Lecture 2. Random Matrix Theory and Phase Transitions of PCA

Yuan Yao

Hong Kong University of Science and Technology

February 26, 2020
Outline

Recall: Horn’s Parallel Analysis of PCA

Random Matrix Theory

Phase Transitions of PCA
How many components of PCA?

- Data matrix: $X = [x_1 | x_2 | \cdots | x_n] \in \mathbb{R}^{p \times n}$

- Centering data matrix: $Y = XH$ where

  $$H = I - \frac{1}{n} \mathbf{1} \cdot \mathbf{1}^T$$

- PCA is given by top left singular vectors of $Y = USV^T$ (called loading vectors) by projections to $\mathbb{R}^p$, $z_j = u_jY$

- MDS is given by top right singular vectors of $Y = USV^T$ as Euclidean embedding coordinates of $n$ sample points

- But how many components shall we keep?

Recall: Horn’s Parallel Analysis of PCA
Recall: Horn’s Parallel Analysis

Data matrix: \( X = [x_1 | x_2 | \cdots | x_n] \in \mathbb{R}^{p \times n} \)

\[
X = \begin{bmatrix}
X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\
X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
X_{p,1} & X_{p,2} & \cdots & X_{p,n}
\end{bmatrix}.
\]

Compute its principal eigenvalues \( \{ \hat{\lambda}_i \}_{i=1,\ldots,p} \)
Recall: Horn’s Parallel Analysis

- Randomly take $p$ permutations of $n$ numbers $\pi_1, \ldots, \pi_p \in S_n$ (usually $\pi_1$ is set as identity), noting that sample means are permutation invariant,

$$X^1 = \begin{bmatrix}
X_{1,\pi_1(1)} & X_{1,\pi_1(2)} & \cdots & X_{1,\pi_1(n)} \\
X_{2,\pi_2(1)} & X_{2,\pi_2(2)} & \cdots & X_{2,\pi_2(n)} \\
\vdots & \vdots & \ddots & \vdots \\
X_{p,\pi_p(1)} & X_{p,\pi_p(2)} & \cdots & X_{p,\pi_p(n)}
\end{bmatrix}.$$

- Compute its principal eigenvalues $\{\hat{\lambda}_i^1\}_{i=1}^p$.

- Repeat such procedure for $r$ times, we can get $r$ sets of principal eigenvalues. $\{\hat{\lambda}_i^k\}_{i=1}^p$ for $k = 1, \ldots, r$.
Recall: Horn’s Parallel Analysis (continued)

For each $i = 1$, define the $i$-th $p$-value as the percentage of random eigenvalues $\{\hat{\lambda}_i^k\}_{k=1,...,r}$ that exceed the $i$-th principal eigenvalue $\hat{\lambda}_i$ of the original data $X$,

$$pval_i = \frac{1}{r} \# \{\hat{\lambda}_i^k > \hat{\lambda}_i : k = 1, \ldots, r\}.$$

Setup a threshold $q$, e.g. $q = 0.05$, and only keep those principal eigenvalues $\hat{\lambda}_i$ such that $pval_i < q$. 

Recall: Horn’s Parallel Analysis of PCA
Example

Let’s look at an example of Parallel Analysis

- Matlab: papca.m
- Python:
How does it work?

- We are going to introduce an analysis based on Random Matrix Theory for *rank-one spike model*.

Recall: Horn's Parallel Analysis of PCA
How does it work?

- We are going to introduce an analysis based on Random Matrix Theory for \textit{rank-one spike model}.

- There is a \textbf{phase transition} in principal component analysis.

Recall: Horn's Parallel Analysis of PCA
How does it work?

▶ We are going to introduce an analysis based on Random Matrix Theory for \textit{rank-one spike model}.

▶ There is a \textbf{phase transition} in principal component analysis:
  - If the signal is strong, principal eigenvalues are beyond the random spectrum and principal components are correlated with signal.

Recall: Horn's Parallel Analysis of PCA.
How does it work?

- We are going to introduce an analysis based on Random Matrix Theory for rank-one spike model.

- There is a phase transition in principal component analysis:
  - If the signal is strong, principal eigenvalues are beyond the random spectrum and principal components are correlated with signal.
  - If the signal is weak, all eigenvalues in PCA are due to random noise.

Recall: Horn's Parallel Analysis of PCA
Outline

Recall: Horn’s Parallel Analysis of PCA

Random Matrix Theory

Phase Transitions of PCA
Let $x_i \sim \mathcal{N}(0, I_p)$ ($i = 1, \ldots, n$) and $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{p \times n}$.

The sample covariance matrix 

$$\hat{\Sigma}_n = \frac{1}{n}XX^T.$$ 

is called Wishart (random) matrix.

When both $n$ and $p$ grow at $\frac{p}{n} \to \gamma \neq 0$, the distribution of the eigenvalues of $\hat{\Sigma}_n$ follows the **Marčenko-Pastur (MP) Law**

$$\mu^{MP}(t) = \left(1 - \frac{1}{\gamma}\right) \delta(x)I(\gamma > 1) + \begin{cases} 
0 & t \notin [a, b], \\
\frac{\sqrt{(b-t)(t-a)}}{2\pi\gamma t}dt & t \in [a, b],
\end{cases}$$

where $a = (1 - \sqrt{\gamma})^2$, $b = (1 + \sqrt{\gamma})^2$. 

Random Matrix Theory

10
Illustration of MP Law

- If $\gamma \leq 1$, MP distribution has a support on $[a, b]$;

- if $\gamma > 1$, it has an additional point mass $1 - 1/\gamma$ at the origin.

Figure: Show by matlab: (a) Marčenko-Pastur distribution with $\gamma = 2$. (b) Marčenko-Pastur distribution with $\gamma = 0.5$. 
Outline

Recall: Horn’s Parallel Analysis of PCA

Random Matrix Theory

Phase Transitions of PCA
Consider the following rank-1 signal-noise model

\[ Y = X + \varepsilon, \]

where

- the signal lies in an one-dimensional subspace \( X = \alpha u \) with \( \alpha \sim \mathcal{N}(0, \sigma_X^2) \);

- the noise \( \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_p) \) is i.i.d. Gaussian.

Therefore \( Y \sim \mathcal{N}(0, \Sigma) \) where the limiting covariance matrix \( \Sigma \) is rank-one added by a sparse matrix:

\[ \Sigma = \sigma_X^2 uu^T + \sigma_\varepsilon^2 I_p. \]
When does PCA work?

▶ Can we recover signal direction $u$ from principal component analysis on noisy measurements $Y$?

▶ It depends on the signal noise ratio, defined as

$$SNR = R := \frac{\sigma_X^2}{\sigma_\varepsilon^2}.$$ 

For simplicity we assume that $\sigma_\varepsilon^2 = 1$ without loss of generality.
Consider the scenario

\[ \gamma = \lim_{p,n \to \infty} \frac{p}{n}. \]  

as in applications, one never has infinite amount of samples and dimensionality.

A fundamental result by I. Johnstone in 2006 shows a phase transition of PCA:
The primary (largest) eigenvalue of the sample covariance matrix satisfies

\[ \lambda_{\text{max}}(\hat{\Sigma}_n) \to \begin{cases} (1 + \sqrt{\gamma})^2 = b, & \sigma_X^2 \leq \sqrt{\gamma} \\ (1 + \sigma_X^2)(1 + \frac{\gamma}{\sigma_X^2}), & \sigma_X^2 > \sqrt{\gamma} \end{cases} \]  

The primary eigenvector (principal component) associated with the largest eigenvalue converges to

\[ |\langle u, v_{\text{max}} \rangle|^2 \to \begin{cases} 0, & \sigma_X^2 \leq \sqrt{\gamma} \\ \frac{1 - \frac{\gamma}{\sigma_X^2}}{1 + \frac{\gamma}{\sigma_X^2}}, & \sigma_X^2 > \sqrt{\gamma} \end{cases} \]
In other words,

- If the signal is strong $SNR = \sigma_X^2 > \sqrt{\gamma}$, the primary eigenvalue goes beyond the random spectrum (upper bound of MP distribution), and the primary eigenvector is correlated with signal (in a cone around the signal direction whose deviation angle goes to 0 as $\sigma_X^2/\gamma \to \infty$);

- If the signal is weak $SNR = \sigma_X^2 \leq \sqrt{\gamma}$, the primary eigenvalue is buried in the random spectrum, and the primary eigenvector is random of no correlation with the signal.
Proof in Sketch

Following the rank-1 model, consider random vectors $y_i \sim \mathcal{N}(0, \Sigma)$ ($i = 1, \ldots, n$), where $\Sigma = \sigma^2_x uu^T + \sigma^2_\varepsilon I_p$ and $u$ is an arbitrarily chosen unit vector ($\|u\|^2 = 1$) showing the signal direction.

The sample covariance matrix is $\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^T = \frac{1}{n} YY^T$ where $Y = [y_1, \ldots, y_n] \in \mathbb{R}^{p \times n}$. Suppose one of its eigenvalue is $\hat{\lambda}$ and the corresponding unit eigenvector is $\hat{v}$, so $\hat{\Sigma}_n \hat{v} = \lambda \hat{v}$.

First of all, we relate the $\hat{\lambda}$ to the MP distribution by the trick:

$$z_i = \Sigma^{-\frac{1}{2}} y_i \rightarrow Z_i \sim \mathcal{N}(0, I_p). \quad (4)$$

Then $S_n = \frac{1}{n} \sum_{i=1}^{n} z_i z_i^T = \frac{1}{n} ZZ^T$ ($Z = [z_1, \ldots, z_n]$) is a Wishart random matrix whose eigenvalues follow the Marčenko-Pastur distribution.
Proof in Sketch

Notice that

\[ \hat{\Sigma}_n = \frac{1}{n} Y Y^T = \Sigma^{1/2} \left( \frac{1}{n} Z Z^T \right) \Sigma^{1/2} = \Sigma^{1/2} S_n \Sigma^{1/2} \]

and \((\hat{\lambda}, \hat{v})\) is eigenvalue-eigenvector pair of matrix \(\hat{\Sigma}_n\). Therefore

\[ \Sigma^{1/2} S_n \Sigma^{1/2} \hat{v} = \hat{\lambda} \hat{v} \Rightarrow S_n \Sigma (\Sigma^{-1/2} \hat{v}) = \hat{\lambda} (\Sigma^{-1/2} \hat{v}) \]  \hspace{1cm} (5)

In other words, \(\hat{\lambda}\) and \(\Sigma^{-1/2} \hat{v}\) are the eigenvalue and eigenvector of matrix \(S_n \Sigma\).

Suppose \(c \Sigma^{-1/2} \hat{v} = v\) where the constant \(c\) makes \(v\) a unit eigenvector and thus satisfies,

\[ c^2 = c \hat{v}^T \hat{v} = v^T \Sigma v = v^T (\sigma_x^2 uu^T + \sigma_\varepsilon^2) v = \sigma_x^2 (u^T v)^2 + \sigma_\varepsilon^2 = R(u^T v)^2 + 1. \]  \hspace{1cm} (6)
Proof in Sketch

Now we have,

$$S_n \Sigma v = \hat{\lambda} v.$$  \hspace{1cm} (7)

Plugging in the expression of $\Sigma$, it gives

$$S_n (\sigma_X^2 uu^T + \sigma_\varepsilon^2 I_p) v = \hat{\lambda} v$$

Rearrange the term with $u$ to one side, we got

$$(\hat{\lambda} I_p - \sigma_\varepsilon^2 S_n) v = \sigma_X^2 S_n u(u^T v)$$

Assuming that $\hat{\lambda} I_p - \sigma_\varepsilon^2 S_n$ is invertible, then multiple its reversion at both sides of the equality, we get,

$$v = \sigma_X^2 \cdot (\hat{\lambda} I_p - \sigma_\varepsilon^2 S_n)^{-1} \cdot S_n u(u^T v).$$  \hspace{1cm} (8)
Multiply (8) by $u^T$ at both side,

$$u^T v = \sigma_X^2 \cdot u^T (\hat{\lambda} I_p - \sigma_\varepsilon^2 S_n)^{-1} S_n u \cdot (u^T v)$$

that is, if $u^T v \neq 0$,

$$1 = \sigma_X^2 \cdot u^T (\hat{\lambda} I_p - \sigma_\varepsilon^2 S_n)^{-1} S_n u$$

(9)
Primary Eigenvalue $\hat{\lambda}$

Assume that $S_n$ has the eigenvalue decomposition $S_n = W\hat{\Lambda}W^T$, where $\Lambda = \text{diag}(\lambda_i : i = 1, \ldots, p)$ and $WW^T = W^TW = I_p$ ($W = [w_1, \ldots, w_p] \in \mathbb{R}^{p \times p}$). Define $\alpha_i = w_i^Tu$ and $\alpha = (\alpha_i) \in \mathbb{R}^p$. Hence $u = \sum_{i=1}^{p} \alpha_i w_i = W^T\alpha$. Now (9) leads to

$$1 = \sigma_X^2 \cdot u^T[W(\hat{\lambda}I_p - \sigma_\varepsilon^2 \Lambda)^{-1}W^T][W\Lambda W^T]u = \sigma_X^2 \cdot \alpha^T(\hat{\lambda}I_p - \sigma_\varepsilon^2 \Lambda)^{-1}\Lambda\alpha$$

which is

$$1 = \sigma_X^2 \cdot \sum_{i=1}^{p} \frac{\lambda_i}{\hat{\lambda} - \sigma_\varepsilon^2 \lambda_i} \alpha_i^2$$

(10)

where $\sum_{i=1}^{p} \alpha_i^2 = 1$.

For large $p$, $\lambda_i \sim \mu^{MP}(\lambda_i)$ and the sum (10) can be approximated by

$$1 = \sigma_X^2 \cdot \frac{1}{p} \sum_{i=1}^{p} \frac{\lambda_i}{\hat{\lambda} - \sigma_\varepsilon^2 \lambda_i} \sim \sigma_X^2 \cdot \int_a^b \frac{t}{\hat{\lambda} - \sigma_\varepsilon^2 t} d\mu^{MP}(t)$$

(11)

where $\sigma_\varepsilon^2 = 1$ by assumption.
Primary Eigenvalue $\hat{\lambda}$

- Using the Stieltjes transform,

\[
1 = \sigma^2 X \cdot \int_a^b \frac{t}{\hat{\lambda} - t} \frac{\sqrt{(b - t)(t - a)}}{2\pi \gamma t} dt
\]

\[
= \frac{\sigma^2 X}{4\gamma} \left[2\hat{\lambda} - (a + b) - 2 \sqrt{|(\hat{\lambda} - a)(b - \hat{\lambda})|}\right].
\]  

(12)

- For $\hat{\lambda} \geq b$ and $R = \sigma^2 X \geq \sqrt{\gamma}$, we have

\[
1 = \frac{\sigma^2 X}{4\gamma} \left[2\hat{\lambda} - (a + b) - 2 \sqrt{(\hat{\lambda} - a)(\hat{\lambda} - b)}\right],
\]

\[
\Rightarrow \hat{\lambda} = \sigma^2 X + \frac{\gamma}{\sigma^2 X} + 1 + \gamma = (1 + \sigma^2 X)(1 + \frac{\gamma}{\sigma^2 X}).
\]
Here we observe the following phase transitions for primary eigenvalue:

- If \( \hat{\lambda} \in [a, b] \), then \( \hat{\Sigma}_n \) has its primary eigenvalue \( \hat{\lambda} \) within \( \text{supp}(\mu^{MP}) \), so it is undistinguishable from the noise.

- So \( \hat{\lambda} = b \) is the phase transition where PCA works to pop up signal rather than noise. Then plugging in \( \hat{\lambda} = b \) in (12), we get,

\[
1 = \sigma_X^2 \cdot \frac{1}{4\gamma} [2b - (a + b)] = \frac{\sigma_X^2}{\sqrt{\gamma}} \iff \sigma_X^2 = \sqrt{\gamma} = \sqrt{\frac{p}{n}} \quad (13)
\]

Hence, in order to make PCA works, we need to let the signal-noise-ratio \( R \geq \sqrt{\frac{p}{n}} \).
From Equation (8), we obtain

\[ 1 = v^T v = \sigma_X^4 \cdot v^T uu^T S_n (\lambda I_p - \sigma^2 \varepsilon S_n)^{-2} S_n uu^T v \]

\[ = \sigma_X^4 \cdot (|v^T u|)[u^T S_n (\lambda I_p - \sigma^2 \varepsilon S_n)^{-2} S_n u](|u^T v|) \]

which implies that

\[ |u^T v|^{-2} = \sigma_X^4 [u^T S_n (\lambda I_p - \sigma^2 \varepsilon S_n)^{-2} S_n u]. \quad (14) \]

Using the same trick as the equation (9), we reach the following Monte-Carlo integration

\[ |u^T v|^{-2} = \sigma_X^4 [u^T S_n (\lambda I_p - \sigma^2 \varepsilon S_n)^{-2} S_n u] \]

\[ \sim \sigma_X^4 \int_a^b \frac{t^2}{(\lambda - \sigma^2 \varepsilon t)^2} d\mu^{MP}(t) \quad (15) \]
For $\lambda \geq b$, from Stieltjes transform introduced later one can compute the integral as

$$|u^T v|^{-2} = \sigma_X^4 \cdot \int_a^b \frac{t^2}{(\lambda - \sigma_\varepsilon^2 t)^2} d\mu^{MP}(t)$$

$$= \frac{\sigma_X^4}{4\gamma} \left( -4\lambda + (a + b) + 2\sqrt{(\lambda - a)(\lambda - b)} + \ldots \right)$$

$$+ \frac{\lambda(2\lambda - (a + b))}{\sqrt{(\lambda - a)(\lambda - b)}}$$

from which it can be computed that (using $\hat{\lambda} = (1 + \sigma_X^2)(1 + \frac{\gamma}{\sigma_X^2})$ obtained above with $R = \sigma_X^2$)

$$|u^T v|^2 = \frac{1 - \frac{\gamma}{\sigma_X^4}}{1 + \gamma + \frac{2\gamma}{\sigma_X^2}}.$$
Primary Eigenvector $\hat{v}$

Now we can compute the inner product of $u$ and $\hat{v}$ that we are really interested in:

$$|u^T \hat{v}|^2 = \left( \frac{1}{c} u^T \Sigma^{1/2} v \right)^2 = \frac{1}{c^2} \left( (\Sigma^{1/2} u)^T v \right)^2$$

$$= \frac{1}{c^2} \left( (\sigma_X^2 uu^T + I_p)^{1/2} u \right)^T v$$

$$\ast = \frac{1}{c^2} \left( (\sqrt{1 + \sigma_X^2} u)^T v \right)^2$$

$$\ast\ast = \frac{(1 + \sigma_X^2)(u^T v)^2}{R (u^T v)^2 + 1}, \quad R = \sigma_X^2,$$

$$= \frac{1 + R - \frac{\gamma}{R} - \frac{\gamma R^2}{R}}{1 + R + \gamma + \frac{\gamma}{R}} = \frac{1 - \frac{\gamma R^2}{R}}{1 + \frac{\gamma}{R}}$$

where the equality (\ast) uses $\Sigma^{1/2} u = \sqrt{1 + R} u$, and the equality (\ast\ast) is due to the formula for $c^2$ (Equation (6) above). Note that this identity holds under the condition that $R \geq \sqrt{\gamma}$ to ensure the numerator above non-negative.
Stieltjes Transform

Define the Stieltjes Transformation of MP-density $\mu^{MP}$ to be

$$s(z) := \int_{\mathbb{R}} \frac{1}{t-z} d\mu^{MP}(t), \ z \in \mathbb{C}$$  \hspace{1cm} (16)

Lemma (Bai-Silverstein’2011, Lemma 3.11)

$$s(z) = \frac{(1-\gamma) - z + \sqrt{(z-1-\gamma)^2 - 4\gamma z}}{2\gamma z}.$$  \hspace{1cm} (17)
Lemma

1. \[
\int_a^b \frac{t}{\lambda - t} \mu^{MP}(t) dt = -\lambda s(\lambda) - 1;
\]

2. \[
\int_a^b \frac{t^2}{(\lambda - t)^2} \mu^{MP}(t) dt = \lambda^2 s'(\lambda) + 2\lambda s(\lambda) + 1
\]
Open Problems

- If one can estimate the noise models, such as the rank-1 model here, then we can use random matrix theory (universality) or by simulations to find the number of principal components.

- Such a random matrix theory can not fully explain why Horn’s Parallel Analysis, whose proof is open.

- In applications, noise models might not be homogeneous $\sigma_\epsilon^2 I_p$. How to deal with heterogeneous noise models is open (Wang-Owen’2015 attacked this problem).

- Distributive PCA can exploit random matrix theory to decide the number of samples in local clients (Fan-Wang et al. 2019).