# Cross-Clustering: A Partial Clustering Algorithm with Automatic Estimation of the Number of Clusters

Paola Tellaroli, Marco Bazzi, Michele Donato,
Alessandra R. Brazzale and Sorin Drăghici

University of Padova and Wayne State University, Detroit

Four of the most common limitations of the many available clustering methods are: i) the lack of a proper strategy to deal with outliers; ii) the need for a good a priori estimate of the number of clusters to obtain reasonable results; iii) the lack of a method able to detect when partitioning of a specific data set is not appropriate; and iv) the dependence of the result on the initialization.

Here we propose Cross-clustering (CC), a partial clustering algorithm that overcomes these four limitations by combining the principles of two well established hierarchical clustering algorithms: Ward's minimum variance and Complete-linkage.

We validated CC by comparing it with a number of existing clustering methods, including Ward's and Complete-linkage. We show on both simulated and real datasets, that CC performs better than the other methods in terms of: the identification of the correct number of clusters, the identification of outliers, and the determination of real cluster memberships.

We used CC to cluster samples in order to identify disease subtypes, and on gene profiles, in order to determine groups of genes with the same behavior. The results show that the method is general enough to be successfully used in such diverse applications.

The algorithm has been implemented in the statistical language R and soon will be freely available from the CRAN contributed packages repository.