

# A Spectral Clustering Method

David Tritchler, Shafagh Fallah and Joseph Beyene

- cluster genes or samples using microarray expression data
- motivated by a multivariate analysis of variance model
- computationally based on eigenanalysis (thus the term “spectral” in the title)
- A leukemia data set is analyzed

## Microarray Technology

- allows investigation of RNA expression of thousands of genes simultaneously
- cluster analysis is used to reduce the dimension of the data and discern meaningful patterns
- eg. partition the samples into groups whose genes express similarly
- eg. partition the genes so the genes in a group covary across samples
- exploratory in nature

## Microarray Data

- $N$  genes  $g_1, \dots, g_N$
- $T$  samples (arrays)  $t_1, \dots, t_T$  which correspond to replicates, cell lines or experimental conditions
- The data is represented as the  $N \times T$  matrix  $X^* = (x_{ij}^*)$ , where the rows correspond to genes and columns to the arrays
- We will work with a transformed matrix  $X = (x_{ij})$  formed by subtracting the row mean from each row of  $X^*$

## Example Microarray Data

- Golub et al. (1999) studied the gene expression of acute leukemia
- 3571 genes on arrays from 11 AML, 19 B-ALL and 8 T-ALL independent tissue samples
- <http://statwww.epfl.ch/davison/teaching/Microarrays/lab/classification.html>

## Motivation for Method

we model the gene expression measurements as a multivariate analysis of variance

$$X = \mathbf{1} \otimes \alpha' + ZB + \text{error}$$

where

- each row of  $X$  (gene) is a multivariate observation of  $T$  components
- $\mathbf{1}$  is a  $N$ -vector of ones
- $\alpha$  is a  $T$ -vector of mean levels for the columns of  $X$
- $Z$  is a  $N \times p$  design matrix whose columns are orthogonal to  $\mathbf{1}$  and to each other
- $B$  is a  $p \times T$  parameter matrix

## Algorithm (in terms of genes)

- construct a design matrix for the gene expression measurements
- The design matrix defines the clusters
- columns of  $Z$  are defined by a stepwise process
- construct the first column  $Z_{(1)}$  of  $Z$  by splitting the genes into two groups
- $Z_{(1)}$  is the *contrast* comparing the groups; by definition  $Z_{(1)}$  is orthogonal and length 1 to 1

- These conditions imply that the dichotomous elements of  $Z_{(1)}$  are given by

$$Z_{(1)i} = \sqrt{\frac{n_2}{n_1 N}} \text{ gene } i \text{ in group 1}$$

$$Z_{(1)i} = -\sqrt{\frac{n_1}{n_2 N}} \text{ gene } i \text{ in group 2}$$

i.e. positive for group 1, negative for group 2

- first row of  $B$  is the least squares regression coefficient estimate  $b'_{(1)}$  corresponding to  $Z_{(1)}$

- We construct the grouping to maximize the magnitude of the regression parameter  $b_{(1)}$
- The second column of  $Z$ ,  $Z_{(2)}$ , is formed by further splitting the genes in the first group of  $n_1$  genes into two smaller groups, maximizing  $b_{(2)}$
- The third column of  $Z$ ,  $Z_{(3)}$ , is formed by further splitting the genes in the second group of  $n_2$  genes into two smaller groups
- and so on: at each step a subset of genes is partitioned
- the result is a set of orthogonal contrasts which determine the clusters

$$X = \begin{bmatrix} -5 & -5 & 10 \\ -5 & -5 & 10 \\ 10 & -5 & -5 \\ 10 & -5 & -5 \\ -5 & 10 & -5 \\ -5 & 10 & -5 \end{bmatrix}$$

$$Z = \begin{bmatrix} \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{4}} \\ \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{4}} \\ \sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{4}} \\ \sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{4}} \\ -\sqrt{\frac{1}{3}} & 0 \\ -\sqrt{\frac{1}{3}} & 0 \end{bmatrix}$$

## Properties of the Algorithm

some definitions:

- At the step where  $Z_{(q)}$  is constructed, denote  $X^{(q)}$  to be the matrix  $X$  adjusted by subtracting the centroid of the genes being split
- Define  $S = (s_{ij})$  to be the covariance matrix corresponding to  $X^{(q)}$
- define  $\bar{g}_1$  and  $\bar{g}_2$  be the mean  $T$ -vectors of the genes in the groups being formed

## Equivalent Local Clustering Criteria

We have the following equivalence:

**Theorem:** *The criteria*

$$C1: \mathbf{b}'_{(q)} \mathbf{b}_{(q)}$$

$$C2: Z'_{(q)} X^{(q)} X^{(q)'} Z_{(q)}$$

$$C3: n \frac{n_1}{n} \frac{n_2}{n} (\bar{\mathbf{g}}_1 - \bar{\mathbf{g}}_2)' (\bar{\mathbf{g}}_1 - \bar{\mathbf{g}}_2)$$

$$C4: n T \frac{n_1}{n} \frac{n_2}{n} \left( \frac{1}{n_1^2} \sum_{i,j \in G_1} s_{ij} + \frac{1}{n_2^2} \sum_{i,j \in G_2} s_{ij} \right. \\ \left. - 2 \frac{1}{n_1 n_2} \sum_{i \in G_1, j \in G_2} s_{ij} \right)$$

*are equivalent.*

## Continuous Approximation

Generating the  $Z_{(q)}$  which maximizes the magnitude of the associated parameter vector  $\mathbf{b}_{(q)}$  is combinatorially hard

We convert the discrete problem to an easily solved continuous one:

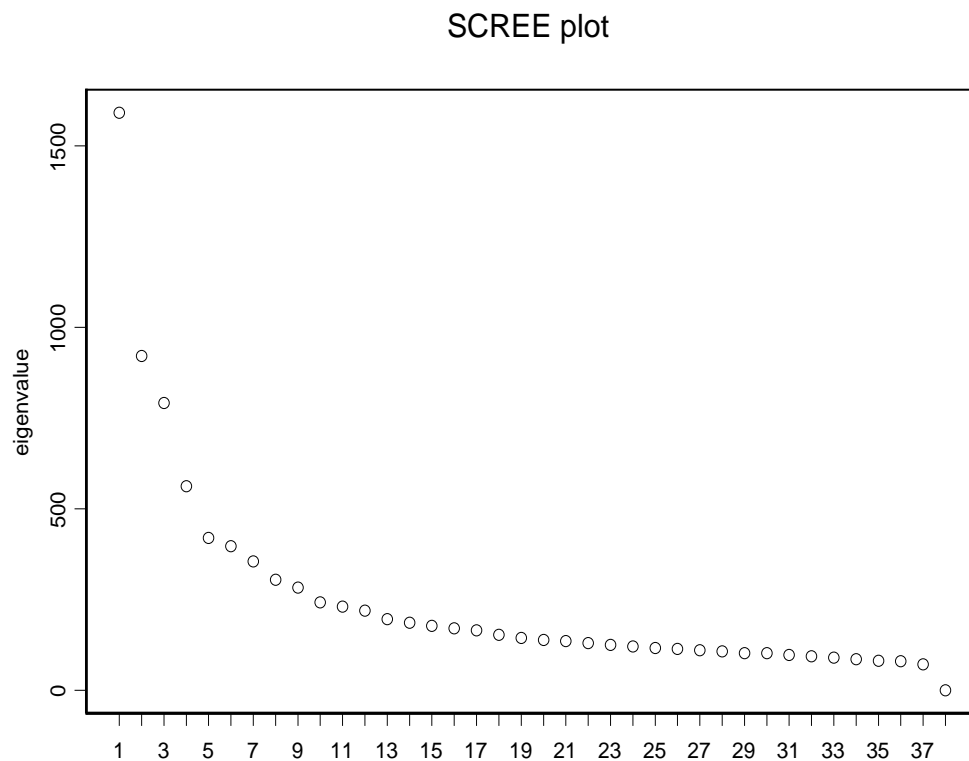
- 

$$\mathbf{b}_{(q)}' \mathbf{b}_{(q)} = Z_{(q)}' X^{(q)} X^{(q)'} Z_{(q)}$$

- the continuous-valued vector  $\mathbf{v}$  which maximizes  $\mathbf{v}' X^{(q)} X^{(q)'} \mathbf{v}$  subject to  $\mathbf{v}' \mathbf{v} = 1$  is the normalized eigenvector corresponding to the largest eigenvalue of  $X^{(q)} X^{(q)'}$
- use  $\mathbf{v}$  as a surrogate for  $Z_{(q)}$  and partition the genes into classes corresponding to positive elements of  $\mathbf{v}$  (group 1) and negative elements of  $\mathbf{v}$  (group 2)

- overall complexity is bounded above by  $kT^2(N + 2^k)$ , using the fact that  $X'X$  has the same nonzero eigenvalues
- The recursive algorithm needs a stopping rule in order to halt. We use the test statistic for the significance of the largest eigenvalue given by Johnstone (2000)

## Clustering Leukemia Samples



The SCREE plot displays the eigenvalues obtained from the leukemia data. The plot suggests that there are four clusters.

Adjusting our threshold to yield four clusters, we obtain clusters consisting of

Cluster 1: 11 AML + 1 B-ALL (AML)

Cluster 2: 10 B-ALL (B-ALL)

Cluster 3: 7 B-ALL (B-ALL)

Cluster 4: 8 T-ALL + 1 B-ALL (T-ALL)

- we share the conjecture of Golub et al. (1999) that there may be an undiscovered subtype of B-ALL
- The three-cluster solution combines the two B-ALL classes into a single cluster

## Predictive Gene Clusters

We can construct gene clusters for which the mean expression is predictive of outcome.

- more meaningful clusters
- useful for diagnosis

note that the regression coefficient vector  $\mathbf{b} = \sqrt{n \frac{n_1}{n} \frac{n_2}{n}} (\bar{\mathbf{g}}_1 - \bar{\mathbf{g}}_2)$

We take  $\mathbf{b}'(\mathbf{y} - \bar{y}\mathbf{1})$  as our criterion for prediction when partitioning the genes

- difference in expression due to group membership varies according to  $y$

To obtain predictive clusters we add a term to the objective function, giving

$$\mathbf{v}'X X'\mathbf{v} + \gamma \mathbf{b}'(\mathbf{y} - \bar{y}\mathbf{1})$$

- $\gamma$  controls the degree of supervision of the clustering by the prediction criterion

the maximizing vector  $\mathbf{v}$  is calculated from the last half of the eigenvector corresponding to the largest eigenvalue of the matrix

$$\begin{bmatrix} X'X & I \\ (\frac{\gamma}{2})^2 X'ss'X & X'X \end{bmatrix}$$

where  $\mathbf{s} = X(\mathbf{y} - \bar{y}\mathbf{1})$

## Predicting Leukemia Classification

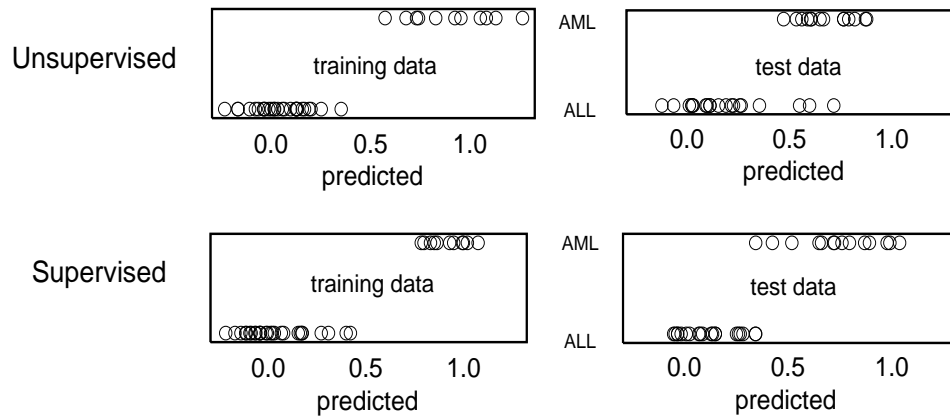
We obtained two 11-cluster solutions

- supervised by ALL/AML classification
- without supervision.
- The level of supervision was set to weight prediction by a factor of 2

For both the supervised and unsupervised case, we fit regressions

- outcome is ALL/AML
- 11 cluster means are predictors
- applied the same equation to a test data set of 20 ALL and 14 AML samples

# Predicting Leukemia Classification



supervision improves the separation of the two sets of predicted values

- perfect separation

## Determining the Number of Clusters

we consider the analysis of an idealized cluster structure yielding a covariance matrix

$$S = \sigma^2 \begin{bmatrix} R^{(11)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & R^{(22)} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & R^{(m-1,m-1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & R^{(mm)} \end{bmatrix},$$

$$R^{(jj)} = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho & \rho & \cdots & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{bmatrix}$$

## Eigenvalues and the Number of Clusters

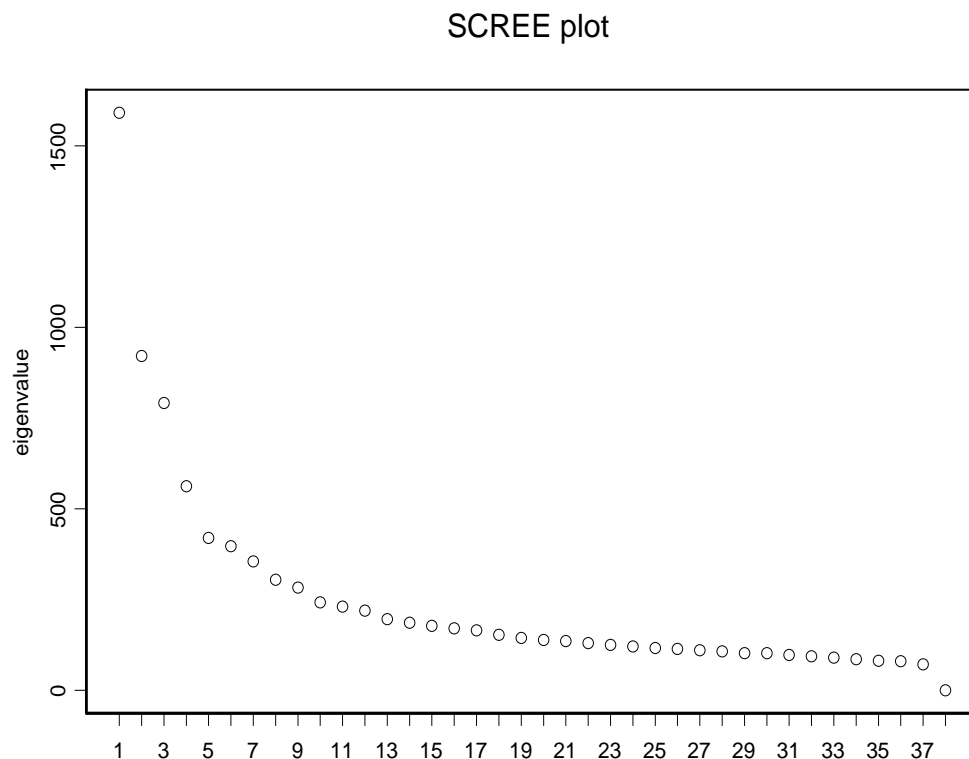
$P$  projects onto subspace  $\perp \mathbf{1}$

for  $m$  clusters of equal size,  $PSP$  has:

- one zero eigenvalue
- $N - m$  eigenvalues  $\sigma^2(1 - \rho)$  "small"
- $m - 1$  eigenvalues  $\sigma^2(1 + \rho(N/m - 1))$  "large"

For unequal but similarly sized clusters, these results should apply approximately, since the eigenvalues vary in a continuous fashion

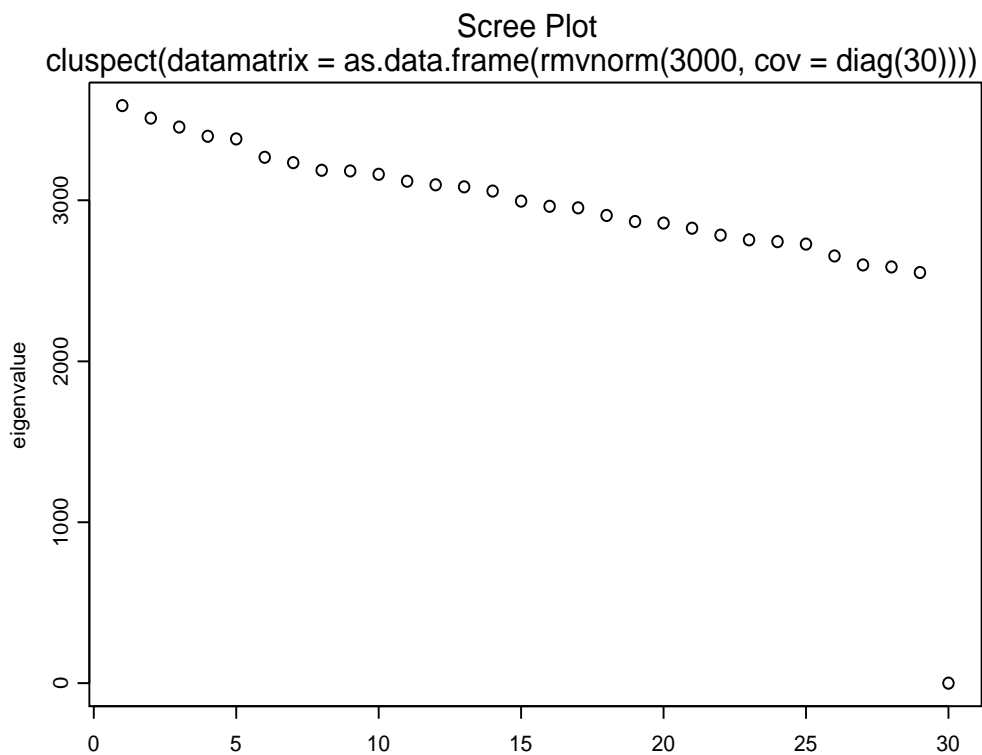
## Leukemia Data Revisited



**Figure 1.** The SCREE plot displays the eigenvalues obtained from the leukemia data. The plot suggests that there are **five** clusters.

## The Null Case

$3000 \times 30$  matrix of i.i.d.  $N(0, 1)$  random variables



**Figure 1.** The SCREE plot for a random matrix. The plot suggests that there is one cluster.

According to our results, five clusters exist in the data, since there are four “larger” eigenvalues

Adjusting our threshold to yield five clusters, we obtain

Cluster 1: 11 AML + 1 B-ALL (AML)

Cluster 2: 10 B-ALL (B-ALL)

Cluster 3: 7 B-ALL (B-ALL)

Cluster 4: 8 T-ALL (T-ALL)

Cluster 5: 1 B-ALL

- The four-cluster solution combined Cluster 5 with Cluster 4, so Cluster 5 appears to be an atypical B-ALL which can be confused with T-ALL
- The six-cluster solution splits up Cluster 1, the AML's