High-dimensional principal component analysis with heterogeneous missingness

Tengyao Wang

London's Global University
E-mail: tengyao.wang@ucl.ac.uk

Abstract: We study the problem of high-dimensional Principal Component Analysis (PCA) with missing observations. In simple, homogeneous missingness settings with a noise level of constant order, we show that an existing inverse-probability weighted (IPW) estimator of the leading principal components can (nearly) attain the minimax optimal rate of convergence. However, deeper investigation reveals both that, particularly in more realistic settings where the missingness mechanism is heterogeneous, the empirical performance of the IPW estimator can be unsatisfactory, and moreover that, in the noiseless case, it fails to provide exact recovery of the principal components. Our main contribution, then, is to introduce a new method for high-dimensional PCA, called primePCA, that is designed to cope with situations where observations may be missing in a heterogeneous manner. Starting from the IPW estimator, primePCA iteratively projects the observed entries of the data matrix onto the column space of our current estimate to impute the missing entries, and then updates our estimate by computing the leading right singular space of the imputed data matrix. It turns out that the interaction between the heterogeneity of missingness and the low-dimensional structure is crucial in determining the feasibility of the problem. We therefore introduce an incoherence condition on the principal components and prove that in the noiseless case, the error of primePCA converges to zero at a geometric rate when the signal strength is not too small. An important feature of our theoretical guarantees is that they depend on average, as opposed to worst-case, properties of the missingness mechanism. Our numerical studies on both simulated and real data reveal that primePCA exhibits very encouraging performance across a wide range of scenarios.