Consistency of principal component scores in visualizations of high-dimensional data

Kristoffer H. Hellton*
Dept. of Biostatistics, University of Oslo, Oslo, Norway - k.h.hellton@medisin.uio.no

Magne Thoresen
Dept. of Biostatistics, University of Oslo, Oslo, Norway - magne.thoresen@medisin.uio.no

Plots of principal component scores are a popular approach to visualize and explore high-dimensional data. However, the inconsistency of high-dimensional eigenvectors prompted the development of sparse principal component analysis (PCA), where sparse eigenvectors can be estimated using different regularization approaches. Still, classical PCA is extensively and successfully used to visualize high-dimensional genetic data. In this work, we try to give an explanation of this paradoxical situation. We show that the visual information in terms of the relative positions of scores will be consistent, if the related signal can be considered to be pervasive, despite of inconsistent eigenvectors. Within the High Dimension Low Sample Size (HDLSS) asymptotic framework, we show that when the spiked population eigenvalues scale linearly with the dimension, the sample principal component scores are scaled and rotated versions of the population principal component scores. Further, we argue that eigenvalues scaling linearly with the dimension can be interpreted as expressions of pervasive signals, specifically given by eigenvectors with an asymptotic non-zero proportion of non-zero coefficients, and we discuss examples of genetic applications where assuming pervasive signals is reasonable.

Keywords: Genomics; HDLSS; PCA; Asymptotic distribution.