Experiments with Maximin Sampling

Omar A. Ibrahim[1], James Keller[1], James C. Bezdek[1], Mihail Popescu[2]
University of Missouri , [1]ECE Dept. / [2]HMI Dept.
Columbia, MO, USA
oai9bc@mail.missouri.edu,{ kellerj, bezdekj, popescum}@missouri.edu

***Abstract*—To apply clustering algorithms to big data, or to build clustering ensembles, it is a standard process to sample the original data set in a way that hopefully spans the original distribution. There are at least six ways to initialize the *Maximin* (MM) sampling algorithm. This paper contains experiments to determine whether samples produced by the six methods differ significantly; and whether they are superior to simple random sampling. Empirical evidence supports two conclusions. First, there is not enough difference in MM samples generated by the six initializations to support using any but the least costly method: viz., using the first sample in the data as the first MM point. Second, unless the input data have subsets (clusters) that are compact and separated in a well-defined sense, random sampling is demonstrably superior to MM sampling for even small data sets.**

Keywords—Maximin sampling, Dunn's index, Cluster Analysis

I. INTRODUCTION

We have a collection of objects, $O=\{o_1,\ldots,o_N\}$ which may be almost anything, e.g., guitars, soccer players, medical treatments, stock market reports, samples of beer, etc. The objects can be represented by numerical data in one of two ways. If each object is associated with a set of p measurements, the data is feature vector data, $X_N=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}\subset\Re^p$. Or, pairs of objects may be represented by a relationship $\rho(o_i,o_j)$ between them, giving us relational data in the form of an $N\times N$ matrix $D_N=[d_{ij}]=[\rho(o_i,o_j)]$. The relation $\rho(o_i,o_j)$ is usually a similarity or dissimilarity relation. Any vector norm $\|\mathbf{x}\|$ on $\Re^p$ can be used to transform X_N to $D_N=\left[\|\mathbf{x}_i-\mathbf{x}_j\|\right]$.

There are many methods for clustering static data sets based on classical [1-4], fuzzy [5], probabilistic [6] and possibilistic models [7]. Suppose that N is so large that clustering the data directly with one of the many available methods is intractable (or impossible). How large is this? No matter how big your computer is, there are data sets of interest that cannot be mounted, much less processed, in the usual manner for literal (or exact) clustering. What to do?

An oft-used approach is based on sampling. Clustering is done on the sample, followed by (non-iterative) extension to the remainder of the data to approximate direct (or literal) clustering in the big data. Random sampling is by far the best known method for this approach. Progressive sampling [8] has also been used [9, 10]. A third approach is based on *Maximin* (MM) sampling (cf. Section II). An extended form of MM sampling called MMRS comprising MM sampling followed by a random sampling step has been applied to the problem of approximate clustering in big data [11].

At least six different ways to initialize the MM sampling algorithm have appeared in the literature. The objective of this article is to study the quality of MM samples produced by the six initializations, with a view towards identifying the "best" way to initialize MM sampling. Towards this end, we use two measures that give meaning to the term "best" MM sample in the context of cluster analysis. Section II presents the MM algorithm. Section III exhibits the relationship of MM sampling to Dunn's cluster validity index. Section IV describes the data sets and quality measures used for the experiments. Section V contains our numerical studies, and Section VI presents our conclusions.

II. THE MAXIMIN ALGORITHM

The idea of (MM) sampling apparently first appeared in 1953 [12], where it is described as a procedure for initializing a set of c prototypes (aka cluster centers). Casey and Nagy [13] provide this summary of how to use the MM algorithm to construct initial prototypes.

> *The first sample in the batch to be processed is designated cluster center number one. The distances of the remaining samples from this one are calculated, and the farthest sample is called center number two. The smaller of the two distances from each sample to these two centers is listed, and the sample having the greatest minimum distance is selected. The remaining centers are chosen in turn to have maximum separation from the existing centers. These initial cluster centers are well scattered over the sample space, an intuitively desirable property.*

Many authors have used the MM sampling scheme to facilitate initialization of a clustering or approximate clustering algorithm. Here is a typical specification of the MM for both object data or relational data.

1	In: metric $d:\Re^p\times\Re^p\mapsto\Re^+$: $X_N=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}\subset\Re^p$ *(or)* $D_N=[d_{ij}]$: c' = desired # of MM samples
2	Initialize: $X_{MM}=\varnothing$ *(or)* $O_{MM}=\varnothing$:
3	If X_N : $\mathbf{x}_{m_0}=\text{rand}\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$:

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

	If $D_N : m_0 = \text{rand}\{1,\ldots,N\}$:
4	$\mathbf{Z}=(z_1,\ldots,z_N)=(d(\mathbf{x}_{m_o},1),\ldots,d(\mathbf{x}_{m_o},N))$ *(or)* $(d_{m_0,1},\ldots,d_{m_0,N})$:
5	For $t \leftarrow 1$ to c' do
6	$\mathbf{Z}=(\min\{z_1,d(\mathbf{x}_{m_{t-1}},\mathbf{x}_1)\},\ldots,\min\{z_N,d(\mathbf{x}_{m_{t-1}},\mathbf{x}_N)\})$ *(or)* $\mathbf{Z}=(\min\{z_1,d_{m_{t-1},1}\},\ldots,\min\{z_N,d_{m_{t-1},N}\})$:
7	$m_t = \underbrace{\arg\max}_{1\le j\le N}\{z_j\}$:
8	$X_{MM} = X_{MM} \cup \{\mathbf{x}_{m_t}\}$ *(or)* $O_{MM} = O_{MM} \cup [o_{m_t}]$:
9	End for
10	Out: c' MM indices $M'=\{m_1,\ldots,m_{c'}\}$: c' MM samples $X_{MM} = \{\mathbf{x}_{m_1},\ldots,\mathbf{x}_{m_{c'}}\} \subset \Re^p$ *(or)* c' MM objects $O_{MM} = \{o_{m_1},\ldots,o_{m_{c'}}\} \subset O_N$

Algorithm 1. Exact Maximin Sampling

Ties in Line 6 are broken arbitrarily. The literature contains a number of ways to initialize MM sampling in Line 3. Here is a (possibly incomplete) list of six ways that exact MM sampling has been initialized in the feature vector case:

$$\mathbf{x}_{m_o} = \text{rand}\{\mathbf{x}_1,\ldots,\mathbf{x}_N\} = \textit{a random point in } X_N \tag{1}$$

$$\mathbf{x}_{m_o} = \mathbf{x}_1 = \textit{the first point in } X_N \tag{2}$$

$$\mathbf{x}_{m_o} = \bar{\mathbf{x}} = \sum_{j=1}^{N} \mathbf{x}_j / N = \textit{the grand mean of } X_N \tag{3}$$

$$\mathbf{x}_{m_o} = \text{the point in } X_N \textit{ furthest from } \bar{\mathbf{x}} \tag{4}$$

$$\{\mathbf{x}_{m_o}, \mathbf{x}_{m_1}\} = \text{the 2 points in } X_N \textit{ furthest apart} \tag{5}$$

$$\mathbf{x}_{m_o} \ni m_0 = \underset{1\le j\le N}{\arg\max}\left\{\|\mathbf{x}_j\|_2^2\right\} \tag{6}$$

Random selection of the starting index at line 3 in Algorithm 1 all but guarantees that repeated runs of the MM algorithm will lead to different sets of MM samples. More generally, using each of the initialization methods (1) to (6) may result in different MM samples. The objective of this study is to determine whether one of the six initialization schemes is superior to the others in terms of good samples for cluster analysis. Method (3) starts at $\bar{\mathbf{x}} \notin X$. This necessitates a slight change in our testing scheme discussed in Section IV.

What are good samples for cluster analysis? We believe the primary requirement is that the cluster proportions in the c' samples from X_N should be as close as possible to the corresponding proportions for each subset in X_N. [If the proportions are regarded as prior probabilities, this amounts to requiring that the sample priors to match the parent priors, where we regard as X_N the parent.] This belief guides our choice of methods for determining which set of samples from a set of sample candidates should be called "best." For comparison, we will include a completely random sample (RS) in each of our tests. Our belief is that as N increases, the quality of MM samples compared to RS will decrease, and at some (data driven) crossover point, RS will be better in terms of distributional quality (RS is clearly superior to all MM methods in terms of CPU time). Next, we relate the quality of MM samples to a theorem concerning their distribution.

III. Labels, Partitions and Dunn's Index

Let c be an integer, $1 < c < n$. The *crisp c-partitions* of n objects are matrices U in $M_{hcn} = \{U \in \Re^{cn} : 0 \le u_{ik} \le 1 \; \forall i,k;$ $\sum_i u_{ik} = 1 \; \forall \; k; \sum_k u_{ik} > 0 \; \forall \; i\}$. An equivalent representation is $X_N = \bigcup_{i=1}^{c} X_i ; X_i \cap X_j = \varnothing \; \forall \; i \ne j$, where $\{X_i\}$ are the crisp subsets comprising the c clusters. We write $U \leftrightarrow \{X_i\}$ .Dunn [14] defined an internal cluster validity index for $U \leftrightarrow \{X_i\}$ based on the geometric rationale that "good" partitions of X_N have compact and separated subsets. To understand this index let S and T be non-empty subsets of $\Re^p$, and let $d : \Re^p \times \Re^p \mapsto \Re^+$ be any metric. The *diameter Δ* of S is $\Delta(S) = \underset{\mathbf{x},\mathbf{y}\in S}{\underbrace{\max}}\{d(\mathbf{x},\mathbf{y})\}$ and the *set distance* δ between S and T is $\delta(S,T) = \underset{\substack{\mathbf{x}\in S\\ \mathbf{y}\in T}}{\underbrace{\min}}\{d(\mathbf{x},\mathbf{y})\}$. For any partition $U \in M_{hcn} \leftrightarrow \{X_i\}$ the separation index of U (universally known as *Dunn's index* (DI)) is

$$DI(U;X) = \underset{1\le i\le c}{\underbrace{\min}}\left\{\underset{\substack{1\le j\le c\\ j\ne i}}{\underbrace{\min}}\left\{\frac{\delta(X_i,X_j)}{\underset{1\le k\le c}{\underbrace{\max}}\{\Delta(X_k)\}}\right\}\right\} \tag{7}$$

Dunn called U *compact and separated* (CS) relative to the distance metric d if and only if : for all s, q and r with q≠r, any pair of points $\mathbf{x},\mathbf{y} \in X_s$ are closer together (with respect to d) than any pair u, v with $\mathbf{u} \in X_q$ and $\mathbf{v} \in X_r$. He then proved that a set X *has* a crisp CS partition relative to d if and only $\underset{U\in M_{hcn}}{\underbrace{\max}}\{DI(U;X)\} > 1$. The following result connects this property of Dunn's index to the MM samples extracted from X_N by Algorithm 1:

Theorem 1. Let c' > c. Suppose there is a CS c-partition of $O = \{o_1,\ldots,o_N\}$. Then MM Algorithm 1 will select at least one object from each of the c clusters.

Proof. Hathaway et al. [15]

Theorem 1 is weak in the sense that most input data sets do NOT have a CS partition, and even if they do, it is not so easy to verify this. But Theorem 1 does assert that in some circumstances, the MM samples at least represent all c clusters in the data. To our knowledge, this is the only result of its kind, and it provides a bit of psychological reassurance that MM sampling doesn't run too far off the rails.

IV. Sample Quality

The data sets in our experiments are labeled, i.e., they have ground truth c-partitions $U \in M_{hcN}$ of X_N. Let n_i be the number of points in the i-th subset, so $N = \sum_{i=1}^{c} n_i$. Define the proportion vector of X_N in $\Re^c$ as

$$\mathbf{V}_N = \left(n_1/N,\ldots,n_c/N\right) \in \Re^c . \qquad (8)$$

Using (1)-(6) at line 3 of Algorithm 1 yields c' MM samples, $X_{MM(k)}$, with corresponding proportion vectors in $\Re^c$

$$\mathbf{V}_{MM(k)} = \left(n_1'/c',\ldots,n_c'/c'\right) \in \Re^c;\ 1 \le k \le 6 . \qquad (9)$$

If $c' < c$ at least one of the $n_i' = 0$ in $\mathbf{V}_{MM(*)}$. If $c' \ge c$, and all c labeled subsets are represented in the sample, then $n_i' \ge 1; 1 \le i \le c$. Therefore, an easy way to determine if all c subsets have been sampled is to examine $\mathbf{V}_{MM(*)}$ for zeroes. Our aim is to determine how well the $\mathbf{V}_{MM(*)}$ match $\mathbf{V}_N$.

Since these are samples from labeled data, we can make histograms that plot numbers of points in each labeled subset against numbers of points in the samples. This affords a visual assessment of the match between proportions in the parent and sample that is independent of N and p. For small c it is easy to make a fairly accurate assessment by visual comparison.

Comparing $\mathbf{V}_{MM(*)}$ to $\mathbf{V}_N$ analytically can be done in several ways. The distance $d(\mathbf{V}_N, \mathbf{V}_{MM}(*))$ in any convenient metric on $\Re^c \times \Re^c$ provides a matching measure: a zero distance corresponds to a perfect match between the proportions of the parent and sample. The two-sample *Kolmogorov-Smirnoff* (KS) test against the null hypothesis that $\mathbf{V}_N$ and $\mathbf{V}_{MM}(*)$ come from the same distribution can also be used. Matlab returns a p-value for the test at any level of significance. We will use the default level $\alpha = 0.05$ for our experiments. Thus, if $p > \alpha = 0.05$, we accept the hypothesis that the sample comes from the same distribution as the parent, and will indicate this by saying simply that the sample passes the KS test. In our experiments, the number of "samples" for the KS test is c, the number of labeled subsets, so the KS test, which is not very accurate for small sample sizes, is not expected to yield very informative results. We will say that a sample "covers" the input data when every labeled subset is sampled at least once.

V. Numerical Experiments

Initialization (3) begins at $\bar{\mathbf{x}}$. For this study we need labels for all of the points, and $\bar{\mathbf{x}}$ is not always in the data, so $X_{MM(3)}$ is obtained by initializing at $\bar{\mathbf{x}}$, and then replacing it after finding $\mathbf{x}_{m_1}$ by renaming $\mathbf{x}_{m_0} \leftarrow \mathbf{x}_{m_1}$. All of our experiments were performed on a CPU with INTEL core I7-8700k and 64 GB memory using MATLAB-2018a for implementation.

Table 1. Data sets

Name	N	p	c
X6	399	2	6
X15	5000	2	15
X31	3100	2	31
WDBC	569	30	2

Table 1 shows the four sets used in our experiments. We use three small data sets: X15 from [16], X31 from [17] and X6 from [18]; and the *Wisconsin Diagnostic Breast Cancer* (WDBC) data [19]. All 4 data sets are subjected to the same analysis, but we won't be able to show all the figures in this short article. A complete set of graphs is available upon request form the first named author.

The visually apparent clusters in X15 (Fig. 2) are drawn from Gaussian distributions with different means and covariance matrices. The cluster size (cardinality) ranges from 300 to 350. Fig. 3 contains 8 histograms for data set X15 for $c' = 10, 20, 50, 1000$. In all views, the histogram of the input data is fixed in the upper left, the random sample (RS, method #7) is lower right, and the MM samples obtained by Algorithm 1 labeled #1 to #6 correspond to the six initializations at equations (1) to (6). There are two values printed on each histogram: ED is the value of $d(\mathbf{V}_N, \mathbf{V}_{MM}(*))$ for d = Euclidean distance; p is the value of the 2 sample KS test returned by Matlab against $\alpha = 0.05$.

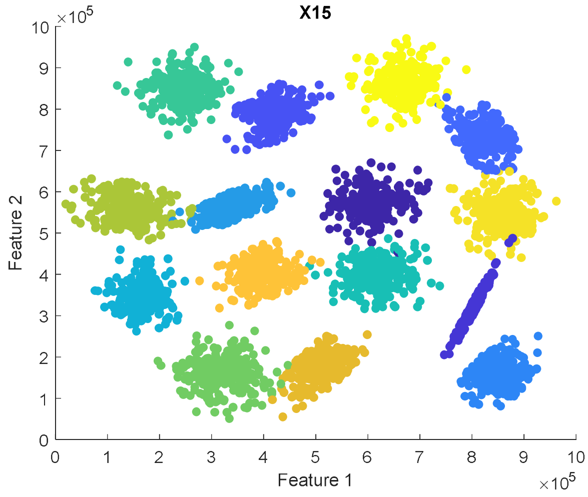

Fig. 2. X15~N=5,000 points in p=2 dimensions, c=15

Fig. 3(a) shows the results of collecting $c'=10$ samples from X15. Since there are 15 clusters, all 7 samples will have 5 or more 0's. Visual examination of the histograms reveals that RS (method #7) collects samples from only 8 of the 15 clusters, whereas all 6 MM samples contain one from each of 10 clusters. The ED metric indicates a slight preference for initialization (2); the KS test for all 7 methods fails. Fig. 3(b) shows the results for $c'=20$. Initialization (2) remains the ED winner, and RS misses points from 4 of the 15 clusters. All seven methods again fail to confirm the hypothesis of the KS test.

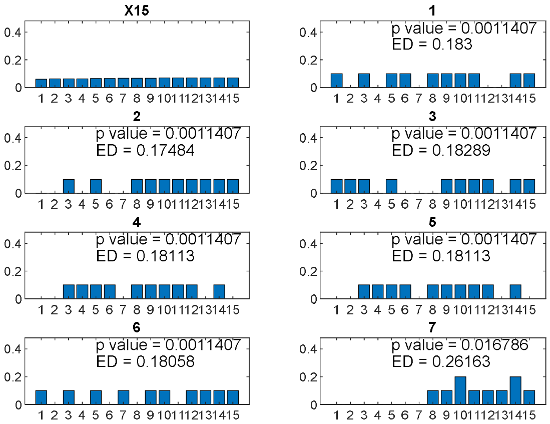

Fig. 3a: $\mathbf{c'=10}$ **MM Samples of X15**

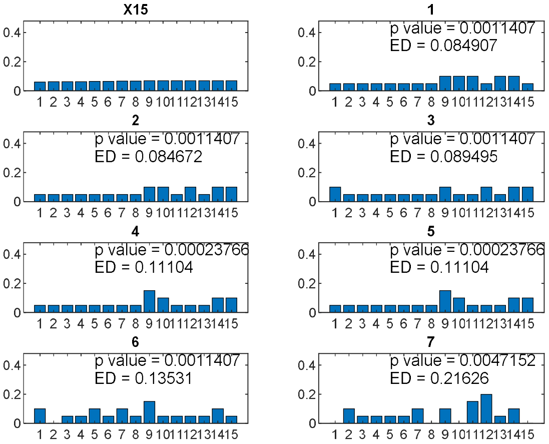

Fig. 3b: $\mathbf{c'=20}$ **MM Samples of X15**

For $c'=50$ in Fig. 3(c), initialization (6) is a slight winner for ED. Methods 2, 4 5 and 7 are accepted by the KS test, but the RS collected by method #7 still fails to collect samples from all 15 labeled subsets. But at $c'=1000$ in Figure 3(d), the story changes. Here RS takes the prize, with the lowest ED value (0.06 vs 0.02) and highest KS acceptance value (0.88 vs. 0.05). Please make a visual comparison of the 7 sample histograms to the input data in Fig. 3(d): method #7 (RS) is clearly a better visual match to the input data than any of the MM samples. So, for N=5000 and c=15, the quality of RS improves enough to be better than all of the MM methods. One last observation: compare the histograms for MM methods (4) and (5): in all 4 views in Fig. 3, these two methods produce identical samples. Do these two methods always produce the same samples? The answer is no. We will demonstrate this using data set X6.

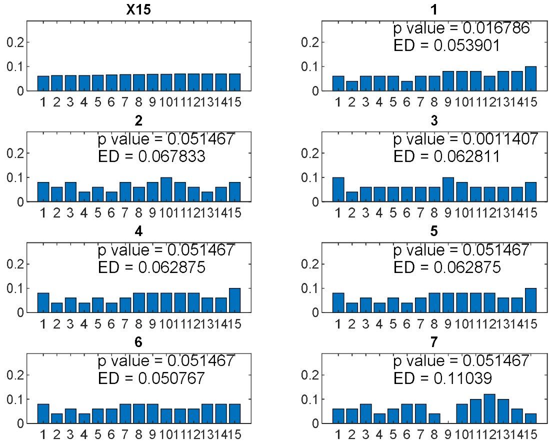

Fig. 3c: $\mathbf{c'=50}$ **Samples of X15**

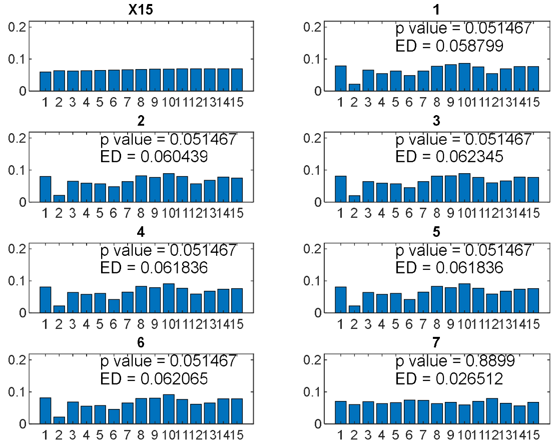

Fig. 3d: $\mathbf{c'=1000}$ **Samples of X15**

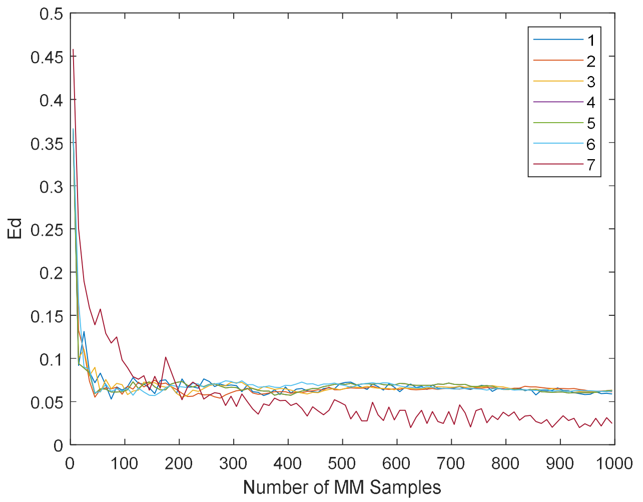

Fig. 4: $\mathrm{ED}(\mathbf{V}_N,\mathbf{V}_{MM}(*))$ **and** $\mathrm{ED}(\mathbf{V}_N,\mathbf{V}_{RS}(*))$ **for samples of X15**

To compare the evolution of samples drawn by the seven methods, consider Figure 4, which graphs $\mathrm{ED}(\mathbf{V}_N,\mathbf{V}_{MM}(*))$ **and**

$\mathrm{ED}(\mathbf{V}_N, \mathbf{V}_{RS}(*))$ for X15 for values of c' up to 1000 samples in increments of 10. For values of c' from 5 to about 50, initialization (2) provides the most favorable match by ED, but after that, all six initializations produce more or less equally good matches. Random sampling of the 15 subsets is erratic until c' reaches 320. For greater numbers of samples, random sampling becomes the method of choice.

Data set X31 (Fig. 5) has 100 points in each of 31 Gaussian clusters. None of the samples covers all 31 subsets for $c'=31$ (not shown). Figure 6(a) shows the histogram for $c'=50$ samples. MM methods 1, 2, 3 and 6 cover all 31 clusters and ED is minimum for methods 2 and 3. MM methods 4 and 5 miss subset #31. The random sample only selects points from 22 of the 31 subsets. At c'=1000 (Fig. 6(b)) all methods cover the data, but the RS still has the greatest ED. Figure 6(c) shows that this trend continues as c' increases: RS produces samples that are about twice as bad with respect to the ED measure as any of the MM methods. The p-values for the KS test are surprisingly small for this experiment: all seven methods fail to pass the test for all choices of c'.

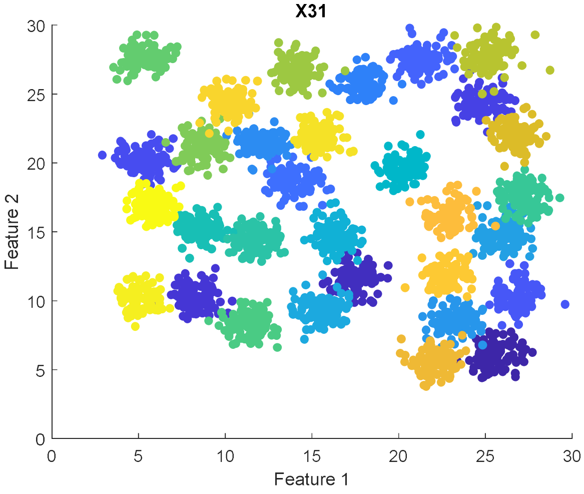

Fig. 5 X31~N=3100 points in p=2 dimensions, c=31

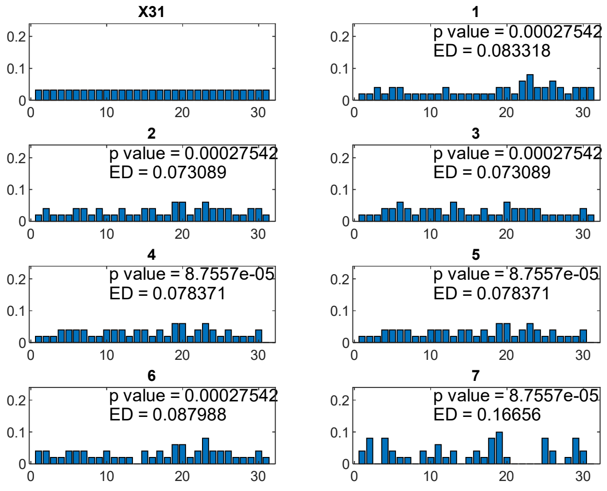

Fig. 6(a): $c'=50$ samples of X31

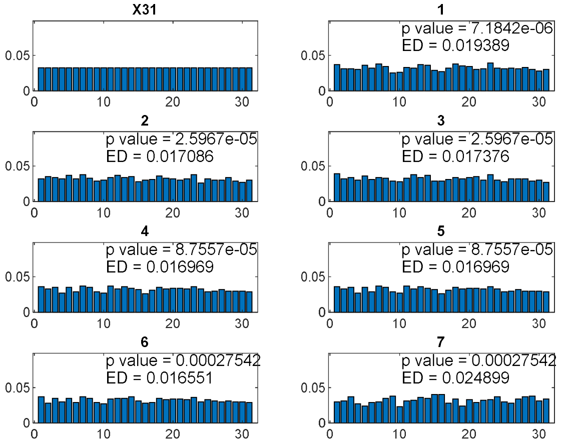

Fig. 6(b): $c'=1000$ Samples of X31

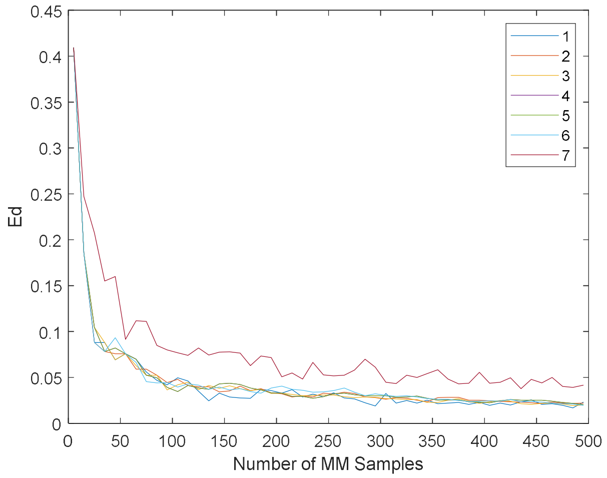

Fig. 6(c): ED values for samples of X31

Data set X6 (Fig. 7) has c=6 labeled subsets. There are 2 Gaussian clusters at the upper left, the right side contains a dense set of dark blue points imbedded in a sparse subset, and the lower left portion of the scatterplot is a "fried egg" set of clusters comprising a central "yolk, bright yellow" surrounded by a ring (the "egg white"). The cardinalities of the 6 subsets are: 50, 92, 38, 45, 158, 16. See the color bar in Fig. 7 to associate the labeled subsets with their sizes.

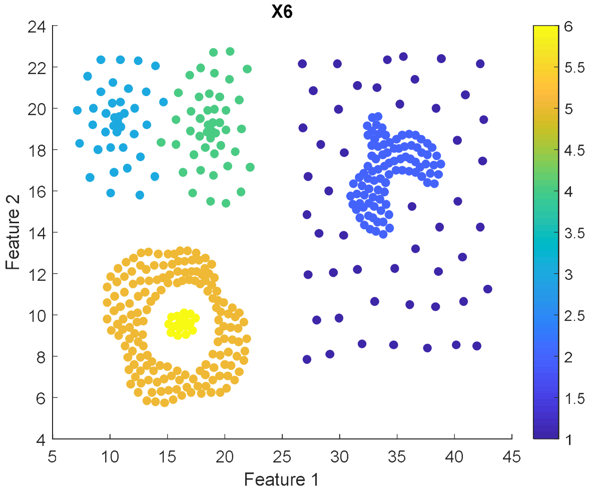

Fig. 7: X6~N=399 points in p=2 dimensions, c=6

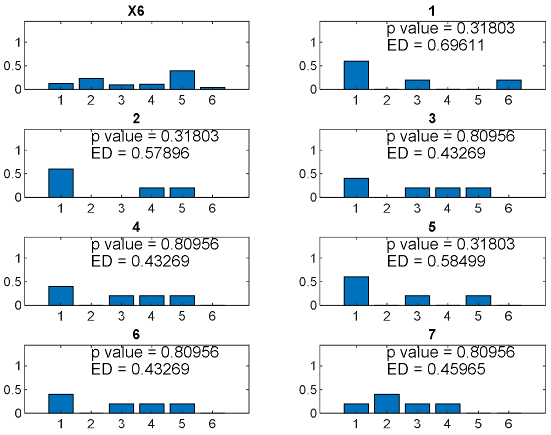

Fig. 8(a): $c' = 5$ samples of X6

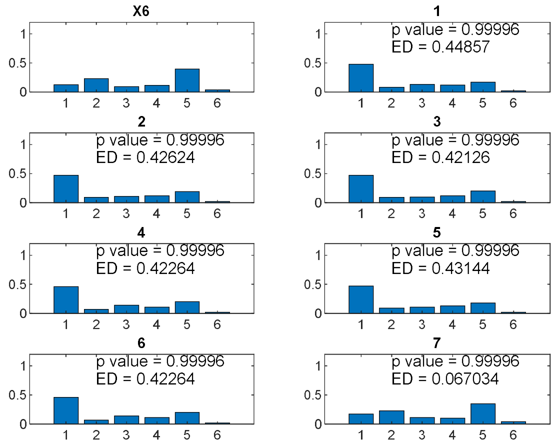

Fig. 8(b): $c' = 100$ samples of X6

Figure 8(a) contains the histograms for c'=5 samples, for which none of the methods can cover c=6 labeled subsets. However, please notice that all 7 samples pass the KS test with values well above 0.05. Figure 8(b) displays the results for c'=100. The KS test is indiscriminant in that it produces the same high value (0.99996) for all seven methods. On the other hand, the ED for RS clearly indicates that the random sample (#7) is the best match to the input distribution, and visual examination of the histograms confirms this. All six MM methods are dominated by samples from (dark blue) subset 1 of X6, which is the subset of widely scattered, least dense points in the data set. Random sampling draws the most points from cluster 5 (the (beige) egg white), because this cluster contains 158 points, which is about 40% of the data. Not many of the points in subset 5 attract MM samples because of their proximity to each other.

Figure 9 shows graphs for the distances $\mathrm{ED}(\mathbf{V}_N, \mathbf{V}_{MM}(*))$ and $\mathrm{ED}(\mathbf{V}_N, \mathbf{V}_{RS}(*))$ for X6 for values of c' up to 200. Method 7 (RS) is clearly superior to all the MM methods for every set of samples beyond c'=10. MM initialization (2) appears to be the most consistent winner among the MM schemes for $c' \le 60$, but all six initializations provide comparable results beyond this. The conclusion to be drawn from our experiment with this very small but interesting data set is that MM sampling is at its best when the condition of Theorem 1 is satisfied ~ viz., that there are CS clusters in the data. X6 clearly does not have them!

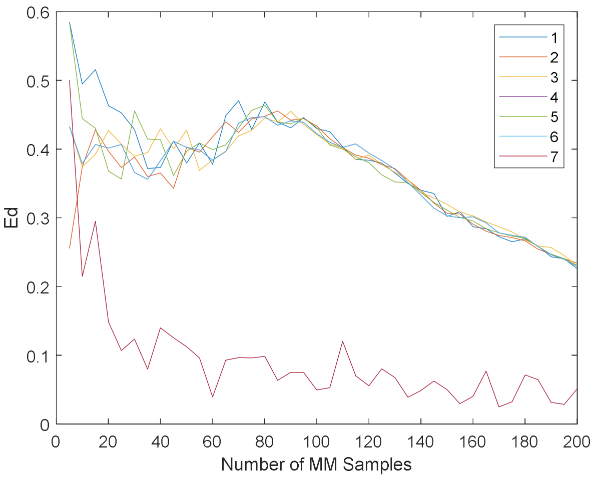

Fig. 9: ED values for samples of X6

Figure 10 shows that initializations (4) and (5) don't always produce the same MM sample. The solid red square in Fig. 9 is the grand mean ($\bar{\mathbf{x}}$) of X6, and the hollow red square is the furthest point from it, chosen by method (4). The two points marked by diamonds on Fig. 9 are the points in the data furthest from each other, method (5). While this shows that (4) and (5) can produce different initializations, they produced identical samples in almost all of our experiments.

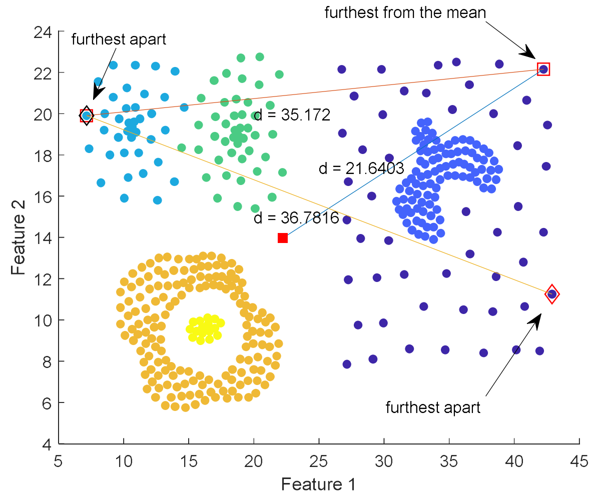

Fig. 10: Initializations (4) and (5) for data set X6

Our last experiment used the Wisconsin Diagnostic Breast Cancer data set. Fig. 11 shows the ED graphs for methods 1-7 on this data set out to 200 MM samples. There is a very striking difference between RS (#7) and the six MM methods. For this experiment, none of the MM methods competes with RS for any number of samples. As the number of samples increases, RS improves (ED decreases), and all the MM methods also improve, albeit slightly. However, in terms of matching proportion vectors, RS is somewhere between 2 and 3 times as effective as any of the MM schemes. Among the MM schemes, method (6) seems most effective, but all six initializations cross over each other, so it's hard to declare any of them as "best."

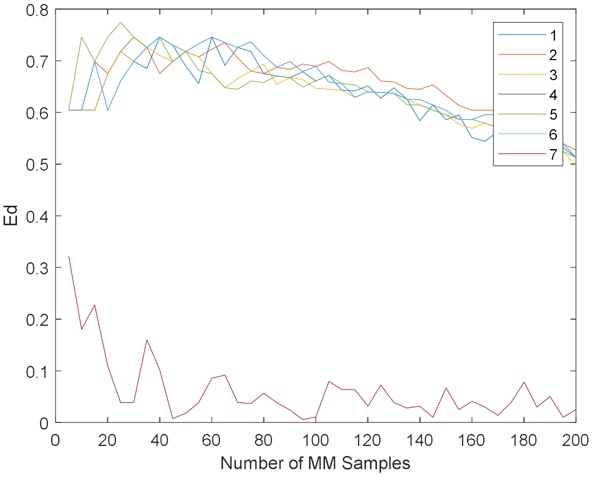

Fig. 11: ED values for samples of WDBC

VI. Conclusions

The experiments presented here are not extensive enough to support any strong conclusions. However, they do suggest that: (i) the six initializations of MM at equations (1) to (6) produce roughly the same samples; (ii) initialization (2), choosing the first point in the data, which is always the fastest MM method, is often also the best of the six methods in terms of Euclidean distance match; (iii) statistical tests such as the KS test to assess sample quality are not reliable for small samples (small values of c), and further, are not applicable for unlabeled data anyway; and (iv) the distribution of subsets in the data is very important ~ MM sampling is at its best for CS data, and RS is much better if the data contain diverse patterns such as those in X6 . Initializations (4) and (5) produce identical samples in almost all of our experiments, but we demonstrated that this is not always the case. Since these two methods are the most costly in terms of CPU time, there is little to recommend either of them.

Our conjecture is that as the number of samples increases, RS will overtake the quality of any of the MM methods for even CS data. Since RS requires no distance calculations it is always superior to MM sampling in terms of the CPU time spent to acquire the samples.

References

[1] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*, Wiley Interscience, NY, 1973.

[2] J. Hartigan. *Clustering Algorithms*, Wiley, NY, 1975.

[3] A. Jain and R. Dubes. *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ, 1988.

[4] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 6th ed., Academic Press, NY, 2009.

[5] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1981.

[6] G. J. McLachlan and K. E. Basford, K. E. *Mixture Models : Inference and Applications to Clustering*, Marcel Dekker, NY, 1988.

[7] J. Keller, D. Liu, and D. Fogel. *Fundamentals of Computational Intelligence: Neural Networks, Fuzzy Systems, and Evolutionary Computation*, Wiley/IEEE Press, Hoboken, NJ, 2016.

[8] F. Provost, D. Jensen and T. Oates. Efficient progressive sampling, *Proc. 5th KDDM*, 23-32, 1999.

[9] R. J. Hathaway and J. C. Bezdek. Extending fuzzy and probabilistic clustering to very large data sets, *Comp. Stat. And Analysis*, 51, 215-234, 2006.

[10] N. R. Pal and J. C. Bezdek. Complexity reduction for large image processing, *IEEE Trans. SMC*, B-32(5), 598-611, 2002.

[11] D. Kumar, J. C. Bezdek, M. Palaniswami, S. Rajasegarar, C. Leckie and T. C. Havens. A hybrid approach to clustering in big data, *IEEE Trans. Cybernetics,* 46(10), 2372 – 2385, 2016.

[12] R. L. Thorndike. Who belongs in the family?, *Psychometrika*, 18(4), 267-278, 1953.

[13] R. G. Casey and G. Nagy. An autonomous reading machine, *IEEE Trans. Computers*, C-17(5), 492-503, 1968.

[14] J. C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *J. Cyberns.,* 3(3), 32-57, 1974.

[15] R. J. Hathaway, J. C. Bezdek and J. M. Huband, Scalable visual assessment of cluster tendency for large data sets, *Patt. Recog.*, 39, 1315-1324, 2006.

[16] P. Fränti and O. Virmajoki, Iterative shrinking method for clustering problems, *Pattern Recognition*, 39 (5), 761-765, 2006.

[17] C.J. Veenman, M.J.T. Reinders, and E. Backer, A maximum variance cluster algorithm. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(9), 1273-1280, 2002.

[18] C.T. Zahn, Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comp.*, 100(1), p. 68-86, 1971.

[19] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).