Tuning of COSA for High-Dimensional Data

Maarten M.D. Kampert*
Mathematical Institute, Leiden University
Leiden
Netherlands
mkampert@math.leidenuniv.nl

Jacqueline J. Meulman
Mathematical Institute, Leiden University
Leiden
Netherlands
jmeulman@math.leidenuniv.nl

Abstract. In high-dimensional data settings, noise can overwhelm the few signals that may be present. Finding these signals is an active and ongoing topic of research. Already in 2004, Friedman and Meulman proposed a procedure called Clustering Objects on Subsets of Attributes (COSA) as a possible solution. The objective of COSA is to cluster the objects in attribute-value data, and its main motivation was to consider a particular kind of data, specifically from genomics, proteomics, and metabolomics, usually subsumed under the name of high-dimensional systems. COSA is an unsupervised algorithm that outputs a "cluster-happy" dissimilarity matrix that one can use for subsequent analysis by a variety of proximity methods.

Of the many publications that cite the COSA paper, only a small number actually use COSA. It is a very complicated algorithm that is difficult to program, and the available software has been out-of-date for quite some time. Furthermore, there is no specific guidance available on how to tune the parameters. In this talk we present the results of two projects. In the first project, we give an introduction to the state-of-the-art software package, rCOSA. The package extends the previous COSA package with additional functions for hierarchical methods, multidimensional scaling, K-groups clustering, and data visualization. The software package and related links are available for free at: https://github.com/mkampert/rCOSA. In the second project we present a clear and specific guidance for tuning the parameters in COSA using the GAP statistic. We will show, optimizing on a small two or three dimensional grid, that one can increase the power of COSA, especially with small group of objects that cluster on small, distinct subsets of attributes. We support our findings by applications on metabolomics.

Keywords
Variable Selection, Clustering, Metabolomics