Mathematics of Bigdata Analysis: An Introduction

S. Lakshmivarahan

School of Computer Science

University of Oklahoma

Norman, OK-73019

varahan@ou.edu

Abundance of data

- Thanks to the advances in technology of
  - Sensors
  - Wireless Communication
  - Mass storage devices
  - Large super computers

- Shift from data sparse to data rich regime – amount of data doubles in every few years.

Data organization

- **Time Series** : Number of daily new covid infection in a city.

- **Spatial**: Number of infected in every country on a given time.

- **Spatial-temporal**: Monthly rain fall in each of 50 states in the US.

- **Data Matrix**: X: $[x_1 , x_2 , x_3 \ldots x_n]$, $x_i \in R^d$, - Represents n points in d- dimensional space.

“Big” in Bigdata

- In the matrix form $x \in R^{d \times n}$: Two variables.
- n - is the number of data (columns).
- d - is the dimension of the space (rows)
- In general: either n or d or both can be large.
- Similar measures apply for other data organization.

Classical Statistics

- In classical mathematical statistics there are a number of asymptotic results obtained by fixing d and letting the number of samples to increase without bound such that the ratio

$$\frac{d}{n} \text{ -> } 0$$

- This asymptotic theory provides the basis for estimation theory.

Examples 1

- **Law of Large Numbers (LLN):** If x_i , $1 \leq i \leq n$ is i.i.d sequence of random variables from, say normal distribution N (m, σ^2) with unknown m.

- $\bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an unbiased estimate.

- LLN: prob [| $\bar{x}$(n) - m| > ε] -> 0 or n -> ∞ ---------> 1

- This is called asymptotic consistency.

- Also known as **measure concentration**.

Examples 2

- **Central Limit Theorem (CLT):**

- In addition to (1), the following stronger result hold:

$$\frac{\sqrt{n}\,(\bar{x}(n) - m)}{\sigma} \xrightarrow[\text{in distribution}]{} N(0,1) \qquad \text{-----------------> (2)}$$

- That is, centered and scaled estimate converges in distribution to a standard normal Gaussian variable.

High dimensional data

- Consider a set of $n = 100$ color images of a human retina with 256 x 256 = 65, 536 pixels in each of the three frames representing Red, Blue and Green with a total of d = 65, 536 x 3 = 196, 608 pixels.

- Here $x \in R^{d \times n}$ where $d >> n$

- In here, $\frac{d}{n} = \alpha > 0$ ------------------------------> (3)

Implications of $\frac{d}{n}$ = α > 0

- Many of the known results from classical statistics when applied to this case , $\frac{d}{n}$ = α > 0 give only "suboptimal" guarantees.

- To address this challenge a new specialty is emerging.

- M.J. Wainwright (2019) High-Dimensional Statistics: A non- asymptotic viewpoint, Cambridge university Press.

- R.Vershynin (2020) High-Dimensional Probability: An Introduction with Application in Data Science Cambridge University Press.

Curse of dimensionality

- Coined by Richard Bellman (1920 – 1984) when developing.
- R.Bellman (1952) “Theory of Dynamic Programming”, Proc of NAS, pp 716-719.
- Finding optimal solution for multistage decision process often require 2^d computation.
- The popular Reinforcement Learning (RL) is based on the theory of Markov Decision Process is an example of the application of DP.

Counter intuitive results in High dimension

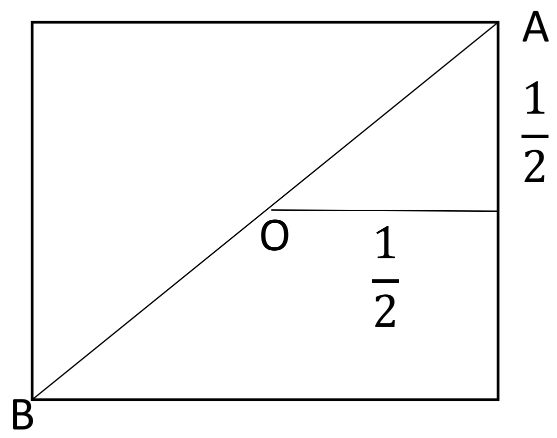
- Empty space – High dimensional geometry.
- Concentration of distances, measures, functions.
- Statistical two class classification.
- Estimation of covariance matrices.

Hyper cube V_c (d,a) in R^d

- V_c (d,a) – hypercube of side “a” in R^d .

- Diagonal AB in V_c (2,1):

$$AB = 2 * OA = 2\left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]^{1/2} = \sqrt{2}$$



V (2,1)

- Diagonal AB in V_c (d,1):

$$AB = 2 * OA = 2\left[\sum_{i=1}^{d}\left(\frac{1}{2}\right)^2\right]^{1/2} = \sqrt{d} \qquad \text{--------------------------->}\ (4)$$

- Diagonal increases as $\sqrt{d}$ while the side of the cube remains constraint as d increases.

Empty space in R^d

- Volume of V_c (d,a) = a^d .

- If we double the side : V_c (d, 2a) = 2^d V_c (d, a) --------------------------> (5)

- Volume of the cube grows exponentially when you double its side.

- Creates a lot of empty space.

Spheres in R^d : V_s (d,r)

- V_s (d,r) – a sphere of radius r in R^d .

- Vol [V_s (d,r)] = $\frac{\Pi^{d/2}}{\Gamma(\frac{d}{2}+1)}\mathrm{r}^{d}$ --------------------------------------> (6)

- For integer k: $\Gamma(k+1)$ = k $\Gamma(k)$ and $\Gamma(k+1)$ = k! ----------> (7)

$\Gamma(1/2)$ = $\sqrt{\prod}$

Unit Sphere : V_s (d,1)

- Vol [V_s (d,1)] = $\frac{\Pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ ->0 as d-> ∞

- Vol [V_s (3,1)] = $\frac{4}{3}\Pi$ = 4.1867

  Vol [V_s (10,1)] =$\frac{\Pi^{10}}{10!}$ = 0.0258

- **<u>Question</u>** : For what values of r, Vol [V_s (d,r)] =1

- Using Strilings approximation to n! :

  n! = $\sqrt{2\Pi n}\left(\frac{n}{e}\right)^n$

- Verify r= O($\sqrt{d}$) for Vol [V_s (d,r)] =1

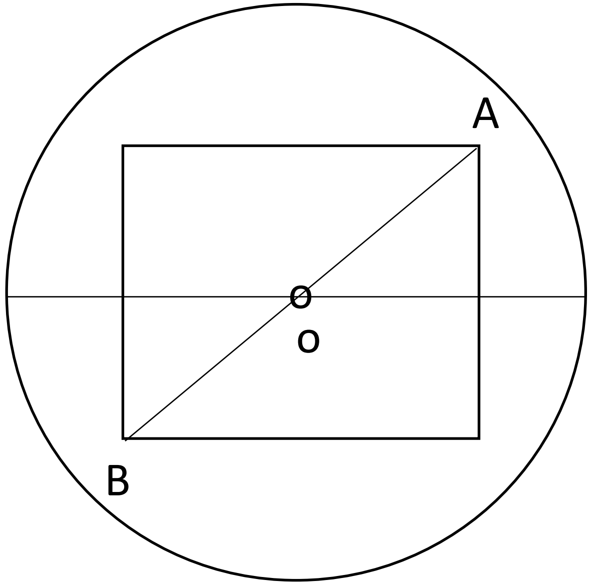
- Empty space syndrome.

Cube inside a cube

- Consider a unit cube inside a concentric unit sphere in R^d .
- Have seen AB = $\sqrt{d}$
- For d < 4, AB < 2 and inside the sphere.

  d = 4, AB =2 and AB is a diameter.

  d > 4, AB > 2 and punches through the sphere.
- For large d, 2^d diagonals get out of the sphere.
- It looks like the picture of the COVID virus.



$V_c\,(d,1) \subseteq V_s\,(d,1)$

Sphere in a Sphere

- Let r < R, concentric spheres of radii r and R.

- $\frac{V_S(d,R) - V_S(d,r)}{V_S(d,R)}$ = 1- $\frac{V_S(d,r)}{V_S(d,R)}$ ---> (8)

 = 1 - $\left(\frac{r}{R}\right)^d$ -> 1 as d increases.

(i.e.) Volume of the sphere reside near the empty space shell.

Sphere in a cube

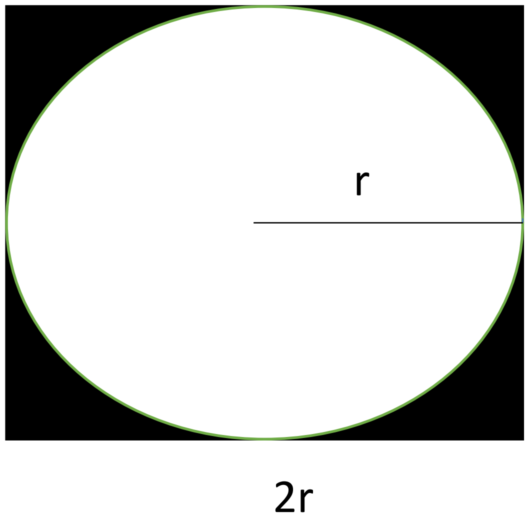
- Ratio $\alpha = \frac{V_s(d,r)}{V_c(d,2r)}$

  $= \frac{\Pi^{d/2} r^d}{\Gamma(\frac{d}{2}+1)} \frac{1}{(2r)^d} = \left(\frac{\Pi}{4}\right)^{d/2} \frac{1}{\Gamma(\frac{d}{2}+1)}$ -> 0

  as d increases ----------------------> (9)

- Fraction of the volume of the cube trapped inside the sphere goes to zero as d increases.

- Empty space at the center and volume of the cube is concentrated at its 2^d corners.



Pairwise distances in R^2

- Consider $\mathrm{V_c}(2,1)$: Generate 1001 independent, identically distributed in $\mathrm{V_c}(2,1)$.

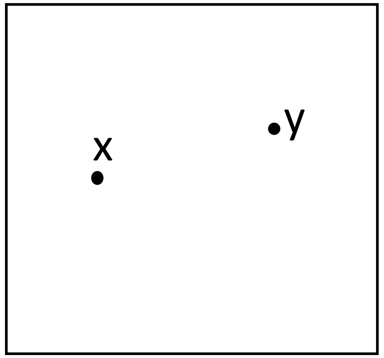
- Fix one of the point and call it x = $(\mathrm{x}_1, \mathrm{x}_2)^T$.

- Compute for each of the rest of 1000 points

$$D^2(\text{x, y}) = [(x_1 - y_1)^2 + (x_2 - y_2)^2\ ]\ (\text{y} \neq \text{x}).$$

- Clearly 0 ≤ D^2(x, y) ≤ 2 for all y ≠ x since $|x_1 - y_1|$ ≤ 1 and $|x_2 - y_2|$ ≤ 1 .

- Histogram of D^2(x, y) is fully supported on [0,2].



$\mathrm{V_c}(2,1)$

Pairwise distances in R^2: d =100

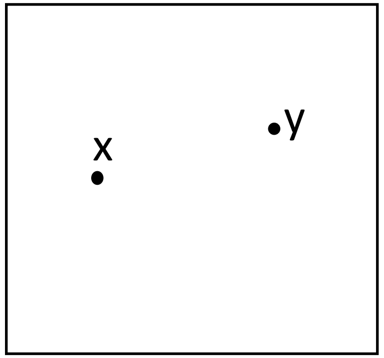
- Repeat the above experiment in $\mathrm{V_c}(\mathrm{d}, 1)$.

- Here $\begin{cases} x = (\mathrm{x}_1, \mathrm{x}_2, \dots \mathrm{x}_\mathrm{d})^T \\ y = (\mathrm{y}_1, \mathrm{y}_2, \dots \mathrm{y}_\mathrm{d})^T \end{cases}$ with $|x_i - y_i| \le 1$

- D^2(x, y) = $\sum_{i=1}^{d}(x_i - y_i)^2$ ---------------------> (10)

- Clearly 0 ≤ D^2(x, y) ≤ 100 .

- A lot more is true – thanks to the law of large numbers.



$\mathrm{V_c}(\mathrm{d}, 1)$

Concentration of distances

- Clearly x_i's and y_i's , $(x_i - y_i)^2$ are i.i.d random variables with finite mean and variance.

- D^2(x, y) = $\sum_{i=1}^{d}(x_i - y_i)^2$ is the sum of i.i.d random variables.

- By the law of large numbers, the distribution of D^2(x, y) is concentrated in the interval [0,100] around the mean.

- For small d, this distribution is spread out in [0,d] but for large d, it gets concentrated.

Gaussian distribution in R^d

- $x \in R^d$, $m \in R^d$ $\sum R^{d \times d}$.

- X ~ N (m, ∑) = $\frac{1}{(2\Pi)^{d/2}\,|\sum|^{1/2}}$ exp [- $\frac{1}{2}$ $(x - m)^T \sum^{-1}$ (x - m)] -----------------------------> (11)

- X ~ N (0, $\sigma^2 I$) = $\prod_{i=1}^{n} \frac{1}{\sqrt{2\Pi}\sigma}$ exp [− $\frac{x_i^2}{2\sigma^2}$] --> (12)

- E [$||x||^2$] = d E(x_1^2) = d σ^2 --> (13)

Since x_i are i.i.d N (0, σ^2).

- For large d, the random variable $||x||^2$ is concentrated about its mean d σ^2.

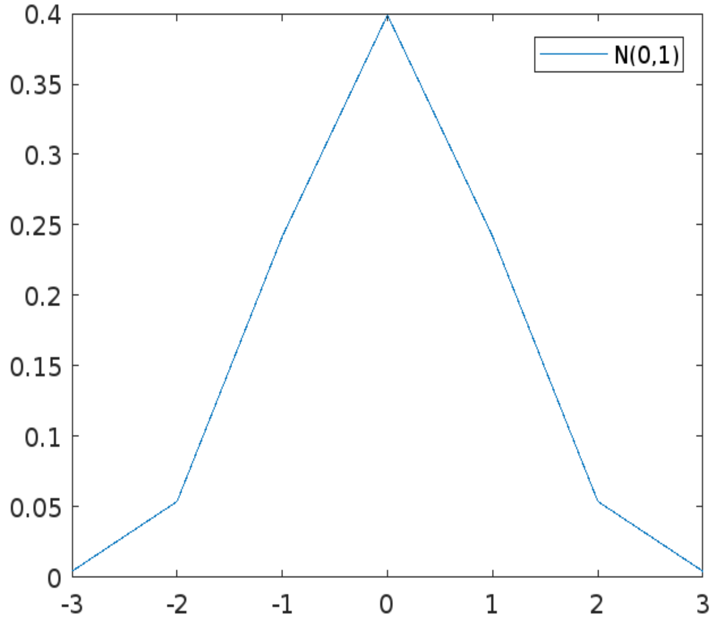
- $\sigma\sqrt{d}$ is called the radius of the Gaussian.

Tail probability of N(0,1) in R^1

- Consider N(0,1)

- Let r(a) = $\frac{1}{\sqrt{2\Pi}}\int_{-a}^{a}\exp\left(\frac{x^2}{2}\right)\ dx$ = Area under N(0,1) between –a and a.

a	r(a)	Tail : 1 – r(a)
1	0.683	0.317
2	0.955	0.045
3	0.997	0.003



Tail probability of N(0,I) in R^d

- Probability that lies outside a sphere of radius 1.

d	1	2	5	10	20	100
P	0.317	0.1353	0.5494	0.9473	0.999	1.0

- N(0, I) still attains its maximum at x =0.

- For large d, tail has more information.

- Probability of N(0,I) contained in a thin annulus around $\|x\|^2 = d$

P[$\sqrt{d}$ - β ≤ $\|x\|^2$ ≤ $\sqrt{d}$ + β] ≥ 1-3 $e^{-\propto\beta^2}$, where β < $\sqrt{d}$ and α > 0 is a constant.

Chi- square distribution of $\|x\|^2$

- Let x ∈ R^k , x_i ~ i.i.d. N(0,1) for 1≤ i ≤ k.

- Y = $\|x\|^2$ = $\sum_{i=1}^{k} x_i^2$ - chi-square distributed with k degrees of freedom given by

- f_Y(y) = $\frac{1}{2^{k/2}\,\Gamma(\frac{k}{2})}\, y^{\frac{k}{2}-1}\, e^{\frac{-y}{2}}$ -------------------------------------> (15)

- Mean of Y = E[$\|x\|^2$] = k -------------------------------------> (16)

- Var of Y = VAR($\|x\|^2$) = 2k -------------------------------------> (16)

Chi- distribution of ||x||

- Let Z = ||x||

- Z said to chi-distributed

f_Z(z) = $\frac{1}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})}\, z^{k-1}\, e^{\frac{-z^2}{2}}$ ---> (17)

- Mean of z = E[||x||] = $\sqrt{2}\,\frac{\Gamma(\frac{k}{2}+1)}{\Gamma(\frac{k}{2})}$ -----------------------------------> (18)

- Var of Z = k - μ^2 ---> (18)

Properties of ||x|| : concentration of ||x||

- Setting n =k +1.

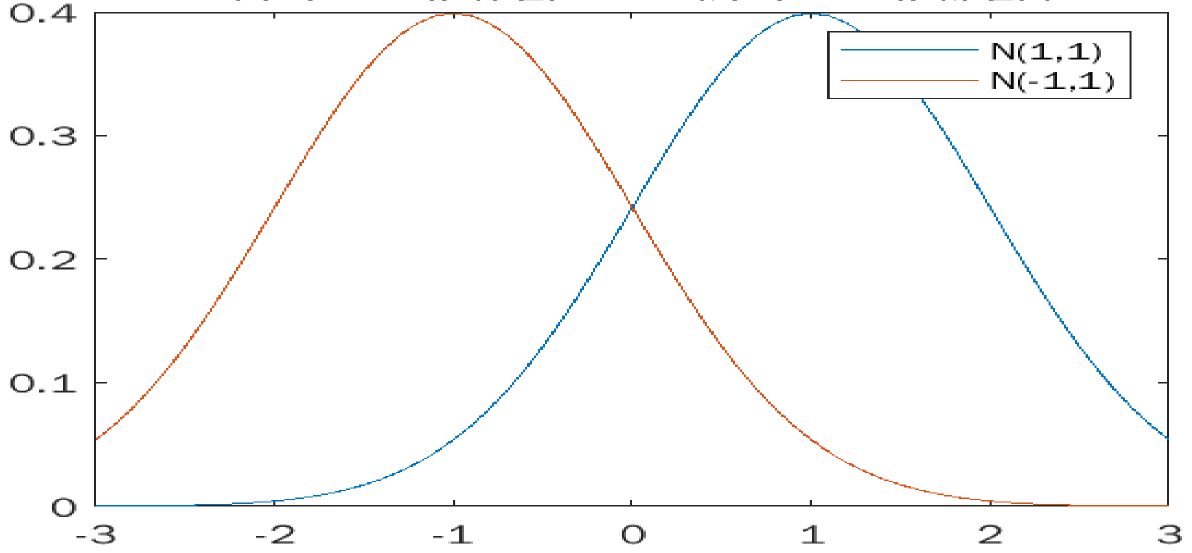
- E[||x||] = $\sqrt{n-1}$ [1 - $\frac{1}{4n}$]

- Var (||x||) = $\frac{n-1}{2n} \approx \frac{1}{2}$ ---> (19)

k	n	E[||x||]	Var (||x||)
10	11	3.09	0.4545
50	51	7.106	0.4902
100	101	10.035	0.4905
500	501	22.35	0.4995

Impact of high dimension in statistics: Linear discriminant analysis : Population based analysis

- Two Gaussian distribution $P_1(x)$ = N (μ_1, Σ) and $P_2(x)$ = N (μ_2, Σ) , $x \in R^d$.



- Mixture : P(x) = $p_1\ P_1(x) + p_2\ P_2(x)$, $p_1 > 0$ and $p_1 + p_2 = 1$.

- A sample is drawn from P(x) and need to identify which class it belongs to.

Standard Algorithm

- Compute L = log $(\frac{P_2(x)}{P_1(x)})$

- L = Ψ (x) = <μ_2 - μ_1, $\sum^{-1}$ (x - $\frac{\mu_2 + \mu_1}{2}$) > ---------------------------> (20)

- Linear statistic.

- Optimum decision rule is based on thresholding Ψ (x).

- When μ_1 = 1 and μ_2 = -1: T =0 is a good threshold.

Error probability

- Set p_1 = p_2 = ½
- Error (Ψ) = ½ [P_1[Ψ (x’) ≤ 0] + P_2[Ψ (x”) > 0]]
- x’ and x’’ are drawn from $P_1(x)$ and $P_2(x)$.
- Error (Ψ) = $\frac{1}{\sqrt{2\Pi}} \int_{-\infty}^{-\frac{r}{2}} e^{-\frac{t^2}{2}}$dt = ϕ (- $\frac{r}{2}$) ---------------------------------------> (21)
- r^2 = ($\mu_1 - \mu_2$) $\sum^{-1}$ ($\mu_1 - \mu_2$) : Mahalanobis Distance.

Sample Counterpart

- We do not know the conditional distributions.

- Given a set of labelled samples: $\{x_1, x_2, \ldots., x_{n_1}\}$ from $P_1(x)$, $\{x_{n_1+1}, x_{n_1+2}, \ldots., x_{n_1+n_2}\}$ from $P_2(x)$

- Sample mean : $\widehat{\mu_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i$ and $\widehat{\mu_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{i+n_1}$

- Pooled sample covariance:
- $\widehat{\Sigma} = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i - \widehat{\mu_1})(x_i - \mu_1)^T + \frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_{i+n_1} - \widehat{\mu_2})(x_{i+n_1} - \mu_2)^T$

Fisher's Linear discriminant function

- $\hat{\Psi}$ (x) = < $\widehat{\mu_1}$ - $\widehat{\mu_2}$, $\widehat{\sum}^{-1}$ (x - $\frac{\widehat{\mu_1} + \widehat{\mu_2}}{2}$) > --> (22)

- Assume n_i > d and $\widehat{\sum}$ is invertible.

- Error ($\hat{\Psi}$) = ½ [P_1[$\hat{\Psi}$ (x') ≤ 0] + P_2[$\hat{\Psi}$ (x'') > 0]] ------------------------------> (23)

where x' and x'' are samples from $P_1(x)$ and $P_2(x)$.

Kolmogorov's analysis (1960's)

- Assume $\sum$ = I and $\widehat{\Psi}_{id}$(x) = < $\widehat{\mu_1}$ - $\widehat{\mu_2}$, x - $\frac{\widehat{\mu_1} + \widehat{\mu_2}}{2}$ >

- When (n_1 = n_2 , d) and grow with out bound with ratios $\frac{d}{n}$ -> α > 0.

- Let || $\widehat{\mu_1} - \widehat{\mu_2}$|| -> a constant ϒ > 0.

Kolmogorov's Analysis Continued

- In this scaling:

Error ($\widehat{\Psi}_{id}$) -> ϕ (- $\frac{r^2}{2\sqrt{r^2+\alpha}}$) in probability ----------------------------------> (24)

- Since $\frac{r^2}{2\sqrt{r^2+\alpha}} < \frac{r}{2}$, Error ($\widehat{\Psi}_{id}$) is larger than when α = 0.

- Clear demonstration of high- dimensional effect and resulting sub optimality.

- When $\frac{d}{n}$ = α = 0 , we get the classical asymptotic result.

Covariance estimation: Effect of high dimension

- Let $\{x_1, x_2, \ldots., x_n\}$ be an i.i.d samples from a distribution with zero mean where $x_i \in R^d$.
- That is, we have n points chosen at random in R^d .
- Let x = $\{x_1, x_2, \ldots., x_n\} \in R^{d \times n}$ – Data matrix.

Estimate Covariance matrix

- Sample Covariance : $\widehat{\sum}$ = $\frac{1}{n}\sum_{i=1}^{n} x_i \, x_i{}^T$ = $\frac{1}{n} x x^T$ ∈ R$^{d \times d}$

- $\widehat{\sum}$ is unbiased : E ($\widehat{\sum}$) = $\sum$.

- $\widehat{\sum}$ -> $\sum$, the population covariance as n -> ∞ when d is fixed – classical convergence.

Measure of distance between $\widehat{\Sigma}$ and Σ

- Matrix norm – spectral norm, can be used $\left\|\widehat{\Sigma} - \Sigma\right\|_2$ = $sup_{\|u\|_2=1} \left\|\left(\widehat{\Sigma} - \Sigma\right)u\right\|_2$ ---------------> (25)

- It can be proved : $\left\|\widehat{\Sigma} - \Sigma\right\|_2$ -> 0 and n -> ∞ .

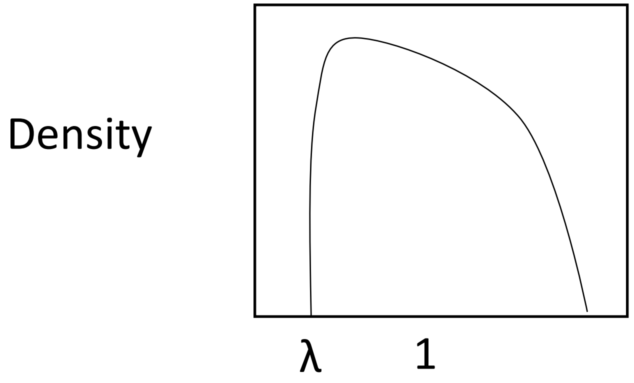
- That is, sample covariance is strongly consistent estimate of Σ is classical setting.

High dimensional effect

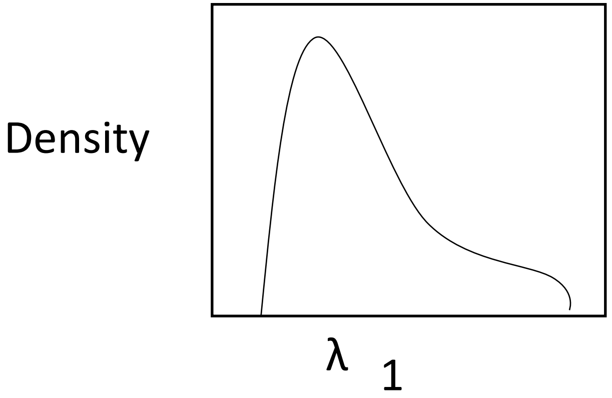
- Let n and d grow, but $\frac{d}{n}$ = α ∈ (0,1).
- Estimate $\widehat{\sum}$ and compute its spectrum.
- Let λ_{max} ($\widehat{\sum}$) = λ_1 ≥ λ_2 ≥ …. ≥ λ_d = λ_{min} ($\widehat{\sum}$) ≥ 0.

Special case $\sum$ = I

- In this special case when $\frac{d}{n}$ = α ∈ (0,1) eigen values λ_i are all dispersed around 1.



n = 4000, d = 800, α = 0.2



n = 4000, d = 2000, α = 0.5

- Empirical distribution of λ 's for α = 0.2 and 0.5.

Marcenko – Pastur law (1967) : Impact of High - dimension

- M-P law : They proved that the density of distribution of λ 's is supported on the interval [t_{min} (α), t_{max} (α)] where t_{min} (α) = $(1\ -\sqrt{\alpha})^2$ and t_{max} (α) = $(1+\sqrt{\alpha})^2$.

- This law allows (d, n) to increase but $\frac{d}{n}$ = α ∈ (0,1) - has a non – classical flavor.

References

[1] M.J. Wainwright (2019) High- Dimensional Statistics: a non – asymptotic viewpoint, Cambridge University Press.

[2] R. Vershynin (2020) High- Dimensional Probability: An introduction with application in Data Science, Cambridge University Press.

[3] A. Blum, J. Hopcroft and R. Kannan (2020) Foundation of Data Science, Cambridge University Press.

[4] V.Shikhman and D. Mueller (2021) Mathematical Foundation of Big Data Analysis, Springer.

[5] J.Wang (2012) Geometric Structure of High- Dimensional Data and Dimensionality reduction, Springer.