Diversions on the minimum spanning tree with mstree

Nicholas Lewin-Koh[1,⋆]

1. Genentech, 1 DNA way, South San Francisco, CA 94080
⋆Contact author: nikko@hailmail.net

Keywords: Multivariate ordering, minimum spanning tree, runt analysis, clustering

The minimum spanning tree (MST), is the tree that covers all points with minimum distance or cost. There are many results on the MST that relate to clustering, ordering of data, and dimension reduction. This package provides several functions for doing analysis on the MST.

The package is designed around a C core set of functions for calculating the MST. A variation on Primm's algorithm is used with a Fibonacci heap based on code by Ben Pfaff. The interface will accept a numeric matrix and calculate Euclidean distance, or a `dist` object and return the MST as a vector of integers as in the similar Splus function `mstree`. The function also returns the linear and radial node orderings of Friedman and Rafsky (1981) as well as the optimal ordering as proven by Shiloach (1979) and Chung (1983).

Further functions calculate the runt distribution of the MST edges. These can be used to prune the cluster tree of a density as outlined in Steutzle 2003. Other functions using the MST may be added in the future.

References

F. R. K Chung (1984). On optimal linear arrangement of trees. *Comp. & Math with Appls.* 10(1), 43–60.

Jerome H. Friedman and Lawrence C. Rafsky (1979). Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two–Sample Tests. *Ann. Statist.* 7, 697–717.

Ben Pfaff (2001). Uniformity testing library
`http://www.stanford.edu/~blp/projects.html`

Yossi Shiloach (1979). A Minimum Linear Arrangement Algorithm for Undirected Trees. *SIAM J. Comput.* 8, 15–32

Werner Stuetzle (2003). Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification* 20, 25–47.