Clustering Leukemia Patients

Katelyn Gao[1], Heather Hardeman[2], Edward Lim[3], Cris Potter[4], Carl Meyer[5], Ralph Abbey[5]

Introduction

- **Clustering** partitions a set of observations into groups where similar observations are grouped together.
- Our Clustering Methods
  - *k*-means
    - $\arg_S \min \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$
    - Initialization → Assignment and Update → Termination
  - Nonnegative Matrix Factorization (NMF)
    - Factor a nonnegative data matrix into two nonnegative matrices

$$\min \| V_{mxn} - W_{mxr} H_{rxn} \|_F^2$$

Update Steps

$$H_{ij} \longleftarrow H_{ij} \frac{(W^T V)_{ij}}{(W^T W H)_{ij}} \qquad W_{ij} \longleftarrow W_{ij} \frac{(V H^T)_{ij}}{(W H H^T)_{ij}}$$

Motivation

Determine the number of clusters present in a data set and use that information to cluster Leukemia patients

Data from the Broad Institute

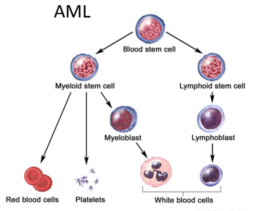
Types of **Leukemia**

- Acute Lymphoblastic Leukemia B cells (ALL - B)
- Acute Lymphoblastic Leukemia T-cells (ALL - T)
- Acute Myeloid Leukemia (AML)

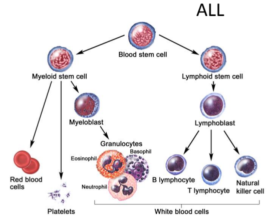
5000 genes x 38 leukemia patients' microarray data

- 1-19: ALL - B
- 20-27: ALL - T
- 28-38: AML

AML



ALL



Methodology

Steps: NMF and K-means → Consensus Matrix → Sinkhorn Knopp → Eigen decomposition

Consensus Matrix

- For Leukemia data set A = 38 x 38 matrix
- A : $a_{ij} = 1$ if i and j are in the same cluster
  $a_{ij} = 0$ if not
- Symmetric
- Consensus matrix: for n runs of a clustering method, $C = \frac{1}{n}\sum_{i=1}^{n} A_i$

Sinkhorn Knopp Algorithm

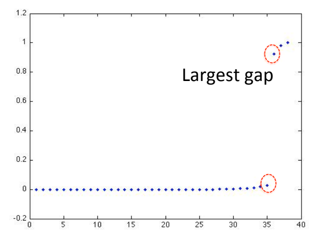
- Create a matrix S by adjusting column sums and row sums of the consensus matrix to 1
- Alternate scaling rows by their row sums and the columns by their column sums
- Preserves the symmetry of a matrix and places the eigenvalues in the interval [0 1]
- Preserves the block structure of a consensus matrix

Results

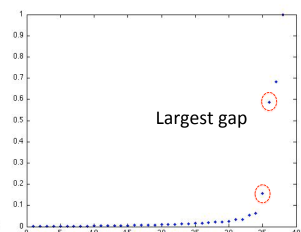
Eigenvalues

We used the location of the largest gap in the eigenvalues of the matrix S to determine the number of clusters *k*. Here the largest gap occurs between the third and fourth, implying *k* = 3

50 runs of NMF k = 3

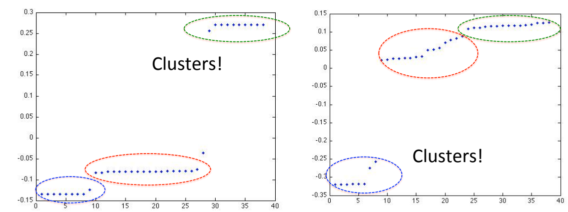


50 runs of K-means k = 2, 3, 4, 5



Eigenvectors

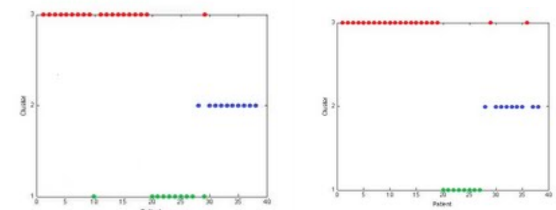
The eigenvectors may reveal information about which cluster each observation belongs.



Conclusions

- The location of the largest gap in the eigenvalues reveals the number of clusters *k.*
- Then we may re-cluster the data based on that information.

NMF with k = 3 | *k*-means with k = 3



- Our results corroborate with those of the Broad Institute of Harvard and MIT.
- Information from the Broad Institute suggests that the "mis-clustering" of patients 10 and 29 is due to misdiagnoses.

Acknowledgements

Dr. Carl D. Meyer, *advisor*
Ralph Abbey, *graduate assistant*
Broad Institute of Harvard and MIT

1. MIT, 2, Univeristy of Montevallo, 3. Johns Hopkins University, 4. East Carolina University, 5. North Carolina State University