# Normalized EM Algorithm for Tumor Clustering using Gene Expression Data

Nguyen Minh Phuong and Nguyen Xuan Vinh

*Abstract*— **Most of the proposed clustering approaches are heuristic in nature. As a result, it is difficult to interpret the obtained clustering outcomes from a statistical standpoint. Mixture model-based clustering has received much attention from the gene expression community due to its sound statistical background and its flexibility in data modeling. However, current clustering algorithms following the model-based framework suffer from two serious drawbacks. First, the performance of these algorithms critically depends on the starting values for their iterative clustering procedures. And second, they are not capable of working directly with very high dimensional data sets whose dimension might be up to thousands. We propose a novel normalized Expectation-Maximization (EM) algorithm to tackle the two challenges. The normalized EM is stable even with random initializations for its EM iterative procedure. Its stability is demonstrated through the performance comparison with other related clustering algorithms such as the unnormalized EM (The conventional EM algorithm for Gaussian mixture model-based clustering) and spherical $k$-means. Furthermore, the normalized EM is the first mixture model-based clustering algorithm that is shown to be stable when working directly with very high dimensional microarray data sets in the sample clustering problem, where the number of genes is much larger than the number of samples. Besides, an interesting property of the convergence speed of the normalized EM with respect to the squared radius of the hypersphere in its corresponding statistical model is uncovered.**

## I. INTRODUCTION

Microarrays is a technological breakthrough in molecular biology, allowing the simultaneous expression measurements of thousands of genes during some biological process [1], [2], [3]. Based on this technology, various microarray experiments have been conducted to give valuable insights into biological processes of organisms, e.g the study of yeast genome [4], [5], [6] and the investigation of human genes [7], [8], [9]. These studies have posed great challenges to elucidate the hidden information given the availability of the genomic-scale data. Applications of microarrays range from the analysis of differentially expressed genes under various conditions to the modeling of gene regulatory networks. One of the main interests in the study of microarray data is to identify naturally occurring groups of genes with similar expression patterns or samples of the same molecular subtypes. Clustering is a basic exploratory tool for investigation of these problems. A variety of clustering methods have been proposed in the microarray literature to analyze the genomic data, including hierarchical clustering [8], [10], [11], self-organizing maps (SOM) [12], k-means and its variants [13], [14], [15], graph-based methods [16], [17] and mixture model-based clustering [18], [19], [20], to name a few.

The authors are with the School of Electrical Engineering and Telecommunications, University of New South Wales, Kensington, NSW 2052, Australia n.m.phuong@student.unsw.edu.au, n.x.vinh@unsw.edu.au

Mixture model-based clustering offers a coherent probabilistic framework for cluster analysis. This approach is based on the assumption that the data points in each cluster are generated by some underlying probability distribution. The performance of model-based clustering greatly depends on the distributional assumption of the underlying parametric models. The most widely-used statistical model for this clustering approach is the mixture of Gaussian distributions. Usually parameters of the model are estimated using the EM algorithm [21]. A serious drawback of Gaussian mixture model-based clustering is that its clustering performance might be heavily affected by the choice of starting values for the EM iterations. Another drawback of the unnormalized EM is its limited capability of working directly with very high dimensional data sets of which dimension is much larger than the number of data points. Usually dimension reduction techniques such as principal component analysis [22] must be pre-applied to resolve this curse of high dimensionality, e.g. McLachlan et al [18] have to resort to a feature selection technique and factor analysis to reduce the dimension of the data before proceeding to the unnormalized EM clustering. A crucial limitation of this approach is that the dimension reduction process may result in the information loss to the original data, e.g. the inherent cluster structure in the original data may not be preserved.

In order to overcome the above-mentioned shortcomings of the popular Gaussian mixture model-based clustering (the unnormalized EM), we propose a novel normalized EM algorithm for clustering gene expression data, in which data points to be clustered are normalized to lie on the surface of a hypersphere. The proposed approach also follows mixture model-based framework but the clustering of the data is performed on a fixed hypersphere. The normalized EM clustering works stably even with very high dimensional microarray data sets, which make use of thousands of genes. Besides, the projection of the data on a hypersphere is shown to eliminate the intrinsic scattering characteristic of the data, thus making the normalized EM work more stably in comparison with the unnormalized EM.

Of particular relevance to our work are the spherical $k$-means algorithm [23], clustering using von-mises Fisher distributions [24] and clustering in a unit hypersphere using the inverse projection of multivariate normal distributions [25]. Spherical $k$-means is similar to $k$-means in nature

except that its clustering of the data is performed in a unit hypersphere. Like $k$-means, spherical $k$-means is fast for high dimensional gene expression data sets. However, the clustering outcomes of spherical $k$-means on the same data set may be significantly different due to the sensitivity of the algorithm to its starting values. In [24] Banerjee et al propose a method to estimate the concentration parameter of the von Mises-Fisher distribution, a statistical distribution for spherical data, and apply it for clustering various types of data including yeast cell cycle gene expression data. An important point to note here is that the clustering approach has difficulty of working on data sets with dimensions up to thousands as it involves the computation of extremely large exponentials. In [25] a new clustering approach is proposed to allow a more flexible description of clusters. However, this approach is not capable of working well with the sample clustering problem where the number of data points is much smaller than the dimension of the data due to either the over-fitting problem or the near singularity of the estimated covariance matrices in its EM iterations. The underlying distribution in our statistical model can be seen as a simplified variant of the von-mises Fisher distribution or of the distribution presented in [25]. Interestingly it is the parsimony that makes our normalized EM work well with very high dimensional microarray data. The normalized EM is stable even with random initializations for its iterative clustering procedure.

To demonstrate the stability and the capability of working with very high dimensional data of the normalized EM, we analyze the algorithm using several microarray data sets and compare the obtained results with the ones produced by the unnormalized EM, spherical $k$-means as well as other related clustering algorithms. Also a detailed analysis of the convergence speed of the normalized EM with respect to the squared radius of the fixed hypersphere is provided, and an interesting result is exposed.

The remaining of this paper is organized as follows. Section II introduces the statistical model of the proposed method and the derivations of the normalized EM algorithm. In section III, the normalized EM is analyzed in detail, and its effectiveness is illustrated using three real microarray data sets. Finally, section IV summarizes the main contributions of this work and briefly discusses possible research directions.

For convenience, some notational conventions used in this paper are provided: $n$ is the number of data points or samples to be clustered; $p$ is the dimension of data points or the number of genes; $\mu$ is the squared radius of the hypersphere; $K$ is the number of clusters in a data set; $\{\mathcal{X}_h\}_{h=1}^K$ is a $K$-cluster partition of the data; $\langle . \rangle$ is the inner product of two vectors; $\|.\|$ is the Euclidean norm.

## II. THE NORMALIZED EM ALGORITHM

A microarray data set is commonly represented by the matrix $G_{n\times p} = [x_1, x_2, ...x_n]$, where $x_j \in \mathbf{R}^p$ is the gene expression profile of sample $j$. Typically the number of genes is much larger than the number of experiments (samples). Our primary goal is to group tumor samples into different molecular subtypes. Specifically, we have to classify the set of samples into $K$ groups $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_K$ such that the samples in the same cluster should have similar gene expression profiles and gene expression patterns of samples in different clusters are as much dissimilar as possible.

We now introduce a new normalized EM algorithm for tumor clustering using gene expression data. First, data points are normalized so that they lie on a hypersphere with predefined radius and then the clustering of the data is performed on this hypersphere only. The statistical model for the normalized EM clustering is described in detail as follows:

First, gene expression profiles $x_i$ are normalized so that they belong to the manifold $S_\mu = \{x : \|x\|^2 = \mu, x \in \mathbf{R}^p\}$ for some $\mu > 0$. In other words, the data points are processed by

$$x_i \leftarrow \sqrt{\mu}\frac{x_i}{\|x_i\|},\ i = 1, 2, ..., n \tag{1}$$

Then these normalized $x_i$'s are treated as samples drawn from a mixture of $K$ exponential distributions

$$p(x|\Theta) = \gamma_\mu \sum_{h=1}^{K} \pi_h e^{-\|x-\mu_h\|^2} \tag{2}$$

where $\Theta = (\pi_1, \mu_1, ..., \pi_k, \mu_k)$, in which the $\pi_h$, $\mu_h$ are mixing proportions and directional mean vectors respectively satisfying the following conditions:

$$\sum_{h=1}^{K} \pi_h = 1, \pi_h \geq 0, \|\mu_h\|^2 = \mu,\ h = 1, 2, ..., K \tag{3}$$

and $\gamma_\mu$ is the normalizing constant

$$\gamma_\mu = \frac{1}{\displaystyle\int_{x\in S_\mu} e^{-\|x-\mu_h\|^2} dx}. \tag{4}$$

Assuming that the data vectors are independent and identically distributed with distribution $p$, then the data likelihood function is

$$\mathcal{L}(\Theta|\mathcal{X}) = p(\mathcal{X}|\Theta) = \prod_{i=1}^{n} p(x_i|\Theta) = \prod_{i=1}^{n} (\gamma_\mu \sum_{h=1}^{K} \pi_h e^{-\|x-\mu_h\|^2}). \tag{5}$$

The maximum likelihood estimation problem is:

$$\max_{\Theta} \{\mathcal{L}(\Theta|\mathcal{X}) : (3)\}. \tag{6}$$

Maximizing the likelihood function (6) is very difficult, thus we employ the EM algorithm to find a local maximizer of the likelihood function ([21]).

Given the current estimate $\Theta^{(\ell)}$ at the $\ell^{th}$ iteration ($\ell \geq 0$) of the EM iterative procedure, for each $h = 1, 2, ..., K$, the

posterior probability $p(h|x_i, \Theta^{(\ell)})$ that $x_i$ is generated by the $h^{th}$ component of the mixture density is defined by

$$\begin{aligned} p(h|x_i, \Theta^{(\ell)}) &= \frac{p(h|\Theta^{(\ell)})p(x_i|h, \Theta^{(\ell)})}{p(x_i|\Theta^{(\ell)})} \\ &= \frac{\pi_h^{(\ell)} e^{2\langle x_i, \mu_h^{(\ell)}\rangle}}{\sum_{h'=1}^{K} \pi_{h'}^{(\ell)} e^{2\langle x_i, \mu_{h'}^{(\ell)}\rangle}}. \end{aligned} \quad (7)$$

The expectation of the marginal log-likelihood function for the observed data over the given posterior distribution is:

$$\begin{aligned} & E[\sum_{i=1}^{n} \log(\gamma_\mu \pi_h e^{-\|x_i - \mu_h\|^2})] \\ = & \sum_{i=1}^{n} E[\log(\gamma_\mu \pi_h e^{-\|x_i - \mu_h\|^2})] \\ = & \sum_{i=1}^{n}\sum_{h=1}^{K} [\log(\gamma_\mu \pi_h e^{-\|x_i - \mu_h\|^2})] p(h|x_i, \Theta^{(\ell)}) \\ = & \sum_{i=1}^{n}\sum_{h=1}^{K} (\log \pi_h - \|x_i - \mu_h\|^2) p(h|x_i, \Theta^{(\ell)}) + n \log \gamma_\mu \\ = & \sum_{i=1}^{n}\sum_{h=1}^{K} (\log \pi_h - 2\mu + 2\langle x_i, \mu_h\rangle) p(h|x_i, \Theta^{(\ell)}) + \\ & + n \log \gamma_\mu \\ = & \sum_{h=1}^{K}\sum_{i=1}^{n} (\log \pi_h + 2\langle x_i, \mu_h\rangle) p(h|x_i, \Theta^{(\ell)}) - 2nK\mu + \\ & + n \log \gamma_\mu. \end{aligned} \quad (8)$$

The maximization step for the normalized EM algorithm is:

$$\begin{aligned} & \max_{\Theta} \{\sum_{h=1}^{K}\sum_{i=1}^{n} (\log \pi_h + 2\langle x_i, \mu_h\rangle) p(h|x_i, \Theta^{(\ell)}) - \\ & - 2nK\mu + n \log \gamma_\mu : (3)\} \\ = & \max_{\Theta} \{\sum_{h=1}^{K}\sum_{i=1}^{n} (\log \pi_h) p(h|x_i, \Theta^{(\ell)}) + \\ & + 2\sum_{h=1}^{K}\sum_{i=1}^{n} \langle x_i, \mu_h\rangle p(h|x_i, \Theta^{(\ell)}) : (3)\} - 2nK\mu + \\ & + n \log \gamma_\mu \\ = & \max_{\{\pi_h\}_{h=1}^{K}} \{\sum_{h=1}^{K}\sum_{i=1}^{n} (\log \pi_h) p(h|x_i, \Theta^{(\ell)}) : \sum_{h=1}^{K} \pi_h = 1, \\ & , \pi_h \geq 0, h = 1, 2, .., K\} + \\ & + 2\sum_{h=1}^{K} \max_{\mu_h} \{\sum_{i=1}^{n} \langle x_i, \mu_h\rangle p(h|x_i, \Theta^{(\ell)}) : \|\mu_h\|^2 = \mu\} - \\ & - 2nK\mu + n \log \gamma_\mu \end{aligned} \quad (9)$$

Solving (9), we obtain the following iterative procedure of the normalized EM:

$$\begin{aligned} \pi_h^{(\ell+1)} &= \frac{1}{n}\sum_{i=1}^{n} p(h|x_i, \Theta^{(\ell)}) \\ &= \frac{1}{n} \frac{\sum_{i=1}^{n} \pi_h^{(\ell)} e^{2\langle x_i, \mu_h^{(\ell)}\rangle}}{\sum_{h'=1}^{K} \pi_{h'} e^{2\langle x_i, \mu_{h'}^{(\ell)}\rangle}} \end{aligned} \quad (10)$$

$$\begin{aligned} \nu_h^{(\ell+1)} &= \sum_{i=1}^{n} x_i p(h|x_i, \Theta^{(\ell)}) \\ \mu_h^{(\ell+1)} &= \frac{\sqrt{\mu}\nu_h^{(\ell+1)}}{\|\nu_h^{(\ell+1)}\|}. \end{aligned} \quad (11)$$

The optimal parameter estimate $\Theta_{opt}$ is obtained when the difference between two observed data log-likelihoods corresponding to two successive iterations is less than a given tolerance threshold. Finally, each data point is assigned to the component with the maximum estimated posterior probability, i.e. a data point $x_i$ is assigned to component $h$ or cluster $\mathcal{X}_h$ if $h = \arg\max_{h'} p(h'|x_i, \Theta_{opt})$.

## III. EXPERIMENTAL RESULTS

The stability and the capability of working directly with high dimensional gene expression data sets of the normalized EM clustering algorithm are demonstrated to three microarray data sets: (1) acute leukemia [26]), (2) colon ([7]), and (3) pediatric acute leukemia [27]. These data sets are popular in the microarray literature. We make attempts to offer illustrations using experimental data sets of significant differences in dimension, the number of samples, the number of underlying clusters and tumor type. We assess the clustering results of the normalized EM on these gene expression data sets with different values of the parameter $\mu$ and compare the obtained results with the ones produced by the unnormalized EM clustering (The EM algorithm for Gaussian mixture model-based clustering), spherical $k$-means and some other related clustering algorithms. It should be noted that the analysis is only provided for the values of $\mu$ in the range from 0 to 350. For $\mu$ bigger than 350 the iterative procedure of the normalized EM involves the difficulty of very large exponential computations.

In this section, the analysis of the convergence speed of the normalized EM is also presented to give a rough idea of which appropriate values of $\mu$ should be chosen to maximize the cluster quality of the normalized EM. The convergence speed of the normalized EM algorithm here is measured through the average number of iterations till the algorithm converges to an optimal solution. For each value of $\mu$, the normalized EM is run several times and the average of the number of iterations of those runs is taken as the convergence speed corresponding with that value of $\mu$.

Additionally, to better characterize the behavior of clustering algorithms for the first two data sets, a cutoff of twenty-five percent of the number of misclassified samples out of all samples is set to determine the distinction between "good" clusterings and "poor" ones, that is, a clustering is good if the number of misclassified samples out of all samples is less than twenty-five percent or poor otherwise.

### *Acute Leukemia Data*

This data set was originally produced and analyzed by Golub et al [26]. The data set utilized here consists of 38 samples $\times$ 5000 genes. These 38 samples are supposed to be categorized into three classes corresponding to three subtypes of leukemias: ALL-B, ALL-T and AML.

Table I shows the clustering results of the normalized EM on this data set. The normalized EM worked stably with $\mu$ in the range from 15 to 350 even with random initializations. It can be seen that within the range of $\mu$ where the normalized EM worked well, the number of misclassified samples were around two. The normalized EM worked best, typically only one misclassified sample, for $17 \leq \mu \leq 25$. The statistics of convergence speed summarized in Table I and Figure 1 show that the "best" values of $\mu$ as just mentioned above occur right after the ones corresponding to the dramatic decrease in the average number of iterations.

TABLE I

CLUSTERING RESULTS OF THE NORMALIZED EM ALGORITHM ON THE ACUTE LEUKEMIA DATA SET (ENCLOSED IN PARENTHESES ARE THE NUMBER OF TIMES OBSERVING THE CORRESPONDING RESULTS AMONG 20 RUNS).

| $\mu$ | Average number of iterations | Average number of misclassified samples | | Minimum number of misclassified samples |
|---|---|---|---|---|
| | | Good | Poor | |
| 15 | 48 | 3(20) | (0) | 3(20) |
| 17 | 41.3 | 1(20) | (0) | 1(20) |
| 20 | 28.1 | 1(20) | (0) | 1(20) |
| 25 | 25.2 | 1(19) | 15(1) | 1(19) |
| 30 | 18.8 | 2.2(18) | 14(2) | 1(6) |
| 40 | 13.2 | 2.4(19) | 16(1) | 1(6) |
| 50 | 12.9 | 2.7(16) | 14(4) | 1(4) |
| 70 | 11.9 | 2.7(18) | 13(2) | 1(4) |
| 90 | 11.7 | 2.4(18) | 14.5(2) | 1(4) |
| 120 | 8.7 | 1.8(17) | 16(3) | 1(7) |
| 150 | 10.1 | 2.3(16) | 13.5(4) | 1(4) |
| 200 | 9.8 | 2.3(16) | 15.8(4) | 1(1) |
| 250 | 9.0 | 2.1(15) | 14(5) | 0(3);1(2) |
| 300 | 9.3 | 2.4(15) | 13.4(5) | 1(4) |
| 350 | 8.7 | 2.2(14) | 15.7(6) | 1(4) |

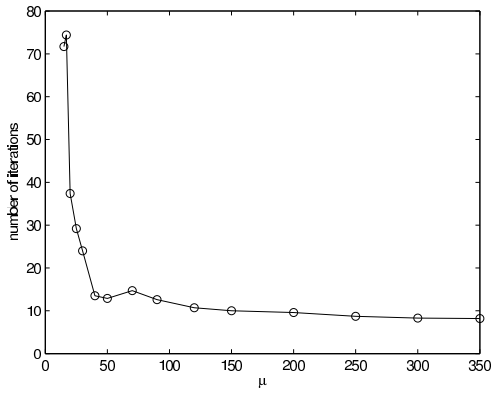


Fig. 1. Convergence speed of the normalized EM on acute leukemia data set.

We next analyze the unnormalized EM clustering on this data set. As the algorithm is unable to work with high dimensional data sets, data reduction techniques must be pre-applied to reduce the dimension of the data. Principal component analysis (PCA) was utilized to reduce the dimension of the data set from 5000 genes to only a few gene components $q$. Table II represents the clusterings of the unnormalized EM on the reduced data set with random initializations. The results tell us that the unnormalized EM might critically depend on the initial values for its own iterative procedure. For a fair comparison, clustering

TABLE II

CLUSTERING RESULTS OF THE UNNORMALIZED EM ON THE REDUCED ACUTE LEUKEMIA DATA SET (ENCLOSED IN PARENTHESES ARE THE NUMBER OF TIMES OBSERVING THE CORRESPONDING RESULTS AMONG 10 RUNS).

| Number of principal components | Average number of misclassified samples | | Minimum number of misclassified samples |
|---|---|---|---|
| | Good | Poor | |
| 3 | 6(10) | (0) | 6(10) |
| 4 | 7.8(6) | 17.75(4) | 6(2) |
| 5 | 7(5) | 15.6(5) | 5(1) |
| 6 | 6.25(4) | 14.5(6) | 3(1) |
| 7 | 5(1) | 15.78(9) | 5(1) |
| 8 | 7(3) | 15(7) | 6(1) |

performance of the normalized EM on the reduced data set of 38 $q$-dimensional samples is also provided (See Table III). Overall we realized that the normalized EM gave better

TABLE III

CLUSTERING RESULTS OF THE NORMALIZED EM ON THE REDUCED ACUTE LEUKEMIA DATA SET (ENCLOSED IN PARENTHESES ARE THE NUMBER OF TIMES OBSERVING THE CORRESPONDING RESULTS AMONG 10 RUNS).

| Number of principal components | Average number of misclassified samples | | Minimum number of misclassified samples |
|---|---|---|---|
| | Good | Poor | |
| 3 | 2(10) | (0) | 2(10) |
| 4 | 2.3(8) | 14.5(2) | 2(7) |
| 5 | 3(9) | 17(1) | 3(9) |
| 6 | 3.6(7) | 14(3) | 3(6) |
| 7 | 3.3(10) | (0) | 3(9) |
| 8 | 3.7(9) | 15(1) | 3(3) |

TABLE IV

CLUSTERING RESULTS OF SPHERICAL $k$-MEANS ON THE ACUTE LEUKEMIA DATA SET (20 RUNS WERE PERFORMED).

| Cluster quality | Average number of misclassified samples | Number of times observing the corresponding results |
|---|---|---|
| Good | 2.93 | 14 |
| Poor | 14.7 | 6 |
| Best | 0 | 1 |

clustering results compared to the combination of PCA and the unnormalized EM. Furthermore, even for the reduced

data set, the normalized EM has been proven to work more stably as well.

Table IV represents clustering results of spherical $k$-means. We find that spherical $k$-means was stable on this acute leukemia data set. However, with the values of $\mu$, e.g. from 17 to 25, where the normalized EM worked best, spherical $k$-means was not comparable to the normalized EM in term of cluster quality.

### *Colon Data*

This data set consists of 62 samples $\times$ 2000 genes. Those 62 samples are supposed to be categorized into two classes: tumor colon tissue samples and normal ones. The 2000 human genes in this data set are those with the highest minimal intensities across samples, which were selected among the total of 6500 genes in the original data set introduced by [7], who produced and also performed cluster analysis on this colon data.

TABLE V

CLUSTERING RESULTS OF THE NORMALIZED EM ON THE SMALL COLON DATA SET (ENCLOSED IN PARENTHESES ARE THE NUMBER OF TIMES OBSERVING THE CORRESPONDING RESULTS AMONG 20 RUNS).

| $\mu$ | Average number of iterations | Average number of misclassified samples | | Mininum number of misclassified samples |
|---|---|---|---|---|
| | | Good | Poor | |
| 15 | 40.1 | 9(15) | 24(5) | 9(15) |
| 20 | 22.9 | 7(17) | 23(3) | 7(17) |
| 30 | 17.2 | 7(16) | 24(4) | 7(16) |
| 33 | 16.0 | 6(16) | 24(4) | 6(16) |
| 40 | 14.4 | 6(16) | 24(4) | 6(16) |
| 70 | 13.4 | 6(17) | 24(3) | 6(17) |
| 100 | 11.5 | 6(15) | 25.2(5) | 6(15) |
| 150 | 10.8 | 6(15) | 26.8(5) | 6(15) |
| 200 | 9.1 | 6(14) | 25.2(6) | 6(14) |
| 250 | 10.9 | 6(14) | 25.9(6) | 6(14) |
| 300 | 9.2 | 6(13) | 24.6(7) | 6(13) |
| 350 | 11.0 | 7.43(14) | 24.3(6) | 6(9) |

With the values of $\mu$ in the range from 50 to 350, the normalized EM was able to produce the results with only 6 misclassified samples, which matched the results produced using supervised classification, e.g by [28]. The 6 misclassified samples here are the three tumor samples (T30, T33, T36) and the other three normal (n8, n34, n36). Note that the samples here were labeled following [7]. In their work, Alon et al also reported their clusterings of this data set with 8 misclassified samples, three normal to tumor class (n8, n12, n34) and five tumor to normal class (T2, T30, T33, T36, T37). It was observed that five among the 8 misclassified samples were misclassified by the normalized EM.

To clearly demonstrate the power of the normalized EM clustering algorithm, we offer the analysis on the small data set of 62 samples $\times$ 500 genes (Genes were selected from the data set of 62 samples $\times$ 2000 genes using t-statistics given known class labels). Table V shows the detailed performance of the normalized EM on the small colon data set. As can be seen, the normalized EM worked stably when $\mu$ was in the range from 20 to 350 with usually only 6 misclassified samples. Also similarly as for the acute leukemia data, from Table V and Figure 2 the values of $\mu$ where the normalized EM worked best ($\mu \geq 33$) follow right after the ones corresponding to the steepest drop in the average number of iterations.

TABLE VI

CLUSTERING RESULTS OF THE UNNORMALIZED EM ON THE REDUCED COLON DATA SET (ENCLOSED IN PARENTHESES ARE THE NUMBER OF TIMES OBSERVING THE CORRESPONDING RESULTS AMONG 10 RUNS).

| Number of principal components | Average number of misclassified samples | | Minimum number of misclassified samples |
|---|---|---|---|
| | Good | Poor | |
| 3 | (0) | 27.5(10) | 25(1) |
| 4 | (0) | 22.4(10) | 18(2) |
| 5 | (0) | 23.9(10)) | 17(2) |
| 6 | (0) | 24.3(10) | 16(1) |
| 7 | (0) | 24.8(10) | 16(1) |
| 8 | (0) | 25(10) | 21(1) |

TABLE VII

CLUSTERING RESULTS OF SPHERICAL $k$-MEANS ON THE COLON DATA SET.

| Cluster quality | Average number of misclassified samples | Number of times observing the result among 20 runs |
|---|---|---|
| Good | 8.1 | 15 |
| Poor | 26.2 | 5 |
| Best | 6 | 4 |

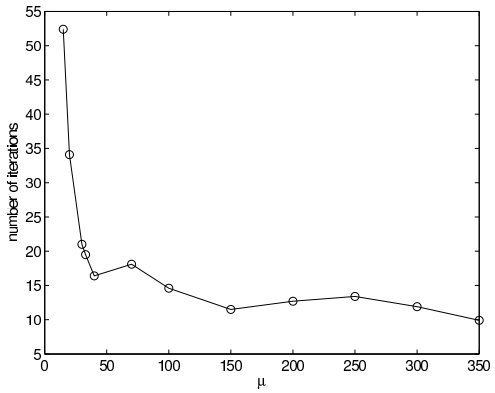


Fig. 2. Convergence speed of the normalized EM on the reduced colon data set

As we already know, the unnormalized EM algorithm for Gaussian mixture model-based clustering is not capable of dealing with the situation where the number of data points is smaller than the dimension of the data, we had to resort to PCA in order to reduce the data set of 62 samples $\times$ 500 genes to the one of 62 samples $\times$ $q$ principal components. The clustering results of the unnormalized EM on the reduced data set of dimension $q$ are shown in Table

VI. As can be observed, the normalized EM with the support of PCA here failed to detect the distinction between tumor and normal tissues in the colon data. The main reason is that PCA was unable to preserve the inherent cluster structure of the data.

On the other hand, spherical $k$-means was able to produce good clusterings on the data set of 62 samples $\times$ 500 genes but still not as stable as the normalized EM in recovering cluster structure of the data (Tables V and VII).

***Pediatric Acute Leukemia Data***

TABLE IX

VI VALUES PRODUCED BY THE UNNORMALIZED EM COUPLED WITH THE SUPPORT OF PCA ON THE PEDIATRIC ACUTE LEUKEMIA DATA SET (10 RUNS WERE PERFORMED FOR EACH VALUE OF $q$)

| $q$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| Average VI | 2.42 | 2.09 | 2.35 | 2.28 | 2.11 | 2.08 | 2.2 |

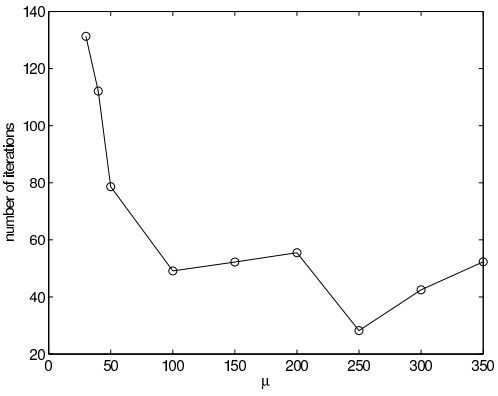


Fig. 3. Convergence speed of the normalized EM on pediatric acute leukemia data set

TABLE X

VI VALUES PRODUCED BY THE NORMALIZED EM ON THE REDUCED PEDIATRIC LEUKEMIA DATA SET

| $q$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| Average VI | 2.14 | 2.01 | 2.06 | 2.01 | 1.92 | 1.77 | 1.78 |

The original data set consists of 327 samples $\times$ 12625 genes and were described and analyzed in [27]. These 327 samples are supposed to be categorized into 7 classes corresponding to 7 leukemia subtypes: BCR-ABL, E2A-PBX1, Hyperdip50, MLL, T-ALL, TEL-AML1 and the other subtypes.

For the purpose of clear analysis, we only took a small subset of the original data, where the relevant genes were selected using feature correlation selection (CFS) as shown publicly at http://www.stjuderesearch.org/data/ALL1. This small data set only consists of 327 samples $\times$ 345 genes. Our analysis and comparison were carried out for the small data set.

To assess the quality of clustering results, the variation of information (VI), which is an information theoretic measure introduced by [29], was utilized. VI is an external index designed to assess the agreement between two partitions of the data, the real clustering $\mathcal{C} = \{\mathcal{X}_h\}_{h=1}^{K}$ and the one induced from predefined class labels $\mathcal{C}^* = \{\mathcal{X}_h^*\}_{h=1}^{K}$. This index is interpreted as the sum of the amount of information left on $\mathcal{C}$ given $\mathcal{C}^*$ and the amount of uncertainty about $\mathcal{C}^*$ given the presence of $\mathcal{C}$. The smaller the value of VI, the better the clustering. In our comparisons, both the partitions had the same number of clusters.

Table VIII shows the clustering results of the normalized EM on the pediatric acute leukemia data set. As shown, values of VI are smallest when $\mu$ is around 50 and Figure 3 indicates that these values of $\mu$ are right after the ones corresponding a notable decrease in the average number of iterations.

Based on the information from Tables VIII and IX, the normalized EM gave better clustering performance compared to the combination of PCA and the unnormalized EM. We also performed clustering on the reduced data set of $q$-dimensional samples obtained after applying PCA on the data set of 327 samples $\times$ 345 genes. As can be seen from Table X, the normalized EM gave smaller corresponding average VI values for each of the selected number of principal components $q$ on the reduced data set.

We next analyze the results produced by spherical $k$-means clustering on the pediatric acute leukemia data set of 327 samples $\times$ 345 genes. Given the results presented in Tables VIII and XI, we find that with the values of $\mu$ where the normalized EM worked best, e.g. $\mu = 50$, it consistently produced smaller values of VI compared to spherical $k$-means.

In the current work on this small data set of 327 samples $\times$ 345 genes [30], the authors utilized average linkage hierarchical clustering to group samples. We again applied the average linkage procedure using Pearson correlation to measure similarity between samples on this pediatric leukemia data set and the value of VI we got is 2.17, bigger than all of the average VI values produced by the normalized EM as shown in Table VIII.

## IV. CONCLUSIONS AND FUTURE WORKS

We have introduced, described and analyzed a new normalized EM algorithm for tumor clustering using gene expression data. It has been demonstrated that the normalized EM algorithm is stable with very high dimensional data sets even with random initializations. Additionally, a detailed analysis of the convergence speed of the normalized EM with respect to different values of $\mu$ has also been provided, and from the analysis an interesting relationship between the convergence speed of the algorithm with the values of $\mu$ at which the normalized EM works best has been presented.

It is left for future works to include unsupervised feature selection methods into our framework so that our approach is able to work with microarray data sets where many noisy or irrelevant genes for clustering are present. Also we will apply this statistical framework to investigate the gene clustering problem.

TABLE VIII

VI VALUES PRODUCED BY THE NORMALIZED EM ON THE PEDIATRIC ACUTE LEUKEMIA DATA (10 RUNS WERE PERFORMED FOR EACH VALUE OF $\mu$).

| $\mu$ | 30 | 40 | 50 | 100 | 150 | 200 | 250 | 300 | 350 |
|---|---|---|---|---|---|---|---|---|---|
| Average VI | 1.46 | 1.37 | 1.32 | 1.39 | 1.43 | 1.49 | 1.54 | 1.45 | 1.46 |
| Average number of iterations | 118.9 | 82.6 | 79.1 | 56.3 | 59.7 | 67.1 | 44.2 | 55.4 | 41.5 |

TABLE XI

CLUSTERING RESULTS OF SPHERICAL $k$-MEANS ON THE PEDIATRIC ACUTE LEUKEMIA DATA SET (20 RUNS WERE PERFORMED).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| VI | 1.64 | 1.42 | 1.78 | 1.88 | 1.41 | 1.6 | 1.13 | 1.73 | 1.64 | 1.64 |
| | 1.64 | 1.42 | 1.78 | 1.88 | 1.41 | 1.6 | 1.13 | 1.73 | 1.64 | 1.64 |
| Average VI | 1.59 | | | | | | | | | |

## REFERENCES


[1] D.J Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and P.O. Brown, "Expression monitoring by hybridization to high density oligonucleotide arrays," *Nature Biotechnology*, vol. 14, pp. 1675–160, 1996.

[2] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative monitoring of gene expression patterns with a DNA microarray," *Science*, vol. 210, pp. 467–470, 1995.

[3] D. Shalon, S.J. Smith, and P.O. Brown, "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization," *Genome Research*, vol. 6, pp. 639–645, 1996.

[4] J.L. DeRisi, V.R. Iyer, and P.O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, pp. 680–686, 1997.

[5] L. Wodicka, H. Dong, M. Mittmann, M.H. Ho, and D.J. Lockhart, "Genome-wide expression minotoring in saccharomyces cerevisiae," *Nature Biotechnology*, vol. 15, pp. 1359–1366, 1997.

[6] R.J. Cho, M.J. Campell, E.A. Winzeler, E.A. Steinmetz, A. Conway, L. Wodicka, T.J. Wolfsberg, A.E. Gabrielan, D. Landsman, D.J. Lockhart, and R.W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, pp. 65–73, 1998.

[7] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences, USA*, vol. 96, pp. 6745–6750, 1999.

[8] V.R Iyer, M.B. Eisen, D.T Ross, G. Schuler, T. Moore, J.C.F Lee, J.M. Trent, L.M. Staudt, J. Hudson, M.S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P.O. Brown, "The transcriptional program in response of human fibroblasts to serum," *Science*, vol. 283, pp. 83–87, 1999.

[9] C.M. Perou, S.S. Jeffrey, M. van de Rijn, C.A. Rees, M.B. Eisen, D.T. Ross, A. Pergamenschikov, C.F. Williams, S.X. Zhu, J.C.F. Lee, D. Lashkari, D. Shalon, P.O. Brown, and D. Botstein, "Distinctive gene expression patterns in human mamary epithelian cells and breast cancers," *Proceedings of the National Academy of Sciences, USA*, vol. 96, pp. 9112–9217, 1999.

[10] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," in *Proceedings of the National Academy of Sciences of the United States of America*, 1998.

[11] X. Wen, S. Fuhrman, D. B. Carr G. S. Michaels, Susan Smith, J. L. Barker, and R. Somogyi, "Large-scale temporal gene expression mapping of central nervous system development," The national academy of sciences, January 1998, pp. 334–339.

[12] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," in *Proceedings of the National Academy of Sciences of the United States of America*, 1999, pp. 2097–2912.

[13] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church, "Systematic determination of genetic network architechture," *Nature Genetics*, vol. 22, pp. 281–285, 1999.

[14] G.C. Tseng, "Penalized and weighted $k$-means for clustering with scattered objects and prior information in high-throughput biological data," *Bioinformatics*, vol. 23, no. 17, pp. 2247–2255, 2007.

[15] F.D Smet, J. Mathys, K. Marchal, G. Thijs, B.D. Moor, and Y. Moreau, "Adaptive quality-based clustering of gene expresion profiles," *Bioinformatics*, vol. 18, no. 5, pp. 735–746, 2002.

[16] R. Sharan and R. Shamir, "CLICK: a clustering algorithm with applications to gene expression analysis," in *Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology(ISMB)*. 2000, pp. 307–316, AAAI Press.

[17] Y. Xu, V. Olman, and D. Xu, "Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees," *Bioinformatics*, vol. 17, no. 4, pp. 309–318, 2001.

[18] G.J. McLachlan, R.W. Bean, and D. Peel, "A mixture model-based approach to the clustering of microarray expression data," *Bioinformatics*, vol. 18, no. 3, pp. 413–422, 2002.

[19] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, vol. 17, pp. 977–987, October 2001.

[20] D. Ghosh and A. M. Chinnaiyan, "Mixture modelling of gene expression from microarray experiments," *Bioinformatics*, vol. 18, no. 2, pp. 275–286, 2002.

[21] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood for incomplete data via the EM algorithm," *Journal of Royal Stastistical Society*, vol. 29, pp. 1–38, 1977.

[22] I. T. Jolliffe, *Principal component analysis*, Springer Series in Statistics, $2^{nd}$ edition, 2002.

[23] I.S. Dhillon and D.S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, no. 1, pp. 143–175, 2001.

[24] A. Banerjee, I.S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *Journal of Machine Learning Research*, vol. 6, pp. 1345–1382, Sept 2005.

[25] J. L. Dortet-Bernadet and N. Wicker, "Model-based clustering on the unit sphere with an illustration using gene expression profiles," *Biostatistics*, vol. 0, no. 0, pp. 1–15, 2007.

[26] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.

[27] E. J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, W. E. Evans, C. Naeve, L. Wong, and J. R. Downling, "Classification, subtype discovery and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, pp. 133–143, March 2002.

[28] T.S. Furey, N. Cristianini, D.W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, pp. 906–914, 2000.

[29] M. Meila, "Comparing clusterings," In COLT, 2003.

[30] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. D. Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature genetics*, vol. 30, pp. 41–47, January 2002.