Econometrics	**July 4, 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

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

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 ν_k
- How to test θ_k

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

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

against the null

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

with the key assumptions:

A1 Gaussian error

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

Under these assumptions, for the $n \times p$ data matrix $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1' & \cdots & \mathbf{X}_n' \end{pmatrix}'$, $\mathbf{X}'\mathbf{X}$ 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

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

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

$$f^{\mathrm{MP}}(x) = \frac{1}{2\pi\gamma x}\sqrt{(b_+ - x)(x - b_-)}$$

where $b_\pm = (1 \pm \sqrt{\gamma})^2$. Then:

- under $H_0 : \mathbf{\Sigma} = \mathbf{I}_p$, we have

  $$n^{2/3}\left(\frac{\lambda_1 - \mu(\gamma)}{\sigma(\gamma)}\right) \xrightarrow{d} \mathrm{TW}_1$$

  where TW_1 is the Tracy-Widom distribution
- under $H_1 : \mathbf{\Sigma} = \mathbf{I}_p + \theta\boldsymbol{\nu}\boldsymbol{\nu}', \theta > 0$, if θ is strong ($\theta \gg \sqrt{\gamma}$), then

  $$n^{1/2}\left(\frac{\lambda_1 - \rho(\theta,\gamma)}{\tau(\theta,\gamma)}\right) \xrightarrow{d} \mathcal{N}(0,1)$$

Here, the largest eigenvalue test is the best test. **But** when the signal is weak ($0 \leq \theta < \sqrt{\gamma}$), the largest eigenvalue under the alternative converges to the same distribution as null:

$$n^{2/3}\left(\frac{\lambda_1 - \rho(\theta,\gamma)}{\tau(\theta,\gamma)}\right) \xrightarrow{d} \mathrm{TW}_1$$

which means that the largest eigenvalue test *fails*. On top of this, **resampling** also fails when p is large.

Next, we develop another test to cope with these problems.

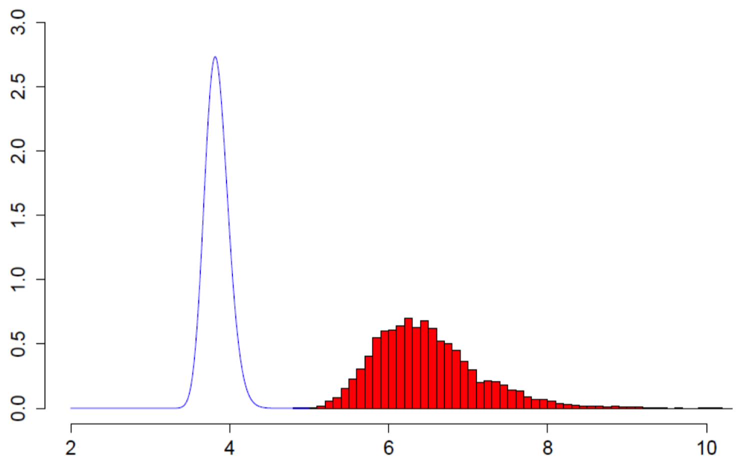

Figure 18.1: Failure of Resampling Test ($n = p = 100$)

18.2 Johnstone and Onatski (2020)

Consider the basic equation of classical multivariate statistics:

$$\det(\mathbf{H} - \mathbf{x}\mathbf{E}) = 0 \tag{18.1}$$

with $p \times p$ matrices

$$n_1\mathbf{H} = \sum_{k=1}^{n_1} \mathbf{x}_k\mathbf{x}_k' \qquad \textit{hypothesis}\text{ SS}$$

$$n_1\mathbf{E} = \sum_{k=1}^{n_1} \mathbf{z}_k\mathbf{z}_k' \qquad \textit{error}\text{ SS}$$

The solution $\mathbf{x}$ is generalized eigenvalues $\{\lambda_i\}_{i=1}^p$, which are the eigenvalue of F-ratio $\mathbf{E}^{-1}\mathbf{H}$. Johnstone and Onatski (2020) summarized 5 topics using $\mathbf{E}^{-1}\mathbf{H}$ relying on the five most common hypergeometric functions[1] ${}_p\mathcal{F}_q$

[1]Hypergeometric functions are:

- scalar inputs

$${}_p\mathcal{F}_q(a, b; x) = \sum_{k=0}^{\infty} \frac{(a_1)_k \cdots (a_p)_k}{(b_1)_k \cdots (b_p)_k} \frac{x^k}{k!}$$

  where $(a_j)_k$ are generalized Pochhammer symbols
- single matrix inputs, where $\mathbf{S}$ is symmetric and usually diagonal

$${}_p\mathcal{F}_q(a, b; \mathbf{S}) = \sum_{k=0}^{\infty} \sum_{\kappa} \frac{(a_1)_\kappa \cdots (a_p)_\kappa}{(b_1)_\kappa \cdots (b_p)_\kappa} \frac{C_\kappa(\mathbf{S})}{k!}$$

  where C_k are the zonal polynomials. Easily, ${}_0\mathcal{F}_0(\mathbf{S}) = e^{\operatorname{tr}(\mathbf{S})}$, ${}_1\mathcal{F}_0(a, \mathbf{S}) = |\mathbf{I} - \mathbf{S}|^{-a}$
- two matrix inputs, where $\mathbf{S}, \mathbf{T}$ are both symmetric

$${}_p\mathcal{F}_q(a, b; \mathbf{S}, \mathbf{T}) = \int_{O(p)} {}_p\mathcal{F}_q(a, b; \mathbf{SUTU}')(d)\mathbf{U}$$

Table 18.1: 5 Statistical Methods

		Statistical method	$n_1\mathbf{H}$	$n_2\mathbf{E}$	Univariate Analog
${}_0\mathcal{F}_0$	PCA	Principal components analysis	$W_p(n_1, \mathbf{\Sigma} + \mathbf{\Phi})$	$n_2\mathbf{\Sigma}$	χ^2
${}_1\mathcal{F}_0$	SigD	Signal detection	$W_p(n_1, \mathbf{\Sigma} + \mathbf{\Phi})$	$W_p(n_2, \mathbf{\Sigma})$	non-central χ^2
${}_0\mathcal{F}_1$	REG_0	Multivariate regression, with known error	$W_p(n_1, \mathbf{\Sigma}, n_1\mathbf{\Phi})$	$n_2\mathbf{\Sigma}$	F
${}_1\mathcal{F}_1$	REG	Multivariate regression, with unknown error	$W_p(n_1, \mathbf{\Sigma}, n_1\mathbf{\Phi})$	$W_p(n_2, \mathbf{\Sigma})$	non-central F
${}_2\mathcal{F}_1$	CCA	Canonical correlation analysis	$W_p(n_1, \mathbf{\Sigma}, \mathbf{\Phi}(\mathbf{Y}))$	$W_p(n_2, \mathbf{\Sigma})$	$\frac{r^2}{1-r^2}$

For ${}_0\mathcal{F}_0$ and ${}_0\mathcal{F}_1$, $\mathbf{E}$ is deterministic, $\mathbf{\Sigma}$ is known, n_2 disppears, otherwise $\mathbf{E}$ is independent of $\mathbf{H}$.

18.2.1 Definitions and global assumptions

Let $\mathbf{Z}$ be an $n \times p$ data matrix with rows (observations) drawn **i.i.d.** from $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, and a deterministic matrix $\mathbf{M}$ of $n \times p$, then for $\mathbf{Y} = \mathbf{M} + \mathbf{Z}$,

- $\mathbf{H} = \mathbf{Y}'\mathbf{Y}$ 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 $\mathbf{\Psi} = \mathbf{\Sigma}^{-1}\mathbf{M}'\mathbf{M}$
- the corresponding central Wishart distribution with $\mathbf{M} = \mathbf{0}$ is $W_p(n, \mathbf{\Sigma})$

Johnstone and Onatski (2020) assume a relative low dimensionality $p \leq \min\{n_1, n_2\}$ where n_1, n_2 are the degrees of freedom as in Table 18.1, where

- $p \leq n_2$ ensures almost sure invertibility of matrix $\mathbf{E}$ in Equation 18.1
- $p \leq n_1$ is not essential, but reduces the number of various situations of consideration.

Under these assumptions, they established a unified statistical problem **symmetric matrix denoising (SMD)** that can be linked to the 5 classes of problems:

PCA

References

Iain M Johnstone and Alexei Onatski. Testing in high-dimensional spiked models. *The Annals of Statistics*, 48(3), 2020.