Econometrics **June 27, 2023**

Topic 18: Eigenvalue and Spike Models

by Sai Zhang

Key points: .

Disclaimer: *The note is built on Prof. Jinchi Lv's lectures of the course at USC, DSO 607, High-Dimensional Statistics and Big Data Problems.*

18.1 Motivation

Consider n independent observations $\mathbf{X}_i \in \mathbb{R}^p$ drawn from a $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, then the covariance can be decomposed into 2 parts, white noise and low rank

$$\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{X}_i) = \mathbf{I} + \sum_{k=1}^{M} \theta_k v_k v_k' = \mathbf{\Sigma}_0 + \mathbf{\Phi}$$

where M denotes the **number of spikes** in the distribution of eigenvalues. The idea is: spikes deviate from a reference model along a **small fixed number** of unknown directions. If $\mathbf{\Phi} = \mathbf{0}$, then none of the sample eigenvalues is separated from the bulk.

Why a spike model is interesting? A spike model can help determine the latent dimension of the data, some examples being

- Principal component analysis (PCA): spikes are related to the directions of the most variations of the data, i.e., the principal components
- Clustering model: M spikes is equivalent to $M + 1$ clusters
- Economic significance: M is related to the number of factor loadings

Then the question is threefold:

- How to determine M
- How to estimate v_k
- How to test θ_k

Under rank one alternative, we would like to test

$$H_0 : \mathbf{\Sigma} = \mathbf{I}_p \qquad\qquad H_1 \mathbf{\Sigma} = \mathbf{I}_p + \theta \boldsymbol{\nu}\boldsymbol{\nu}', \theta > 0$$

References