Eigen-Structure of Sample Covariance

김성민

서울대학교
통계학과, 베이즈통계 연구실

2022. 10. 05

CONTENTS

CONTENTS

Material

- El Karoui, N. Spectrum estimation for large dimensional covariance matrices using random matrix theory.
  The Annals of Statistics 36, 6 (2008), 2757–2790 [1]
- Paul, D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model.
  Statistica Sinica (2007), 1617–1642 [3]
- Ledoit, O., and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices.
  Journal of multivariate analysis 88, 2 (2004), 365–411 [2]

CONTENTS

Eigenvalue of Covariance

- Principal component analysis (PCA)
- Low-dimensional approximation to the data by projecting the data on the "best" possible k-dimensional subspace.

Setting

- We observed iid random vectors $X_1, \ldots, X_n$ in $\mathbb{R}^p$.
- Assume that covariance of X_i is Σ_p
- Sample covariance : $S_p = \dfrac{1}{n-1}(X - \bar{X})^T (X - \bar{X})$

Fixed p, large n

- It is well known that eigenvalues of S_p are good estimators of that of Σ_p.
- Let l_i be the ordered eigenvalues of S_p ($l_1 > l_2 > \cdots$) and λ_i be the ordered that of Σ_p ($\lambda_1 > \lambda_2 > \cdots$).

$$\sqrt{n}(l_i - \lambda_i) \xrightarrow{d} N(0, \lambda_i^2)$$

where X_i are normally distributed.

Large n, large p

- Let us consider the simplest case where $\Sigma_p = I_p$.
- If X_i are iid and have a fourth moment, and if $\frac{p}{n} \to \gamma$, then

$$l_1 \to (1 + \sqrt{\gamma})^2 \quad \text{a.s.}$$

- Note that if $n = p$, then l_1 goes to 4.

Large n, large p

- [3] focuses on Eigenvector of Sample Covariance.
- $\Sigma_p = diag(l_1, l_2, \ldots, l_M, 1, \ldots, 1)$
- If $p/n \to \gamma \in (0,1)$, then the sample eigenvectors can be inconsistency according to true eigenvalues.
  - If $l_v > 1 + \sqrt{\gamma}$,

$$|< p_v, e_v >| \to \sqrt{\left(1 - \frac{\gamma}{(l_v - 1)^2}\right) \Big/ \left(1 + \frac{\gamma}{l_v - 1}\right)}.$$

  - If $l_v \leq 1 + \sqrt{\gamma}$,

$$|< p_v, e_v >| \to 0.$$

Recent work on covariance estimation

- There is some work on shrinkage of eigenvalues to improve covariance estimation.
- [2] proposed to estimate Σ_p by $(1-\rho)S_p + \rho I_p$.
- The estimator can be viewed as maintaining eigenvector and linearly shrinkaging the eigenvalues.

CONTENTS

From vectors to measures

- There are some issues that arise when estimating vectors (set of eigenvalues) of high dimension.
- Propose to associate high-dimensional vectors probability measures.
  - Allow us to look into the structure of the population eigenvectors.
  - Practical benefits of the measure estimation approach.

From vectors to measures

- Suppose we have a eigenvalue $(\lambda_1, \ldots, \lambda_p)$ of Σ_p.
- Define a measure with p point mass with equal weight, and denote H_p as the **population spectral distribution**.

$$dH_p(x) = \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i}(x).$$

- Equivalently, define **empirical spectral distribution** F_p as

$$dF_p(x) = \frac{1}{p} \sum_{i=1}^{p} \delta_{l_i}(x).$$

Example of spectral distribution

- Suppose $dH_p = (1 - \frac{1}{p})\delta_1 + \frac{1}{p}\delta_2$.

  $\Longrightarrow$ It means that the Σ_p has one eigenvalue that is equal to 1, and $(p-1)$ that are equal to 2.

- Clearly, H_p weakly converges to H_∞, with $dH_\infty = \delta_1$.

Goal

- H_p : **Population spectral distribution**, F_p : **Empirical spectral distribution**,
- $F_p \to F_\infty$.
- $H_p \to H_\infty$.
- Some theorem connects F_∞ and H_∞.
- Our goal is estimating H_p by F_p.

CONTENTS

Stieltjes transform of measures

- Stieltjes transform of a measure G on $\mathbb{R}$ is defined as

$$m_G(z) = \int \frac{dG(x)}{x - z} \quad \text{for } z \in \mathbb{C}^+,$$

where $\mathbb{C}^+ = \mathbb{C} \cup \{z : Im(z) > 0\}$.

Properties of Stieltjes transforms on $\mathbb{R}$

Properties

- If G is a probability measure, $m_G(z) \in \mathbb{C}^+$ if $z \in \mathbb{C}^+$ and $\lim_{y\to\infty}(-iy) \cdot m_G(iy) = 1$.
- If F and G are two measures, and if $m_F(z) = m_G(z)$, for all $z \in \mathbb{C}^+$, then $G = F$, a.s.
- And so on....

Stieltjes transform of the spectral distribution

- Stieltjes transform of the spectral distribution Γ_p of a $p \times p$ matrix A_p is

$$m_{\Gamma_p}(z) = \frac{1}{p} trace((A_p - zI_p)^{-1}).$$

Definition

We will call v_{F_p} the function defined by

$$v_{F_p}(z) = (1 - \frac{p}{n})\frac{-1}{z} + \frac{p}{n} m_{F_p}(z).$$

Theorem

Theorem

Suppose the data matrix X can be written $X = Y\Sigma_p$, where Σ_p is a $p \times p$ positive definite matrix and Y is an $n \times p$ matrix whose entries are i.i.d. (real or complex), with $E(Y_{ij}) = 0,\ E(|Y_{ij}|^2) = 1$ and $E(|Y_{ij}|^4) < \infty$. Assume that H_p converges weakly to a limit denoted H_∞. Then, when $p, n \to \infty$, and $p/n \to \gamma,\ \gamma \in (0, \infty)$:

- $v_{F_p}(z) \to v_\infty(z)$ a.s., where $v_\infty(z)$ is a deterministic function.
- $v_\infty(z)$ satisfies the Marcenko-Pastur equation

$$-\frac{1}{v_\infty(z)} = z - \gamma \int \frac{\lambda dH_\infty(\lambda)}{1 + \lambda v_\infty(z)} \quad \forall z \in \mathbb{C}^+.$$

- The previous equation has one and only one solution which is the Stieltjes transform of a measure.

CONTENTS

Strategy

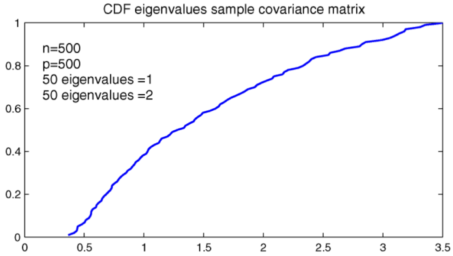
1. Estimating the measure H_∞ appearing in the Marcenko-Pastur equation.
2. Estimating λ_i as the ith quantile of $\hat{H}_\infty$.
3. Since we are considering fixed distribution asymptotics $(H_p = H_\infty)$, $\hat{H}_\infty$ will serve as estimate of H_p

CONTENTS

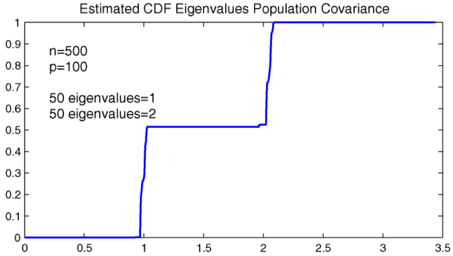
Examples

- $n = 500, \ p = 100$
- CASE1 : Consider the Σ_p which has 50% of its eigenvalues equal to 1 and 50% equal to 2.
- CASE2 : Consider the Toeplitz matrix Σ_p with entries $0.3^{|i-j|}$.

CASE1

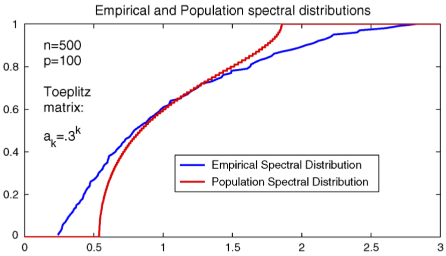

CDF eigenvalues sample covariance matrix
n=500
p=500
50 eigenvalues =1
50 eigenvalues =2
0
0.2
0.4
0.6
0.8
1
0
0.5
1
1.5
2
2.5
3
3.5

(b)

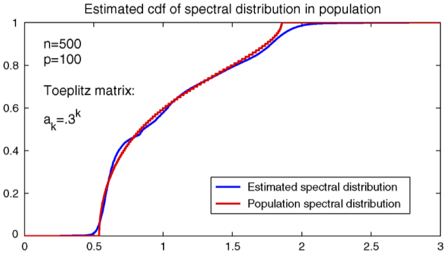

Estimated CDF Eigenvalues Population Covariance
n=500
p=100
50 eigenvalues=1
50 eigenvalues=2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5
3
3.5

(c)

CASE2


Empirical and Population spectral distributions
n=500
p=100
Toeplitz
matrix:
$a_k = .3^k$
Empirical Spectral Distribution
Population Spectral Distribution
0
0.2
0.4
0.6
0.8
1
0
0.5
1
1.5
2
2.5
3

(b)


Estimated cdf of spectral distribution in population
n=500
p=100
Toeplitz matrix:
$a_k = .3^k$
Estimated spectral distribution
Population spectral distribution
0
0.2
0.4
0.6
0.8
1
0
0.5
1
1.5
2
2.5
3

(c)

References I

[1] El Karoui, N. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics 36*, 6 (2008), 2757–2790.

[2] Ledoit, O., and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis 88*, 2 (2004), 365–411.

[3] Paul, D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica* (2007), 1617–1642.