Sparse Principal Component Analysis Incorporating Stability Selection

Martin Sill*

German Cancer Research Center, Heidelberg, Germany
m.sill@dkfz.de

Principal component analysis (PCA) is a popular dimension reduction method that approximates a numerical data matrix by seeking principal components (PC), i.e. linear combinations of variables that captures maximal variance. Since each PC is a linear combination of all variables of a data set, interpretation of the PCs can be difficult, especially in high-dimensional data. In order to find 'sparse' PCs that are linear combinations of only a subset of possibly relevant variables and therefore easier to interpret, several sparse PCA approaches have been proposed in the recent years. Typically, these methods use the singular value decomposition (SVD) to calculate PCs. Sparsity is attained by relating the SVD to linear regression and perform a variable selection using penalty terms similar to those in penalized linear regression models. Our approach combines such a regularized SVD with stability selection. Stability selection is a general approach that combines variable selection methods, e.g. penalized regression models, with resampling techniques to control the error of falsely selecting irrelevant variables. Thus our new approach is able to find sparse PCs that are linear combinations of subsets of variables selected with respect to Type I error control. The performance of the proposed method will be compared with other sparse PCA approaches by a simulation study. Application of the method will be demonstrated using high-dimensional molecular data.

Keywords: SVD, high-dimensional, penalization, resampling, variable selection