Applications of random matrix theory to principal component analysis(PCA)

Jun Yin

IAS, UW-Madison

IAS, April-2014

Joint work with A. Knowles and H. T Yau.

Basic picture:

Let H be a Wigner (symmetric) random matrix:

$$H = (H_{ij})_{1 \leqslant i,j \leqslant N}, \quad H = H^*, \quad H_{ij} = N^{-1/2} h_{ij}$$

is a random matrix, whose upper right entries h_{ij}'s are independent random variables with mean 0 and variance 1.

$$\mathbb{E}\, H_{ij} = 0, \quad \mathbb{E}\,|H_{ij}|^2 = \frac{1}{N}, \quad 1 \leqslant i, j \leqslant N$$

Let $A = A_{N \times N}$ be a (full rank) deterministic symmetric matrix. Most of the e.values of A are O(1).

What can we say about

$$H + A \quad ?$$

[Knowles and Y, 2011-12]: on the rank $A = O(1)$ case.

Example: $A = \sum_{k=1}^{N} d_k \mathbf{v}_k \mathbf{v}_k^*$, where $d_1 = 10$, $d_k = 1$ for $2 \leqslant k \leqslant N/2$ and other $d_k = 2$.

Some basic questions:

- Limiting spectral distributions: ρ_{HA} Voiculescu 1986, Speicher 1994 (Free probablity)

- Local density or rigidity.

$$|\lambda_k - \gamma_k| \leqslant N^{\varepsilon} \left|\gamma_k - \gamma_{k\pm 1}\right|, \qquad \int_{-\infty}^{\gamma_k} \rho_{HA} = k/N$$

  holds with $1 - N^{-D}$ for any small $\varepsilon > 0$ and $D > 0$.

- Delocalization of e.vectors (in any direction). Let $\mathbf{u}_k$ be ℓ^2 normalized eigenvectors of $H + A$, $(1 \leqslant k \leqslant N)$, and $\mathbf{w}$ be deterministic vector

$$\max_k \langle \mathbf{u}_k, \mathbf{w} \rangle^2 \leqslant N^{-1+\varepsilon}$$

  holds with $1 - N^{-D}$ for any small $\varepsilon > 0$ and $D > 0$.

Some basic questions:

- Behaviors of outlier.
- Behaviors of the e.vector of outlier.
- Joint distribution of the largest k non-outlier eigenvalues.
- k-point correlation functions of $H + A$ in bulk.

Wigner matrix has two important properties:

1. Independent entries.

2. Isotropic rows/columns (without diagonal entries): let $\mathbf{h}_k = \{H_{kl}\}_{l \neq k}$ and $\mathbf{w} \in \mathbb{R}^{N-1}$ be deterministic vector, then

$$\langle \mathbf{h}_k, \mathbf{w} \rangle \sim \mathcal{N}(0, N^{-1}\|\mathbf{w}\|_2^2)$$

Why is 2 important, think about a growing matrix

Clearly $H + A$ does not have the second property.

Two exceptional cases:

- A is diagonal
- H is GOE.

Sample covariance matrix

Let XX^* be a sample covariance matrix:

$$X = (X_{ij})_{1 \leqslant i \leqslant M', 1 \leqslant j \leqslant N}, \quad X_{ij} = (M'N)^{-1/4} x_{ij}$$

is a random matrix, where x_{ij}'s are independent random variables with mean 0 and variance 1.

Let $T = T_{M \times M'}$ $(M' \geqslant M)$ be a deterministic matrix, $TT^* - I$ has full rank and most of the e.values of TT^* are $O(1)$.

What can we say about

$$TXXT^* \quad ?$$

Furthermore, let $\mathbf{e}$ be $M^{-1/2}(1, 1, \cdots, 1, 1)$. How about

$$TX(\mathbf{1} - \mathbf{ee}^*)XT^*$$

Real life question:

Let $\mathbf{y} = (y(1), y(2), \cdots, y(M))$ be some random vector with unknown distribution. For example: price changes of M stocks.

How can one get the covariance matrix:

$$\left\{ \text{Cov}(y(i),\ y(j)) \right\}_{i,j=1}^{M}$$

Stat 101: Mesure $\mathbf{y}$ N times independently: $\mathbf{y}_1$, $\mathbf{y}_2$, $\cdots, \mathbf{y}_N$

$$\text{Cov}(y(i),\ y(j)) = \lim_{N \to \infty} \frac{1}{N-1} \sum_{\alpha=1}^{N} (y_\alpha(i) - \bar{y}(i))\,(y_\alpha(j) - \bar{y}(j))$$

where

$$\bar{y}(i) = \frac{1}{N} \sum_{\alpha=1}^{N} y_\alpha(i)$$

Model: we assume that $\mathbf{y}$ is linear mix of some fundamental independent random variable $\mathbf{x} = (x(1), x(2), \cdots, x(M'))$

$$\mathbf{y} = T\mathbf{x}, \quad T_{M \times M'}$$

If you measure $\mathbf{y}$ N times then

$$Y = (\mathbf{y}_1, \mathbf{y}_2 \cdots, \mathbf{y}_N) = TX = T(\mathbf{x}_1, \mathbf{x}_2 \cdots, \mathbf{x}_N)$$

We know

$$\mathrm{Cov}(y(i),\ y(j)) = \lim_{N\to\infty} \frac{1}{N-1} TX(1 - \mathbf{e}\mathbf{e}^*)XT^*$$

For example: Let $\mathbf{v}$ is a fixed vector, $\tilde{\mathbf{x}}$ is random vector and ξ is random variable, and they are independent.

$$\mathbf{y} = \tilde{\mathbf{x}} + \xi\mathbf{v}$$

Then

$$\mathbf{y} = (I, \mathbf{v})\binom{\tilde{\mathbf{x}}}{\xi} = T\mathbf{x}$$

Model:

$$\mathbf{y} = T\mathbf{x}, \quad T_{M\times M'}, \quad \mathbf{x} = (x(1), x(2), \cdots, x(M'))$$

$$\text{Cov}(y(i),\ y(j)) = \lim_{N\to\infty} \frac{1}{N-1} TX(\mathbf{1} - \mathbf{e}\mathbf{e}^*)XT^*, \quad (*)$$

Without loss of generality, we assume that $\mathbb{E}x(i) = 0$ (by defining $x'(i) = x(i) - \mathbb{E}x(i)$). And we assume $\mathbb{E}|x(i)|^2 = 1$ (by rescaling T). With this setting

$$\text{Cov}(y(i),\ y(j)) = (TT^*)_{ij}$$

Recall the previous example:

$$\mathbf{y} = \widetilde{\mathbf{x}} + \xi\mathbf{v}, \quad \mathbb{E}|\widetilde{x}(i)|^2 = 1, \quad \mathbb{E}|\xi|^2 = 1$$

Here $\|\mathbf{v}\|_2 \gg 1$,

$$\mathbf{y} = (I, \mathbf{v})\binom{\widetilde{\mathbf{x}}}{\xi} = T\mathbf{x}, \quad TT^* = I + \mathbf{v}\mathbf{v}^*$$

As we can see TT^* has a large e. value $(1 + \|\mathbf{v}\|_2^2)$ with e. vector $\mathbf{v}$, and they are related to "signals"

Though

$$\text{Cov}(y(i),\ y(j)) = (TT^*)_{ij} = \lim_{N\to\infty} \frac{1}{N-1} TX(\mathbf{1} - \mathbf{ee}^*)XT^*$$

in most case it does not work very well, since N needs to be very large (like $N \gg M$).

The basic idea is estimating the principal component (large e.values and their e.vectors) of the matrix

$$TT^*$$

with those of the matrix

$$TX(\mathbf{1} - \mathbf{ee}^*)XT^*$$

or just

$$TXXT^*$$

PCA model:

- $TT^* = \sum d_\beta \mathbf{v}_\beta \mathbf{v}_\beta^*$, $\quad |\text{signal}| = O(1)$

$$TT^* = \sum_{\beta \in \text{signal}} d_\beta \mathbf{v}_\beta \mathbf{v}_\beta^* + \sum_{\beta \in \text{noise}} d_\beta \mathbf{v}_\beta \mathbf{v}_\beta^*,$$

- $T_{M \times M'}$, $\quad M' = M + O(1)$

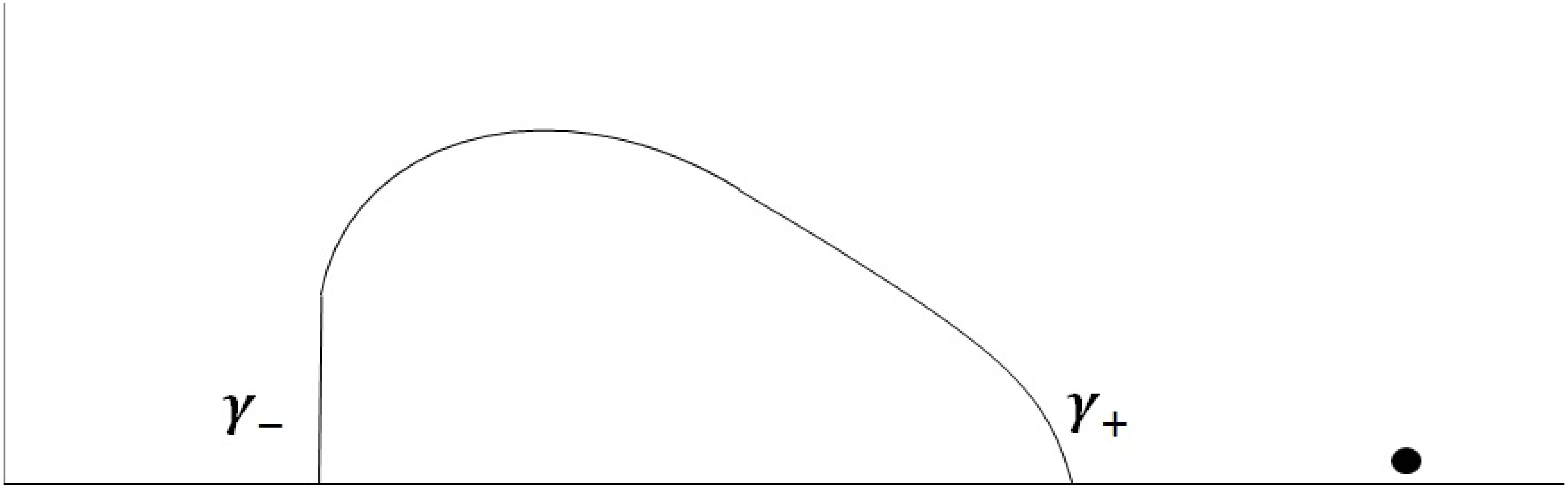
- Noise are comparable

$$c \leqslant d_\beta \leqslant C, \qquad d_\beta \in \text{noise}$$

- $\log N \sim \log M$.

- d_β and $\mathbf{v}_\beta$ could depend on N or M

Basic picture ($\lambda_{\text{noise}} = 1$ and only one λ_{signal} case)



Let $d := (N/M)^{1/2}\left(\lambda_{\text{signal}} - \lambda_{\text{noise}}\right)$. Outlier appears when $d > 1$ and outlier μ satisfies:

$$\mu = \gamma_+ - 2 + d + d^{-1} + \text{error}.$$

Detection of $\mathbf{v}_{\text{signal}}$.

Let $\mathbf{u}$ be the eigenvector of μ (outlier), then for **any** fixed normalized $\mathbf{w}$, we have

$$(\mathbf{w},\mathbf{u})^2 = f_\mu(\mathbf{w},\mathbf{v}_{\text{signal}})^2 + \text{error}$$

Distribution of $\mathbf{u}$?

1. $\theta(\mathbf{v}_{\text{signal}},\mathbf{u}) = \arccos\sqrt{f_\mu} + \text{error}$

2. Delocalization in any direction orthogonal to $\mathbf{v}_{\text{signal}}$, i.e., if we have $(\mathbf{w},\mathbf{v}_{\text{signal}}) = 0$, then $(\mathbf{w},\mathbf{u}) \leqslant M^{-1/2+\varepsilon}$.

Briefly speaking, $\mathbf{u} - (\mathbf{u},\mathbf{v}_{\text{signal}})\mathbf{v}_{\text{signal}}$ is random and isotropic.

Two outliers cases: see graph.

Application of delocalization

Assume we know $\mathbf{v}_{\text{signal}} \in \mathbb{R}^M$ only has $\widetilde{M}$ non-zero components, $\widetilde{M} \ll M$ and

$$\mathbf{v}_{\text{signal}}(i) \sim (\widetilde{M})^{-1/2}, \quad if \quad \mathbf{v}_{\text{signal}}(i) \neq 0$$

Then

1. if $\mathbf{v}_{\text{signal}}(i) = 0$, delocalization property shows $|\mathbf{u}(i)| \leqslant M^{-1/2+\varepsilon}$

2. if $\mathbf{v}_{\text{signal}}(i) \neq 0$, parallel property shows $|\mathbf{u}(i)| \geqslant (\widetilde{M})^{-1/2-\varepsilon}$

Using this method, we can know that which components of $\mathbf{v}_{\text{signal}}$ are non-zero.

Some previous results: Eigenvalues:

$TXX^*T, \quad T_{M\times M'}, \quad TT^* = \sum_{\beta\in\text{signal}} d_\beta \mathbf{v}_\beta \mathbf{v}_\beta^* + \sum_{\beta\in\text{noise}} d_\beta \mathbf{v}_\beta \mathbf{v}_\beta^*$

Baik and Silverstein (2006): $p = q$, $\lambda_{\text{noise}} = 1$, for fixed d, they obtain the limit of μ.

Bai and Yao (2008): $p = q$, $\lambda_{\text{noise}} = 1$, T is of the form (where A is $O(1)\times O(1)$ matrix) $T = \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix}$ for fixed d, they obtain the CLT of μ.

Bai and Yao (2008): $p = q$, T is symmetric matrix of the form (where A is $O(1)\times O(1)$ matrix) $T = \begin{pmatrix} A & 0 \\ 0 & \widetilde{T} \end{pmatrix}$, for fixed d, they obtain the limit of μ.

Nadler (2008): The spiked covariance model. $T = \begin{pmatrix} I & \lambda \mathbf{e}_1 \end{pmatrix}$

Eigenvector:

Then for any fixed $\mathbf{w}$, we have $(\mathbf{w}, \mathbf{u})^2 = f_d(\mathbf{w}, \mathbf{v}_{\text{signal}})^2 + \text{error}$

Paul (2007): $p = q$, $\lambda_{\text{noise}} = 1$, T is diagonal and X_{ij} is Gaussian.

Shi (2013): $p = q$, $\lambda_{\text{noise}} = 1$, T is diagonal.

Benaych-Georges, Nadakuditi (2010), $p = q$, $\lambda_{\text{noise}} = 1$, T is random symmetric matrix, T is independent of X, either X_{ij} is Gaussian or T is isotropic.

Benaych-Georges, Nadakuditi (2012): $T = \begin{pmatrix} I & \lambda \mathbf{v} \end{pmatrix}$ with random isotropic $\mathbf{v}$.

Results: limit of $(\mathbf{u}, \mathbf{v}_{\text{signal}})^2$, except the first one.

Main results

1. Rigidity of e.values (including outliers): (up to N^{ε} factor)

$$\lambda_i - \gamma_i = \text{error}$$

2. Delocalization of e.vectors of non-outliers.

3. Direction of the e.vectors of outliers. $\mathbf{u} - (\mathbf{u}, \mathbf{v}_{\text{signal}})\mathbf{v}_{signal}$ is random and isotropic.

4. Some eigenvector information can be detected even if $d = 1 - o(1)$

5. TW distribution of the largest k non-outliers. El Karoui (2007): Gaussian case.

6. Isotropic law of $(H + A)$ or $TXXT^*$.

7. Bulk universality with 4 moment match (for $(H + A)$ or $TXXT^*$).

Strategy:

With $V_{M\times M}$, $U_{M'\times M'}$ and diagonal $D_{M\times M}$

$$T = VD(I_M, 0)U', \quad TT^* = \sum_{\beta\in\text{signal}} d_\beta \mathbf{v}_\beta \mathbf{v}_\beta^* + \sum_{\beta\in\text{noise}} d_\beta \mathbf{v}_\beta \mathbf{v}_\beta^*$$

Define

$$S = S^*, \quad SS^* = \sum_{\beta\in\text{signal}} 1 \mathbf{v}_\beta \mathbf{v}_\beta^* + \sum_{\beta\in\text{noise}} d_\beta \mathbf{v}_\beta \mathbf{v}_\beta^*$$

Note: S has no signal. Represent

$$G := \Big(TX(1-ee^*)X^*T^* - z\Big)^{-1},$$

with

$$G_S, \quad G_S X, \quad XG_S X, \quad etc$$

where $G_S = \Big(SXX^*S - z\Big)^{-1}$.

Question:

Let A be a matrix with only one non-zero entry and $X' = X + A$. Then

$$(XX^* - z)^{-1} \to (X'X'^* - z)^{-1}, \quad ?$$

Isotropic law:

Wigner: Let H be a Wigner matrix, $G = (H-z)^{-1}$, then for any fixed $\mathbf{w}$ and $\mathbf{v}$, we have

$$(\mathbf{w}, (H-z)^{-1}\mathbf{v}) = m(z) + O((N\eta)^{-1}(\log N)^C), \quad \eta = \operatorname{Im} z$$

and $m(z) = \int \rho_{sc}(x)(x-z)^{-1}\mathrm{d}z$. Knowles and Y (2011).

PCA: For fixed $\mathbf{w}$ and $\mathbf{v}$, what are the behaviors of

$$(\mathbf{w}, G_S\mathbf{v}), \quad (\mathbf{w}, G_S X\mathbf{v}), \quad (\mathbf{w}, X^*G_S X\mathbf{v}), \quad G_S = (SXX^*S - z)^{-1}$$

Bloemendal, Erdos, Knowles, Yau and Y (2013): $S = I_{M\times M}$ case.

Isotropic law for general S or general A

$$(SXX^*S^* - z)^{-1}, \quad (H + A)^{-1}$$

Knowles and Y (2014):

Let $A = UDU^*$ with $D = \text{diag}(d_1, d_2, \cdots, d_N)$. Here $|d_i| \leqslant C$. Define

$$m_i = (d_i - z + m)^{-1}, \quad m := \frac{1}{N}\sum_j m_j$$

Then for fixed $\mathbf{w}$ and $\mathbf{v} \in \mathbb{R}^N$,

$$(\mathbf{w}, (H-z)^{-1}\mathbf{v}) = (\mathbf{w}, (A - z + m)^{-1}\mathbf{v}) + error$$

Based on this result: rigidity, delocalization, TW law. (Capitaine, Peche 2014: GOE+A)

Basic idea of proving the isotropic law of $H + A$.

1. The isotropic law of $GOE + A$. (polynomialization method) Bloemendal, Erdos, Knowles, Yau and Y (2013)

2. Compare $(H + A)^{-1}$ with $(GOE + A)^{-1}$ with Newton method.

Let ν be the distribution density of H_{ij} and ν^G be the distribution density of the entries of GOE. Let H^t be the Wigner matrix whose entries having distributions

$$\nu^t = t\nu + (1 - t)\nu^G$$

Continuous bridge between H and GOE.

Recall

$$\mathbb{E}F(H) = \int_1^0 \partial_t \, \mathbb{E}F(H^t) \mathrm{d}t$$

Let $H^{t,k,l,0}$ be the Wigner matrix whose entries having the same distribution as H^t except that (k,l) entry of $H^{t,k,l0}$ has the distribution of ν.

Let $H^{t,k,l,1}$ be the Wigner matrix whose entries having the same distribution as H^t except that (k,l) entry of $H^{t,k,l,1}$ has the distribution of ν^G.

Then

$$\partial_t \, \mathbb{E}F(H^t) = \sum_{kl} \Big(\mathbb{E}F(H^{t,k,l,1}) - \mathbb{E}F(H^{t,k,l,0}) \Big)$$

Note: $H^{t,k,l,0}$ and $H^{t,k,l,1}$ are very close to H^t.

For example:

$$F_{i,j,p}(H) = \left[(H-z)^{-1}_{ij}\right]^{2p}$$

Goal: Create a self-consistent differential equation of

$$\left(\mathbb{E}F_{i,j,p}(H^t)\right)_{ij=1}^{N}$$

which is stable.

Thank you