StatsM254 Statistical Methods in Computational Biology Lecture 5 - 04/15/2014

Lecture 5

Lecturer: Jingyi Jessica Li *Scribe: Artur Jaroszewicz*

1 Introduction: How to choose a proper *k* for *k-means* and *k-medoid* clustering

We would like to find w_k as a function of k. Define the following:

- Data: $x_1, ..., x_n \in \mathbb{R}^p$, where p is defined as the number of samples, and each vector represents one gene.
- Assignment: $c_1, ..., c_k$ where c_r denotes the index of observations in cluster r, and $n_r =\mid c_r \mid$.
- Distance metric: $d_{ii'}$, e.g., $d_{ii'} = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$ or $d_{ii'} = \frac{1-corr(x_i,x_{i'})}{2} \in [0,1]$.
- Within-cluster variance: $w_k = \sum_{r=1}^{k} \frac{1}{2n_r} \sum_{i,i' \in c_r} d_{ii'}$.

To choose k either:

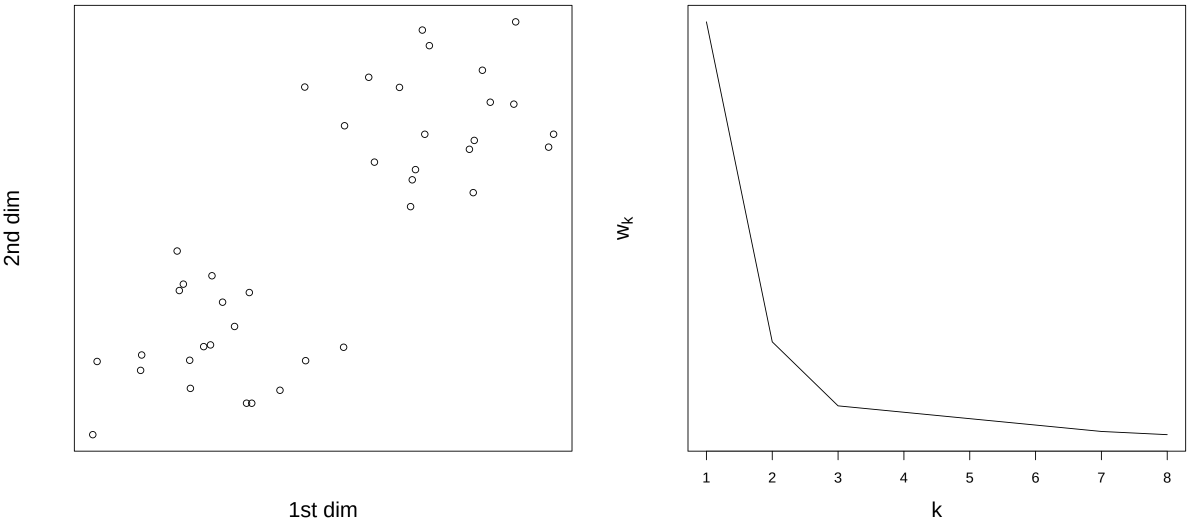
1. Plot $log(\frac{w_k}{w_{k+1}})$ as a function of k (see notes on lecture 4), OR
2. Use the Gap Statistic [1].

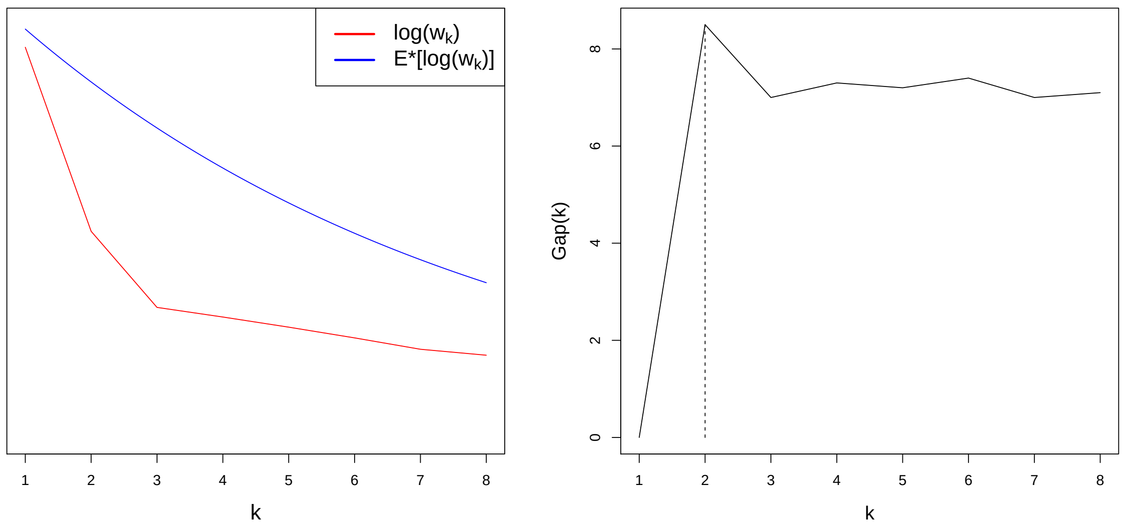
2 Gap Statistic

Define the Gap Statistic:

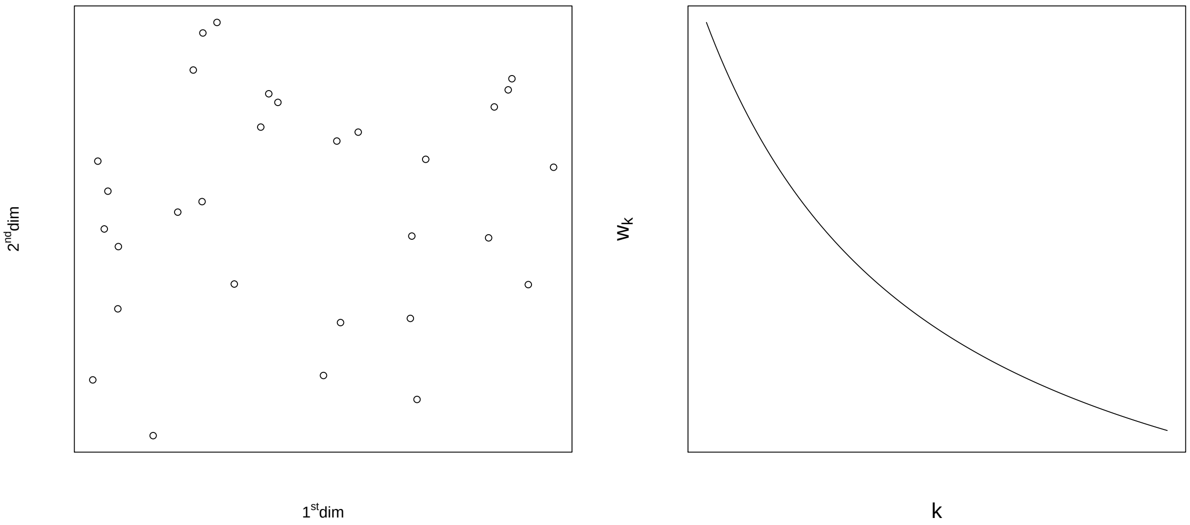
$$Gap_n(k) = E_n^*[\log(W_k)] - \log(w_k)$$

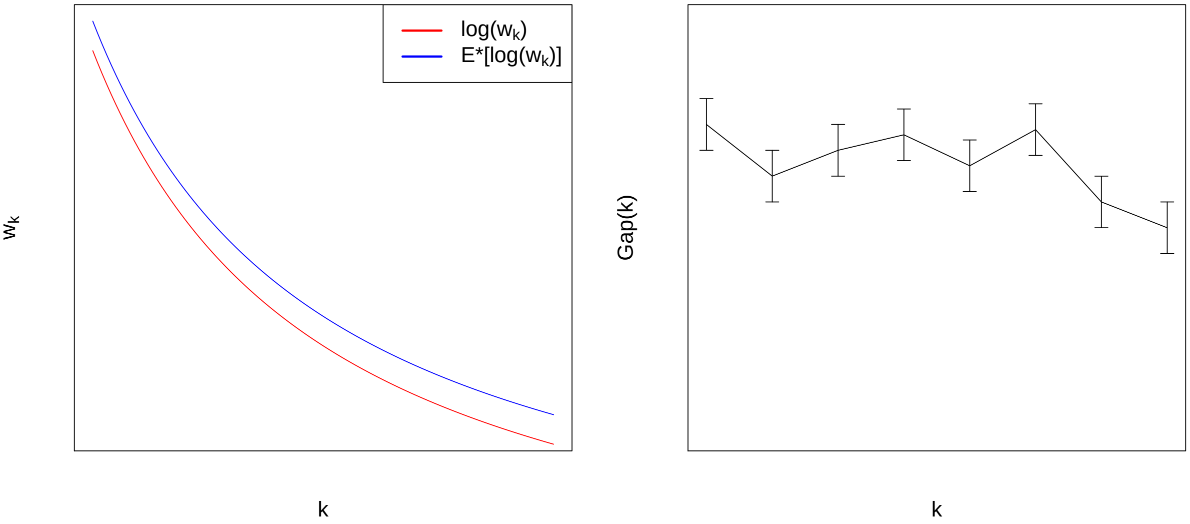
and look for the largest difference between observed and expected within-cluster variance.





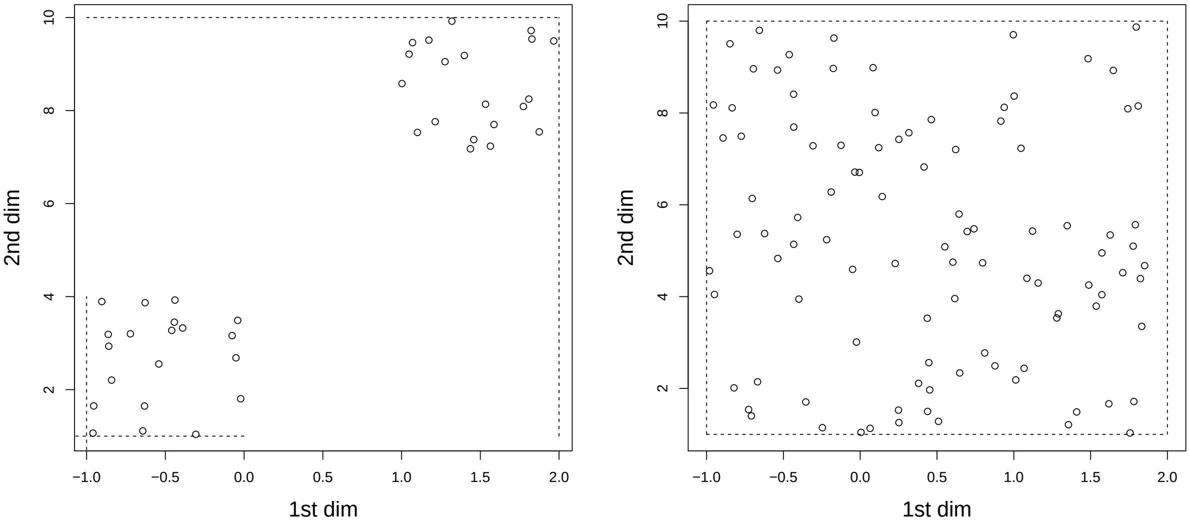
However, we cannot simply choose the largest gap, as we must have a penalty for creating too many clusters and account for some degree of random noise.



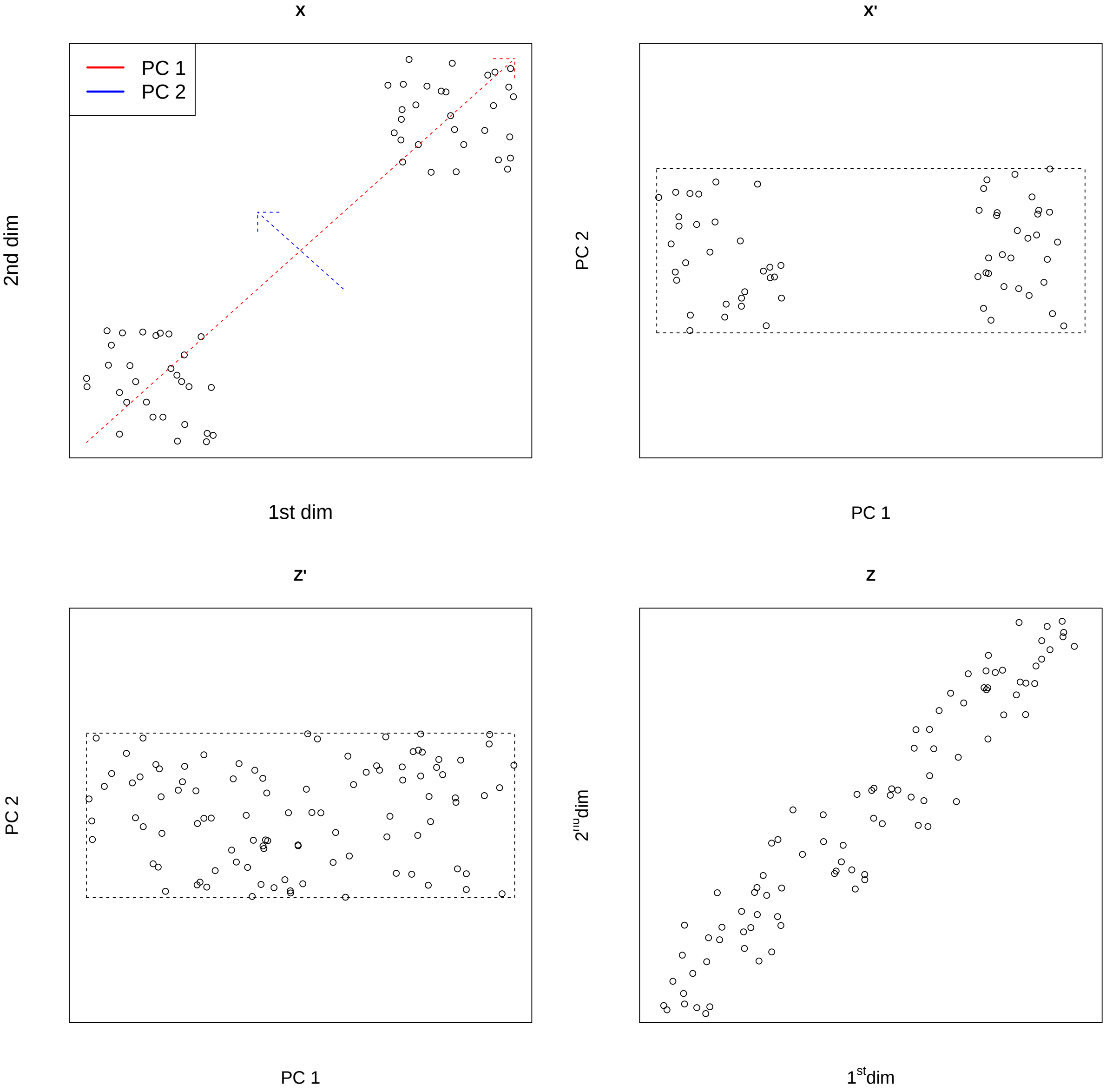


Thus, we add error bars to account for noise. For a reference distribution to calculate $E_n^*[\log(W_k)]$, we consider two choices:

a Generate each reference feature (e.g., sample) uniformly over the range of observed values for that feature.



b Generate the reference features from a uniform distribution over a box aligned with the principal components of the data.
More specifically, if X is our $N \times P$ data matrix, assume that the columns (e.g., samples) have mean 0 and compute the singular value decomposition (SVD) such that $\mathbf{X} = \mathbf{VDV}^\top$.
We then transform $\mathbf{X}' = \mathbf{XV}$ and draw uniform features $\mathbf{Z}'$ over the ranges of the columns of $\mathbf{X}'$. Finally, we transform back via $\mathbf{Z} = \mathbf{Z}'\mathbf{V}^\top$ to give our reference data $\mathbf{Z}$.



3 Algorithm

Note: R package available [2].

1. Cluster the observed data $X_1, X_2, ..., X_n$. Vary numbers of clusters from $k = 1, .., K$ (where K is the upper bound), resulting in w_k, $k \in \{1, ..., K\}$.

2. Generate B reference data sets using the uniform prescription **a** or **b** above, and cluster each dataset under each k, resulting in w^*_{kb}, $b = 1, ..., B$, $k = 1, ..., K$. Compute the (estimated) gap statistic

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B} \log(w^*_{kb}) - \log(w_k),$$

where $\frac{1}{B}\sum_{b=1}^{B}\log(w^*_{kb})$ is the estimator for $E^*_n[\log(W_k)]$.

3. Let $\bar{l} = \frac{1}{B}\sum_{b=1}^{B}\log(w^*_{kb})$. Compute the standard deviation

$$sd_k = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\log(w^*_{kb}) - \bar{l})^2}$$

and define $s_k = sd_k\sqrt{1+\frac{1}{B}}$. (Note: we use logs to make estimates more robust to outliers if we assume it is logarithmically concave as normal distributions). Finally, choose the number of clusters via $\hat{k}$ = smallest k such that

$$Gap(k) \geq Gap(k+1) - s_{k+1}$$

.

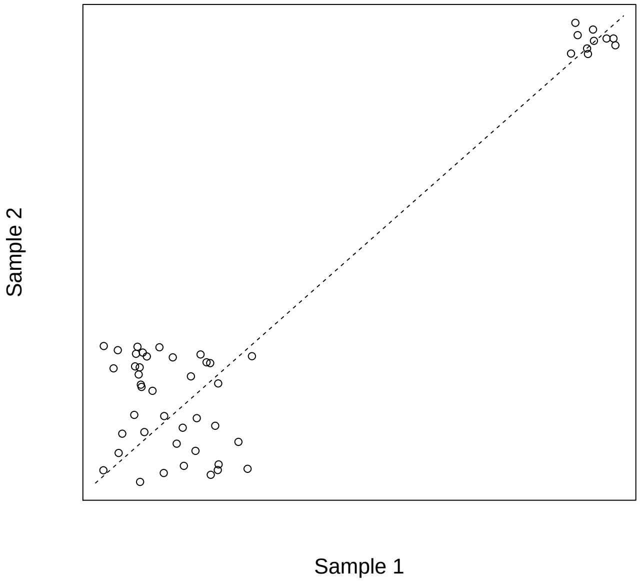
4 Practical Issues

1. Apply some filtering criteria before clustering genes to avoid housekeeping gene bias, e.g., via *Coefficient of Variation*: $CV = \frac{\sigma}{\mu}$, or

$$CV(i) = \sqrt{\frac{\frac{1}{p-1}\sum_{j=1}^{p}(x_{ij} - \bar{x_i})^2}{(\bar{x_i})^2}}, \bar{x_i} = \frac{1}{p}\sum_{j=1}^{p} x_{ij}.$$

Afterward, filter out genes with low CV (e.g., 20%).

2. Distance metric (in comparison of two samples):



May have high Pearson correlation, but may not mean the two samples are good replicates. As a solution, try either:

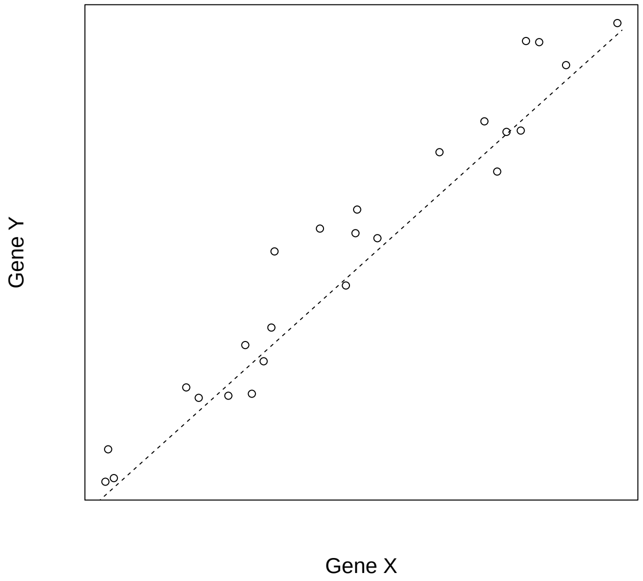
i) log transformation, or

ii) rank correlation.

5 Liquid Association

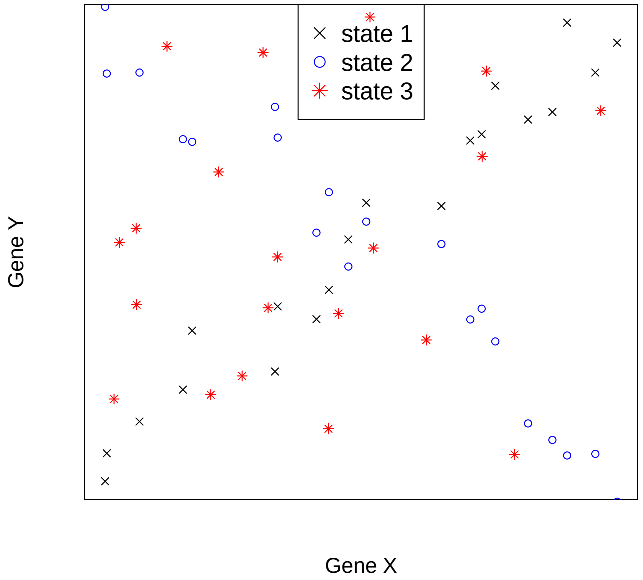
To measure dynamic correlation between datasets, we can use Liquid Association (LA) [3].

1. Static similarity between the expression profiles / patterns of two genes X and Y



   will always be highly correlated.

2. Dynamic correlation between X and Y, depending on the cellular state



   supposes the cellular state is positively correlated with a third gene Z.

5.1 Definition of Liquid Association

Suppose X, Y, and Z all have mean *0* and variance *1*. Then

$$LA(X, Y|Z) = E[g\prime(Z)],$$

where

$$g(z) = corr(X, Y|Z = z) = E[XY|Z = z].$$

Then if $Z \sim N(0, 1)$, then using Stein's Lemma,

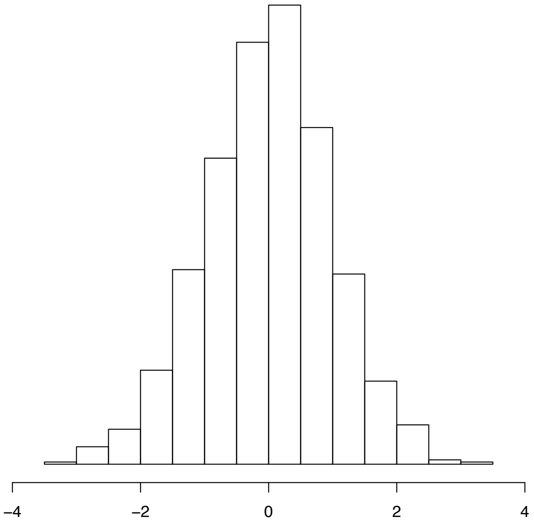
$$E[g\prime(Z)] = E[g(Z)Z] = E[E[XY|Z]Z] = E[XYZ].$$

5.2 Calculation of LA score

1. Standardize each gene expression profile $(g_1, ..., g_n)$ with a normal score transformation. Record the ranks of the n values as $R_1, ..., R_n$ and obtain the transformed profile:

$$\Phi^{-1}(\frac{R_1}{n+1}), ..., \Phi^{-1}(\frac{R_n}{n+1}).$$

   We transform the gene pattern to a normal distribution by ranking the values and sampling to a normal distribution.



2. Compute the average product of the three transformed profiles

$$\frac{X_1Y_1Z_1 + ... + X_nY_nZ_n}{n}.$$

5.3 Statistical Significance

Randomly permute the expression profile of genes $z = (z_1, ..., z_n)$ after transformation and for each permuted profile z^*, compute the LA score of X and Y. For a significance estimate, calculate how often $LA(X, Y|Z^*) \geq LA(X, Y|Z)$.

References

[1] R. Tibshirani, G. Walther and T. Hastie, “Estimating the Number of Data Clusters via the Gap Statistic”, *J.R. Statist. Soc. B*, vol. 63, Part 2, pp. 411–423, 2001.

[2] M Maechler, “Gap Statistic for Estimating the Number of Clusters”, *Seminar for Statistics*, Swiss Federal Institute of Technology Zurich, 2014.

[3] K. Li, “Genome-wide coexpression dynamics: Theory and application”, *PNAS*, vol. 99, no. 26, pp. 16876–16880, 2002.