

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 $\mathcal{N}(\mathbf{0}, \Sigma)$, then the covariance can be decomposed into 2 parts, white noise and low rank

$$\Sigma = \text{Cov}(\mathbf{X}_i) = \mathbf{I} + \sum_{k=1}^M \theta_k \mathbf{v}_k \mathbf{v}_k' = \Sigma_0 + \Phi$$

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 $\Phi = \mathbf{0}$, 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 \mathbf{v}_k
- How to test θ_k

Under rank one alternative, we would like to test the hypothesis

$$H_1 : \Sigma = \mathbf{I}_p + \theta \mathbf{v} \mathbf{v}', \theta > 0$$

against the null

$$H_0 : \Sigma = \mathbf{I}_p$$

with the key assumptions:

A1 Gaussian error

A2 large p : $p \leq n$ but allows $p/n \rightarrow \gamma \in (0, 1)$

Under these assumptions, for the $n \times p$ data matrix $\mathbf{X} = (\mathbf{X}'_1 \cdots \mathbf{X}'_n)'$, $\mathbf{X}'\mathbf{X}$ has a p -dimensional **Wishart** distribution $W_p(n, \mathbf{\Sigma})$ with the degree of freedom n and covariance matrix $\mathbf{\Sigma}$, which is a *random matrix*.

If $\mathbf{Y} = \mathbf{M} + \mathbf{X}$, that is, the sum of the *random matrix* \mathbf{X} and a *deterministic matrix* \mathbf{M} (also $n \times p$), then $\mathbf{Y}'\mathbf{Y}$ has a p -dimensional Wishart distribution $W_p(n, \mathbf{\Sigma}, \mathbf{\Psi})$ with n degrees of freedom, covariance matrix $\mathbf{\Sigma}$ and non-centrality matrix $\mathbf{\Psi} = \mathbf{\Sigma}^{-1}\mathbf{M}'\mathbf{M}$.

Definition 18.1.1: Density of Wishart Distribution

The PDF of Wishart distribution is defined as

$$f(\mathbf{X}) = \frac{1}{2^{np/2} \Gamma_p\left(\frac{n}{2}\right) |\mathbf{\Sigma}|^{n/2}} |\mathbf{X}|^{(n-p-1)/2} \exp\left(-\frac{1}{2} \text{tr}(\mathbf{\Sigma}^{-1}\mathbf{X})\right)$$

where \mathbf{X} is a symmetric positive semidefinite and $\Gamma_p\left(\frac{n}{2}\right)$ is a multivariate gamma function such that

$$\Gamma_p\left(\frac{n}{2}\right) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^p \Gamma\left(\frac{n}{2} - \frac{j-1}{2}\right)$$

Notice that the sample covariance matrix $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ is just a scaled version of Wishart distribution

$$n\mathbf{S} = \mathbf{X}'\mathbf{X} \sim W_p(n, \mathbf{\Sigma})$$

For $\mathbf{\Sigma} = \mathbf{I}_p$, the empirical distribution fo eigenvalues converges to Marcenko-Pastur distribution

$$f^{\text{MP}}(x) = \frac{1}{2\pi\gamma x}$$

References