Mathematical Statistics
Stockholm University
Master Thesis **2020:5**
http://www.math.su.se

When graph theory meets unsupervised learning: the statistical properties 'spectral clustering'

Fanny Bergström*

June 2020

Abstract

Spectral clustering treats clustering as a graph partitioning problem, where clusters are constructed based on the *commute time distance* (CTD) between the nodes of the graph. The CTD combines the local connections between nearby points to establish the global connections among remote points, allowing us to detect shapes and intrinsic manifold structures buried in high dimensional data. Furthermore, the CTD is simply the Euclidean distance in the space spanned by the eigenvectors of the graph Laplacian. This implies that clusters can be easily detected using a classical clustering algorithm (e.g., k-means) when data is represented in this space. This thesis aims to scrutinize the statistical principles of this nonparametric clustering method, which is robust and manages to capture both local and global geometrical structures in the data. Properties of the CTD and its relations with the clustering structures of the data are investigated extensively.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: fannybergstrom@gmail.com. Supervisor: Chun-Biu Li.