STATISTICAL PERTURBATION THEORY FOR SPECTRAL CLUSTERING WITH APPLICATION TO THE ANALYSIS OF GENE EXPRESSION DATA

Zhivko Stoyanov, Alastair Spence
Department of Mathematical Sciences, University of Bath,
Bath BA2 7AY, UK,
e-mail: mapzvs@maths.bath.ac.uk, as@maths.bath.ac.uk

Keywords: statistical perturbation, eigenvectors, singular vectors, clustering, gene expression data

Abstract

In biology, microarrays generate huge gene expression data sets that require reliable algorithms for their analysis. An important problem is to cluster genes (or samples) into groups, and this is often achieved by spectral clustering, for example, using the Fiedler vector of the Laplacian matrix of a network or a singular vector of the adjacency matrix of a network. A major difficulty is that the biological data is subject to many random errors and the clustering results can be very sensitive to perturbation.

In this talk we first discuss some spectral perturbation theory for a symmetric matrix where the perturbation parameter is a random variable. Second, we extend this to the statistical perturbation of singular vectors of a rectangular matrix. We illustrate the theory with numerical results obtained from gene expression data arising from the analysis of certain cancer tissue samples.