Lecture 10: Distribution of Eigenvalues of Random Matrices

Nikola Zlatanov

Innopolis University

Advanced Statistics

10-th of April to 17-th of April, 2023

Simple Model to Understand the Complex Process

- In order to understand more of whether PCA works with estimated covariance matrices, we need to understand how the eigenvalues of the estimated covariance matrices behave.
- Now this is a complex task.
- The best thing when dealing with such complex tasks is to start with the simplest model in order to gain intuition.
- As we shell see, this turns out to be the right approach since using this approach we will understand many things about the eigenvalues of the estimated covariance matrix and thereby understand when PCA works and when it doesn't.

Symmetric Random Matrix

- One of the most simplest models of a symmetric random matrix (only symmetric matrices have real eigenvalues, i.e., $\lambda_i \in \mathbb{R}$, $\forall i$) is the following.
- Let $\boldsymbol{W}$ be a symmetric random matrix created as follows.
- Each elements on the main diagonal and above the main diagonal of $\boldsymbol{W}$ are obtained i.i.d. from some distribution with mean zero and unit variance.
- Next, the elements below the main diagonal of $\boldsymbol{W}$ are a symmetrical copy of the elements above the main diagonal, where the main diagonal acts as the axis of symmetry.
- An example:

$$W = \begin{bmatrix} 0.6 & 1.2 & -1.1 \\ 1.2 & -0.3 & 0.75 \\ -1.1 & 0.75 & 0.4 \end{bmatrix}$$

Asymptotic Distribution of the Eigenvalues of a Randomly Generated Symmetric Matrix With i.i.d. Zero-Mean Unit-Variance Entries

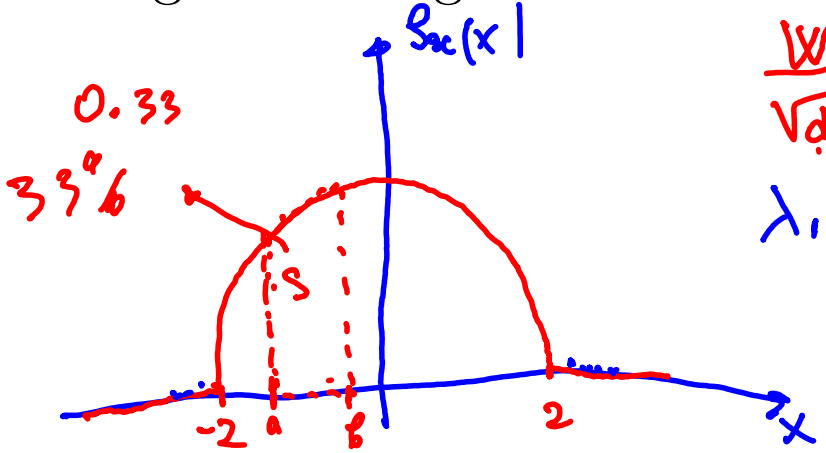
- Thm (Wigner's Semicircle Law): Let $\boldsymbol{W}$ be a $d \times d$ symmetric random matrix (generated as described on the previous slide). Then, as $d \to \infty$, the eigenvalues of $\boldsymbol{W}/\sqrt{d}$ converge to the following distribution

$$\rho_{\text{sc}}(x) = \frac{1}{2\pi}\sqrt{4 - x^2}, \quad \text{for } x \in [-2, 2], \tag{1}$$

  and $\rho_{\text{sc}}(x) = 0$, for $x \notin [-2, 2]$.
- We will not prove it!

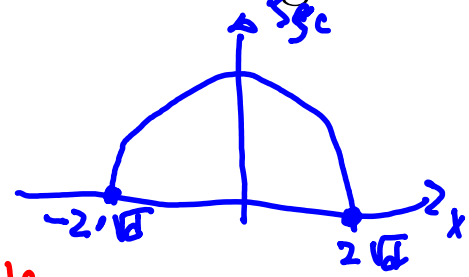
Asymptotic Distribution of the Eigenvalues of a Randomly Generated Symmetric Matrix With i.i.d. Zero-Mean Unit-Variance Entries

- Figure of Wigner's semicircle distribution and its meaning:



$\frac{W}{\sqrt{d}}$, when $d \to \infty$

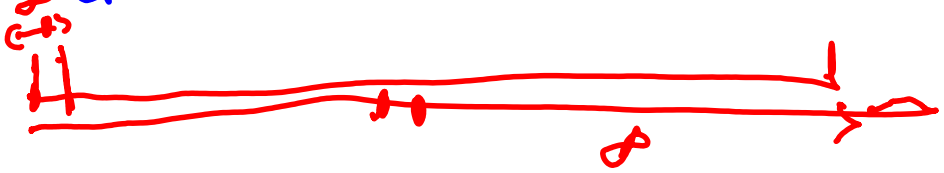
$\lambda_1, \lambda_2 \ldots \lambda_d$



$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$

$P(A) = 0$ if n_A is finite

$\lim_{d \to \infty} \frac{\varepsilon}{d} = 0$

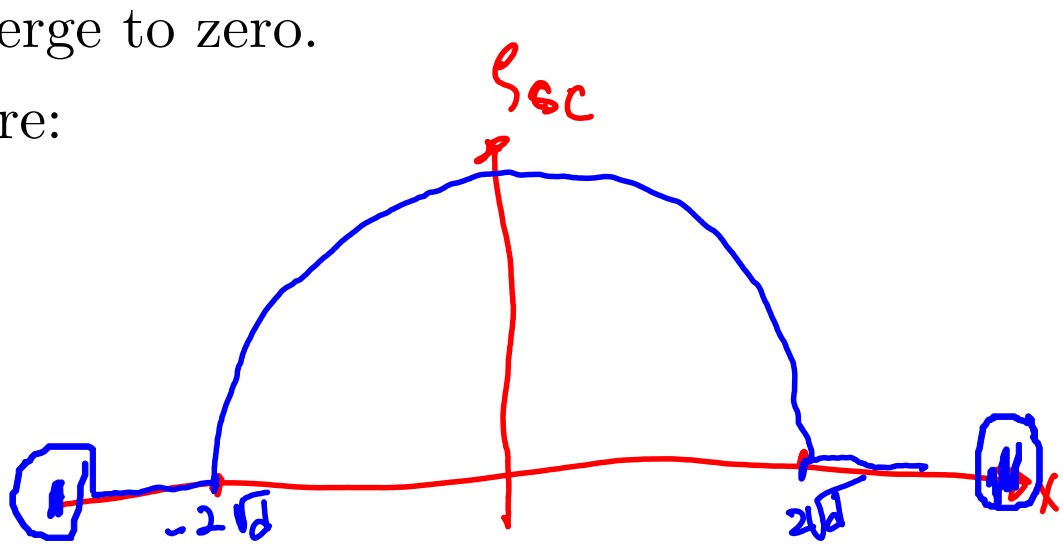


$\lim_{n \to \infty} \frac{n_A}{n} = 0$

$\lim_{n \to \infty} \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}} \approx 0$

Asymptotic Distribution of the Eigenvalues of a Randomly Generated Symmetric Matrix With i.i.d. Zero-Mean Unit-Variance Entries

- One of the most unexpected things that we observe from this result is that in Wigner's random matrices, as $d \to \infty$, the fraction of eigenvalues that are larger than $2\sqrt{d}$ and smaller than $-2\sqrt{d}$ converge to zero.
- Figure:

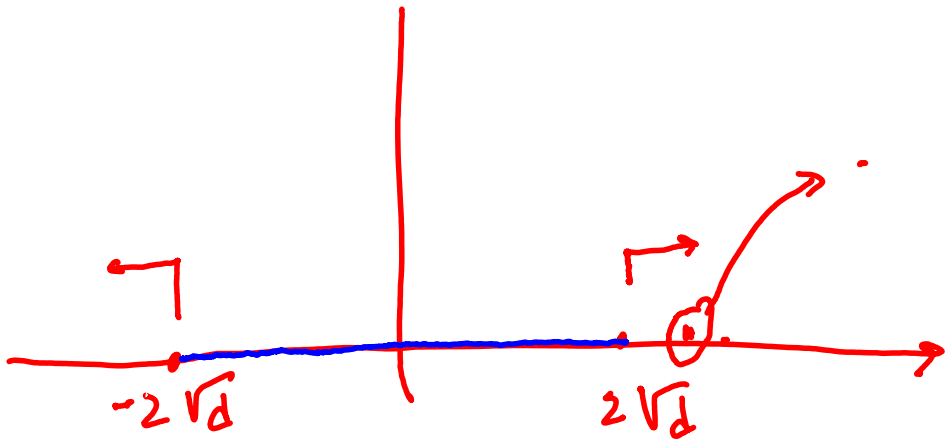


Asymptotic Distribution of the Eigenvalues of a Randomly Generated Symmetric Matrix With i.i.d. Zero-Mean Unit-Variance Entries

- Now consider a matrix $\boldsymbol{R} = \boldsymbol{S} + \boldsymbol{W}$, where $\boldsymbol{S}$ is our signal/information matrix and $\boldsymbol{W}$ is the noise Wigner's matrix that corrupts our signal $\boldsymbol{S}$.
- Assume that we only have access to $\boldsymbol{R}$ and do not know $\boldsymbol{W}$ nor $\boldsymbol{S}$.
- However, we would like to know what $\boldsymbol{S}$ is by observing only $\boldsymbol{R}$.
- Then, what this Wigner's Semicircle Law tells us is the following.
- As d increases, all the eigenvalues of the noise matrix are located within the interval $-2\sqrt{d}$ and $2\sqrt{d}$.
- Therefore, we should 'look' for our signal/information as the eigenvalues (and their corresponding eigenvectors) that are located outside the range $-2\sqrt{d}$ and $2\sqrt{d}$.

Asymptotic Distribution of the Eigenvalues of a Randomly Generated Symmetric Matrix With i.i.d. Zero-Mean Unit-Variance Entries

- Figure of the distribution of a signal matrix plus Wigner's noise matrix:



- What if the Estimated Covariance Matrix is generated from pure noisy data? This is the topic on our next slide.

Asymptotic Distribution of the Eigenvalues of the Estimated Covariance Matrix Generated From Zero-Mean Unit-Variance Gaussian Noise

Thm (Marchenko-Pastur Law): Let $\boldsymbol{X}_i \in \mathbb{R}^d$, for $i = 1, 2, ..., N$, be N i.i.d. vectors generated from $\mathcal{N}(0, \boldsymbol{I}_d)$. Let $\hat{\boldsymbol{\Sigma}}_N$ be the estimated covariance matrix, obtained as

$$\hat{\boldsymbol{\Sigma}}_N = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^T. \tag{2}$$

Let $d/N \to r$ as $N \to \infty$. Then, the distribution of the eigenvalues of $\hat{\boldsymbol{\Sigma}}_N$ converge to the following distribution

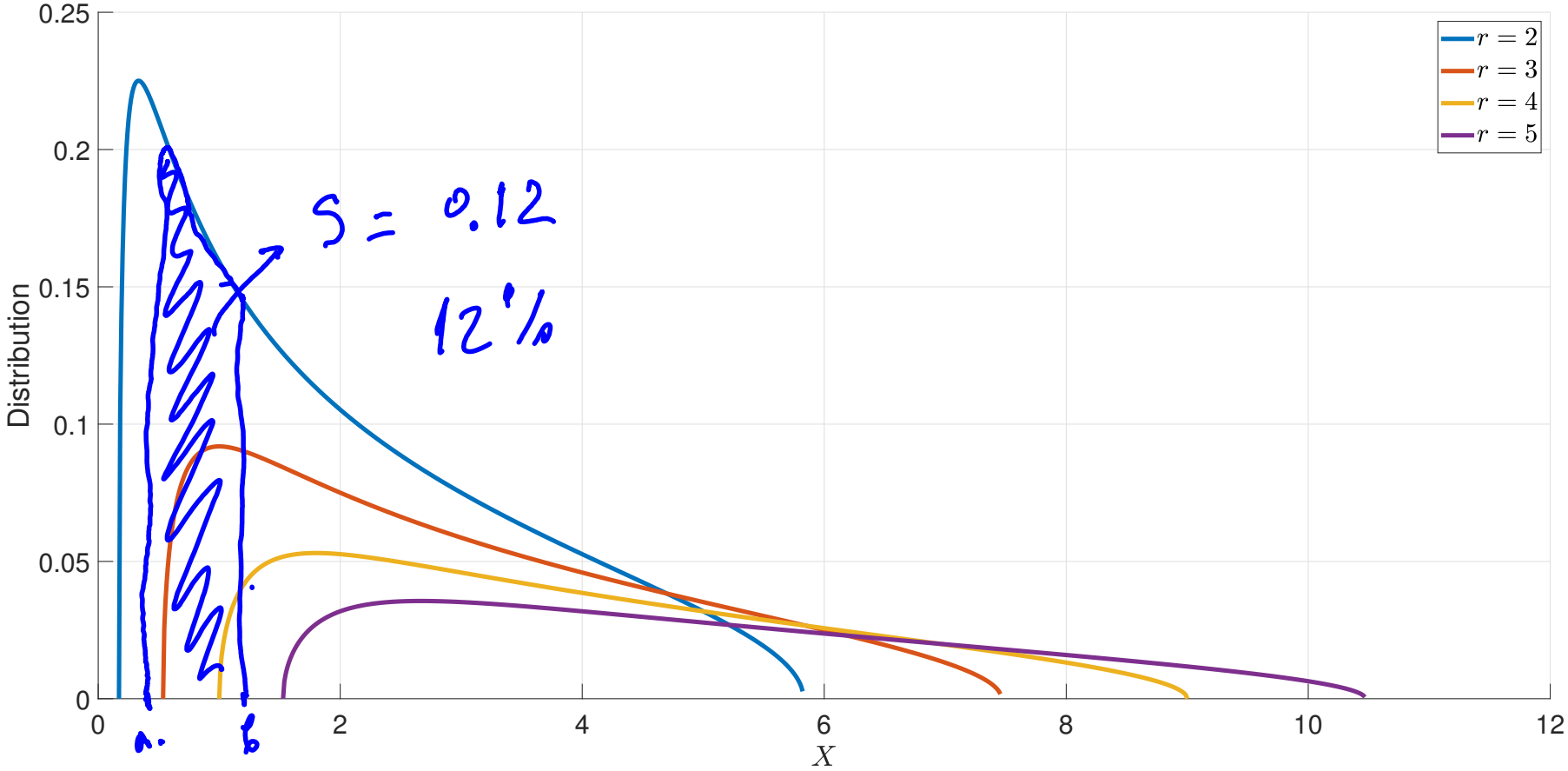
$$\rho_{\text{mp}}(x) = \frac{1}{2\pi r} \frac{1}{x} \sqrt{\left(x - \left(1 - \sqrt{r}\right)^2\right)\left(\left(1 + \sqrt{r}\right)^2 - x\right)}, \tag{3}$$

$$\text{for } x \in \left[\left(1 - \sqrt{r}\right)^2, \left(1 + \sqrt{r}\right)^2\right]$$

and $\rho_{\text{mp}}(x) = 0$ for $x \notin \left[\left(1 - \sqrt{r}\right)^2, \left(1 + \sqrt{r}\right)^2\right]$. No proof given!

Asymptotic Distribution of the Eigenvalues of the Estimated Covariance Matrix Generated From Zero-Mean Unit-Variance Gaussian Noise

Figure of the Marchenko-Pastur distribution for different r:

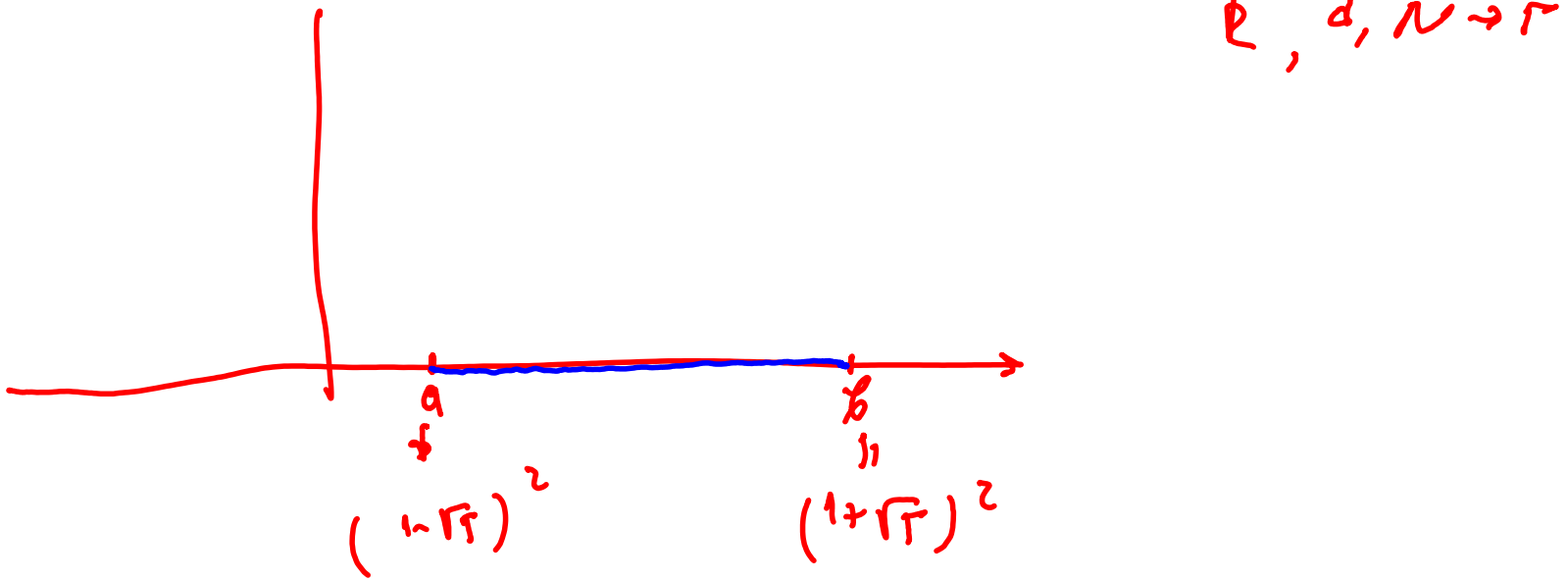


Asymptotic Distribution of the Eigenvalues of the Estimated Covariance Matrix Generated From Zero-Mean Unit-Variance Gaussian Noise

- What we observe from Marchenko-Pastur Law is that in the estimated covariance matrix of data generated from Gaussian noise, the fraction of eigenvalues that are outside of the interval $\left[\left(1-\sqrt{r}\right)^2, \left(1+\sqrt{r}\right)^2\right]$ converges to zero as $N \to \infty$.
- Hence, if we again consider the signal plus noise matrix $\boldsymbol{R} = \boldsymbol{S} + \hat{\boldsymbol{\Sigma}}_N$, then as N increases, all the eigenvalues of the noise are located within the interval $\left[\left(1-\sqrt{r}\right)^2, \left(1+\sqrt{r}\right)^2\right]$.
- Therefore, we should 'look' for our signal/information as the eigenvalues (and their corresponding eigenvectors) located above $\left(1+\sqrt{r}\right)^2$.
- Why not bellow $\left(1-\sqrt{r}\right)^2$?

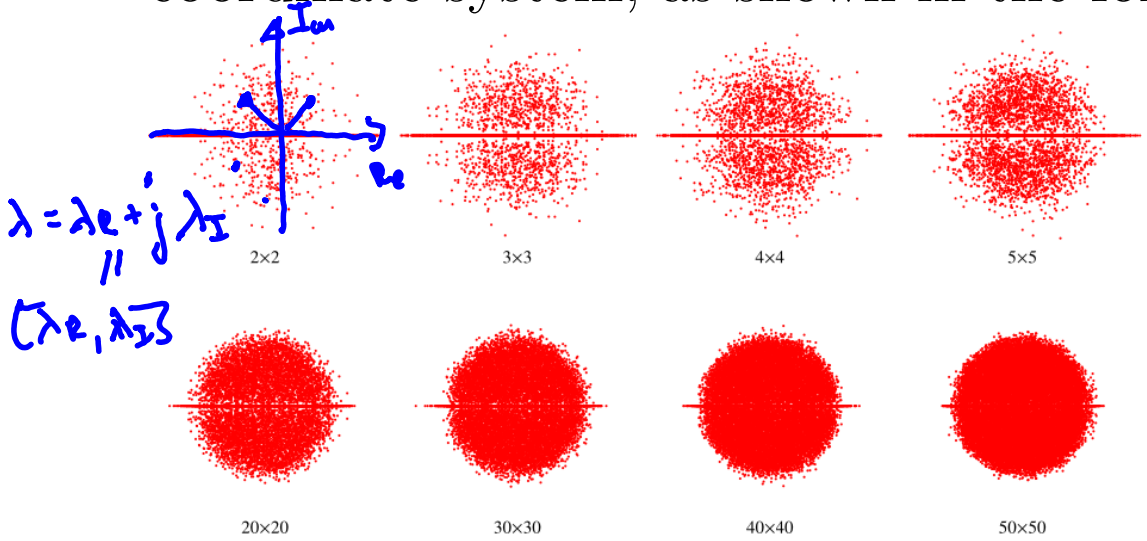
Asymptotic Distribution of the Eigenvalues of the Estimated Covariance Matrix Generated From Zero-Mean Unit-Variance Gaussian Noise

- Figure of the distribution of a signal matrix plus estimated covariance matrix of data generated from Gaussian noise:



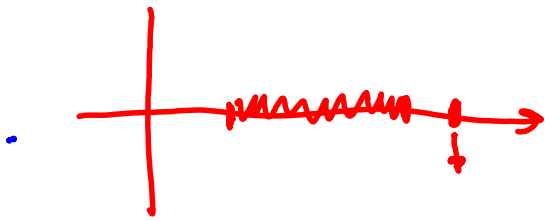
Asymptotic Distribution of a Non-Symmetric Random Matrix

- Is what we observed limited to only those two matrices?
- How about we choose all elements of a matrix $\boldsymbol{M}$ to be randomly generated i.i.d. from some zero-mean unit-variance distribution.
- Then, the matrix $\boldsymbol{M}$ is non-symmetric and as a result some or all of its eigenvalues are complex valued.
- It turns out that then the eigenvalues of $\boldsymbol{M}/\sqrt{d}$, as $d \to \infty$, are located uniformly within the unit-circle, formed on the $\Im$ vs $\Re$ coordinate system, as shown in the following figure:



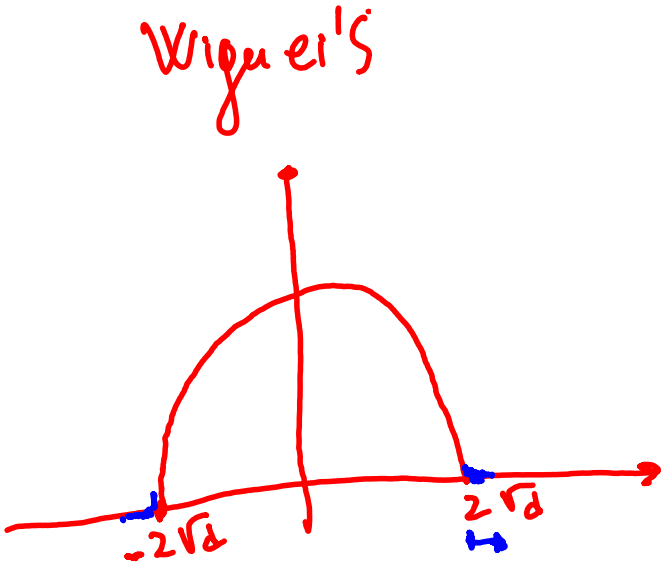
Outlier Eigenvalues

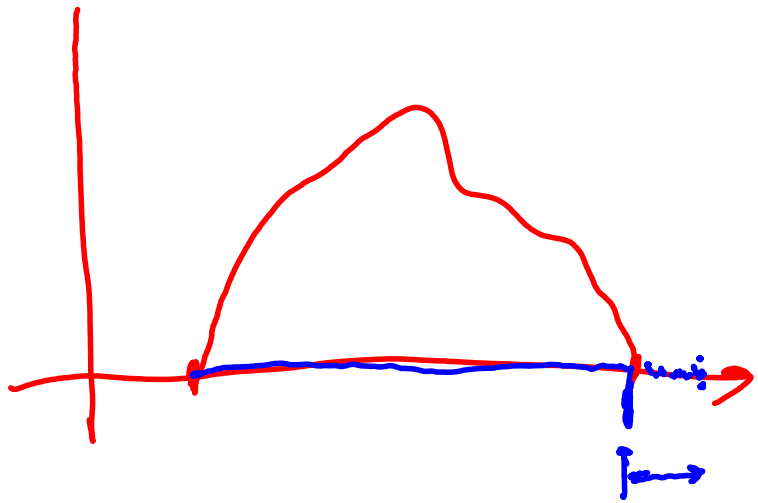
- The above results do not tell us about the outlier (i.e., largest and smallest) eigenvalues when d and N are finite.
- Why? Because the distributions we saw tell us only about the fraction of eigenvalues within a range, when the total number of eigenvalues goes to infinity due to $d \to \infty$ and/or $N \to \infty$.
- But when d and N are limited, the number of eigenvalues will be limited and in that case there might be some finite number of outlier eigenvalues.
- The questions is: For finite d and N, are the largest and smallest eigenvalues also limited to the given ranges in the above slides?
- Why do we care about outliers? Well, because PCA depends on them. If there are few large outliers, then PCA would not work.



Outlier Eigenvalues

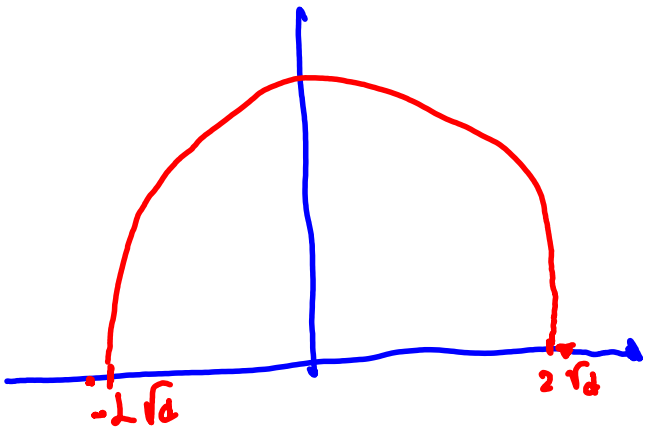
- Examples of outlier eigenvalues for Wigner's and MP laws:





Laws for Extreme Eigenvalues in Wigner's Matrices

- For Wigner matrices, the largest and the smallest eigenvalues satisfy $\lambda_1(\boldsymbol{W}) = 2\sqrt{d} + O(d^{-1/6})$ and $\lambda_d(\boldsymbol{W}) = -2\sqrt{d} - O(d^{-1/6})$, respectively.
- Figure:



Laws for Extreme Eigenvalues in Covariance Matrices Generated From Gaussian Noise

- For covariance matrices generated from Gaussian noise, the largest and the smallest eigenvalues satisfy

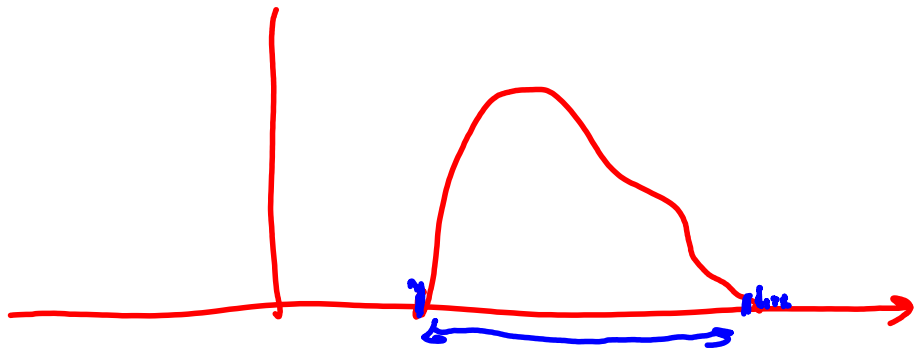
$$\lambda_1(\hat{\mathbf{\Sigma}}_N) = (1+\sqrt{r})^2 + O\big(d^{-1/6}(1+\sqrt{r})^{-2/3}\big)$$

and

$$\lambda_d(\hat{\mathbf{\Sigma}}_N) = (1-\sqrt{r})^2 + O\big(d^{-1/6}(1-\sqrt{r})^{-2/3}\big),$$

respectively.

- Figure:



Signal Plus Noise - Wigner Noise

- So far, we have only investigated the behaviour of matrices generated from pure noise data.
- How about if we have signal matrix plus a noise matrix. Then, how will the signal eigenvalues behave
- Let

$$\boldsymbol{R} = \gamma \boldsymbol{s}\boldsymbol{s}^T + \frac{1}{\sqrt{d}}\boldsymbol{W}, \tag{4}$$

where $\boldsymbol{s} \in \mathbb{R}^d$ is a unit vector, $\boldsymbol{s}\boldsymbol{s}^T$ is the unit-rank signal matrix, $\gamma \geq 0$ is the strength of the signal (or signal-to-noise ratio (SNR)), and $\boldsymbol{W}$ is the Wigner's noise matrix.

- Now by observing $\boldsymbol{R}$ we would like to know whether we can detect the signal $\boldsymbol{s}\boldsymbol{s}^T$ and what are the conditions for this detection to be successful

Signal Plus Noise - Wigner Noise

- For the largest eigenvalue of $\boldsymbol{R}$, given by (4), we have the following convergence

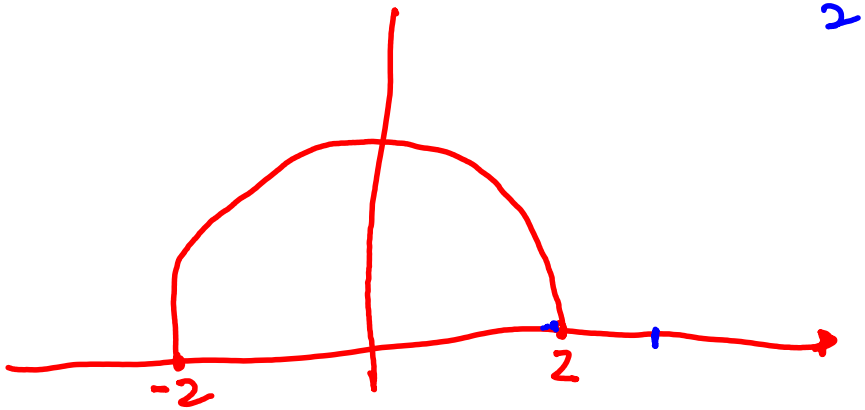
$$\lambda_1(\boldsymbol{R}) \to \begin{cases} 2, & \text{if } \gamma \leq 1 \\ \gamma + \frac{1}{\gamma}, & \text{if } \gamma \geq 1 \end{cases} \tag{5}$$

6ed

$1.5 + \frac{1}{1.5} > 1$

$\frac{3}{2} + \frac{1}{\frac{3}{2}} = \frac{3}{2} + \frac{2}{3} = \frac{9+4}{6} = \frac{13}{6} > 2$

- What does it mean:



$V_1 \approx S$

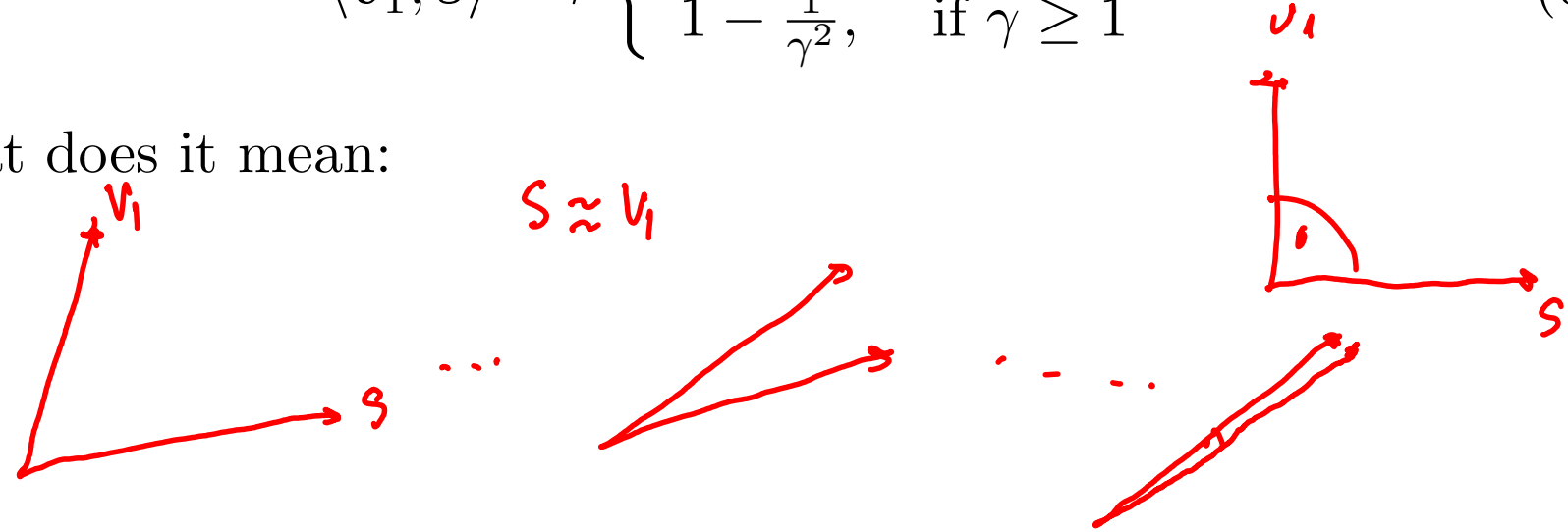
$S \cdot S^T \approx V_1 \cdot V_1^T$

Signal Plus Noise - Wigner Noise

- Next, let's measure the correlation between the eigenvector associated with $\lambda_1(\boldsymbol{R})$, denoted by $\boldsymbol{v}_1$, and the eigenvector of, $\boldsymbol{s}$, using their inner product

$$\langle \boldsymbol{v}_1, \boldsymbol{s} \rangle^2 \to \begin{cases} 0, & \text{if } \gamma \leq 1 \\ 1 - \frac{1}{\gamma^2}, & \text{if } \gamma \geq 1 \end{cases} \tag{6}$$

- What does it mean:



Signal Plus Noise - Estimated Covariance Noise

- Let

$$\boldsymbol{R} = \gamma \boldsymbol{s}\boldsymbol{s}^T + \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^T = \gamma \boldsymbol{s}\boldsymbol{s}^T + \hat{\boldsymbol{\Sigma}}_N \tag{7}$$

  where $\boldsymbol{s} \in \mathbb{R}^d$ is a unit vector, $\boldsymbol{s}\boldsymbol{s}^T$ is the unit-rank signal matrix, $\gamma \geq 0$ is the strength of the signal (or signal-to-noise ratio (SNR)), and $\boldsymbol{X}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$.

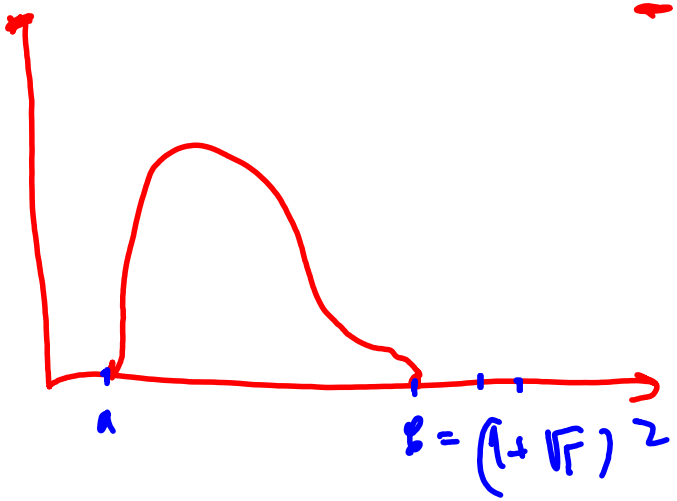
- Now by observing $\boldsymbol{R}$ we would like to know whether we can detect the signal $\boldsymbol{s}\boldsymbol{s}^T$ and what are the conditions for this detection to be successful

Signal Plus Noise - Estimated Covariance Noise

- For the largest eigenvalue of $\boldsymbol{R}$, given by (7), we have the following convergence

$$\lambda_1(\boldsymbol{R}) \to \begin{cases} (1+\sqrt{r})^2, & \text{if } \gamma \leq \sqrt{r} \\ (1+\gamma)\left(1+\frac{r}{\gamma}\right), & \text{if } \gamma \geq \sqrt{r} \end{cases} \tag{8}$$

- What does it mean:

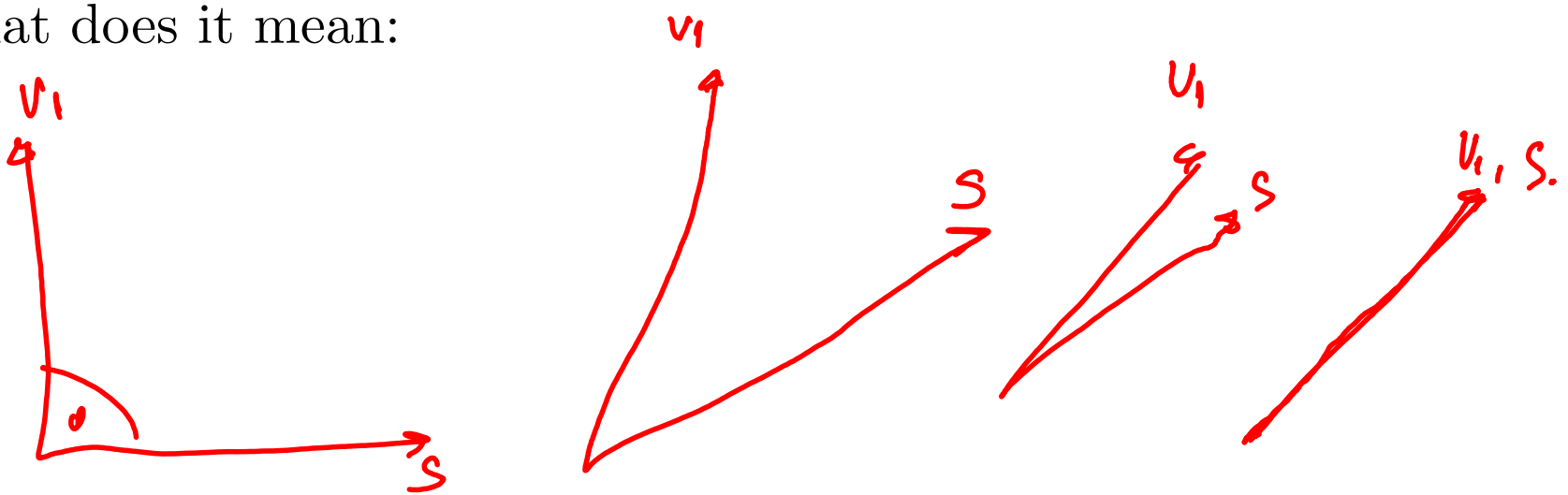


Signal Plus Noise - Estimated Covariance Noise

- Next, let's measure the correlation between the eigenvector associated with $\lambda_1(\boldsymbol{R})$, denoted by $\boldsymbol{v}_1$, and the eigenvector of, $\boldsymbol{s}$, using their inner product

$$\langle \boldsymbol{v}_1, \boldsymbol{s} \rangle^2 \to \begin{cases} 0, & \text{if } \gamma \leq \sqrt{r} \\ \frac{1-r/\gamma^2}{1+r/\gamma^2}, & \text{if } \gamma \geq \sqrt{r} \end{cases} \tag{9}$$

- What does it mean:

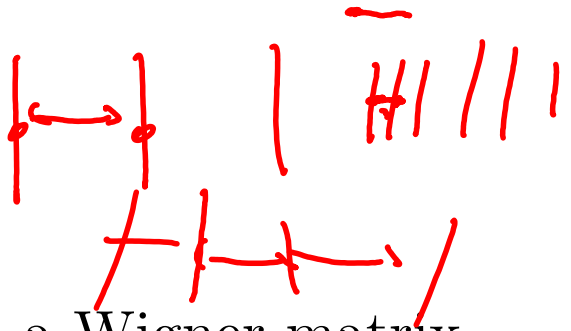


Joint Eigenvalue Distribution

- So far, we have not said anything about the eigenvalues in one matrix are jointly dependent or independent?
- This information can be seen if we have the joint distribution of the eigenvalues.
- It turns out, even the joint distribution of the eigenvalues is computable.
- Let $\boldsymbol{W}$ be created as

$$\boldsymbol{W} = \frac{\boldsymbol{A} + \boldsymbol{A}^T}{2},$$

where $\boldsymbol{A}$ has i.i.d. entries from $N(0, 1)$.



- Then, $\boldsymbol{W}$ is a symmetric random matrix, i.e., a Wigner matrix.
- Then, the joint distribution of the eigenvalues of this matrix is given by

$$f_{\gamma_1,\gamma_2,\ldots,\gamma_d}(x_1, x_2, \ldots, x_d) \sim \exp\left(-\frac{1}{2}\sum_{i=1}^{d} x_i^2\right) \prod_{i=1}^{d}\prod_{j=i+1}^{d} |x_i - x_j|$$

- What does it mean?

Universality Of The Results

- So far, all of the results for the estimated covariance matrix

$$\hat{\boldsymbol{\Sigma}}_N = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^T \tag{10}$$

  were derived for $\boldsymbol{X}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$.
- Are these result only limited to Gaussian vectors?
- It turns out no.
- It turns out that this is a broad phenomenon that holds for all distributions.
- Thm (Universality): All of the results we derived so far also hold when $\boldsymbol{X}_i$ in (10) are generated from any other distribution, not just Gaussian.
- Again, we will not prove it.