CLUSTERING SINGLE CELL EXPRESSION PROFILES USING MODIFIED GAUSSIAN MIXTURE MODEL

BILAL AHMED

ROLL NO: 17MBI006

Department of Computer Science
Faculty of Natural Science
Jamia Millia Islamia
May 2019

CLUSTERING SINGLE CELL EXPRESSION PROFILES USING MODIFIED GAUSSIAN MIXTURE MODEL

BILAL AHMED

Enrolment No: 17-0720

Submitted

In partial fulfilment of the requirements of the award of the degree of

Master of Science (Bioinformatics)

Under the Supervision of

Dr Debarka Sengupta

Department of Computer Science

Faculty of Natural Science
Jamia Millia Islamia
May 2019

Declaration

I, Bilal Ahmed, student of M.Sc. (Bioinformatics) hereby declare that the project report entitled "**CLUSTERING SINGLE CELL EXPRESSION PROFILES USING MODIFIED GAUSSIAN MIXTURE MODEL**" which is submitted by me to the Department of Computer Science, Jamia Millia Islamia, New Delhi, in partial fulfilment of the requirement of the degree of **M.Sc. (Bioinformatics),** have not been submitted in part or full to any other university or institute for the award of any degree or diploma.

(Bilal Ahmed)

Roll No.:17MBI006

Enrollment No.: 17-0720

Certificate

On the basis of declaration made by the student **Bilal Ahmed,** I hereby certify that the project report entitled **"CLUSTERING SINGLE CELL EXPRESSION PROFILES USING MODIFIED GAUSSIAN MIXTURE MODEL"** submitted by **Bilal Ahmed** to the Department of Computer Science, Jamia Millia Islamia, New Delhi, for the partial fulfilment of the requirements of the degree of **M.Sc. (Bioinformatics)**, is carried out by him under my guidance and supervision. The report has reached the requisite standards for submission.

Dr. S.M.K Quadri
Prof. (HOD), DCS
Jamia Millia Islamia

(Internal Supervisor)

Dr. Debarka Sengupta
Asst. Prof, CCB
IIIT-Delhi

(External Supervisor)

Acknowledgements

First of all, I would like to thank my supervisor, Dr Debarka Sengupta for providing me such an opportunity, I consider myself fortunate enough to work under one of the pioneers of scRNA analysis in India. He never fails to amaze me for his guidance and support throughout my time as a trainee. His vision and easy-to-go-attitude are the main cause for the successful completion of this work. I would also like to thank Dr Abhik Ghosh, my co-guide, His achievements and conceptual knowledge always inspire me, I am also thankful to Prof. S.M.K Quadri, Internal Supervisor for his time in guiding and humble nature to help with evaluating this report before submission.

I would like to thank Dr Rafat. Parveen and our Course coordinator, Dr Khalid Raza. Teaching Faculty of the department had always helped, guided, and educated us. I thank Dr Munazzah Tasleem who guided me throughout my time at Jamia. Other office staff members were at all times there on the need. They had always provided me with the best services.

I would like to express humble gratitude to, Chitrita Goswami, a PhD scholar at IIIT, New Delhi. Her philosophies have always nurtured and guided me in the project. Many thanks go to Lab mates and friends, especially Dinesh, Yash, Khushnuma and the group for the motivation when needed and for keeping the coffee flowing.

I would like to thank my aunt S.Parveen for the warm hospitality and all those tea-break at midnights which eventually helps me to complete this.

I would like to thank my family members for their constant support, backup, and encouragements, and for putting up with me over the past years and for bringing joy and spirit into my life. At last but not least, I thank Almighty Allah who has given this beautiful life, guided us to follow the path of truth and gave enough courage to complete this dissertation.

This thesis is dedicated to Ammi and Papa.

Bilal Ahmed

TABLE OF CONTENTS

List of Figures

Abstract

With the advent of recent technologies in biology, automated medical data is growing at an unprecedented rate and analysis of these voluminous data turned out to be one of the major complexities for the biologists. Biologists have always attempted their hands on Clustering, to give each form of life a meaningful definition. scRNA Sequencing, on the contrary to Bulk-RNA Sequencing, gives a deeper insight into processing at cellular level. However, a normal ScRNA assay could produce an enormous amount of high-dimensional data which is not possible to visualize unless dimension reduction techniques are applied to it.
Machine learning researchers have employed many such algorithms to solve and analyse the complexity in heterogeneity and functional diversity among cells population. Among those algorithms, Model-based algorithms are what we're most interested in, GMM based density-based clustering has been extensively used in wide ranges, However the existing methods to calculate the MLE (maximum Likelihood Estimators), EM is very sensitive to initialization.

In this thesis, we're presenting two loosely related parts: Initialization of EM algorithm for GMM and Visualization of ScRNA sub-population detection by PCA. Model-Based Clustering method known as Density or Finite Mixture Models which models the density that generates the data, a special case where each mixture component is a multivariate Gaussian often used known as GMM is quite easy to interpret.

We'll be using the EM algorithm to predict class labels of scRNA data assuming it as coming from the mixture models and will visualize it by dimension technique as PCA. Our work shows that GMM greatly improves the clustering results by calculating adjusted rand score (91.32%), as well as the visualization and interpretability of the data.

Introduction

1

In this Chapter, we introduce Clustering, along with an overview of different type of techniques that are available for clustering (Section 1.1) and we'll discuss scRNA sequencing and its significance in section 1.2

Clustering, it is one of the practises which humans are using naïvely from the pre-dawn era, the early-civilization, early domestication of animals, cereals, all are the illustrations of human minds to select the best, or to sorting items for making life well-co-ordinated. There is no such universally accepted definition of a cluster. Through clustering, we can make a large dataset easy to understand for humans. It is a useful tool for summarizing the data and could reduce the intricacy of a dataset.

On a primary level, it may seem that clustering and classification work alike, but they are way too different than the norm. In Classification, each class is well defined by labels in training, which is not provided in clustering. Therefore, Clustering relies on the structure of the data to determine clusters, instead of learning from the sample labels. Clustering is the classification of objects into different groups or subsets, or the partitioning of the dataset into subsets, the clustering (Jain et al., 1999) is the problem of dividing a given set {x1. . . xN} of N data points into several non-overlapping homogenous groups. Each such group or cluster should contain similar data items and data items from different groups should not be similar.

Many different approaches to clustering problem have been developed, some works on data characterized by their coordinate in a feature space and some on a matrix of pairwise similarities between data points, to give a short overview of different types of methods, we categorize them into three groups.

We'll discuss about the clustering approaches from the traditional to advance ones in the following section.

1.1 Background:

1.1.1 Hierarchical Clustering: As the name suggests, it forms a hierarchy of clusters. The levels of hierarchy contain data and at each succeeding level, it splits in two. The last level contains all data in individual clusters, it could be based on the pairwise similarity between data points and is assembled either by a top-down or bottom-up approach.

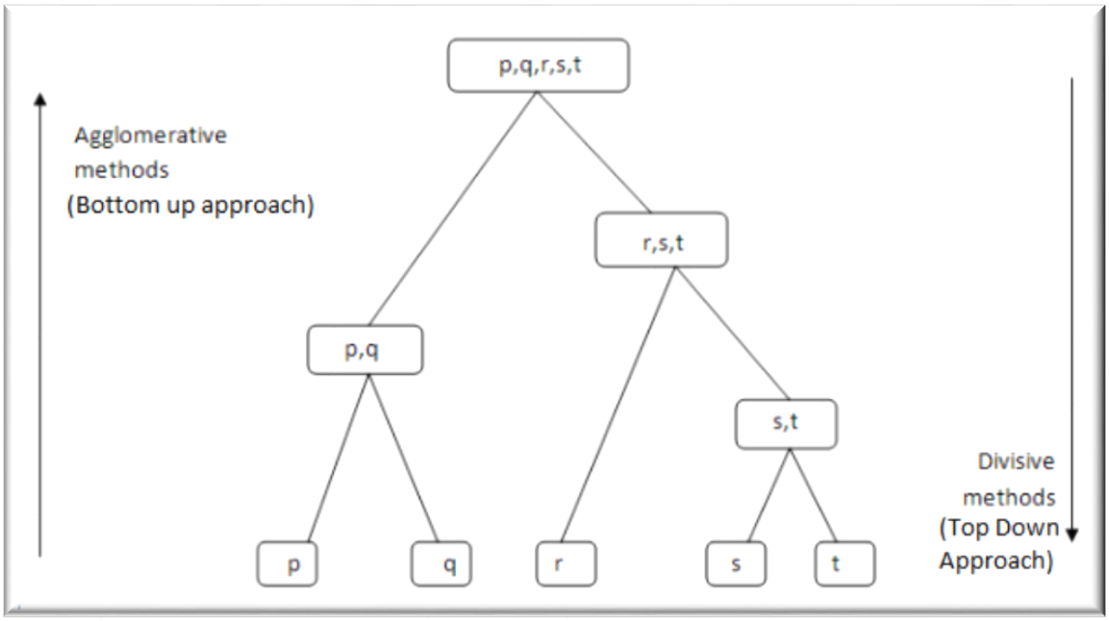

Fig 1. A visual to compare Top-down and Bottom-Up approaches (Perceptive analytics, 2018)

1.1.2 Partitional Clustering:

The most traditional approaches to clustering come under partitional clustering, they form single clustering with a distinct number of clusters. They don't operate on the basis of pairwise similarity but with the data in feature space. In this method, Parameters are initialized randomly and tries to maximize or minimize a function using an iterative algorithm to upgrade the initial parameters. Partitional clustering approaches are much more efficient than Hierarchical clustering methods but they also have their limitations i.e., Spherical shapes clusters etc. we'll be discussing this in section 1.2

K-means

Also Known as Lloyd algorithm or Gray algorithm (1992), is the one of the most commonly applied unsupervised algorithm for partitioning a given data set into a set of k groups, it aims to reduce the sum-of-squared function, it improves its clustering by

$$W(C_k) = \sum_{x^i \in C_k} (xi - u_k)^2$$

x^i Is a data point belonging to cluster C_k

u_k Is the mean of points assigned to the cluster C_k

It could be summarized as:

- Specify the no. of clusters to be created.
- Select randomly k objects from data, set them as initial cluster centres or means.
- Assign each observation to the nearest centroid, based on Euclidean distance between centroid and object.
- Update the centroid of the cluster by calculating new mean values assigned to that cluster
- Iteratively, Minimize the total sum of the square, and stops until it gets converged.

1.1.3 Model-Based Clustering:

An alternative is Model-Based Clustering, which assumes data follows a distribution that is a mixture of two or more clusters, it incorporates the soft assignment, where each point has the probability to belong more than a cluster.

Concept of Model-based Clustering:

GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities, are commonly used as a model of the probability distribution of continuous measurements or features. The component priors can be view as a weight in the

output layer. Parameters are estimated from training data using the iterative Expectation-Maximization (E-M) Algorithm.

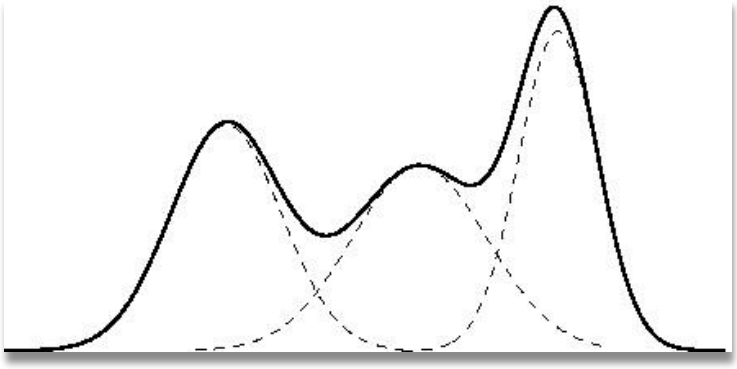

Fig.2 Image showing multimodal distributions.

Each component (i.e. cluster) k is modelled by Normal or Gaussian distribution which is characterized by the parameters:

μk: mean vector,

$\sum_k$: Covariance matrix

Ω: Weights, associated probability in the mixture (Each point has a probability of belonging to each cluster.

$$f(x) = \frac{1}{\sigma\sqrt{(2\pi)}} e \frac{-(\mu - x)^2}{2\sigma^2}$$

Here, The parameters are:

1. Mean

2. Variance and Covariance in 1-D and 2-D respectively.

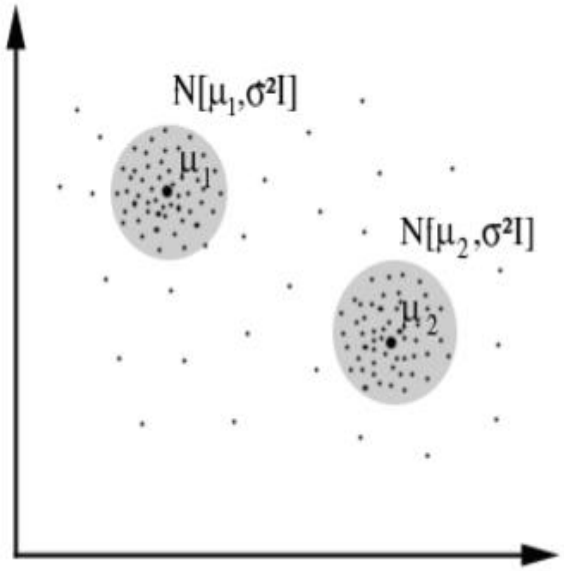

Fig.3 Mixture Model-based Clustering

This representation of each cluster with a parametric representation like Gaussian. The assumption is that entire dataset is modelled by a mixture (a weighted sum) of these distributions. Each distinct distribution used for modelling a cluster is referred to as a component distribution.

GMM estimates its parameters with an iterative algorithm, EM which is described in following part

EM Algorithm:

Expectation- Maximization algorithm was proposed to calculate the maximum likelihood estimator (MLE) when data is missing. The general form of EM can be put as:

Let **Y** be a set of observed data, **Z** a set of unobserved data, and **θ** parameters.

The aim of EM algorithm is to find the MLE, i.e., the maximum of the following function known as observed data likelihood.

$$L(\boldsymbol{\theta}|\mathbf{Y}) = E_z[p(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\theta})]$$

Where,

$$L(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}) = p(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\theta})$$

Is often referred to as the complete data likelihood.

The EM algorithm proceeds by iterating between the following two steps:

Expectation step (E-step): It calculates the expectation of the likelihood with respect to the conditional distribution of **Z** given **X** and the current estimate of the parameter $\boldsymbol{\theta}_{[t]}$

Maximization step (M-step): It updates the parameter.

$$\boldsymbol{\theta}[t+1] = \text{argmax}_{\theta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}[_t])$$

The heuristic for this algorithm is to replace the unobserved variables with their conditional expectations given observed data and the current guess of the parameters (E-Step) and to then update the parameter by maximizing the conditional likelihood function (M-Step). The simplicity of this algorithm is an important reason why it gained so popularity.

1.2 Motivation of Work

ScRNA-Sequencing

Rapid advancement in the development of NGS has enabled to look deeper into complex biological systems, now the shift has been observed on the characterization of individual cells. It will allow the researcher to reveal new hidden insights. Single-cell RNA sequencing (ScRNA-seq), for example, can uncover complex and rare cell detections [1].

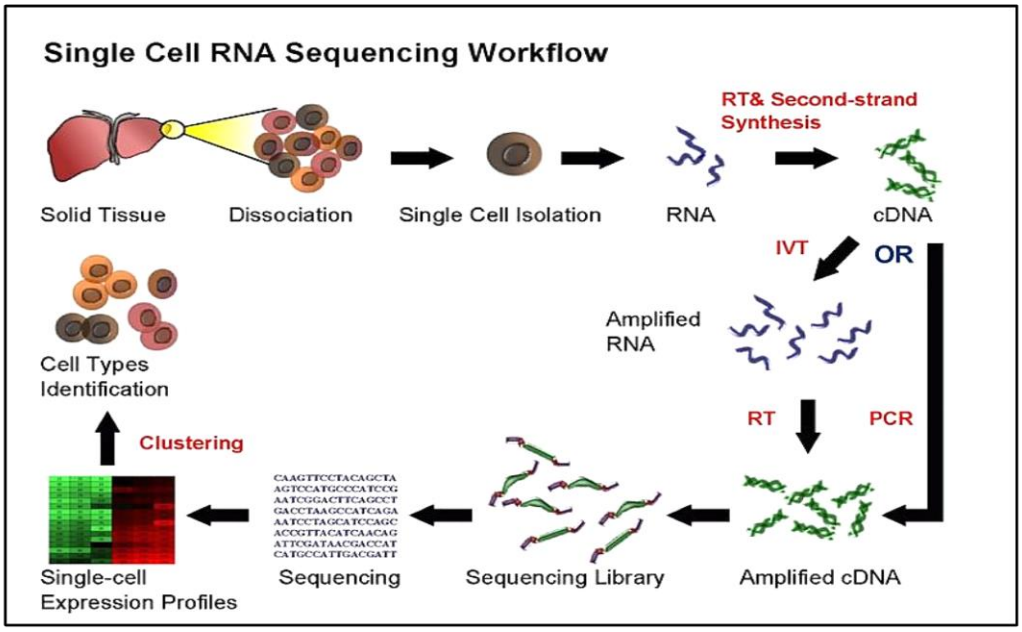

Fig.4 ScRNA workflow (10X Genomics)

In the last decade, ScRNA seq data is almost generating at an exponential rate. Recent studies have demonstrated that de-novo cell type discovery and Identification of distinct cell lines are possible via unbiased analyses of all transcriptomic information provided by ScRNA data. Therefore, unsupervised clustering of individual cells using ScRNA seq is critical to developing new biological insight and validating prior knowledge.

This has driven the development and application of a wide range of clustering methods based on various underlying algorithms. However, there are already multiple algorithms to cluster ScRNA high-dimensional data, but most of them were either computationally expensive or unsophisticated. Zheng et.al used a partitional clustering algorithm (k-means) as a method to cluster ScRNA-seq data. It is quite fast but also suffers from its own limitations.

Drawbacks of K-means:

1. Prerequisite knowledge to specify the number of clusters.
2. Works mostly for spherical shapes (K-mean has no built-in way of accounting oblong or elliptical clusters)
3. Ambiguity in the real-dataset.

To overcome these above shortcomings, we're using Density-based mixture models, which can be seen as the extension of the ideas behind k-means but can also be a powerful tool for estimation beyond simple clustering. It is more mathematically sound algorithm which

operates on the probabilistic approach rather than the Euclidean distance and allows the uncertainty, it could serve the purpose of finding clusters incorporating soft assignments which is true for any real-world dataset.

High dimension dataset cannot be just mentioned without talking dimension reduction, in the following part, we'll be discussing the curse of dimension and the technique PCA to overcome the problem.

The curse of dimensionality

This Term was coined by Bell-man (1961), it refers to the phenomenon that arises while analysing high-dimension data. As they said, excessive information is a curse anyway. Principle Component Analysis (Pearson 1901), is a linear feature extraction technique that can be driven from different perceptions. PCA minimized the total squared distance between the original data and its reconstruction from the extracted features. PCA is a method to find the linear feature with maximum variance, by projecting the data on Principal component, we discard the dimensions of data with less variance and if dimensions are not so important one can discard to reduce the dimensionality. In General, classifying methods for larger dimensions need numerous examples to estimate the accuracy, it almost becomes impossible to handle them practically. Clustering provides a compressed representation of the data by mapping each data point to a discrete cluster index whereas dimension reduction is used to find a compact representation of data by mapping each point to lower dimension vector.

To understand this, consider a basic approach to Classification and regression, where we partition the input dimension into y cells. For D input Dimensions, No. of cells are $\mathbf{y^D}$ and to estimate the function of output we would need at least $\mathbf{y^D}$ examples/

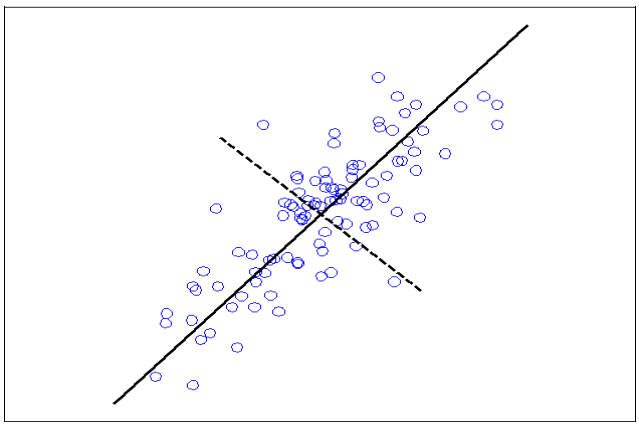

Fig. 5 Example of PCA, data are plotted as circles, PC and direction with least variance are plotted.

1.3 Objectives

The objectives of our studies are as follows:

- To develop a modified algorithm (Mixture Model) to cluster ScRNA seq-data
- Visualization and analysis of ScRNA seq data.
- To evaluate the adjusted rand_score of our algorithm

1.4 Organization of Chapters

The first chapter of the thesis includes introduction along with the fundamentals of Clustering algorithms and ScRNA sequencing technology, almost all major algorithms and their shortcomings are covered in this portion under the heading of background.

The second chapter is on Literature review, which is an extended version of the background discussed in the introduction part of the thesis, it is an ample collection of the literature related to ScRNA and their clustering's results along with the E-M part.

The third Chapter, the methodology includes the step by step details, involved in this work, there are further subdivisions of the chapter on the basis of computational and statistic tools involved in the series. We'll begin with Data collection, to its re-processing and final cleansing of data, then its dimension reduction to the further algorithm to cluster. An algorithm was initialized for simulated data and we try to fit random Gaussian's over the data to maximize the likelihood function.

The fourth chapter is on Results, It is a crucial chapter consisting of the results, which were obtained through the comparative evaluation of ScRNA Clustering and their adjusted rand score for the quantification of the clustering results.

The last chapter of the thesis comprises all the discussions related to the results obtained, along with the limitations that we discovered in the work, and also the future work-idea to enhance our clustering results, in this work, we have also mentioned our future work too.

2 Literature Review

ScRNA Seq

ScRNA (single-cell RNA) Sequencing has enabled the wide profiling of cell transcriptomes in a single run, ScRNA is reasonably good at capturing gene expression at cellular levels. A single experiment could lead to 250k single cell expression profile. (Zheng et. al, 2017).

There are many advantages of ScRNA seq over its contrary Bulk RNA sequencing, Bulk RNA sequencing gives the average expression profile of the cells, while the single cell is equipped to allow the heterogeneity in the expression profile, there is a good analogy of ScRNA and Bulk RNA sequencing with Salad and Smoothie. The distinctness in the former allows researchers to look into deeper insights at the individual cellular level.

While there are multiple protocols to perform ScRNA sequencing, i.e., Microfluidic based method, Plate method and Droplet-Based Method, this latter method represented by 10x Genomics Chromium has received great attention in the recent times because of its high efficiency, relatively lower cost and high throughput (Zheng et. al)

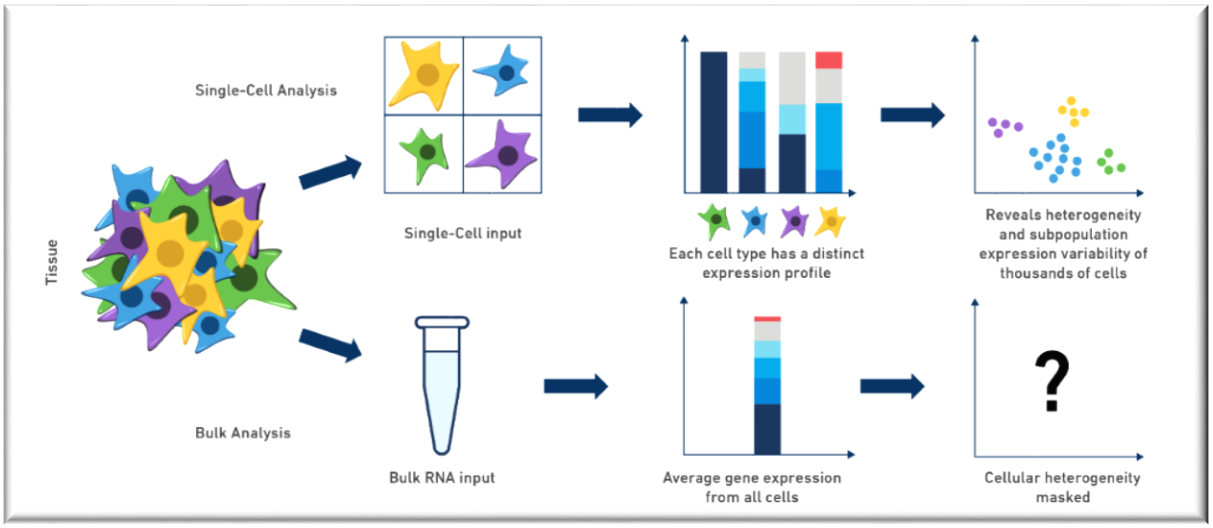

Fig.6 Comparison of ScRNA vs Bulk RNA Sequencing (10X genomics.com)

ScRNA Seq has already been used to study several different tissues and organs. These studies include various regions of the brain (Darmanis et al., 2015; Karlsson and Linnarsson,2017; Liu et al., 2016; Tasic et al., 2016; Zeisel et al., 2015), retina((Baron et al., 2016; Jaitin et al., 2014; Macosko et al., 2015; Zheng et al., 2017)), pancreas (Baron et al., 2016; Segerstolpe et al., 2016;Wang et al., 2016), immune cells (Jaitin et al., 2014; Villani et al.,2017), early embryonic development (Biase et al., 2014; Goolamet al., 2016; Xue et al., 2013) and in haematopoiesis (Velden et al.,2017; Wilson et al., 2015). Below is a visual showing the ScRNA scaling done in the last decade.

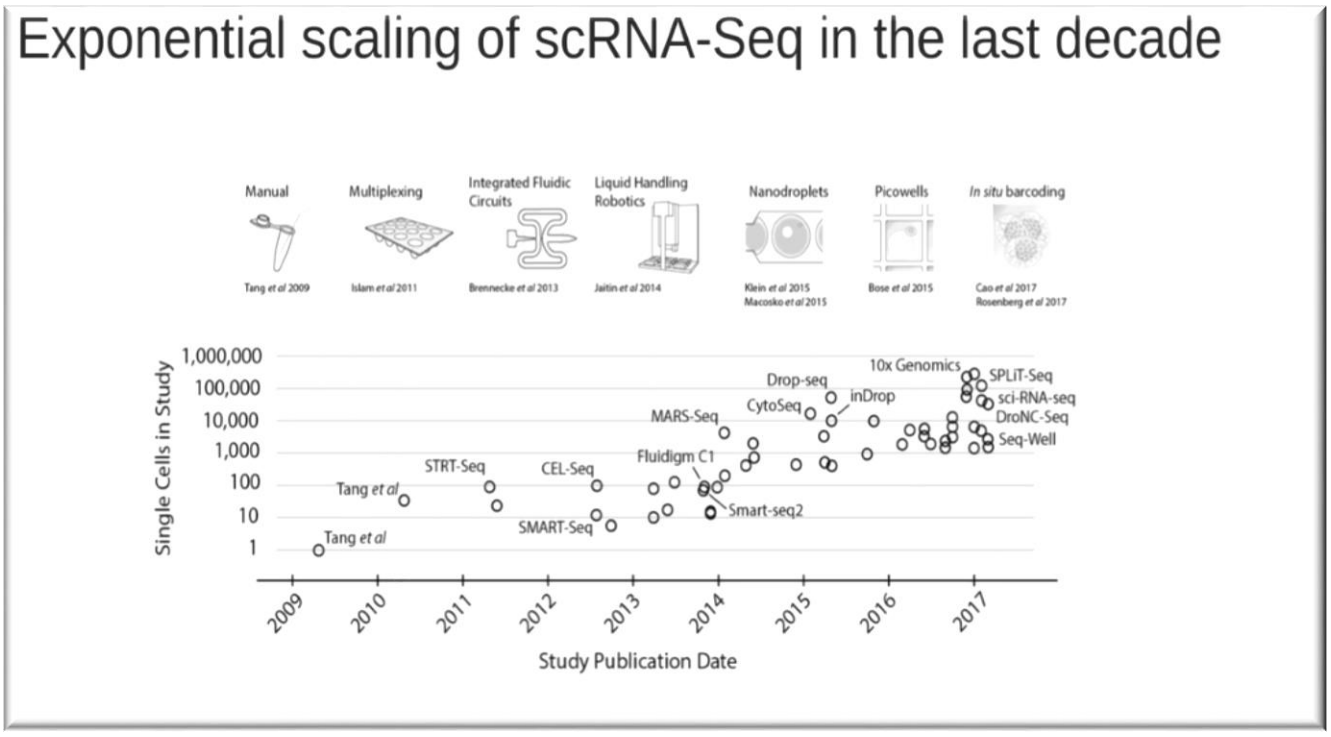

Fig 7. Exponential scaling of single-cell RNA sequencing in the last decade.

(Valentine Svensson, Roser Vento-Tormo, Sarah A. Teichman)

Recent attraction in ScRNA seq work leads to an enormous amount of data production which is not a trivial task to analyse without using the mathematically sophisticated algorithm. One of the most crucial tasks of ScRNA seq data is to find clusters of single cells, to accomplish this objective, many existing single-cell algorithms employ statistical methods that have been developed initially for analysis of bulk RNA-Seq generated data.

These methods fail to address the characteristics that make single cell expression data, especially challenging to analyse: outlier cell population, transcript noise and biological effects.

Unsupervised methods such as K-means clustering, is commonly employed clustering algorithm for ScRNA cell analysis (Burns et al., 2015; Grün et al., 2015; Kiselev et al., 2017; Muraro et al., 2016; Tsang et al., 2015).K-means is a very fast method however had major shortcomings However k-means with outlier detection methods, e.g. Race ID (Grun et al, 2015) can identify rare cell populations.

Hierarchical can be applicable after normalization. It is a general purpose method, Different variants may lead to different assumptions, but the most common ones, Ward (ward, 1963) It is quite slower but has the upper edge for determining the relationship between different clusters as a result, can be viewed as a dendrogram.

Density-based Clustering, (Campbell et al., 2017; Jiang et al., 2016; Macosko et al., 2015) requires a large number of samples to accurately estimate densities. DBSCAN (Ester at al.) combined with dimension reduction in Novel Seurat (Macosko et al.) is one of the dominant method used for ScRNA seq clustering.

There are ScRNA-seq tailored methods that have been designed and proposed especially for ScRNA-seq data such as SIMLR, Dropclust, SC3, and TSCAN.

Dropclust (Debajyoti et.al), operates on Locality Sensitive Hashing (LSH) to find the nearest neighbour of distinct transcriptomics. It outperformed the existing best algorithms in terms of computational time, accuracy and detection of cell sub-types.

SIMLR (Single-cell Interpretation via Multikernel Learning) adopted a similarity measure from ScRNA seq data in order to perform dimension reduction, clustering and visualization.

SC3 also known as consensus clustering, uses a combined multiple clustering techniques through a consensus approach which achieves high accuracy and robustness.

TSCAN, they have employed pseudo-temporal path to arrange cells based on the transition of their transcriptome to study gene expression in the heterogeneous cell population. SCAN (Tools for Single-cell Analysis) is developed to support in-silico pseudo time reconstruction in ScRNA Analysis. It uses an MST (Minimum Spanning Tree) approach to arrange cells. It contains GUI to support data visualization and user interaction.

GMM based clustering, it has been extensively implemented for automatic speech recognition and sub-population detection. Market Analyst has applied them to analyse personalized marketing campaigns [M. Wedel et. al]. GMM is a way to construct vector quantizer, after clustering data using GMM, each observation can be mapped to the mean vector of its corresponding Gaussian. But it also fails to achieve a global maximum of the log-likelihood function, therefore, a better initialization scheme is requiring for the task.

Dimensionality reduction:

A variant of PCA has been developed (Andrews and Hemberg, 2016) which explicitly deals with a large number of zeroes in ScRNA seq data but the zero-inflation model may not fit all the datasets. Recently, (Rissoet al. 2017) proposed a method similar to PCA based on a zero-inflated negative binomial model instead of Gaussian mode.

Materials & Methods 3

3.1 Workflow

The framework consists of three major stages, the first stage of the work includes the pre-processing of the scRNA seq data. The initial part of the second stage is related to the algorithm development, the latter part of this stage is focused on the testing on the simulated data. In the final stage of the workflow, clustering is made. A schematic representation of the workflow has been made in Figure

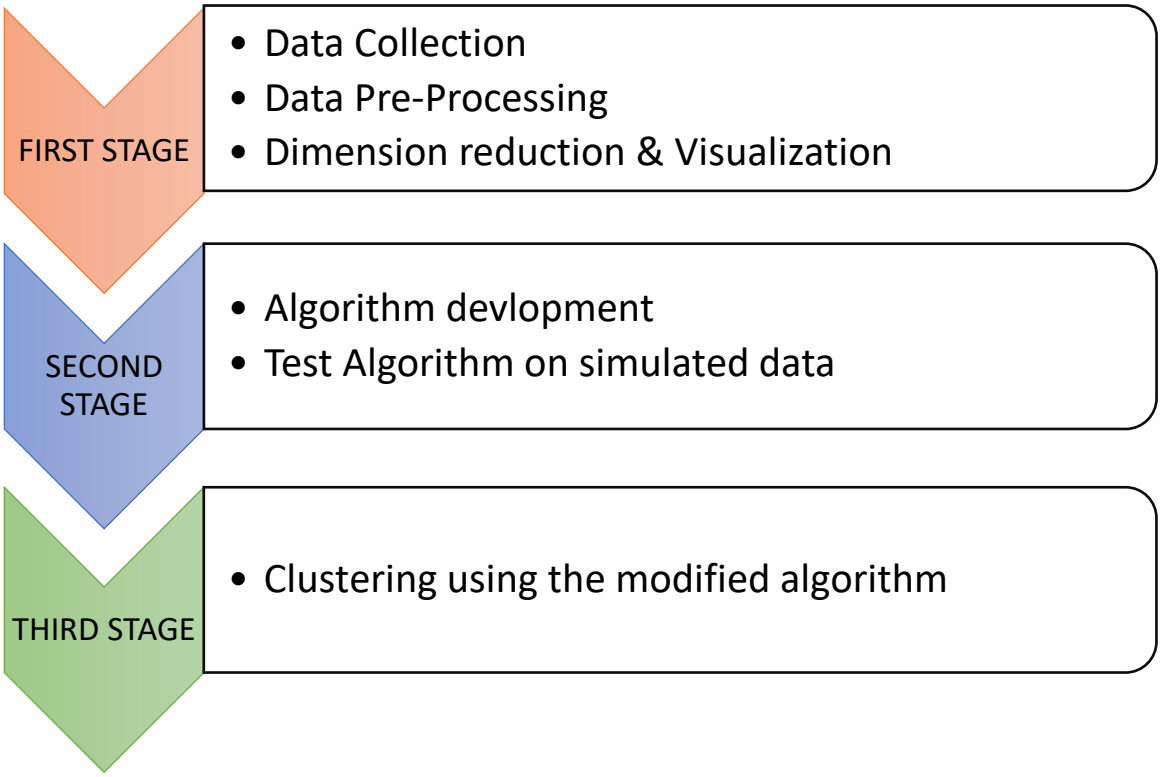

Fig.8 Workflow showing the Stages involved in the Clustering.

The stage workflow discussed in the previous section of this chapter can be further resolved to the following steps:

1. **Data collection**: The Jurkat (T-Lymphocytes) data was collected.
2. **Data Pre-Processing**: Filtering out cells showing poor gene expression and genes having 0 standard deviations.
3. **Data Normalization**: With the help of TMM, Data is normalized.
4. **Dimension Reduction**: PCA was applied to the data
5. **Data Visualization**: PC_1 Data was visualized with provided annotations.
6. **Algorithm Development**: A modified GMM is constructed
7. **Simulated data generation**: A data having multi-modal data points were generated.
8. **Testing on Simulated data**: Algorithm performance was evaluated on testing data.
9. **Clustering**: using Modified algorithm on scRNA data.
10. **Performance analysis**: measurement of accuracy.

For a better understanding of the concepts and knowledge regarding the process, the next section of the chapter will be focused primarily on the detailed elucidation of the implemented steps, and at the end will be a comprehensive description of the mainstay algorithm used in the work i.e. GMM and EM.

3.2 Data collection

1. Jurkat – ScRNA Sequencing data a feature matrix of the dimension of **32738 X 3388** genes and cells was used. This data was adopted from Sengupta Lab.

2. Jurkat Annotations – Labels (From Species1, Species2) were also provided from the above source.

Description of the data:

Jurkat cells are immortalized cell lines of human T lymphocytes that are used to study for acute T cell leukaemia, T cell signalling, and the expressions of various

receptor, particularly, HIV (Abraham et. al), the data comprises of two species (Homo sapiens and Mus-Musculus) in a particular ratio to find out the clusters.

3.3 Data Normalization and Processing

In the case of Jurkat cell-lines data, the expressions profile are normalized using the following method.

Following are the steps of Normalization and pre-processing:

1. Cells having less than 5 gene expressions were filtered out and discarded.
2. Genes having 0 standard deviations were filtered out.
3. Then, we applied TMM Normalization to the data

TMM Normalization: Trimmed Means of M-values, it is one of the simplest and effective methods for estimating relative RNA production level from RNA-Seq data. It estimates scale factors between samples that can be incorporated into the statistical method for differential expression analysis.

After, pre-processing and the normalization of the data, 3292 genes over 7773 cells were obtained as a matrix

```
In [4]: Data
```

Out[4]:

	0	1	2	3	4	5	6	7	8	9	...	7763	7764	7765	7766	7767
0	0.000000	0.527305	3.420503	0.000000	0.000000	1.216425	0.527305	0.000000	0.000000	1.216425	...	5.347350	5.128254	5.930923	2.031621	0.000000
1	0.000000	0.000000	0.000000	2.916216	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	4.765246	4.765246	5.252058	1.670332	0.000000
2	0.000000	1.262268	0.000000	1.262268	1.262268	1.262268	0.000000	1.262268	0.000000	1.262268	...	4.152565	5.377229	5.168647	0.000000	1.262268
3	0.000000	1.389065	3.421497	2.083444	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	3.421497	4.462586	5.129319	2.902335	0.000000
4	0.000000	1.239980	1.896816	3.396936	0.000000	0.749299	0.000000	0.000000	0.000000	0.000000	...	3.935507	4.920388	4.983857	2.965279	0.000000
5	0.000000	0.995082	3.313063	0.000000	0.000000	0.000000	0.000000	1.578401	0.000000	0.000000	...	5.034841	4.450028	5.513924	2.991382	0.000000
6	0.000000	1.360986	2.048659	2.512434	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	3.380125	5.212490	5.212490	2.048659	0.000000
7	0.000000	2.538000	0.000000	2.809371	0.000000	0.000000	1.138787	1.138787	0.000000	0.000000	...	5.164174	5.211742	5.164174	0.000000	1.138787
8	0.000000	1.228229	1.228229	2.330093	1.228229	1.228229	0.000000	0.000000	0.000000	0.000000	...	4.653648	5.270461	4.799908	1.228229	0.000000
9	1.524609	0.954973	1.524609	0.000000	0.000000	0.000000	0.954973	0.000000	0.000000	0.000000	...	3.616184	4.235140	5.081043	1.931922	0.000000
10	0.000000	1.972698	4.155331	0.791817	0.000000	1.300128	1.300128	1.675256	0.000000	1.300128	...	5.725714	5.440046	5.929857	2.776110	0.000000
11	0.000000	1.169738	3.087166	2.247662	0.000000	0.000000	0.000000	1.169738	0.000000	0.000000	...	3.285100	5.218841	5.118614	2.584682	0.000000
12	0.000000	1.558904	1.558904	1.970651	0.000000	0.000000	0.000000	0.980500	0.000000	0.000000	...	3.286636	4.547353	4.662703	0.980500	0.000000
13	0.000000	1.742627	1.742627	0.000000	0.000000	0.000000	0.000000	1.119834	0.000000	1.119834	...	5.130246	4.922697	6.356710	1.119834	0.000000
14	0.840208	0.840208	3.770293	0.000000	0.000000	0.000000	0.000000	2.057011	0.000000	0.000000	...	4.716434	4.716434	5.753568	2.057011	0.000000
15	0.000000	1.044267	3.538934	1.044267	0.000000	1.044267	1.044267	0.000000	1.044267	0.000000	...	5.214080	4.784398	5.407028	0.000000	0.000000
16	0.000000	2.090258	3.104767	0.000000	0.000000	0.000000	0.000000	2.090258	0.000000	0.000000	...	4.869777	4.815166	5.508724	1.665466	0.000000
17	0.000000	2.043258	2.043258	0.000000	0.000000	0.000000	0.000000	1.623477	0.000000	2.043258	...	4.698997	5.728715	5.811238	2.043258	0.000000
18	0.000000	0.000000	0.000000	3.724675	0.000000	0.000000	0.000000	2.712518	0.000000	0.000000	...	3.724675	5.574678	4.901309	1.077878	0.000000
19	0.000000	0.000000	3.203642	2.567469	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	5.651823	5.214761	4.933862	0.000000	0.000000
20	0.000000	1.038410	0.000000	1.038410	0.000000	1.038410	1.038410	1.635952	0.000000	0.000000	...	3.977782	5.321807	4.531894	1.635952	1.038410

Fig .9 Normalized/Cleaned data after pre-processing.

3.4 Dimension Reduction

After dimension reduction, we'll get the following Principle component, with the usage of sklearn imports, we applied dimension reduction on our data,

PC1 denotes the maximum variance and could be used for further analysis of our data.

```
In [745]: X_pca.head()
```

Out[745]:

	PC1	PC2
0	17.229558	-1.475836
1	-14.279237	2.968064
2	-15.736484	2.135867
3	-16.947541	-0.060144
4	-15.657918	-0.170140

Fig.10 PC_1 component on Jurkat Dataset

We try to visualize our PC_1 component and try to model this under a single Gaussian, but that won't fit the data perfectly.

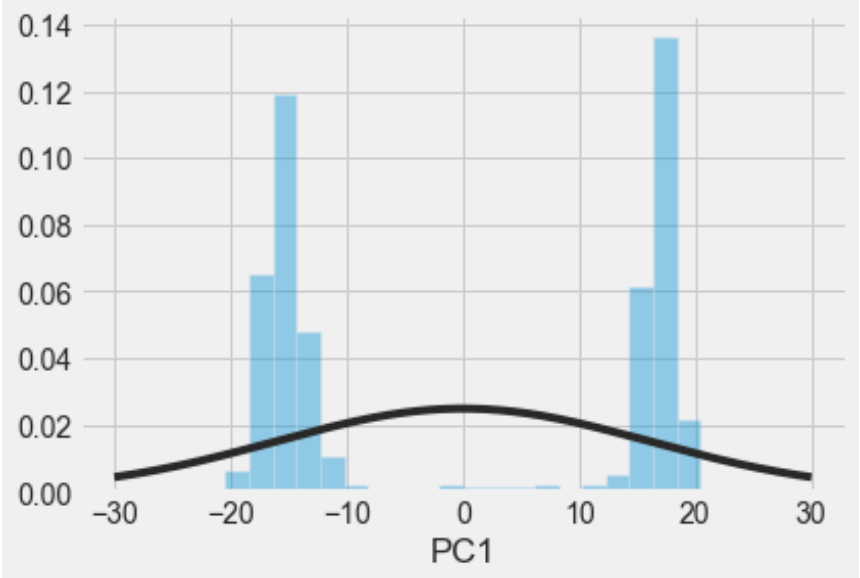

Fig.11 Fitting a single Gaussian over the PC_1 of Jurkat data.

As we can clearly notice that a single Gaussian won’t be enough to fit the PC_1 of Jurkat data, this is more of bimodal data distribution so, we are implementing a mixture model of two Gaussian s on this data to estimate the parameters for the labels.

Method:

Algorithm:

We propose a Gaussian Mixture Model to explicitly characterize the different source of variability in scRNA seq dataset. Our aim is to perform simultaneously clustering for all cells.

A Modified Gaussian Mixture model is constructed with minor changes to maximize the cluster differences and to make the clusters more clear and distinct.

Suppose there are two Gaussians G1 and G2 and the priors are equal for now,

P (G2) = P (G1) = 0.5.

Step 1: We start with calculating pdf that point x_i is coming from the Gaussian-1:

PDF for a univariate normal distribution could be calculated as,

$$P(x_i|\mathrm{G1}) = \frac{1}{\sigma_{G1}\sqrt{(2\pi)}} e^{\frac{-(x-\mu_i)^2}{2{\sigma_{G1}}^2}}$$

This gives the probability that point x_i coming from the G-1,

Step 2: Calculate the Bayesian Posterior, Here we incorporate both the probability that x_i could belong to both the Gaussian 1 and Gaussian 2.

$$G1_i = \mathrm{P}\ (\mathrm{G1}|x_i) = \frac{P(x_i|\mathrm{G1}) \times \mathrm{P}(\mathrm{G1})}{P(x_i|\mathrm{G1}).\mathrm{P}(\mathrm{G1}) + P(x_i|\mathrm{G2}).\mathrm{P}(\mathrm{G2})}$$

Similarly, for the Gaussian-2

$$G2_i = \mathrm{P}\ (\mathrm{G2}|x_i) = 1\text{-}G1_i$$

The only difference is that these $G1_i$ and $G2_i$ will not equal be {0, 1}, It is the soft assignment, It would be [0, 1] basically the probability or mass to Gaussians, It would be maximum to those gaussian which seems to be the source of the data point.

The modified algorithm is:

We'll try to minimize the posterior so the clusters would be more accurate and distinct. We squared the posterior probability and normalize it with respect to each other.

Mathematically, it is feasible to minimize any value by squaring the component of it and dividing it by the total sum of squares of that component.

Here, we square the probability that x_i belongs to G-1 and divide it by the sum of square of probabilities of each Gaussian corresponding to that particular point.

$$Squaring\ the\ posteriors\ = (G1_i)^2$$

And normalizing it by dividing from the sum of squares of posteriors.

$$Normalized\ value\ of\ G1_i{*} \quad = \frac{(G1_i)^2}{(G1_i)^2 + (G2_i)^2}$$

$$Normalized\ value\ of\ G2_i{*} = \quad \frac{(G2_i)^2}{(G1_i)^2 + (G2_i)^2}$$

We'll be using these new values in the following steps to re-estimate mean and variance for the different Gaussians.

Step 3: Re-Estimation of Gaussian-1 and Gaussian-2 parameters.

This is done by the multiplying data points with posterior and dividing them with posteriors.

$$\mu_{G1} = \frac{G1_{1\times}x_1 + G2_i \times x_2 + \cdots\ldots + G1_n \times x_n}{G1_1 + G1_2 + \cdots + G1_n}$$

And,

$$\sigma_{G1} = \frac{G1_i(x_i - \mu_{G1})^2 + .. + G1_n(x_{n-} \mu_{Gn})^2}{G1_1 + G1_2 + \cdots + G1_n}$$

We can do the same for Gaussian-2

$$\mu_{G2} = \frac{G2_1 \times x_1 + G2_2 \times x_2 + \cdots \ldots + G2_n \times x_n}{G2_1 + G2_2 + \cdots + G2_n},$$

$$\sigma_{G2} = \frac{G2_i(x_i - \mu_{G2})^2 + .. + G2_n(x_{n-} \mu_{Gn})^2}{G2_1 + G2_2 + \cdots + G2_n}$$

Now after iterations, we want to maximize our parameters,

We estimate the priors:

$$\text{P (G1)} = \frac{(G1_1 + G1_2 + \cdots + G1_n)}{n}$$

And for the Gaussian-2

$$\text{P (G2)} = 1\text{-P (G1)}.$$

With the simulated data, we try to mimic the Normal Distribution of data and Initialize the parameters of Mixture model randomly and with the help of E-M Iterative algorithm, we try to maximize the likelihood our estimated parameters.

First, we have to decide how many clusters(c) we want to fit our data by looking at data visualization or Histogram or some other techniques.

Then, we have to initialize the parameters of Our Gaussians:

Mean **μc**,

Covariance **Σc**,

And fraction_per_class πc per cluster c (Weight)

The above algorithm could be resolved as follows:

Expectation Step:

We've calculated the probability density function (pdf) for each data point x(i) that does it comes from cluster c:

It could be different for univariate and multivariate cases though the parameter remains same just the variance becomes covariance matrix in the case of multivariate.

Maximization Step:

For each cluster c, we calculated the weight (Fractions of total data points allocated to that cluster) and update the parameters.

Iteratively, Repeat the E and M step until the model converges.

Following are the algorithms proceeding over simulated data:

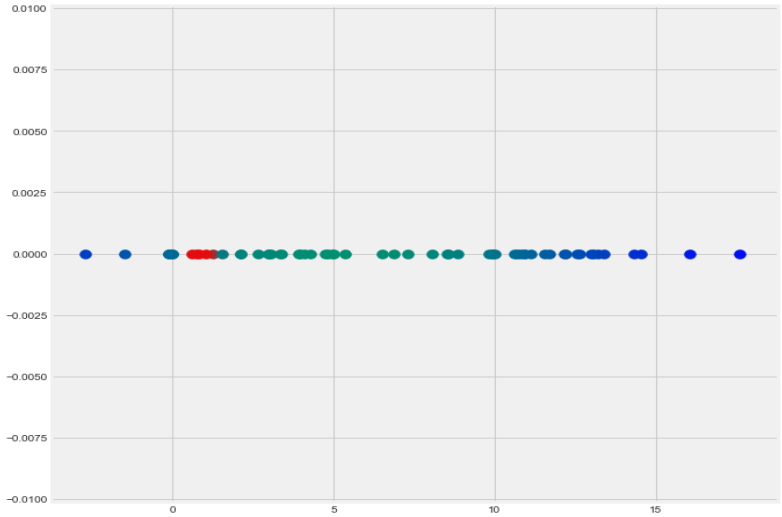

Fig.12 Simulated generated data in 1-d.

In these steps, we try to fit random Gaussians to our simulated data and tries to obtain the MLE-function, Initialization of random parameters and then at each step we re-calculate the parameters and iterate this process until there is no major change in our modelling, this could be achieved with the help of MLE.

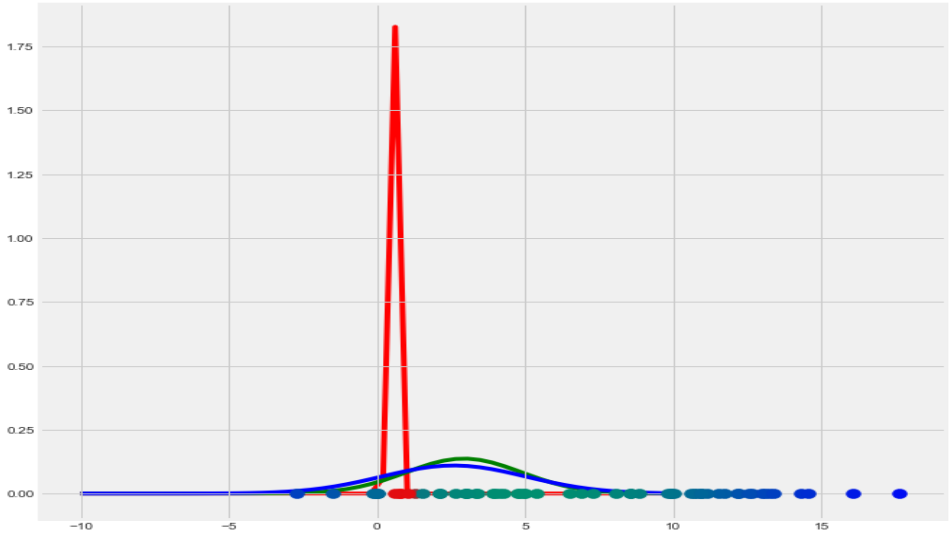

Fig.13 Initialization of Random Parameters over simulated data

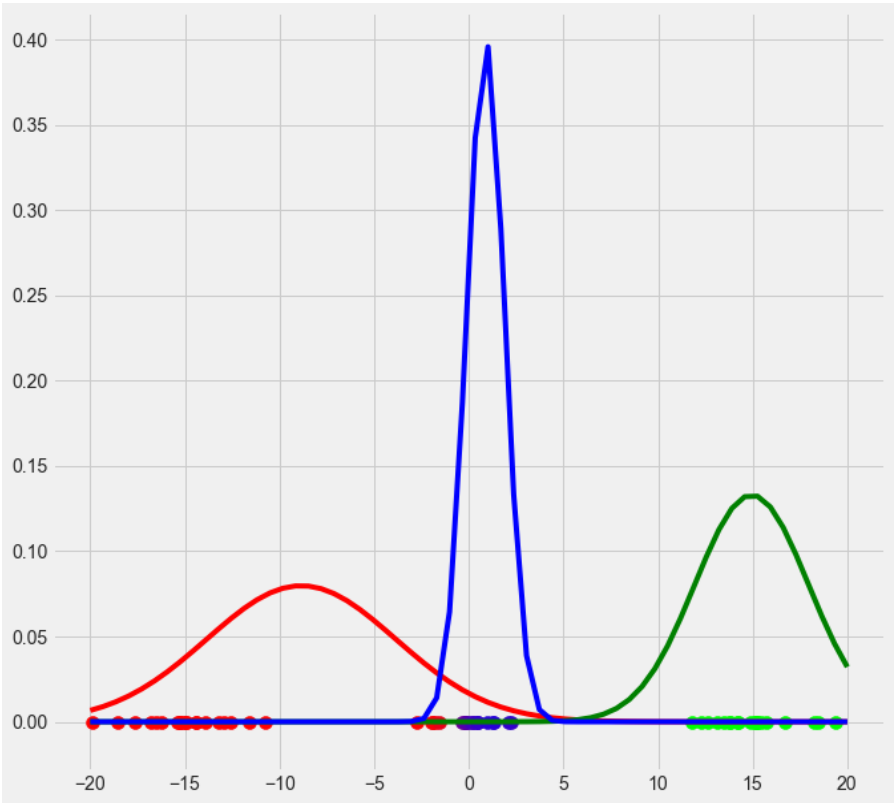

Fig. 14 Algorithm tries to fit data over 5-iterations

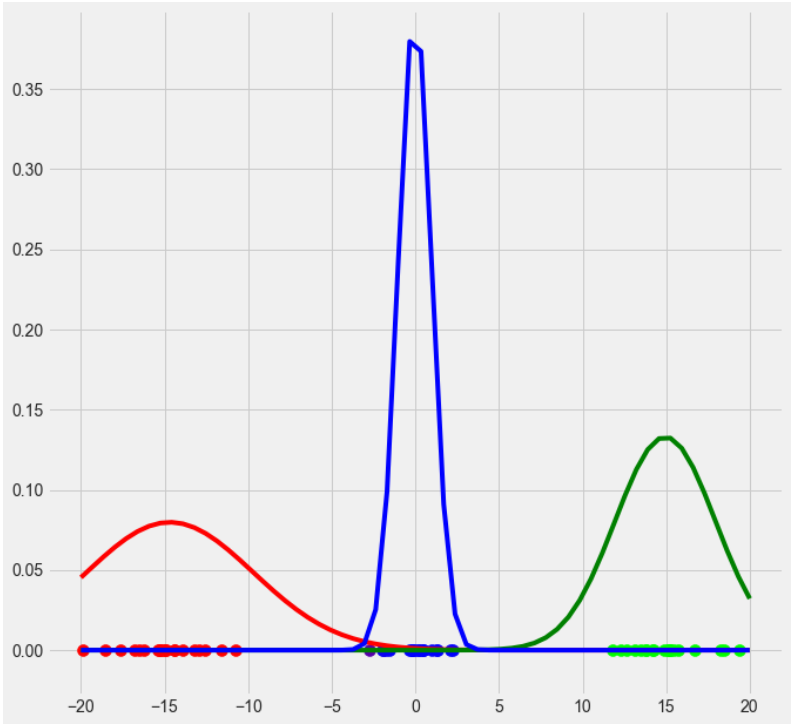

Fig.15 Fitting over 10 iterations

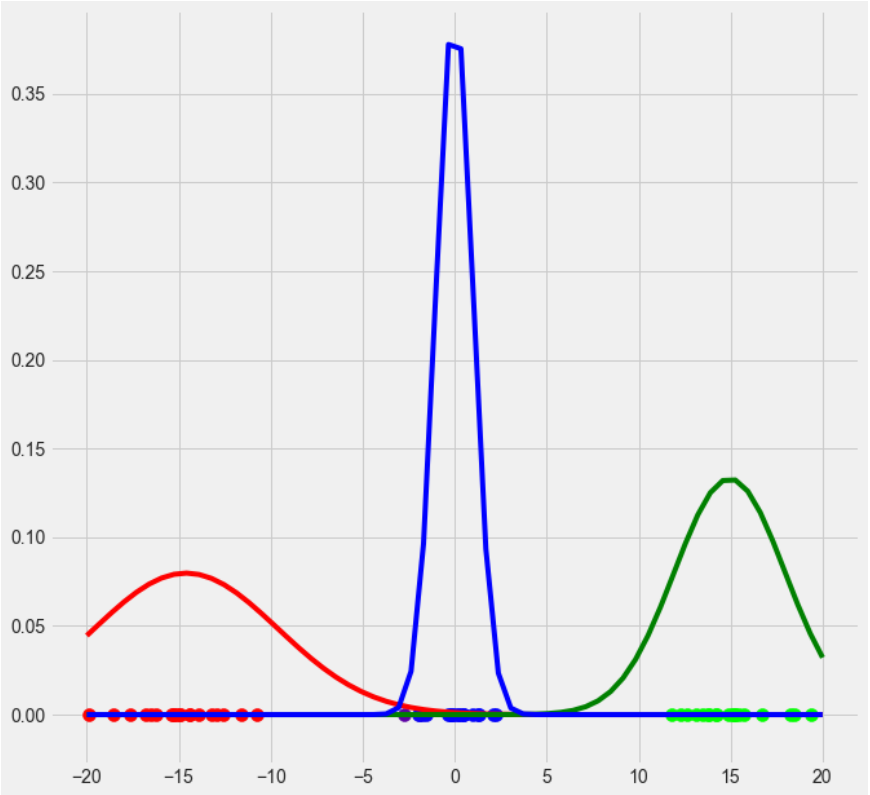

Fig.16 Fitting data over 20 iterations

We tries to maximize MLE-function, once the iteration achieves maximum likelihood, after checking the convergence, we stopped and applies the same algorithm for real dataset.

3.4 Tools and Software

Jupytr notebook was used as IDE,

Various Python Libraries such as

1.) Pandas
2.) NumPy
3.) Scipy
4.) Matplotlib

Were used for programming for different purposes like creating Gaussians Mixture Model algorithm to calculating probabilities

Results & Discussion

4

Overview of the Algorithm

Here we highlight the main ideas in our approach underlying GMM, and we've provided full details in **Materials and Methods.**

Given an N×M gene expression matrix X, with N cells and M genes (N <M) as an input, After Pre-Processing and Normalization of the matrix, we performed PCA(Principal Component Analysis) with giving annotations to visualize the clusters and report it as standard level(when Labels were provided).

We introduce a GMM framework that learns itself the labels after certain iterations of Expectation-Maximization algorithm and once it converges we compare our result of PCA visualization with standard PCA. The cell to cell similarity values in unsupervised approach can be used to create an embedding of the data in 2-D or 3-D for visualization, as well as the projection of data into a latent space of arbitrary dimensions to further identify classes/groups of cells that are alike. This is the standard PCA results after providing class labels of the data.

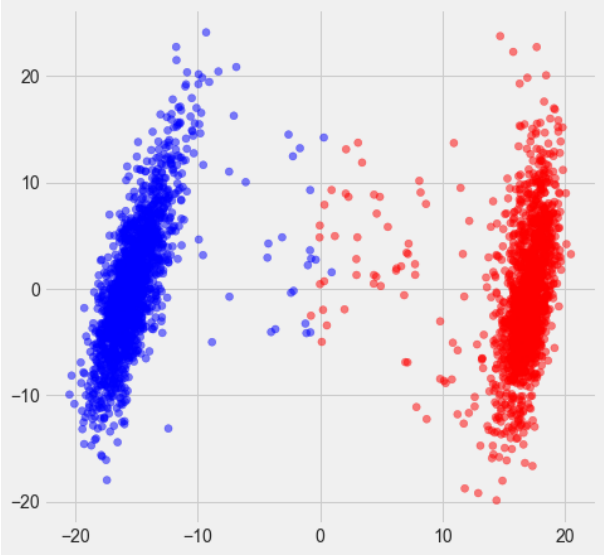

Fig.17 Standard PCA results on Jurkat data

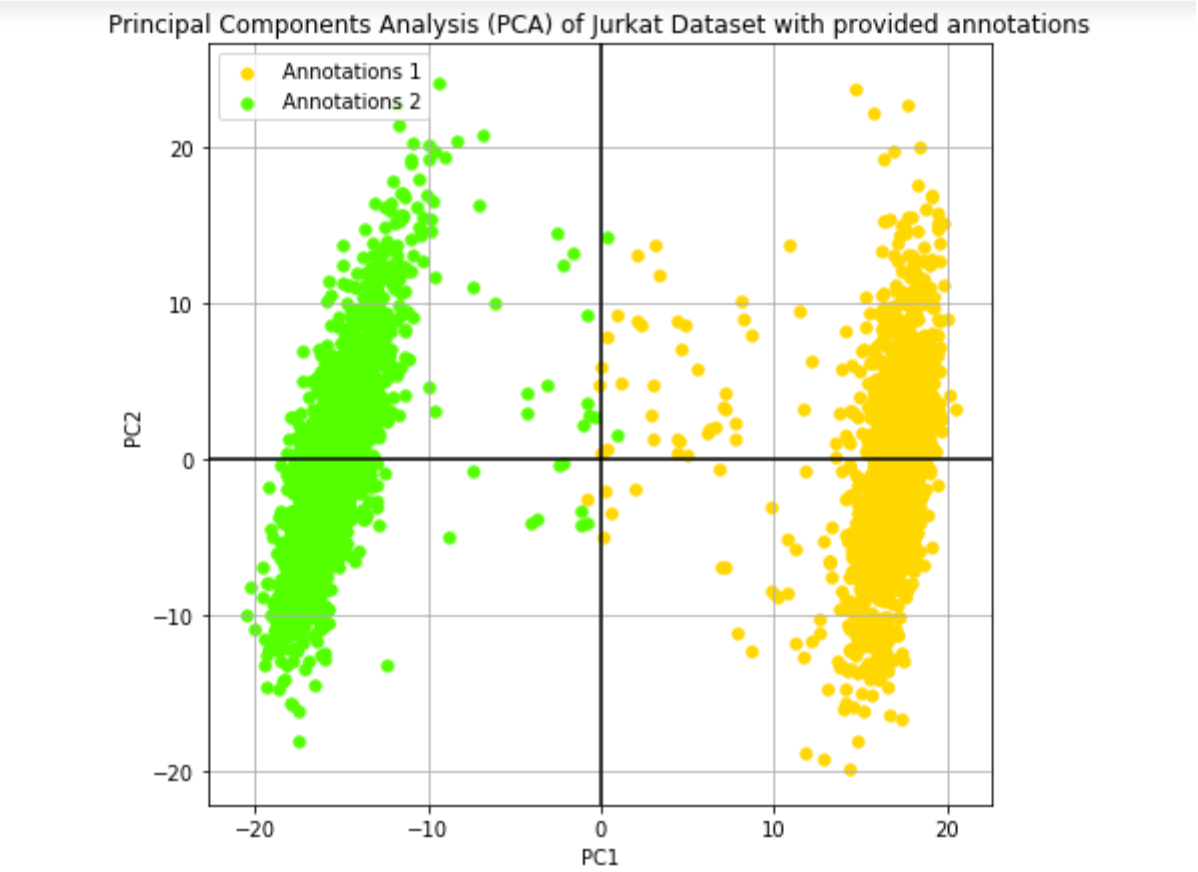

Fig.18 Standard PCA with provided annotations

This denoted the Principal component analysis of Jurkat dataset which comprises of two different species, Mus-Musculus and Homo sapiens. These two clusters denote that the variance between the data, and it could be of two different species.

We try to find the same result from our Modified GMM algorithm where we minimizes the prior distribution further to make the clusters distinct, and we run our algorithm on the first component analysis of the Jurkat data,.PCA transforms the original co-ordinates system, new co-ordinates are called principal components,the first PC points is in the direction of highest variance, the second PC is in the direction of the second highest variance … and so on.

Now, we implemented our modified algorithm on the PC_1 of the data and tries to find the best result,

We got the following results:

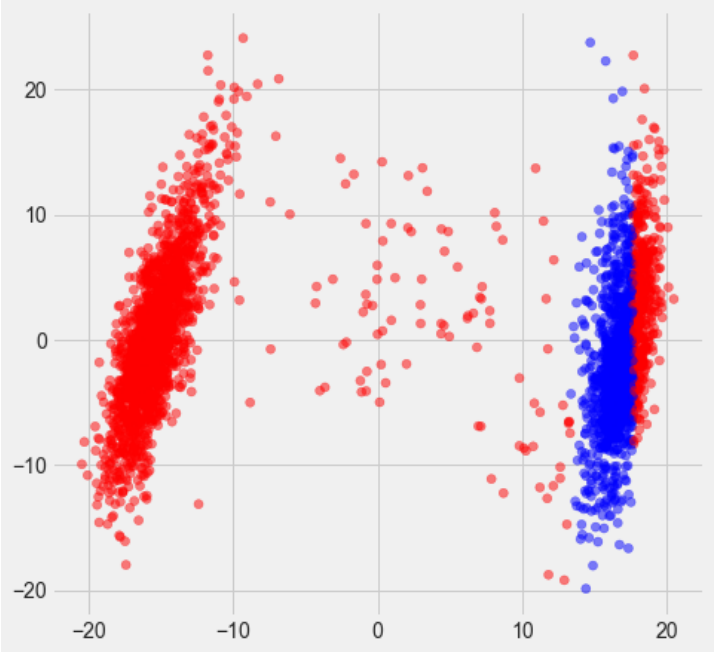

Fig. 19 Visualization of clusters after 1st-Iteration

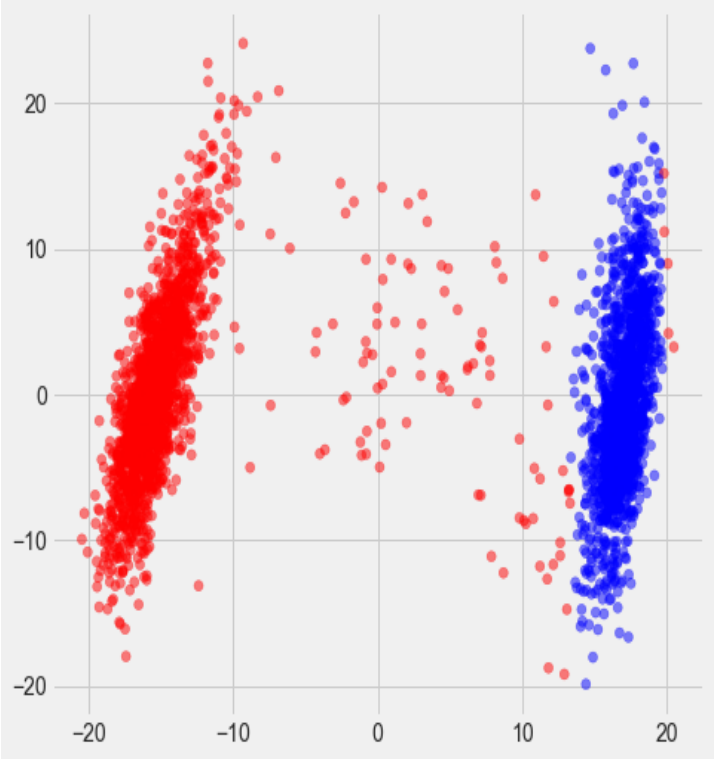

Fig.20 Visualization of clusters after 2nd iteration

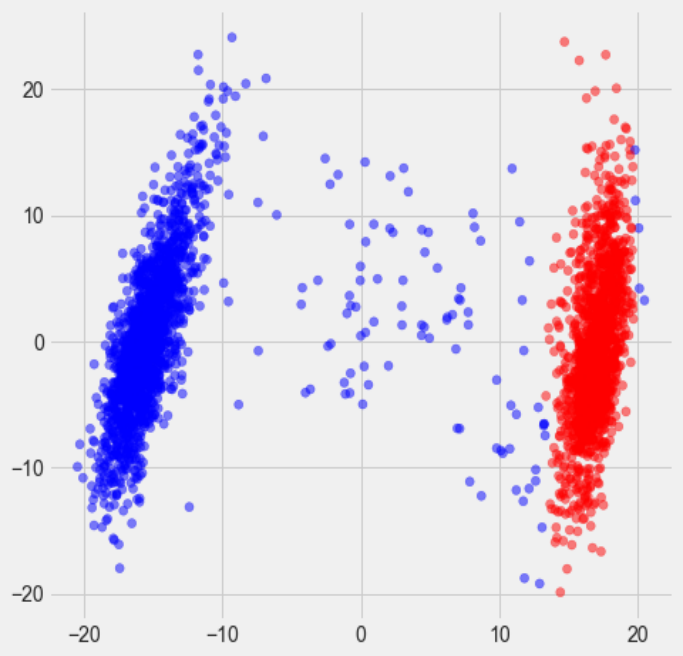

Fig. 21 Visualization of clusters after the convergence

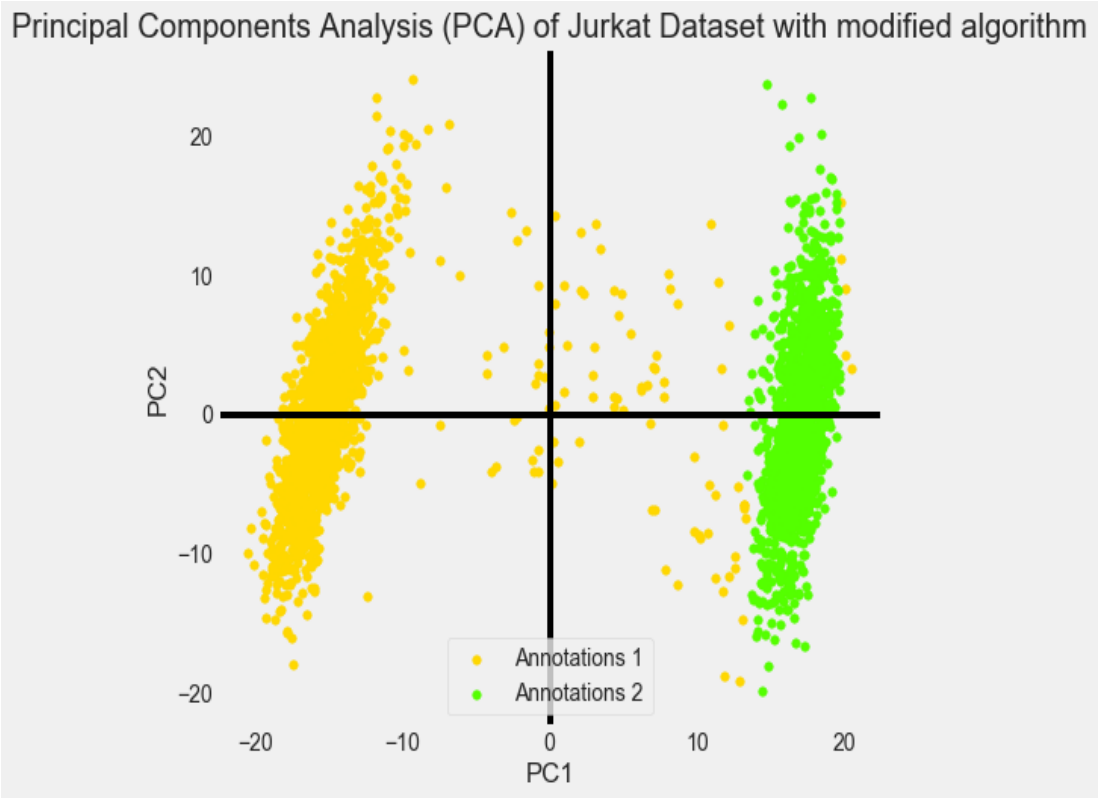

Fig. 22 Visualization of data with the modified algorithm.

As, we find our algorithm is capable to find the good cluster, and it finds two different clusters while estimating parameters for the annotations of the dataset.

We find that the clustering score comes near to 91.20% which is quite good for clustering.

It successfully finds the two different species and almost gives true labels via unsupervised approach.

5 Conclusion and Future Work

Last decade has shown tremendous growth in scRNA-seq research due to the fact that these have been found to be very crucial constituent involved in detailed gene expression profile, characterization of the rare cell including the study of different cancer types and neurological disorders. Therefore, it has become very important for the biologists to know more about scRNA. Their pattern of clustering which could be identified by unsupervised clustering only. This has also witnessed multiple algorithms to analyse ScRNA seq data such as SC3, Seurat, Dropclust and others. Unsupervised clustering of high dimensional-data is computationally expensive. Hence, the researchers started looking forward for more sophisticated computational approaches for solving these problems.

Our results suggest that the Gaussian Mixture Model could serve as a great tool for ScRNA seq clustering, though it has some of its limitations.

- The flexibility of density shape is a bane, it causes unboundedness of the likelihood function.
- Supremum of the Likelihood function is equal to infinity
- Excessive sensitivity to outliers

Therefore, we need better clustering algorithm to overcome above shortcoming or we could modify algorithm in such a way that it becomes robust to handle outliers, and don't provide the weight component to any such outliers.

To enhance our work, some of the future directions are as follows:

(1) To estimate robust parameters in the Normal Mixture Model.

(2) For rigorous testing of the clustering result, we can consider multiple data sources.

(3) Boundedness of the likelihood function.

This work can further be extended in future in identifying rare cell detection and could provide more robustness to outliers for a more reliable and more efficient clustering.

References

1. Single-cell RNA sequencing technologies and bioinformatics pipelines, Byungjin Hwang, Ji Hyun Lee & Duhee Bang Experimental & Molecular Medicine volume 50, Article number: 96 (2018)
2. dropClust: efficient clustering of ultra-large ScRNA-seq data Debajyoti Sinha, Akhilesh Kumar, Himanshu Kumar**, Sanghamitra Bandyopadhyay, * and Debarka Sengupta** Nucleic Acids Research, 2018, Vol. 46, No. 6 e36
3. "Single-cell sequencing-based technologies will revolutionize whole-organism science," E. Shapiro, T. Biezuner, and S. Linnarsson, Nat. Rev. Genet., vol. 14, no. 9, pp. 618–630, 2013.
4. SC3: consensus clustering of single-cell RNA-seq data Kiselev., Kirschner,K., Schaub,M.T., Andrews,T., Yiu,A., Chandra,T., Natarajan,K.N., Reik,W., Barahona,M., Green,A.R. et al. (2017). Nat. Methods, 14, 483–486.
5. Massively parallel single-nucleus RNA-seq with DroNc-seq, **Habib et al.** *Nature Methods* volume14, pages955–958 (2017)
6. Dempster, A.P., Laird,N.M.,Rubin,D.B.,1977.Maximum likelihood from incomplete data via the EMalgorithm. J. Roy. Statist. Soc. B 39, 1–38
7. Tanay,A. and Regev,A. (2017) Scaling single-cell genomics from phenomenology to mechanism. Nature, 541, 331–338.
8. Robust estimation in the normal mixture model, Hironori Fujisawa∗, Shinto Eguchi, Journal of Statistical Planning and Inference Volume 136, Issues 11, Pages 3989-4011 (2006).
9. McLachlan, G., Peel, D., 2000. Finite Mixture Models.Wiley, NewYork
10. Robust and efficient estimation by minimising a density power divergence**. Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C**., 1998. Biometrika 85, 549–559

11. A joint finite mixture model for clustering genes from independent Gaussian and beta distributed data **Xiao Feng Dai*, Timo Erkkilä, Olli Yli-Harja and Harri Lähdesmäki,** BMC Bioinformatics 10:165, 2009.
12. Jones, M.C., Hjort,N.L.,Harris,I.R.,Basu,A.,2001.A comparison of related density-based minimum divergence estimators. Biometrika 88, 865–873.

Books:

1. Bishop - Pattern Recognition And Machine Learning - Springer 2006
2. Alboukadel Kassambara -Practical Guide To Cluster Analysis in R Edition 1