Th. 4.1. (Marchenko–Pastur, 1967)

Let $X_1, \dots, X_n \in \mathbb{R}^p$, i.i.d. r.vectors with $E(X_i) = 0$ and $\mathrm{Cov}(X_i) = \sigma^2 I_p$. $X = (X_1, \dots, X_n) \in \mathbb{R}^{p \times n}$,

$$S_n = \frac{1}{n} X X^T = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \in \mathbb{R}^{p \times p},$$

$\lambda_1, \dots, \lambda_p$ eigenvalues of S_n.

Let $p, n \to \infty$ such that $\frac{p}{n} \to y \in (0, 1]$ $(n \to \infty)$.

Then the sample distribution of $\lambda_1, \dots, \lambda_p$ (the histogram) converges a.s. to the density

$$f_y(u) = \frac{1}{2\pi\sigma^2 u y} \sqrt{(b-u)(u-a)}, \quad a \le u \le b$$

with $a(y) = \sigma^2 (1 - \sqrt{y})^2$ $b(y) = \sigma^2 (1 + \sqrt{y})^2$. ⌋

Remark: If $y > 1$ there will be a mass point at zero.

Conclusions: even in the i.i.d. uncorrelated case there is a wide spectrum of eigenvalues.

What happens, if there is a low-dimensional structure in the data?

4.1.5. Spike models

Model assumptions: $X_1, \dots, X_n \in \mathbb{R}^p$, i.i.d.,

$$\mathrm{Cov}(X_i) = \Sigma = I_p + \beta v v^T$$

for some $v \in \mathbb{R}^p$, $\|v\| = 1$, $\beta \geq 0$

Interpretation:

$$X_i = U_i + \sqrt{\beta}\, V_i\, v$$

$U_i \sim N(0, I_p)$ noise

$V_i \sim N(0,1)$ signal, independent from U_i, multiplied by a fixed vector $\sqrt{\beta}\, v$.

Then $\mathrm{Cov}(X_i) = \mathrm{Cov}(U_i) + \beta\, \mathrm{Var}(V_i)\, v v^T = I_p + \beta v v^T$.

Example: $p = 500$, $n = 1000$, $v = e_1$, $\beta = 1.5$

$$\Sigma = I_p + 1.5\, e_1 e_1^T$$

$$\lambda_{\max}(\Sigma) = 1 + 1.5 = 2.5$$

all other eigenvalues 1

Question: Is there a threshold for β above which we will one eigenvector popping out.

Th 4.2. (BBP transition, Baik, Ben Arous, Péché [2005])

Assume $X_1, \dots, X_n \in \mathbb{R}^p$ r.vectors, with $E(X_i) = 0$

$\mathrm{Cov}(X_i) = I_p + \beta v v^T$, $\beta \geq 0$, $v \in \mathbb{R}^p$, $\|v\| = 1$.

$S_n = \frac{1}{n} X X^T$. $n, p \to \infty$, $\frac{p}{n} \to \gamma$.

If $\beta \leq \sqrt{\gamma}$ then $\lambda_{\max}(S_n) \to (1+\sqrt{\gamma})^2$.

and $(v_{\max}^T v)^2 \to 0$

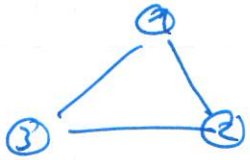
If $\beta > \sqrt{\gamma}$ then $\lambda_{\max}(S_n) \to (1+\beta)(1+\frac{\gamma}{\beta}) > (1+\sqrt{\gamma})^2$.

and $(v_{\max}^T v)^2 \to \frac{1-\gamma/\beta^2}{1-\gamma/\beta}$ ⊣

Proof. Bandera, references therein.

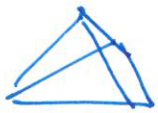
4.2. Multidimensional Scaling

Given pairwise distances d_{ij} between points.



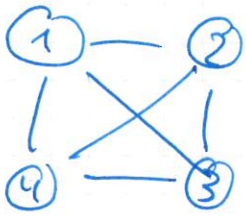
$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Euclidean embedding in dim. 2 ? yes!



$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

Euclidean embedding in dim 2? no!
in dim 3? yes!



$$\begin{pmatrix} 0 & 1 & \sqrt{2} & 1 \\ 1 & 0 & 1 & \sqrt{2} \\ \sqrt{2} & 1 & 0 & 1 \\ 1 & \sqrt{2} & 1 & 0 \end{pmatrix}$$

Euclidean embedding in dim. 2? yes!

Given n objects $O_1, \dots, O_n$ and pairwise dissimilarities δ_{ij} between object i and j.

Assume that $\delta_{ij} = \delta_{ji} \geq 0$ and $\delta_{ii} = 0$. $\forall i,j = 1, \dots, n$.

Define $\Delta = (\delta_{ij})_{1 \leq i,j \leq n}$ as the dissimilarity matrix.

and

$$\mathcal{U}_n = \left\{ \Delta = (\delta_{ij})_{1 \leq i,j \leq n} \mid \delta_{ij} = \delta_{ji} \geq 0,\ \delta_{ii} = 0\ \forall i,j \right\}$$

the set of dissimilarity matrices.

Objective: Find n points $x_1, \dots, x_n$ in a Euclidean space, typically $\mathbb{R}^k$, such that the distances $\|x_i - x_j\|$ fit the dissimilarities δ_{ij} best.

Notation: $X = [x_1, \dots, x_n]^T \in \mathbb{R}^{n \times k}$

$d_{ij}(X) = \|x_i - x_j\|$ distances

$D(X) = (d_{ij}(X))_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}$

$\Delta^{(q)} = (\delta_{ij}^q)_{ij}$ and $D^{(q)}(X) = (d_{ij}^q(X))_{ij}$.

Optimization problem: given $q \geq 1$

$$\min_{X \in \mathbb{R}^{n \times k}} \| \Delta^{(q)} - D^{(q)}(X) \| \qquad (*)$$

typically $q = 2$

Consider the case that $(*) = 0$.

4.2.1. Characterizing Euclidean Distance Matrices

$\Delta = (\delta_{ij}) \in \mathcal{U}^n$ is called Euclidean distance matrix or if has a Euclidean embedding in $\mathbb{R}^k$, if there are $x_1, \dots, x_n \in \mathbb{R}^k$ such that $\delta_{ij}^2 = \|x_i - x_j\|^2 \quad \forall i,j$.
where $\|\cdot\|$ denotes the Euclidean norm $\|y\| = \sqrt{\sum_{i=1}^{k} y_i^2}$.

That means opt.-problem (*) has a solution with value 0.

Projection $E_n = I_n - \frac{1}{n} 1_n 1_n^T$ plays an important role.

$$E_n = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{pmatrix}$$