Clustering of Multidimensional Data Sets with

Applications to Spatial Distributions of Ribosomal Proteins

UMBC REU Site: Interdisciplinary Program in High Performance Computing

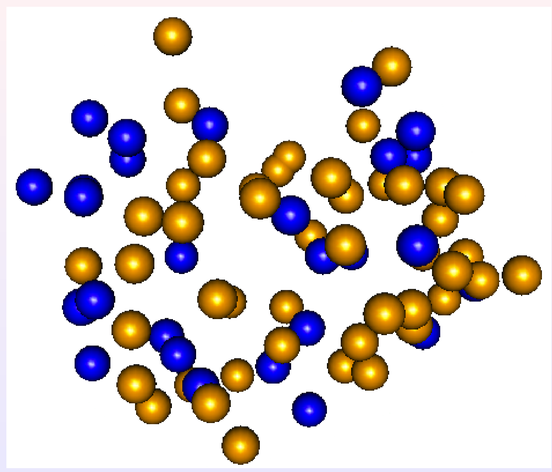
Nil Mistry[1], Jordan Ramsey[2], Benjamin Wiley[3], and Jackie Yanchuck[4]
Graduate Research Assistants: Xuan Huang[2] and Andrew Raim[2]
Faculty Mentors: Matthias K. Gobbert[2] and Nagaraj K. Neerchal[2], Client: Philip Farabaugh[5]
[1]University of Connecticut, [2]Mathematics and Statistics, UMBC, [3]University of New Mexico, [4]Seton Hill University, and [5]Biology, UMBC

Problem Statement

The objective is to consider ribosome samples of two groups, phenotype and non-phenotype, and investigate whether the distribution of the locations of the two groups of proteins are the same. To do this, the Mahalanobis distance is computed between each pair, and the optimal coupling minimizes the sum of the within-pair distances. The number of cross-matched pairs will determine whether or not the distributions of the locations of proteins are the same.

Biology

- Ribosomes are composed of RNA molecules and ribosomal proteins.
- Two different types of proteins: phenotype (blue) and non-phenotype (yellow).
- Object is to determine if the distribution of the locations of the proteins in both groups are the same.



Non-bipartite Matching

- Matching sort algorithm for N data points.
- Optimal pairing minimizes sum of within-pair distances.
- Within-pair distance: $\beta_{ij} = \max_{b,c}(\delta_{bc}) - \delta_{ij}$,
  where δ_{ij} is the distance between two points.
- Total number of possible pairings: $\binom{N}{2}$ where $N = 76$.

Mahalanobis Distance

- Non-Euclidean distance calculated with respect to the pooled variance-covariance matrix
- Mahalanobis distance from i^{th} to the j^{th} group (Rosenbaum):

$$\delta_{ij}^2 = \delta_{ji}^2 = (R_i - R_j)^T \, S^{-1} \, (R_i - R_j)$$

- R_k: rank of observation k
- S: pooled variance-covariance matrix

Method and Results

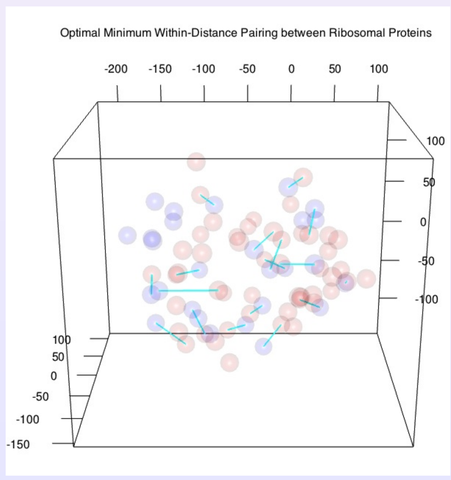
- The test statistic used is the number of cross-matched pairs, a_1, from the non-bipartite matching (Rosenbaum).
- The null distribution for the number of cross-matched pairs a_1 is

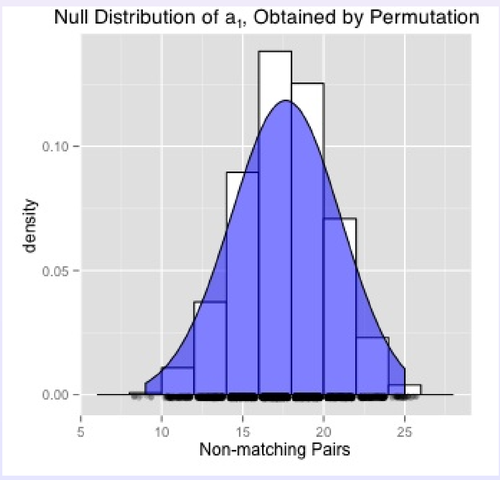
$$Pr(A_1 = a_1 | Y) = \frac{2^{a_1} \, I!}{\binom{N}{n} \, a_0! \, a_1! \, a_2!} = \pi_{a_1},$$

  where n is the number of phenotype proteins,

$$a_2 = \frac{n - a_1}{2}, \text{ and } a_0 = I - \frac{n + a_1}{2}.$$

- For our data, $a_1 = 17$ with associated p-value $= 0.547$. Therefore, we conclude that the distribution of the protein locations are statistically similar.





Highlights of Our R Code

- Determines rank of each attribute
- Computes Mahalanobis distance between all pairs
- Finds optimal matching based on Mahalanobis distances
- Determines number of cross-matched pairs
- Calculates test statistic based on number of cross-matched pairs

We implement these methods in an R code that calls a C-program for the non-bipartite matching.

References and Acknowledgments

- For full technical report, see HPCF–2013–10 at `www.umbc.edu/hpcf` > Publications
- Paul R. Rosenbaum, *J. R. Statist. Soc. B*, 2005

REU site: `www.umbc.edu/hpcreu`, NSF, NSA, UMBC, HPCF, CIRC