Spare Simultaneous Factor Analysis in Bioclustering

許英麟
中興大學應用數學系

Abstract

One important research question in genomic data analysis is to identify subgroups for appropriate medical treatment. Biclustering is an effective clustering method to help classify subgroups by forming similar genomic patterns. Various approaches have been employed in biclustering, including spare factor modeling which assuming both factor loading and factor scores are both sparse. However, existing spare factor modeling in biclustering involves the crucial issue of estimation in both factor loading and factor scores, especially in high-dimensional data. A new estimation approach, spare simultaneous factor analysis (SSFA), is proposed to address this concern. Specifically, to deal with the sparseness assumption and to estimate the parameters simultaneously in the factor model, the loss function includes two L_1 penalty terms that are associated with both factor loadings and factor scores, as well as one alternative least-squares algorithm for estimating parameters. In addition, sparse singular value decomposition is utilized to simultaneously estimate sparse factor loading and sparse factor scores in the iterative process. Simulation studies show that the proposed approach yields a smaller bias and variance than other common sparse modeling methods, such as factor analysis for bicluster acquisition, penalized factor analysis, and sparse singular value decomposition. Microarray data from a breast cancer study is used to illustrate the potential application of the method in genomic research with high-dimensional data.