Discussion of large covariance estimation by thresholding principal orthogonal complements

Hanwen Huang, Yufeng Liu, J.S. Marron, Dan Shen, and Haipeng Shen
University of North Carolina at Chapel Hill
Correspondence email: marron@unc.edu

We congratulate the authors on a very interesting contribution, which takes the fundamentally important field of covariance matrix estimation in some important new directions. We agree that now is a good time to be studying asymptotic contexts, where the first $K$ eigenvalues of $\Sigma$ grow quickly. The asymptotic mode of the sample size tending to infinity, with an exponentially growing dimension can be improved by taking the dimension as the asymptotic driver, with the sample size growing at a logarithmic rate. This makes it clear that this setting is very close to the high dimension low sample size (HDLSS) setting with fixed sample size. Hall et al. (2005). Shen et al. (2012a) studied another notion of PCA consistency, in a wide range of such settings. Shen et al. (2012b) studied another approach to sparsity under a growing eigenvalue assumption, establishing a new characterization of the boundary between regions of consistency and strong inconsistency for sparse PCA, in HDLSS settings. Can similar results be established for POET?

Another reason we are excited about these results, is that covariance estimation is a critical component of SigClust, which is very useful for testing statistical significance of clusters in high dimensional contexts, Liu et al. (2008) and Huang et al. (2012). This motivated us to compare POET with the approaches used in SigClust. A key step of the SigClust analysis is to estimate the eigenvalues of the covariance matrix of the null multivariate Gaussian distribution. The latter paper proposed a likelihood based soft thresholding approach for estimating the covariance eigenvalues which gave a large improvement relative to the hard thresholding approach of the former paper. Figure A shows estimates of the eigenvalue spectrum for two simulated HDLSS examples with sample size $n = 50$ and dimension $d = 1000$. Gaussian data are simulated with mean 0 and covariance matrix $\Lambda = \text{diag}(\lambda, \cdots, \lambda, 1, \cdots, 1)$. Figure A (left) shows that in situations where the number of the spikes is larger than the sample size, the POET method gives a major improvement. Figure A (right) shows that in situations with few large spikes, the soft method works better than POET due to better background noise estimation. Ultimately, a combination of POET with existing SigClust methods may work better.

References


Fig. A. Plots of the estimated covariance matrix eigenvalues based on the hard, soft, and POET methods for two simulated data sets with $d = 1000$ and $n = 50$. The left panel displays the results for spike size $\lambda = 5$ and $w = 200$ spike entries. The right panel displays the results for $\lambda = 100$ and $w = 10$. This shows POET works better in the left example, while the soft thresholding is better in the right hand case.