Statistically Guided Divide-and-Conquer for Sparse Factorization of Large Matrix

Abstract

The sparse factorization of a large matrix is fundamental in modern statistical learning. In particular, the sparse singular value decomposition and its variants have been utilized in multivariate regression, factor analysis, bi-clustering, vector time series modeling, among others. The appeal of this factorization is owing to its power in discovering a highly interpretable latent association network, either between samples and variables or between responses and predictors. However, many existing methods are either ad hoc without a general performance guarantee, or are computationally in- tensive, rendering them unsuitable for large-scale studies. We formulate the statistical problem as a sparse factor regression and tackle it with a divide-and-conquer approach. In the first stage of division, we consider both sequential and parallel approaches for simplifying the task into a set of co-sparse unit-rank estimation (CURE) problems and establish the statistical underpinnings of these commonly adopted and yet poorly understood deflation methods. In the second stage of division, we innovate a contended stagewise learning technique, consisting of a sequence of simple incremental updates, to efficiently trace out the whole solution paths of CURE. Our algorithm has a much lower computational complexity than alternating convex search, and the choice of the step size enables a flexible and principled tradeoff between statistical accuracy and computational efficiency. Our work is among the first to enable stagewise learning for non-convex problems, and the idea can be applicable in many multi-convex problems. Extensive simulation studies and an application in genetics demonstrate the effective- ness and scalability of our approach.

Keywords: biconvex; boosting; singular value decomposition; stagewise estimation.