Random Matrix

Spectrum estimation for large dimensional covariance matrices using random matrix theory

Yuheng Ma Xinyi Wu Keyu Zhang

August 23, 2020

Content

Background introduction

For i.i.d random vector $X_1 \ldots, X_n \in R^p$,and that the covariance of X_i is Σ_p. We call X the data matrix whose rows are the X_i's. How to estimate the eigenvalues of the population covariance matrix?
When p is fixed, n goes to ∞:
(Anderson,1963):Eigenvalues of the sample covariance matrix $S_p = (X - \overline{X})'(X - \overline{X})/(n-1)$ are good estimators of eigenvalues of Σ_p
when comes to to “large n, large p”:
For example, we consider p/n $\rightarrow$ r, $\Sigma_p = Id_p$, if X_i i.i.d and have a forth moment, we know that

$$l_1 \rightarrow (1+\sqrt{r})^2 a.s.$$

Random Matrix: From vectors to measures

Suppose we have a vector $(y_1, \dots, y_p)$ in R_p. We can associate to it the following measure:

$$dG_p(x) = \frac{1}{p} \sum_{i=1}^{p} \delta_{y_i}(x)$$

We denote by H_p the spectral distribution of the population covariance matrix Σ_p, i.e the measure associated with the vector of eigenvalues of Σ_p.
Similarly, we denote by F_p the measure associated with the eigenvalues of the sample covariance matrix S_p.

Random Matrix: Convergence

Notion of convergence: Weak convergence of probability measures.
The **Stieltjes transform** of a measure G on R is defined as

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

Properties:
If G_n is a sequence of probability measures and $m_{Gn}(z)$ has a (pointwise) limit m(z) for all $z \in C^+$, then there exists a probability measure G with Stieltjes transform $m_G = m$ if and only if $\lim_{y\to\infty} -iym(iy) = 1$. If it is the case, G_n converges weakly to G.

Random Matrix: the Marčenko-Pastur equation

X:n×p data matrix. $S_p = X'X/n$ H_p: the population spectral distribution
m_{F_p} :the Stieltjes transform of the spectral distribution, F_p, of S_p.
$v_{F_p}(z) = (1 - p/n)\frac{-1}{z} + \frac{p}{n} m_{F_p}(z)$,which is actually the Stieltjes transform of the spectral distribution XX'/n

Theorem

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

Random Matrix: the Marčenko-Pastur equation

Then, when $p, n \to$, *and* $p/n \to r$, $r \in (0, \infty)$,
1.$v_{F_p}(z) \to v_\infty(z)$ *a.s, where* $v_\infty(z)$ *is a deterministic function.*
2. $v_\infty(z)$ *satisfies the equation*

$$-\frac{1}{v_\infty(z)} = z - r\int \frac{\lambda dH_\infty(\lambda)}{1+\lambda v_\infty(z)}, \forall z \in C^+$$

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

Formulation of the estimation problem

Goal:
estimate the population eigenvalues
Strategy:
1. use the Marčenko-Pastur equation to estimate the measure $H_{\infty}.(\hat{H})$
2. estimate λ_i as the i-th quantile of our estimated distribution $\hat{H}_{\infty}$.

Fact:
since we are considering fixed distribution asymptotics, our estimate of H_{∞} will serve as our estimate of H_p, so $\hat{H}_{\infty} = \hat{H}_p$.

Formulation of the estimation problem

$$-\frac{1}{v_\infty(z)} = z - r\int \frac{\lambda dH_\infty(\lambda)}{1+\lambda v_\infty(z)}, \forall z \in C^+$$

Step 1:Estiamte H_∞ from F_p.

Compute eigenvalue of $S_p \to F_p \to v_{F_p}(z) \xrightarrow{\{z_j\}_{j=1}^{J_n}} \{v_{F_p}(z_j)\}_{j=1}^{J_n}$

$$\hat{H}_\infty = \arg\min_H L(\{\frac{1}{v_{F_p}(z_j)} + z_j - \frac{p}{n}\int \frac{\lambda dH(\lambda)}{1+\lambda v_{F_p}(z_j)}\}_{j=1}^{J_n})$$

Discretization

Step 2:Discretization.
Naturally, dH can be simply approximated by a weighted sum of point masses:

$$dH(x) \approx \sum_{k=i}^{K} w_k \delta_{t_k}(x)$$

where $\{t_k\}_{k=1}^{K}$ is a grid of points, chosen by us, and w_k's are weights, which satisfies

$$\sum_{k=1}^{K} w_k = 1, w_k \geq 0$$

Discretization

Hence finding a measure that approximately satisfies Equation (M-P) is equivalent to finding a set of weights $\{w_k\}_{k=1}^{K}$,for which we have

$$-\frac{1}{v_{F_\infty}(z_j)} \approx z_j - \frac{p}{n}\sum_{k=1}^{K}\frac{w_k t_k}{1+t_k v_{F_\infty}(z_j)}, \forall j$$

Replace v_∞ by v_{F_p}. Our problem is thus to find $\{w_k\}_{k=1}^{K}$ such taht,

$$-\frac{1}{v_{F_p}(z_j)} \approx z_j - \frac{p}{n}\sum_{k=1}^{K}\frac{w_k t_k}{1+t_k v_{F_p}(z_j)}, \forall j$$

Convex Optimization formulation

Approximation errors

$$e_j = \frac{1}{v_{F_p}(z_j)} + z_j - \frac{p}{n}\sum_{k=1}^{K} \frac{w_k t_k}{1 + t_k v_{F_p}(z_j)}$$

Step 3:Formulating our problem as a convex optimization problem.
"L_∞" version: Find w_k's to

$$Minimize \max_{j=1,\ldots,J_n} max\{|Re(e_j)|, |Im(e_j)|\}$$

Convex Optimization formulation

The “translation” of the problem into a convex optimization problem is

$$\min_{(w_1,\dots,w_K,u)} u$$

$$\forall j, -u \leq Re(e_j) \leq u$$

$$\forall j, -u \leq Im(e_j) \leq u$$

$$\sum_{k=1}^{K} w_k = 1$$

$$w_k \geq 0, \forall j$$

Consistency

Our estimate is consistent in L^∞ sense. The general result is as follows.

Theorem

Under setups in previous thm, $H_p \Rightarrow H_\infty$ *and* $\frac{p}{n} \to \gamma$. $J_1 \cdots \in \mathbb{Z}$ *with limit* ∞ *and* $z_1, \cdots \in \mathbb{C}^+$ *bounded and convergent,* $\hat{H}_p$ *is the solution to*

$$\widehat{H}_p = \arg\min_H \max_{j \le J_n} \left| \frac{1}{v_{F_p}(z_j)} + z_j - \frac{p}{n} \int \frac{\lambda dH(\lambda)}{1 + \lambda v_{F_p}(z_j)} \right|$$

Then

$$\hat{H}_p \Rightarrow H_\infty \quad a.s.$$

Same thing holds for our estimation which is made over a mixture of diracs.

Idea of Proof

$$\Delta = \frac{1}{v_{F_p}(z_j)} + z_j - \frac{p}{n} \int \frac{\lambda dH(\lambda)}{1 + \lambda v_{F_p}(z_j)}$$

Recall Thm1, when $H_p \Rightarrow H_\infty$

- $v_{F_p} \to v_\infty$ a.s. (Thm 1)
- $v_{F_p} \to v_\infty$ uniformly (Prop 1)
- $Im(v_{F_p})$ bounded away from 0 (Prop 2)
- $|\Delta| \downarrow 0$ (Prop 3)
- $\hat{H}_p \Rightarrow H_\infty$ (Lemma)
- $\hat{H}_p$ is fine to be sum of atoms (Cor)

Lemma

Find convergent $\hat{H}_p$ over constraint over convergent sequence of z_i.

Lemma

With $\epsilon_n \downarrow 0$ *and*

$$\forall j \leq J_n \quad \left| \frac{1}{v_{F_p}(z_j)} + z_j - \frac{p}{n} \int \frac{\lambda d\tilde{H}_p(\lambda)}{1 + \lambda v_{F_p}(z_j)} \right| < \varepsilon_n$$

Also, $v_{F_p}(z_j) \to v_\infty(z_j)$ *and both analytic as well as bounded away from reals, then we have convergence of* $\tilde{H}_p(\lambda)$ *to* H_∞.

Proposition

Proposition

Stieltjes transform S_H is Lipschitz by $\frac{1}{u_{min}^2}$ on $\mathbb{C}^+ \cap \{Im(z) > u_{min}\}$

Proposition

Population spectral distribution H_p has a limit H_∞, where all spectra uniformly bounded, then for a ball in $\mathbb{C}^+$,

$$\exists N \quad n > N \quad \Rightarrow \quad \inf_{n, z \in B(z_0, r)} \operatorname{Im}\left(v_{F_p}(z)\right) = \delta > 0$$

where v_{F_p} is the Stieltjes transform of F_p.

Proposition

Proposition

If $\exists$ *N s.t. any n>N,* $|v_{F_p}(z) - v_\infty(z)| < \epsilon$*, and* $|Im(v_\infty(z))| > u_{min}$*, then for* $\epsilon < \frac{u_{min}}{2}, \exists N' \in \mathbb{N}, \forall z \in B\left(z_0, r\right), \forall n > N'$.

$$\left| \frac{1}{v_{F_p}(z)} + z - \frac{p}{n} \int \frac{\lambda dH_\infty(\lambda)}{1 + \lambda v_{F_p}(z)} \right| < 2\varepsilon \frac{1 + 2\gamma}{u_{\min}^2}$$

Consistency of Proposed Algorithm

Only need to show Proposition 3 holds for sum of atoms.

Corollary

Restrict $\hat{H}_p$ over measures which are sums of atoms, the locations of which are restricted to belong to a grid (depending on n) whose step size is going to 0 as $n \to \infty$. Then everything holds.

Simulations

Case 1:Identity covariance:Σ=Id

Case 2:"Two spike" spectrum: Σ has d/2 eigenvalues equal to 1 and d/2 equal to 2

Case 2:covariance $\Sigma = T$ where $T_{i,j} = 0.3^{|i-j|}$is a d×d Toeplitz matrix

eigenvalues of sample covarience matrix

(a) Case1 (b) Case2 (c) Case3

Case1

(d) n=500,p=100

(e) n=500,p=250

Case1

(f) n=500,p=500

(g) n=500,p=1000

Case2

(h) n=500,p=100

(i) n=500,p=250

Case2

(j) n=500,p=500

(k) n=500,p=1000

Case3

(l) n=500,p=100

(m) n=500,p=250

Case3

(n) (o)

n=500,p=500 n=500,p=1000

Conclusion

Thanks for watching