LECTURE 20

Last Class: Covariance estimation problem for $X \in \mathbb{R}^d$, $\mathbb{E}X = 0$.

- (Population) covariance matrix: $\Sigma = \mathbb{E}XX^T$
- Sample covariance matrix: $\Sigma_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$, where X_i are iid copies of X

$$\Sigma_n \overset{?}{\approx} \Sigma \quad \text{in the operator norm}$$

- $\|\Sigma_n - \Sigma\| = \max_{v \in S^{d-1}} |v^T(\Sigma_n - \Sigma)v|$ (HW)
- $\exists$ ε-net $\mathcal{N} \subset S^{d-1}$, $|\mathcal{N}| \overset{\text{last class}}{\leq} \left(\frac{2}{\varepsilon}+1\right)^d = 9^d$ if we choose $\boxed{\varepsilon = 1/4}$
- $\|\Sigma_n - \Sigma\| \leq \frac{1}{1-2\varepsilon} \max_{v \in \mathcal{N}} |v^T(\Sigma_n - \Sigma)v|$ (Last class & HW)

$$= 2 \cdot \max_{v \in \mathcal{N}} \underbrace{\left| \frac{1}{n}\sum_{i=1}^n \langle X_i, v\rangle^2 - \mathbb{E}\langle X, v\rangle^2 \right|}_{\| \; S(v)} \leq \; ? \qquad (*)$$

$S(v)$ = sum of indep. r.v's. Use:

Thm (Bernstein's inequality – lec. 11)

i.e. $\mathbb{P}\{|Z_i| \geq t\} \leq 2\exp\left(-\frac{t}{M}\right)$ $\forall t$, smallest $M = \|Z_i\|_{\psi_1}$

If Z_i are independent subexponential r.v's, then

$$\mathbb{P}\left\{\sum_{i=1}^n (Z_i - \mathbb{E}Z_i) \geq t\right\} \leq 2\exp\left[-c \cdot \min\left(\frac{t^2}{\sigma^2}, \frac{t}{K}\right)\right]$$

where $\sigma^2 = \sum_{i=1}^n \|Z_i\|_{\psi_1}^2$, $K = \max_i \|Z_i\|_{\psi_1}$

If all Z_i are independently distributed $\Rightarrow \sigma^2 = K^2 n$

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}Z\right| \geq \delta\right\} \overset{t = \delta n}{\leq} 2\exp\left[-c \cdot \min\left(\frac{\delta^2 n^2}{K^2 n}, \frac{\delta n}{K}\right)\right] = 2\exp\left[-c \cdot \min\left(\frac{\delta^2}{K^2}, \frac{\delta}{K}\right) n\right] \qquad (**)$$

ALL TOOLS ARE READY TO PROVE:

THM (Covariance Estimation) If $X_i \sim N(0,\Sigma)$ and $n \geq Cd$,

then $\|\Sigma_n - \Sigma\| \leq 0.1\,\|\Sigma\|$

with probability at least $1 - 2e^{-cn}$.

Proof W.L.O.G. $\|\Sigma\| = 1$.

① Fix any $v \in S^{d-1}$ ← unit sphere of $\mathbb{R}^d$.

X = mean 0 normal r.vector $\Rightarrow \langle X, v\rangle \sim$ mean 0 normal r.vector

$\Rightarrow \langle X, v\rangle \sim N(0,\sigma^2)$.

↑ ? Let's compute it.

v is unit

$$\sigma^2 = \mathbb{E}\langle X,v\rangle^2 = v^T(\mathbb{E}XX^T)v = v^T\Sigma v \leq \|\Sigma\| = 1.$$

$$\Rightarrow \mathbb{P}\{|\langle X,v\rangle| \geq t\} \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right) \leq 2\exp(-t^2/2)$$

↑ gaussian tail (lec. 4)

$$\Rightarrow \mathbb{P}\{\langle X,v\rangle^2 \geq s\} = \mathbb{P}\{|\langle X,v\rangle| \geq \sqrt{s}\} \leq 2\exp(-s/2)$$

$\Rightarrow \langle X,v\rangle^2$ is subexponential, $\|\langle X,v\rangle^2\|_{\psi_1} \leq 2$ ↑

- Apply Bernstein (**p.1) for $Z_i = \langle X_i,v\rangle^2$, $K=2$, $\delta = 0.01$ $\Rightarrow$

$$\mathbb{P}\left\{\underbrace{\left|\frac{1}{n}\sum_{i=1}^n \langle X_i,v\rangle^2 - \mathbb{E}\langle X,v\rangle^2\right|}_{\| \; S(v)} \geq 0.01\right\} \leq 2\exp(-c'n).$$

② $$\mathbb{P}\left\{\max_{v\in\mathcal{N}} |S(v)| \geq 0.01\right\} \leq \underbrace{|\mathcal{N}|}_{\| \; 9^d \text{ (p.1)}} \cdot 2\exp(-c'n) \leq 2\exp\left(-\frac{c'n}{2}\right)$$

if $\boxed{d < c''n}$ ☺

③ The complement holds w/prob. $\geq 1 - 2\exp(-c'n/2)$. When it holds,

$$\|\Sigma_n - \Sigma\| \overset{(*)\text{p.1}}{\leq} 2\cdot\max_{v\in\mathcal{N}}|S(v)| \leq 2\cdot 0.01 \leq 0.1.$$

QED

PERTURBATION THEORY

From $\Sigma_n \approx \Sigma$ to $\lambda_i(\Sigma_n) \approx \lambda_i(\Sigma)$, $v_i(\Sigma_n) \approx v_i(\Sigma)$:

↓

Thm (Weyl's inequality) $\forall d \times d$ symmetric matrices A, B:

$$\max_i |\lambda_i(A) - \lambda_i(B)| \leq \|A - B\|$$

Eigenvectors: (a) up to sign
(b) spectral gap needed: if $\lambda_k(A) \approx \lambda_{k+1}(A)$, perturbation can swap the order of eigenvectors.

Thm (Davis-Kahan inequality)

Let A, B be $d \times d$ symmetric matrices, $1 \leq k \leq d$,

$P_A :=$ orthogonal projection onto $\text{span}\{v_1(A), \dots, v_k(A)\}$ and
$P_B :=$ orthogonal projection onto $\text{span}\{v_1(B), \dots, v_k(B)\}$. Then

$$\|P_A - P_B\| \leq \frac{\|A - B\|}{\lambda_k(A) - \lambda_{k+1}(A)}.$$

- Combine Cov. Est. Thm (P.2) with Weyl & Davis-Kahan

⇓

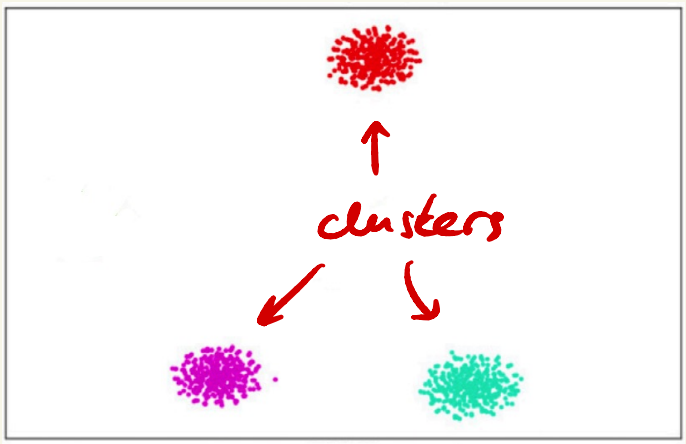
SUMMARY $\text{PCA}_k(\text{sample}) \approx \text{PCA}_k(\text{population})$ as long as:

(a) sample size n is linear in dim d (or larger)

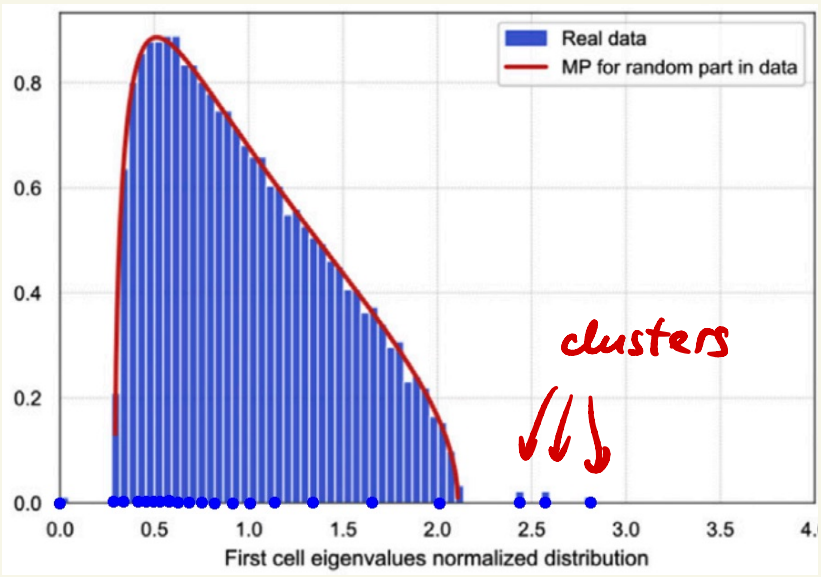
(b) there is an eigengap: the top k eigenvalues of Σ are separated from the rest of the spectrum.

HW-state?

Example: single-cell data from [Aparicio et al. 2020]



→



noise
"Bulk"

"signal"
"outliers"

$X \sim N(0, I_n)$

$\Sigma = I_n$

REMARKS : ① Linear sample size $n = O(d)$ is optimal (HW)

② Normal distribution $X \sim N(0, \Sigma)$ can be generalized to $\forall$ subgaussian (simple - see Book) HW?

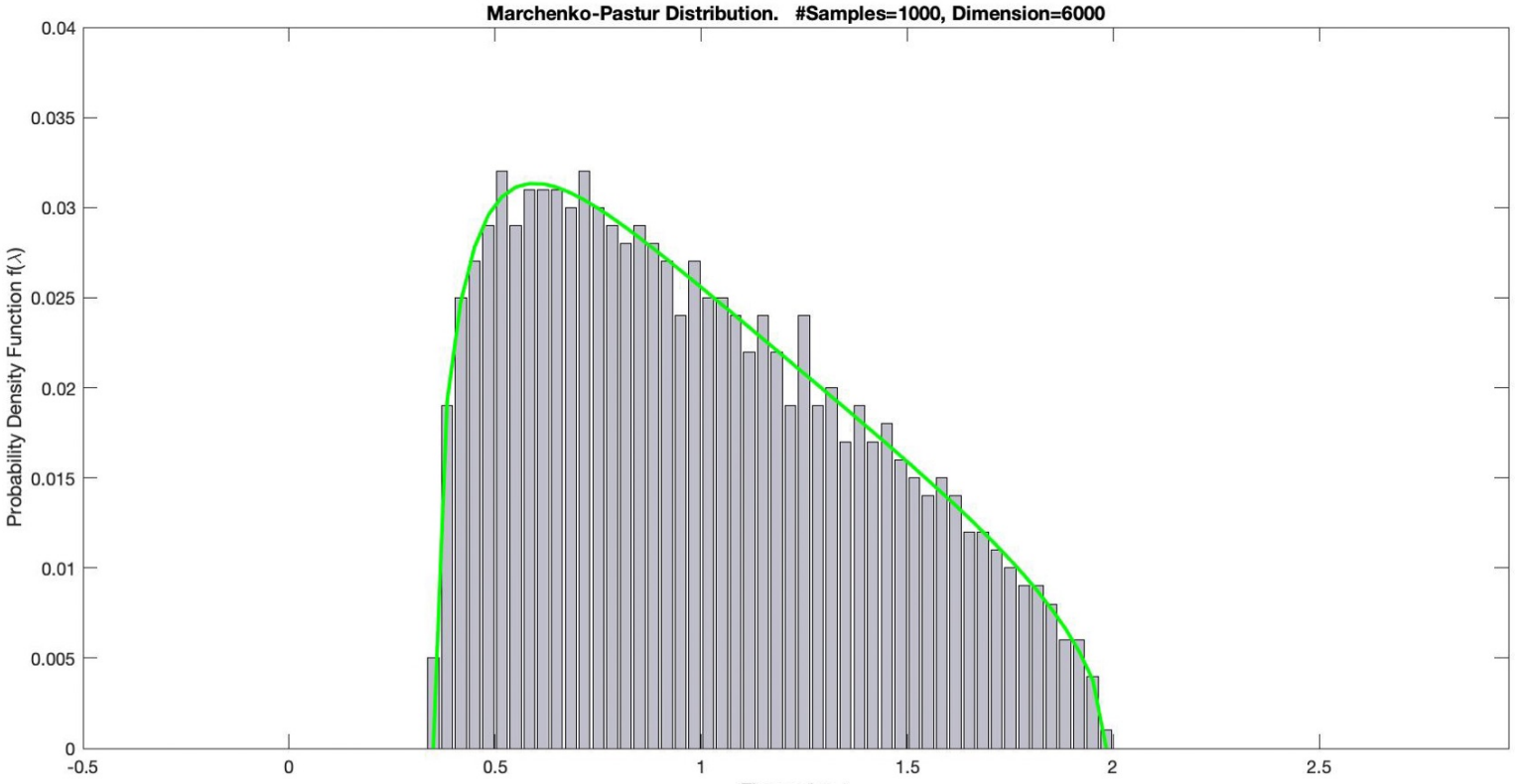
③ Structured data $\Rightarrow$ smaller sample size n?

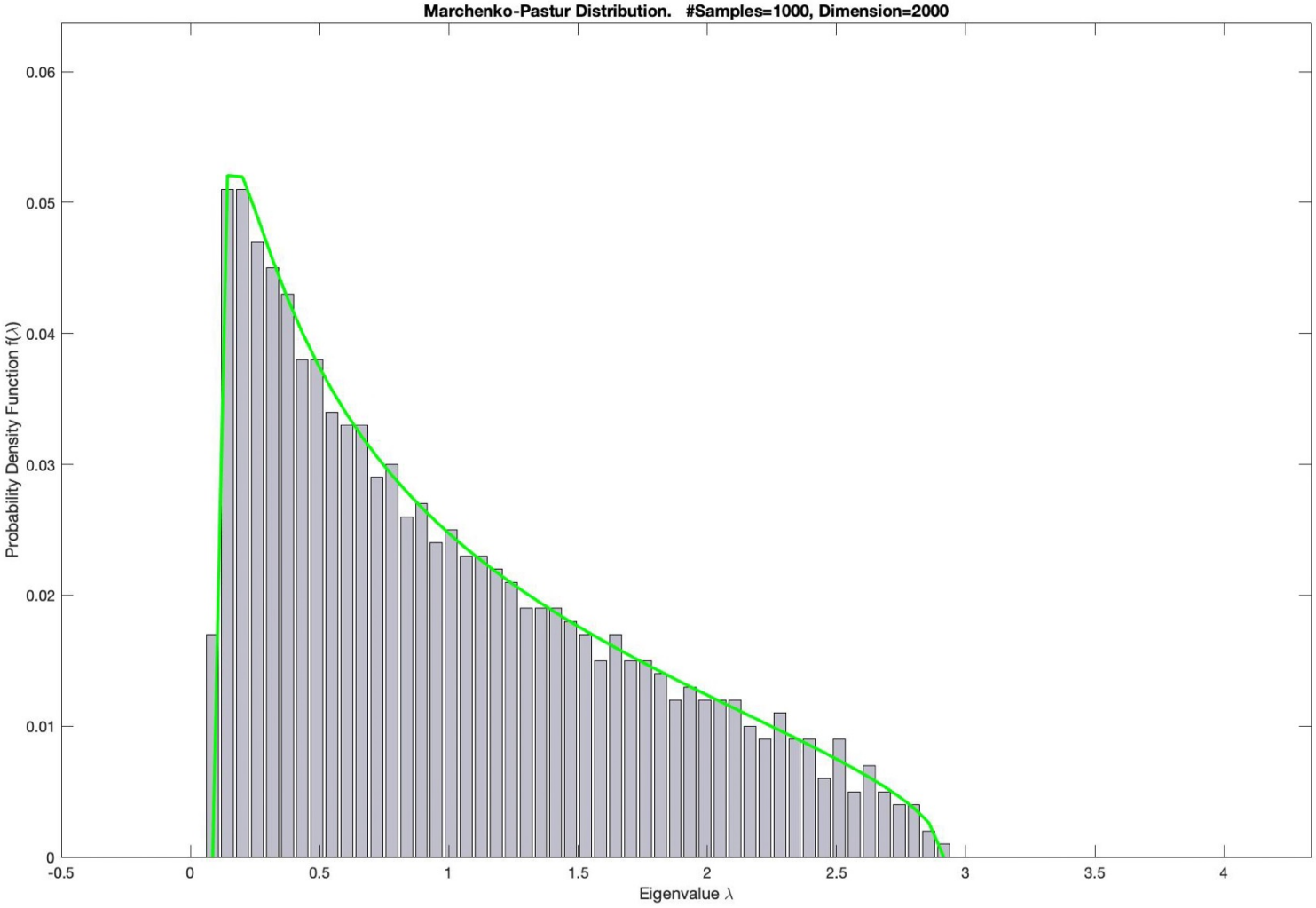
Yes : (a) trivially, if $X = (X_1, X_1, \ldots, X_1)$ then represent $X \in \mathbb{R}^1 \Rightarrow n = O(1)$.

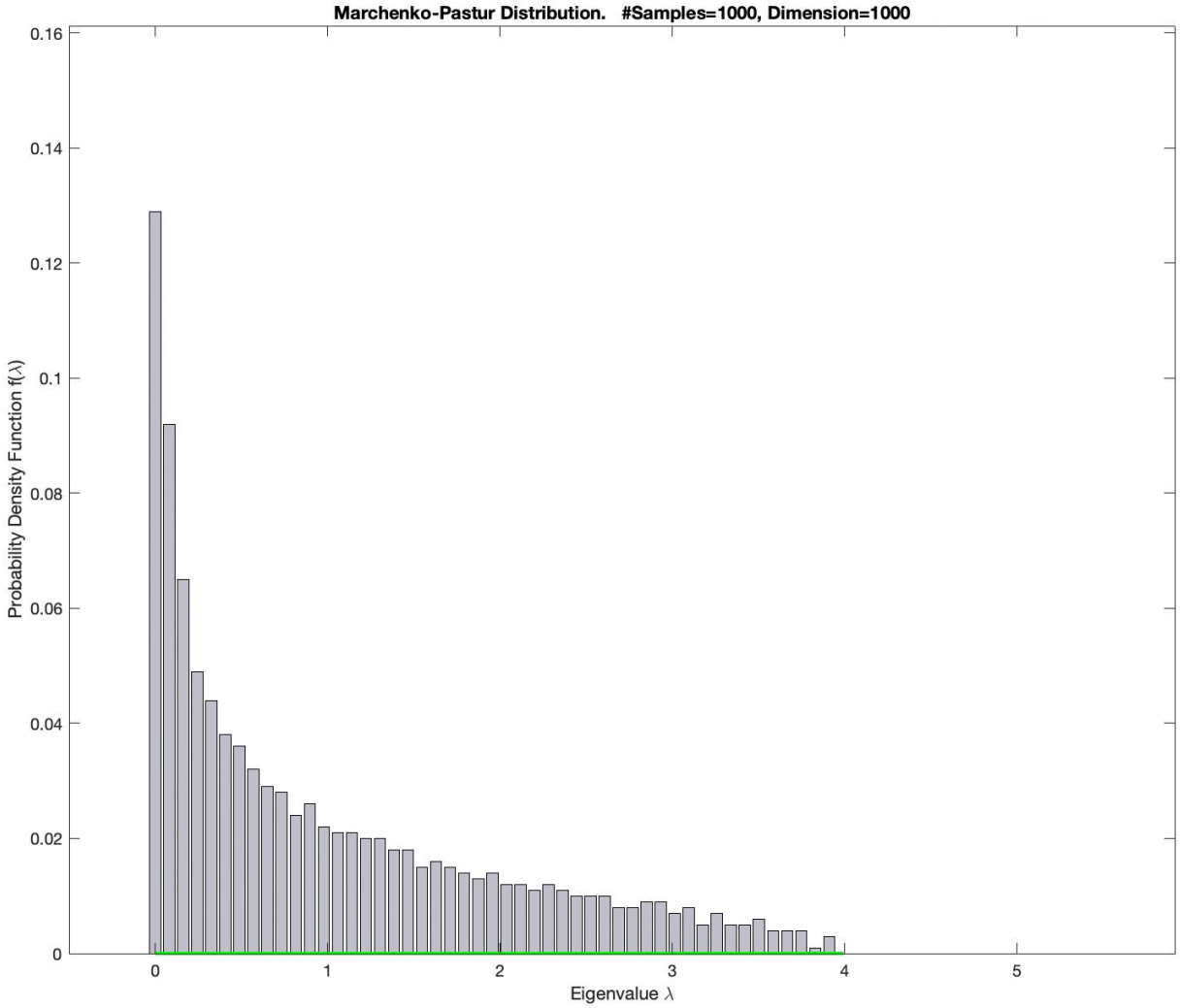
(b) more generally, if data is approximately k-dimensional, $k = O(n)$ is enough.

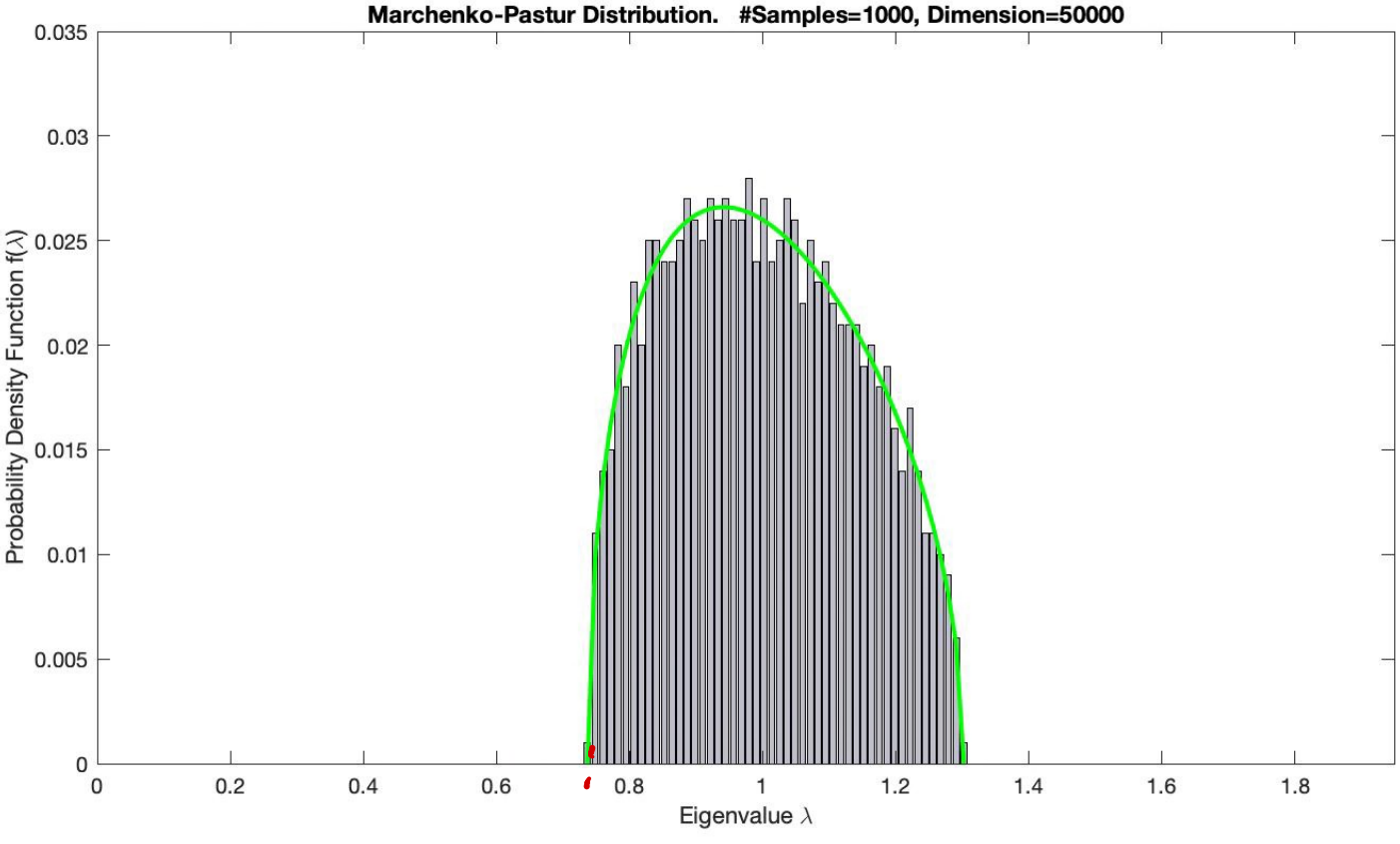
[Koltchinskii-Lounici 2017] = Thm. 9.2.4 in the Book.

Will probably cover later.


Marchenko-Pastur Distribution. #Samples=1000, Dimension=6000
Probability Density Function f(λ)
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
-0.5
0
0.5
1
1.5
2
2.5

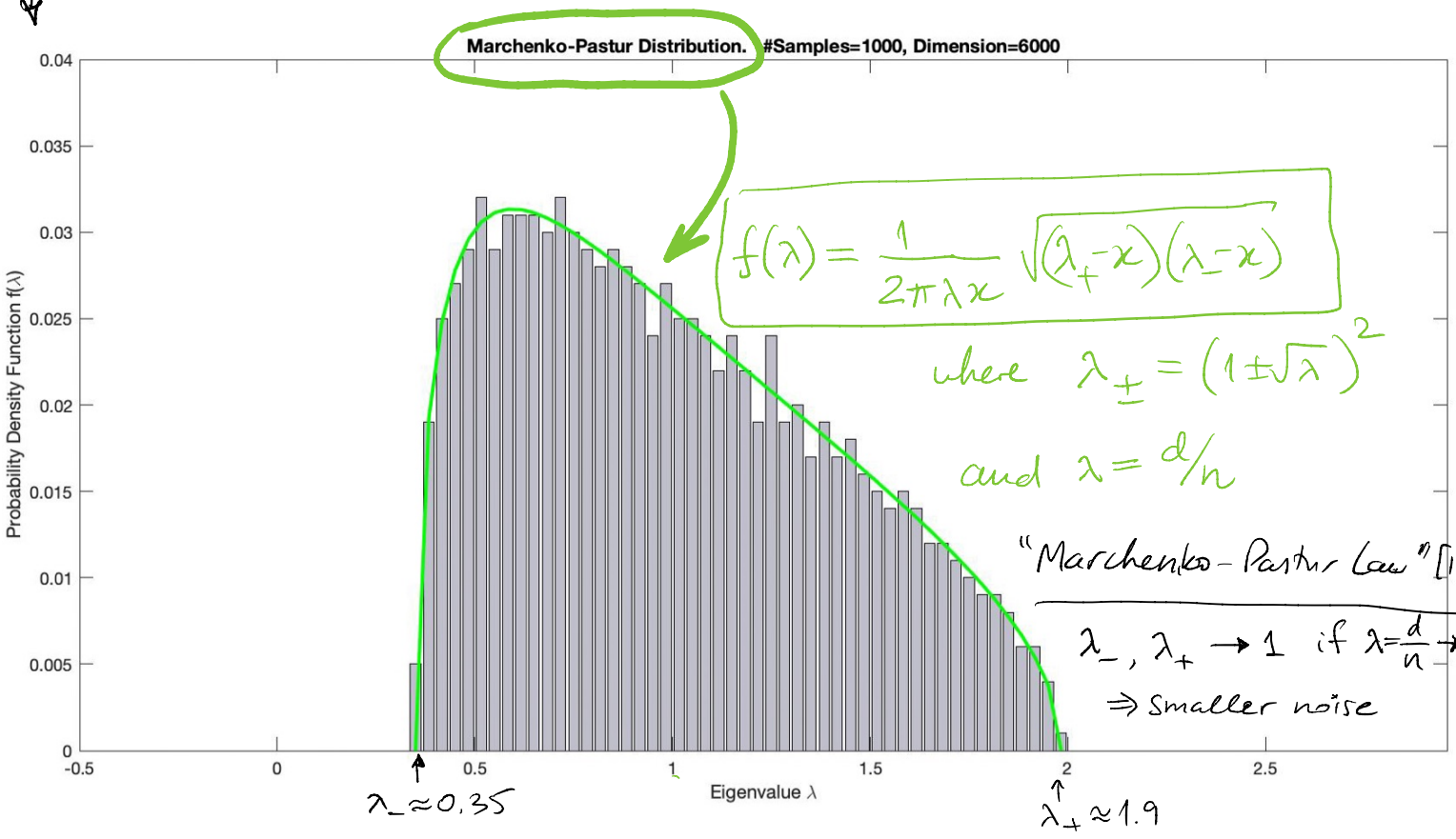

Marchenko-Pastur Distribution. #Samples=1000, Dimension=2000
Probability Density Function f(λ)
0.06
0.05
0.04
0.03
0.02
0.01
0
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4
Eigenvalue λ


Marchenko-Pastur Distribution. #Samples=1000, Dimension=1000
Probability Density Function f(λ)
Eigenvalue λ
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
1
2
3
4
5


Marchenko-Pastur Distribution. #Samples=1000, Dimension=50000
Probability Density Function f(λ)
Eigenvalue λ
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8

- histogram $\text{eigs}(\Sigma_n)$ seems → density as $d, n \to \infty$, $\boxed{\frac{d}{n} \to \lambda}$

compactly supported! No outliers (good for PCA)



$$f(\lambda) = \frac{1}{2\pi\lambda x}\sqrt{(\lambda_+ - x)(\lambda_- - x)}$$

where $\lambda_\pm = (1 \pm \sqrt{\lambda})^2$

and $\lambda = d/n$

"Marchenko-Pastur Law" [1967]

$\lambda_-, \lambda_+ \to 1$ if $\lambda = \frac{d}{n} \to \infty$

⇒ smaller noise

$\lambda_- \approx 0.35$

$\lambda_+ \approx 1.9$