Statistical and computational trade-offs in estimation of sparse principal components

Richard Samworth*
University of Cambridge, Cambridge, United Kingdom – r.samworth@statslab.cam.ac.uk

Tengyao Wang
University of Cambridge, Cambridge, United Kingdom – t.wang@statslab.cam.ac.uk

Quentin Berthet
California Institute of Technology – qberthet@caltech.edu

In recent years, Sparse Principal Component Analysis has emerged as an extremely popular dimension reduction technique for high-dimensional data. The theoretical challenge, in the simplest case, is to estimate the leading eigenvector of a population covariance matrix under the assumption that this eigenvector is sparse. An impressive range of estimators have been proposed; some of these are fast to compute, while others are known to achieve the minimax optimal rate over certain Gaussian or subgaussian classes. In this paper we show that, under a widely-believed assumption from computational complexity theory, there is a fundamental trade-off between statistical and computational performance in this problem. More precisely, working with new, larger classes satisfying a restricted Covariance Concentration condition, we show that no randomised polynomial time algorithm can achieve the minimax optimal rate. On the other hand, we also study a (polynomial time) variant of the well-known semidefinite relaxation estimator, and show that it attains essentially the optimal rate among all randomised polynomial time algorithms.

Keywords: Restricted covariance concentration condition, Sparse PCA, statistical and computational trade-offs.