rotation degree required to remap is increased but still less than 1. That is because the density level will be increased when the size of cluster turns large. Then the case that shape of cluster is decided by some points in the margin area could be avoided to a large extent. To summarise, we could get a similar new generated dataset even the size is small.

| Distributions | Measures | Mean | S.D |
|---|---|---|---|
| 5*Normal (M1) | $E_d$ | 0.047 | 0.029 |
| | $E_r$ | 0.03 | 2.559 |
| | $E_s$ (100% box) | [0.078, 0.089, 0.061, 0.062 ] | 0.048 |
| | $E_s$ (80% box) | [0.091, 0.109, 0.072, 0.06 ] | 0.051 |
| | $E_s$ (40% box) | [0.068, 0.078, 0.052, 0.048] | 0.037 |
| 5*Normal (M2) | $E_d$ | 0.137 | 0.106 |
| | $E_r$ | 1.67 | 2.96 |
| | $E_s$ (100% box) | [0.055, 0.053, 0.123, 0.123 ] | 0.073 |
| | $E_s$ (80% box) | [0.262, 0.482, 0.199, 0.204 ] | 0.131 |
| | $E_s$ (40% box) | [0.207, 0.388, 0.178, 0.171] | 0.111 |
| 5*Uniform (M1) | $E_d$ | 0.045 | 0.027 |
| | $E_r$ | 0.100 | 4.270 |
| | $E_s$ (100% box) | [0.038, 0.04, 0.036, 0.035] ] | 0.025 |
| | $E_s$ (80% box) | [0.055, 0.075, 0.04, 0.047 ] | 0.036 |
| | $E_s$ (40% box) | [0.075, 0.078, 0.056, 0.06] | 0.042 |
| 5*Uniform (M2) | $E_d$ | 0.095 | 0.064 |
| | $E_r$ | 0.668 | 4.656 |
| | $E_s$ (100% box) | [0.036, 0.043, 0.086, 0.086] ] | 0.047 |
| | $E_s$ (80% box) | [0.095, 0.08, 0.104, 0.098 ] | 0.067 |
| | $E_s$ (40% box) | [0.125, 0.136, 0.1, 0.115] | 0.078 |

TABLE II: Measures varied with underlying distributions

| N | Measures | Mean | S.D |
|---|---|---|---|
| 5*50 | $E_d$ | 0.071 | 0.039 |
| | $E_r$ | -0.014 | 4.468 |
| | $E_s$ (100% box) | [0.12, 0.243, 0.114, 0.14] | 0.116 |
| | $E_s$ (80% box) | [0.129, 0.189, 0.121, 0.126 ] | 0.091 |
| | $E_s$ (40% box) | [0.121, 0.139, 0.137, 0.121] | 0.077 |
| 5*100 | $E_d$ | 0.051 | 0.026 |
| | $E_r$ | 0.06 | 2.98 |
| | $E_s$ (100% box) | [0.076, 0.144, 0.075, 0.086 ] ] | 0.066 |
| | $E_s$ (80% box) | [0.126, 0.11, 0.082, 0.076 ] | 0.087 |
| | $E_s$ (40% box) | [0.08, 0.076, 0.071, 0.07] | 0.049 |
| 5*200 | $E_d$ | 0.039 | 0.020 |
| | $E_r$ | 0.78 | 1.98 |
| | $E_s$ (100% box) | [0.079, 0.099, 0.065, 0.057] | 0.049 |
| | $E_s$ (80% box) | [0.098, 0.087, 0.075, 0.053] | 0.048 |
| | $E_s$ (40% box) | [0.065, 0.06, 0.06, 0.057] | 0.032 |

TABLE III: Measures varied with number of data points

## III. Measurement with Different Percentiles

In this section, we investigate how different percentiles for the bounding boxes affect the measurement. Data points are generated from different distribution but with a fixed standard deviation. Table IV shows that performance achieved by different percentiles. Comparing the first four experiments, the overall performance achieved by percentile arrays [0.0,0.1,0.3 0.7,0.9,1.0] and [0.0,0.1,0.3,0.7,0.9,1.0] are a little better. The rotation degree by [0.0,0.2,0.4,0.6,0.8,1.0] is higher and the standard deviation is higher as well. Comparing the shift distance of the $100\%$ bounding boxes, we could know shift distance for [0.0,0.05,0.3,0.7,0.95,1.0] is smaller than the others, followed by [0.0,0.1,0.3,0.7,0.9,1.0]. However, the rotation degree required is higher. The most likely reason lies in percentile value we choose. In this paper, we use the percentile [0.0, 0.1,.0.3, 0.7,0.9, 1.0] to describe the nested bounding boxes. The PCs are mainly effected by the boundary points, which located inside the $100\%$ bounding box but outside the second inner bounding boxes.

In the later two rows, we increase the number of bounding boxes. We see that the rotation degree, shift distance of centroids, shift distance of bounding boxes are worse when we increase the number. There appears to be little advantage in generating more bounding boxes in the summaries.

| Percentile | Measures | Mean | S.D |
|---|---|---|---|
| [0.0,0.1,0.3 0.7,0.9,1.0] | $E_d$ | 0.057 | 0.03 |
| | $E_r$ | -0.472 | 2.67 |
| | $E_s$ (100% box) | [0.073, 0.096, 0.063, 0.065] ] | 0.043 |
| | $E_s$ (80% box) | [ 0.099, 0.092, 0.071, 0.073 ] | 0.06 |
| | $E_s$ (40% box) | [0.079, 0.08, 0.065, 0.062] | 0.043 |
| [0.0,0.05,0.3 0.7,0.95,1.0] | Shift $E_d$ | 0.054 | 0.03 |
| | $E_r$ | -0.692 | 3.37 |
| | $E_s$ (100% box) | [0.048, 0.077, 0.049, 0.051] ] | 0.044 |
| | $E_s$ (90% box) | [ 0.184, 0.231, 0.11, 0.084] | 0.082 |
| | $E_s$ (40% box) | [0.078, 0.067, 0.066, 0.06] | 0.04 |
| [0.0,0.2,0.4, 0.6,0.8,1.0] | Shift $E_d$ | 0.038 | 0.022 |
| | $E_r$ | 0.329 | 2.72 |
| | $E_s$ (100% box) | [0.08, 0.096, 0.051, 0.049] | 0.064 |
| | $E_s$ (60% box) | [0.162, 0.157, 0.061, 0.078] | 0.042 |
| | $E_s$ (20% box) | [0.074, 0.061, 0.041, 0.044] | 0.025 |
| [0.0,0.3,0.4, 0.6,0.7,1.0] | Shift $E_d$ | 0.042 | 0.019 |
| | $E_r$ | 1.02 | 3.87 |
| | $E_s$ (100% box) | [0.092, 0.082, 0.054, 0.061] | 0.054 |
| | $E_s$ (40% box) | [0.113, 0.166, 0.068, 0.064 ] | 0.03 |
| | $E_s$ (20% box) | [ 0.052, 0.071, 0.047, 0.04] | 0.023 |
| [0.0,0.1,0.3,0.4 0.6,0.7,0.9,1.0] | Shift $E_d$ | 0.217 | 0.032 |
| | $E_r$ | 0.519 | 3.51 |
| | $E_s$ (100% box) | [0.091, 0.179, 0.203, 0.21 ] ] | 0.054 |
| | $E_s$ (80% box) | [ 0.15, 0.445, 0.204, 0.279] | 0.072 |
| | $E_s$ (40% box) | [0.1, 0.385, 0.204, 0.25] | 0.037 |
| [0.0,0.1,0.2,0.4 0.6,0.8,0.9,1.0] | Shift $E_d$ | 0.259 | 0.046 |
| | $E_r$ | -0.43 | 2.69 |
| | $E_s$ (100% box) | [0.119, 0.148, 0.24, 0.245 ] ] | 0.061 |
| | $E_s$ (80% box) | [ 0.128, 0.438, 0.245, 0.3] | 0.079 |
| | $E_s$ (60% box) | [ 0.144, 0.447, 0.244, 0.294] | 0.046 |

TABLE IV: Measures varied with percentiles

## IV. Possible issue caused by regeneration

Based on the previous sections, regenerating data from the bounding box descriptions does seem to be feasible, and would allow an agent to generate similar clusters of data. But there may be an issue if this is done repeatedly. If a bounding box extends to exactly the bounds of the extreme points, when we regenerate points inside the box, we are unlikely to generate points on the edges, and so the spread of the cluster shrinks after each regeneration, and repeated regeneration will make it worse. In order to address this issue, we considered four potential solutions:

1) $P_1$: do nothing.
2) $P_2$: impose 4 data points on the extremes of the outer box during regeneration.
3) $P_3$: adjust the bounding boxes on creation by expanding the inner boxes to midway to the next relevant point.
4) $P_4$: combine step 3 and 4 above.

Figure 7 shows that the proposed algorithm without any make-up plan ($P_1$) and with bounding line extension outperforms other schemes ($P_3$). Comparing to $P_1$, there is little difference in convergence time, but the moving distance of three bounding boxes and rotation increase, and accuracy decreases when we make up the shrink of outer box only ($P_2$). If we extend the inner boxes to midway to the next relevant point ($P_3$), the convergence time decreases. Although there is slightly increase in rotation (but with a higher standard deviation), the accuracy still remains high. When step 3 and 4 combined ($P_4$), the rotation increases with a higher standard deviation and accuracy decrease. The most possible reason lies in the re-generated data. Without any shrink make-up plan, data points of different clusters are tender to be more separated from each other, that is the reason why the clustering accuracy is high by this scheme.

Figure 7 shows that proposed algorithm without any plan to make up the shrink outperforms that with bounding line extension when clusters are well separated from each other (the first two rows). However, when there are overlaps among clusters, the later scheme outperforms the previous scheme ( see Figure 8. In some practical applications, there are slightly overlaps, even huge overlaps between clusters, we used the scheme that extending the bounding line to half way to the closet points to address the shrink up problem.

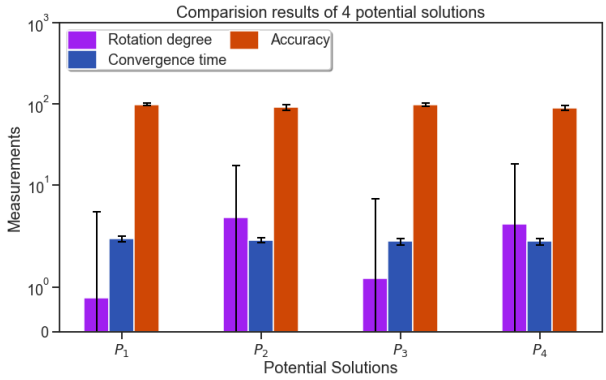


Fig. 7: Comparison results on elliptical well separated clusters

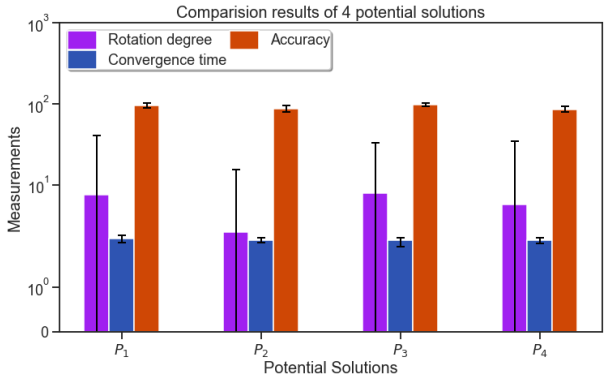


Fig. 8: Comparison results on overlapped clusters

## V. Summary

To conclude, we could generate a similar cluster based on its summary description generated by our method. Experiments show that, descriptions for clusters that are elliptical are more stable than those for circular clusters. The rotation degree required to remap the new PCs from old PCs is heavily relying on the boundary points. One possible approach to improve the quality of generated cluster is choosing an appropriate percentile. However, there is a shrink problem about the bounding boxes of regenerated data points. This problem can be addressed by extending the bounding line of inside boxes to half way to the closest points.