A Mathematical Introduction to Data Science

Lecture 1: PCA/MDS and High Dimensionality

Yuan Yao

Peking University

2015.5.5.

Short Course Information

- Instructor: 姚远
- Email: yuany@math.pku.edu.cn
- PKU-Course Website: http://www.math.pku.edu.cn/teachers/yaoy/Fall2014
- Ebanshu public lectures: http://www.ebanshu.com/
- Time & Venue:
  - Lecture 1, 2, 3: May 5, 6, 7; 3-5pm, 107 Hung-Ching Bldg, NCU

Main Content

- Lecture 1: Sample mean and Covariance (PCA/MDS): Fisher's Principle of Maximum Likelihood Estimate, yet things might go wrong --
  - Stein's phenomenon and shrinkage
  - Random matrix theory and failure of PCA
- Lecture 2: Generalized PCA/MDS
  - Random projections and compressed sensing
  - PCA/MDS with uncertainty
  - Nonlinear manifold learning
- Lecture 3: Topological and geometric structures of data
  - From graphs to simplicial complexes
  - Persistent homology, exterior calculus, cohomology, etc.

Data Representation: Geometric Embedding

- Data Science is the study of generalizable extraction of Knowledge from data [Wikipedia]
- A fundamental problem is data representation
- “unstructured data” => Euclidean space
- a.k.a. “feature learning” (e.g. kernel method, deep learning)
- speech, text, image, video …
- Sparsity structure lies in the core of high dimensional data analysis
  - Low dimensional vector spaces
  - Low rank matrices, tensors, etc.

Multidimensional Scaling

- Given pairwise distances between data points, can we find a system of Euclidean coordinates for those points whose pairwise distances meet given constraints?

		1 BOST	2 NY	3 DC	4 MIAM	5 CHIC	6 SEAT	7 SF	8 LA	9 DENV
1	BOSTON	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	2786	1771
3	DC	429	233	0	1075	671	2684	2799	2631	1616
4	MIAMI	1504	1308	1075	0	1329	3273	3053	2687	2037
5	CHICAGO	963	802	671	1329	0	2013	2142	2054	996
6	SEATTLE	2976	2815	2684	3273	2013	0	808	1131	1307
7	SF	3095	2934	2799	3053	2142	808	0	379	1235
8	LA	2979	2786	2631	2687	2054	1131	379	0	1059
9	DENVER	1949	1771	1616	2037	996	1307	1235	1059	0

Algorithm 1: Classical MDS Algorithm

Input: A squared distance matrix $D^{n\times n}$ with $D_{ij} = d_{ij}^2$.

Output: Euclidean k-dimensional coordinates $\widetilde{X}_k \in \mathbb{R}^{k\times n}$ of data.

1 Compute $B = -\frac{1}{2} H \cdot D \cdot H^T$, where H is a centering matrix.

2 Compute Eigenvalue decomposition $B = U\Lambda U^T$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$;

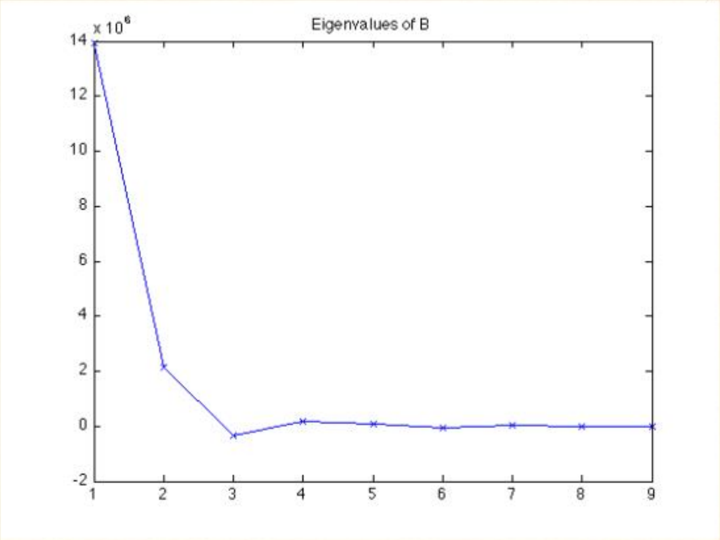
3 Choose top k nonzero eigenvalues and corresponding eigenvectors, $\widetilde{X}_k = U_k \Lambda_k{}^{\frac{1}{2}}$ where

$$U_k = [u_1, \ldots, u_k], \quad u_k \in \mathbb{R}^n,$$
$$\Lambda_k = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$$

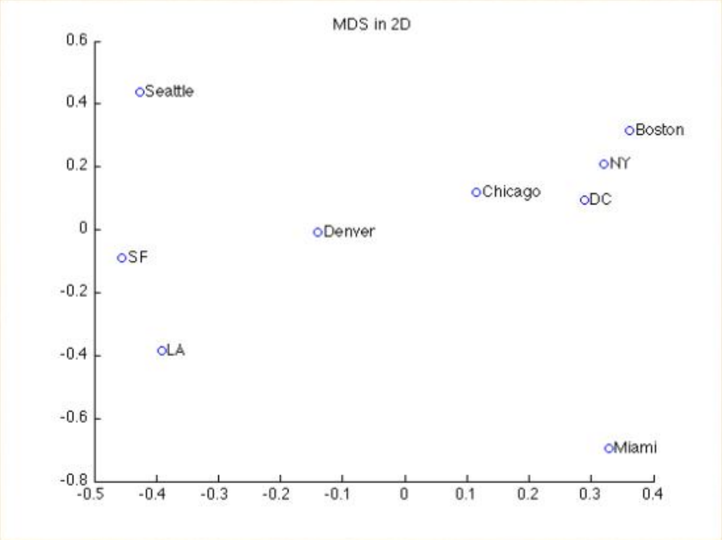
with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$.

		1 BOST	2 NY	3 DC	4 MIAM	5 CHIC	6 SEAT	7 SF	8 LA	9 DENV
1	BOSTON	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	2786	1771
3	DC	429	233	0	1075	671	2684	2799	2631	1616
4	MIAMI	1504	1308	1075	0	1329	3273	3053	2687	2037
5	CHICAGO	963	802	671	1329	0	2013	2142	2054	996
6	SEATTLE	2976	2815	2684	3273	2013	0	808	1131	1307
7	SF	3095	2934	2799	3053	2142	808	0	379	1235
8	LA	2979	2786	2631	2687	2054	1131	379	0	1059
9	DENVER	1949	1771	1616	2037	996	1307	1235	1059	0

(a)



(b)



(c)

Inverse problem: D->X?

Given a set of points $x_1, x_2, ..., x_n \in \mathbb{R}^p$, let

$$X = [x_1, x_2, ..., x_n]^{p \times n}.$$

The distance between point x_i and x_j is

$$d_{ij}^2 = \|x_i - x_j\|^2 = (x_i - x_j)^T (x_i - x_j) = {x_i}^T x_i + {x_j}^T x_j - 2{x_i}^T x_j.$$

General ideas of classic (metric) MDS is:

(1) transform squared distance matrix D to an inner product form;
(2) compute the eigen-decomposition for this inner product form.

Below we shall see how to do this given D.

From inner product to squared distance

Let K be the inner product matrix

$$K = X^T X,$$

with $k = \mathrm{diag}(K_{ii}) \in \mathbb{R}^n$. So

$$D = (d_{ij}^2) = k \cdot \mathbf{1}^T + \mathbf{1} \cdot k^T - 2K.$$

where $\mathbf{1} = (1, 1, ..., 1)^T \in \mathbb{R}^n$.

Centered the data

Define the mean and the centered data

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\cdot X \cdot \mathbf{1},$$

$$\widetilde{x}_i = x_i - \widehat{\mu}_n = x_i - \frac{1}{n}\cdot X \cdot \mathbf{1},$$

$$\widetilde{X} = X - \frac{1}{n} X \cdot \mathbf{1} \cdot \mathbf{1}^T.$$

$$\begin{aligned}
\tilde{K} &\triangleq \tilde{X}^T \tilde{X} \\
&= \left(X - \frac{1}{n} X \cdot \mathbf{1} \cdot \mathbf{1}^T\right)^T \left(X - \frac{1}{n} X \cdot \mathbf{1} \cdot \mathbf{1}^T\right) \\
&= K - \frac{1}{n} K \cdot \mathbf{1} \cdot \mathbf{1}^T - \frac{1}{n} \mathbf{1} \cdot \mathbf{1}^T \cdot K + \frac{1}{n^2} \cdot \mathbf{1} \cdot \mathbf{1}^T \cdot K \cdot \mathbf{1} \cdot \mathbf{1}^T
\end{aligned}$$

Let

$$B = -\frac{1}{2} H \cdot D \cdot H^T$$

where $H = I - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^T$. H is called as a *centering matrix.*

$$B = -\frac{1}{2} H \cdot (k \cdot \mathbf{1}^T + \mathbf{1} \cdot k^T - 2K) \cdot H^T$$

Since $k \cdot \mathbf{1}^T \cdot H^T = k \cdot \mathbf{1}(I - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^T) = k \cdot \mathbf{1} - k(\frac{\mathbf{1}^T \cdot \mathbf{1}}{n}) \cdot \mathbf{1} = 0$, we have $H \cdot k\, \mathbf{1} \cdot H^T = H \cdot \mathbf{1} \cdot k^T \cdot H^T = 0$.

Therefore,

$$\begin{aligned} B &= H \cdot K \cdot H^T = (I - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^T) \cdot K \cdot (I - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^T) \\ &= K - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1} \cdot K - \frac{1}{n} \cdot K \cdot \mathbf{1} \cdot \mathbf{1}^T + \frac{1}{n^2} \cdot \mathbf{1}(\mathbf{1}^T \cdot K\mathbf{1}) \cdot \mathbf{1}^T \\ &= \tilde{K}. \end{aligned}$$

Inner product matrix!

$$B = -\frac{1}{2} H \cdot D \cdot H^T = \tilde{X}^T \tilde{X}.$$

Note that often we define the covariance matrix

$$\widehat{\Sigma}_n \triangleq \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \widehat{\mu}_n)(x_i - \widehat{\mu}_n)^T = \frac{1}{n-1} \widetilde{X} \widetilde{X}^T.$$

PCA

- PCA is given by the top k eigenvector of covariance matrix

$$\widehat{\Sigma}_n = \frac{1}{n-1} \widetilde{X} \cdot \widetilde{X}^T$$

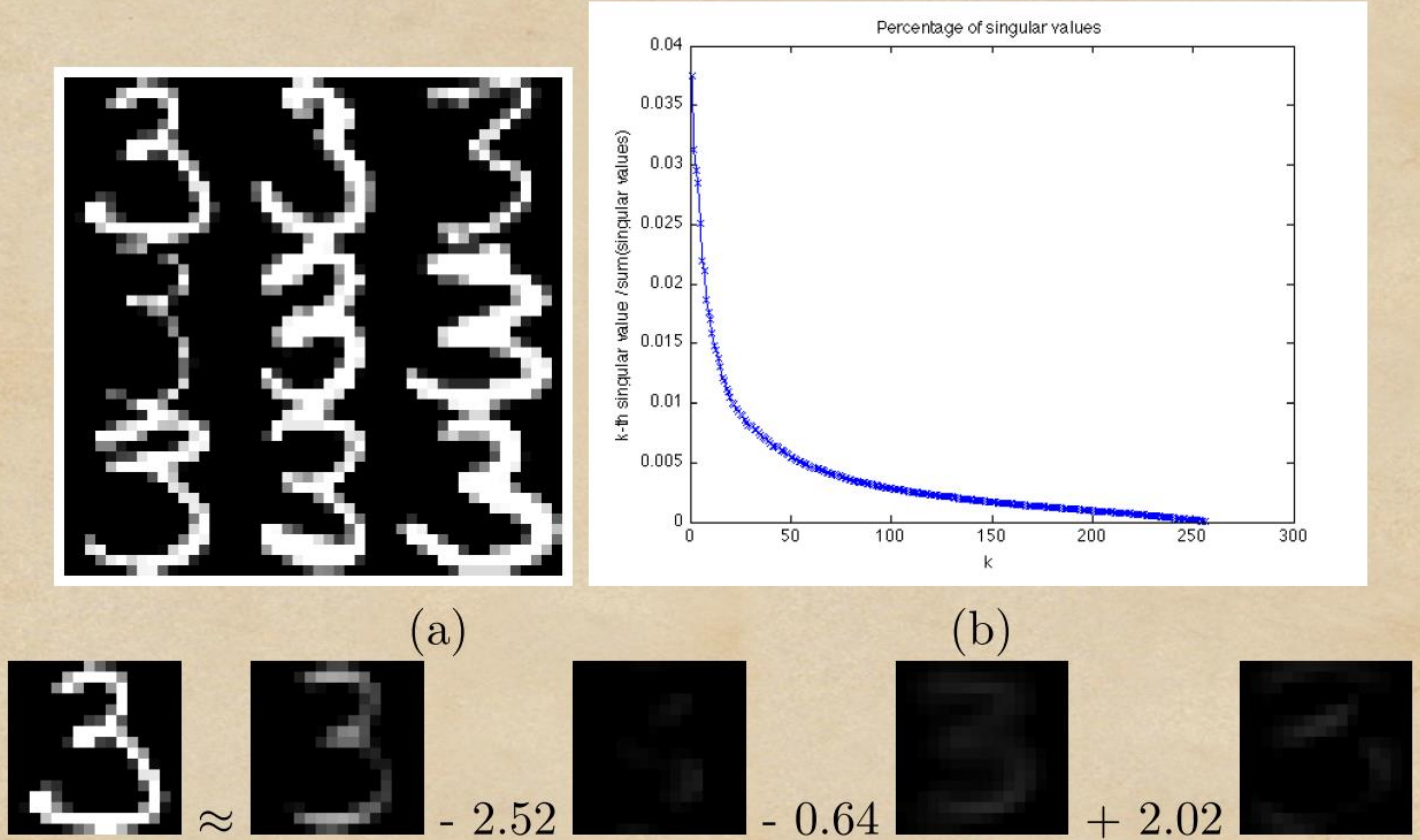
Both MDS and PCA are given by SVD of centered data matrix.

MDS and PCA≈SVD

(SVD) of $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n\times p}$ in the following sense,

$$Y = X - \frac{1}{n}\mathbf{1}\mathbf{1}^T X = \tilde{U}\tilde{S}\tilde{V}^T, \quad \mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$$

- top k left singular vectors give MDS (Kernel spectrum)
- top k right singular vectors give PCA (Covariance spectrum)



Percentage of singular values
k-th singular value /sum(singular values)
k
(a)
(b)
≈ - 2.52 - 0.64 + 2.02

(c)

Any principles underlying these tricks?

- Dimensionality reduction formed in geometry and optimization
- Statistics: Fisher’s maximal likelihood estimate

Dimensionality Reduction

Find low dimensional embedding

$$\min_{Y_i \in \mathbb{R}^k} \sum_{i,j} (\|Y_i - Y_j\|^2 - d_{ij}^2)^2$$

take the derivative w.r.t $Y_i \in \mathbb{R}^k$:

$$\sum_{i,j}(\|Y_i\|^2 + \|Y_j\|^2 - 2Y_i^T Y_j - d_{ij}^2)(Y_i - Y_j) = 0$$

which implies $\sum_i Y_i = \sum_j Y_j$. For simplicity set $\sum_i Y_i = 0$, *i.e.*putting the origin as data center.

Use a linear transformation to move the sample mean to be the origin of the coordinates, *i.e.* define a matrix $B_{ij} = -\frac{1}{2}HDH$ where $D = (d_{ij}^2)$, $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, then, the minimization (1) is equivalent to find $Y_i \in \mathbb{R}^k$:

$$\min \|Y^T Y - B\|_F^2$$

then the row vectors of matrix Y are the eigenvectors corresponding to k largest eigenvalues of $B = \widetilde{X}^T \widetilde{X}$, or equivalently the top k *right singular vectors* of $\widetilde{X} = USV^T$.

B is Gram matrix or kernel matrix

Geometry of PCA

$$\text{Let } X = [X_1|X_2|\cdots|X_n] \in \mathbb{R}^{p\times n}.$$

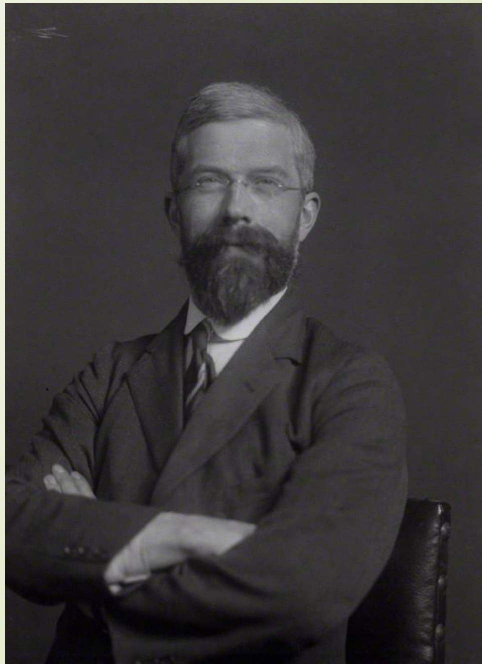
$$\min_{\beta,\mu,U} I := \sum_{i=1}^{n} \|X_i - (\mu + U\beta_i)\|^2 \qquad (2)$$

where $U \in \mathbb{R}^{p\times k}$, $U^T U = I_p$, and $\sum_{i=1}^{n} \beta_i = 0$ (nonzero sum of β_i can be repre-

Best k-affine space approximation of data

Fisher’s Principle of Maximum Likelihood Estimate

Ronald Fisher: On the Mathematical Foundations of Theoretical Statistics, 1921



Consider the statistical model $f(X|\theta)$ as a conditional probability function on $\mathbb{R}^p$ with parameter space $\theta \in \Theta$. Let $X_1, ..., X_n \in \mathbb{R}^p$ are independently and identically distributed (i.i.d.) sampled according to $f(X|\theta_0)$ on $\mathbb{R}^p$ for some $\theta_0 \in \Theta$. The likelihood function is defined as the probability of observing the given data as a function of θ,

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta),$$

and a maximum likelihood estimator is defined as

$$\hat{\theta}_n^{MLE} \in \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} f(X_i|\theta)$$

which is equivalent to

$$\arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log f(X_i|\theta).$$

1.1. Example: Multivariate Normal Distribution. For example, consider the normal distribution $\mathcal{N}(\mu, \Sigma)$,

$$f(X|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left[-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu)\right],$$

where $|\Sigma|$ is the determinant of covariance matrix Σ.

$$\max_{\mu,\Sigma} P(X_1, ..., X_n|\mu, \Sigma) = \max_{\mu,\Sigma} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi|\Sigma|}} \exp[-(X_i-\mu)^T \Sigma^{-1} (X_i-\mu)]$$

$$\Rightarrow \mu^* = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\mu}_n$$

$$\Sigma^* = \frac{n-1}{n}\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^T$$

$$\hat{\mu}_n^{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\Sigma}_n^{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^T.$$

Asymptotic properties of MLE

A. (Consistency) $\hat{\theta}_n^{MLE} \to \theta_0$, in probability and almost surely.

B. (Asymptotic Normality) $\sqrt{n}(\hat{\theta}_n^{MLE} - \theta_0) \to \mathcal{N}(0, I_0^{-1})$ in distribution, where I_0 is the Fisher Information matrix

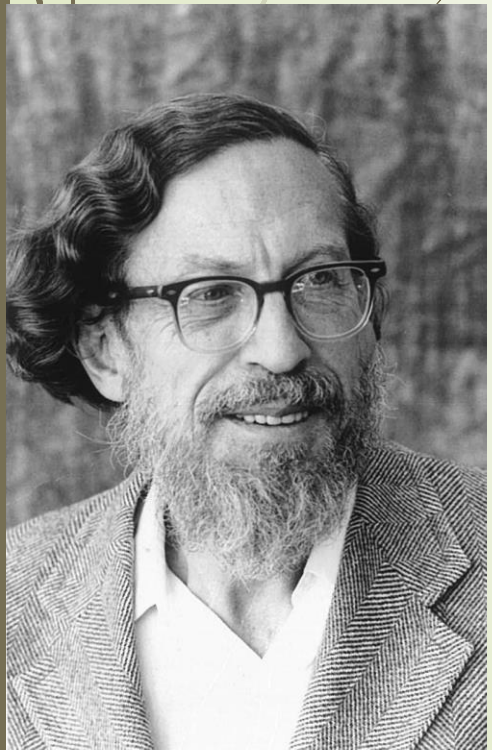
$$I(\theta_0) := \mathbb{E}[(\frac{\partial}{\partial\theta} \log f(X|\theta_0))^2] = -\mathbb{E}[\frac{\partial^2}{\partial\theta^2} \log f(X|\theta_0)].$$

C. (Asymptotic Efficiency) $\lim_{n\to\infty} \mathrm{cov}(\hat{\theta}_n^{MLE}) = I^{-1}(\theta_0)$. Hence $\hat{\theta}_n^{MLE}$ is the Uniformly Minimum-Variance Unbiased Estimator, i.e. the estimator with the least variance among the class of unbiased estimators, for any unbiased estimator $\hat{\theta}_n$, $\lim_{n\to\infty} \mathrm{var}(\hat{\theta}_n^{MLE}) \leq \lim_{n\to\infty} \mathrm{var}(\hat{\theta}_n^{MLE})$.

- Yet with a finite sample, MLE might not be good!

Two curses of high dimensionality for MLE

- Stein's Phenomenon: sample mean may have more 'risks' (mean square error) than some biased estimators (e.g. shrinkage)
- Random Matrix Theory: PCA via sample covariance might tell you nothing more than noise



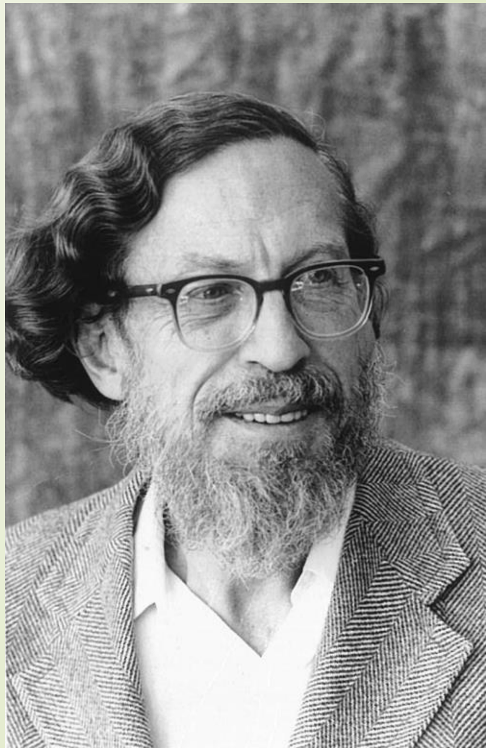
Stein's Phenomenon

Risk of Estimators and Inadmissibility

Linear Estimators (MLE is a special case)

James-Stein Estimators

Soft- and Hard-Thresholding Estimators



Risk and Bias-Variance Decomposition

To measure the performance of an estimator $\hat{\mu}_n$, one may look at the following so-called *risk*,

$$R(\hat{\mu}_n, \mu) = \mathbb{E}L(\hat{\mu}_n, \mu)$$

where the loss function takes the square loss here

$$L(\hat{\mu}_n, \mu) = \|\hat{\mu}_n - \mu\|^2.$$

The mean square error (MSE) to measure the risk enjoys the following *bias-variance decomposition*, from the Pythagorean theorem.

$$\begin{aligned} R(\hat{\mu}_n, \mu) &= \mathbb{E}\|\hat{\mu}_n - \mathbb{E}[\hat{\mu}_n] + \mathbb{E}[\hat{\mu}_n] - \mu\|^2 \\ &= \mathbb{E}\|\hat{\mu}_n - \mathbb{E}[\hat{\mu}_n]\|^2 + \|\mathbb{E}[\hat{\mu}_n] - \mu\|^2 \\ &=: Var(\hat{\mu}_n) + Bias(\hat{\mu}_n)^2 \end{aligned}$$

Example 1. For the simple case $Y_i \sim \mathcal{N}(\mu, \sigma^2 I_p)$ $(i = 1, \dots, n)$, the MLE estimator satisfies

$$Bias(\hat{\mu}_n^{MLE}) = 0$$

and

$$Var(\hat{\mu}_n^{MLE}) = \frac{p}{n}\sigma^2$$

In particular for $n = 1$, $Var(\hat{\mu}^{MLE}) = \sigma^2 p$ for $\hat{\mu}^{MLE} = Y$.

Example 2. MSE of Linear Estimators. Consider $Y \sim \mathcal{N}(\mu, \sigma^2 I_p)$ and linear estimator $\hat{\mu}_C = CY$. Then we have

$$Bias(\hat{\mu}_C) = \|(I - C)\mu\|^2$$

and

$$Var(\hat{\mu}_C) = \mathbb{E}[(CY - C\mu)^T(CY - C\mu)] = \mathbb{E}[\mathrm{trace}((Y-\mu)^T C^T C(Y-\mu))] = \sigma^2 \mathrm{trace}(C^T C).$$

In applications, one often consider the diagonal linear estimators $C = \mathrm{diag}(c_i)$, e.g. in Ridge regression

$$\min_{\mu} \frac{1}{2}\|Y - X\beta\|^2 + \frac{\lambda}{2}\|\beta\|^2.$$

For diagonal linear estimators, the risk

$$R(\hat{\mu}_C, \mu) = \sigma^2 \sum_{i=1}^{p} c_i^2 + \sum_{i=1}^{p} (1 - c_i)^2 \mu_i^2.$$

$$\inf_{c_i} \sup_{|\mu_i| \le \tau_i} R(\hat{\mu}_C, \mu) = \sum_{i=1}^{p} \frac{\sigma^2 \tau_i^2}{\sigma^2 + \tau_i^2}.$$

From here one can see that for those sparse model classes such that $\#\{i : \tau_i = O(\sigma)\} = k \ll p$, it is possible to get smaller risk using linear estimators than MLE.

MLE (sample mean + covariance) might not be the best!

In general, is it possible to introduce some *biased* estimators which significantly reduces the *variance* such that the total risk is smaller than MLE uniformly for all μ? This is the notion of inadmissibility introduced by Charles Stein in 1956 and he find the answer is YES by presenting the James-Stein estimators, as the shrinkage of sample means.

Definition (Inadmissible)**.** An estimator $\hat{\mu}_n$ of the parameter μ is called **inadmissible** on $\mathbb{R}^p$ with respect to the squared risk if there exists another estimator μ_n^* such that

$$\mathbb{E}\|\mu_n^* - \mu\|^2 \leq \mathbb{E}\|\hat{\mu}_n - \mu\|^2 \qquad \text{for all } \mu \in \mathbb{R}^p,$$

and there exist $\mu_0 \in \mathbb{R}^p$ such that

$$\mathbb{E}\|\mu_n^* - \mu_0\|^2 < \mathbb{E}\|\hat{\mu}_n - \mu_0\|^2.$$

In this case, we also call that μ_n^* **dominates** $\hat{\mu}_n$. Otherwise, the estimator $\hat{\mu}_n$ is called **admissible**.

Stein (1956) [Ste56] found that if $p \geq 3$, then the MLE estimator $\hat{\mu}_n$ is inadmissible. This property is known as ***Stein's phenomenon***. This phenomenon can be described like:

For $p \geq 3$, there exists $\hat{\mu}$ such that $\forall \mu$,

$$R(\hat{\mu}, \mu) < R(\hat{\mu}^{\mathrm{MLE}}, \mu)$$

which makes MLE inadmissible.

A typical choice is the *James-Stein estimator* given by James-Stein (1961),

$$\tilde{\mu}_n^{JS} = \left(1 - \frac{\sigma^2(p-2)}{\|\hat{\mu}_n^{MLE}\|}\right) \hat{\mu}_n^{MLE}, \qquad \sigma = \varepsilon.$$

- E.g. assume the sample has unit-variance $Y \sim N(\mu, I)$

$$R(\hat{\mu}^{\mathrm{JS}}, \mu) = \mathbb{E}U(Y) = p - \mathbb{E}_\mu \frac{(p-2)^2}{\|Y\|^2} < p = R(\hat{\mu}^{\mathrm{MLE}}, \mu)$$

Linear Estimators are mostly inadmissible

$$Y \sim \mathcal{N}(\mu, \sigma^2 I)$$

$$\hat{\mu}_C(Y) = Cy$$

$$R(\hat{\mu}_C, \mu) = \|(I - C(\lambda))Y\|^2 - p\sigma^2 + 2\sigma^2 \mathrm{trace}(C(\lambda))$$

Theorem 3.3 (Lemma 2.8 in Johnstone's book (GE)). $Y \sim N(\mu, I)$, $\forall \hat{\mu} = CY$, $\hat{\mu}$ is admissable iff

(1) C is symmetric.
(2) $0 \leq \rho_i(C) \leq 1$ (eigenvalue).
(3) $\rho_i(C) = 1$ for at most two i.

Stein's Unbiased Risk Estimate (SURE)

- W.L.G. assume unit variance, $Y \sim N(\mu, I)$

Lemma 3.2. (Stein's Unbiased Risk Estimates (SURE)) Suppose $\hat{\mu} = Y + g(Y)$, g satisfies [1]

(1) g is weakly differentiable.
(2) $\sum_{i=1}^{p} \int |\partial_i g_i(x)| \mathrm{d}x < \infty$

then

$$R(\hat{\mu}, \mu) = \mathbb{E}_{\mu}(p + 2\nabla^T g(Y) + \|g(Y)\|^2) \tag{18}$$

where $\nabla^T g(Y) := \sum_{i=1}^{p} \frac{\partial}{\partial y_i} g_i(Y)$.

Examples of $g(x)$: For James-Stein estimator

$$g(x) = -\frac{p-2}{\|Y\|^2} Y$$

and for soft-thresholding, each component

$$g_i(x) = \begin{cases} -\lambda & x_i > \lambda \\ -x_i & |x_i| \leq \lambda \\ \lambda & x_i < -\lambda \end{cases}$$

Both of them are weakly differentiable. But Hard-Thresholding:

$$g_i(x) = \begin{cases} 0 & |x_i| > \lambda \\ -x_i & |x_i| \leq \lambda \end{cases}$$

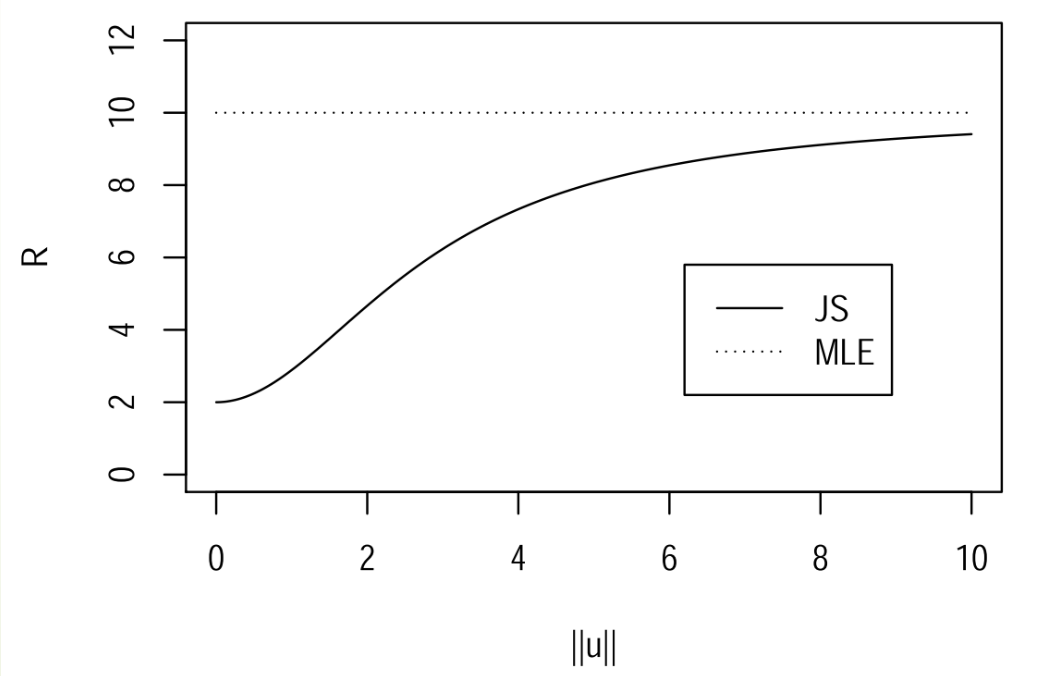
which is not weakly differentiable!

- Note that soft-thresholding solves LASSO while hard-thresholding solves L0-penalization (nonconvex)

Risk of James-Stein Estimator

Proposition 3.4 (Upper bound of MSE for the James-Stein Estimator). $Y \sim \mathcal{N}(\mu, I_p)$,

$$R(\hat{\mu}^{\mathrm{JS}}, \mu) \leq p - \frac{(p-2)^2}{p-2+\|\mu\|^2} = 2 + \frac{(p-2)\|\mu\|^2}{p-2+\|\mu\|^2}$$



Risk of Soft-Thresholding

$$\hat{\mu}(x) = x + g(x). \quad \frac{\partial}{\partial i} g_i(x) = -I(|x_i| \leq \lambda)$$

We then have

$$\begin{aligned} \mathbb{E}_\mu \|\hat{\mu}_\lambda - \mu\|^2 &= \mathbb{E}_\mu \left(p - 2\sum_{i=1}^{p} I(|x_i| \leq \lambda) + \sum_{i=1}^{p} x_i^2 \wedge \lambda^2 \right) \\ &\leq 1 + (2\log p + 1) \sum_{i=1}^{p} \mu_i^2 \wedge 1 \quad \text{if we take } \lambda = \sqrt{2\log p} \end{aligned}$$

By using the inequality

$$\frac{1}{2} a \wedge b \leq \frac{ab}{a+b} \leq a \wedge b$$

we can compare the risk of soft-thresholding and James-Stein estimator as

$$1 + (2\log p + 1) \sum_{i=1}^{p} (\mu_i^2 \wedge 1) \quad \lesseqgtr \quad 2 + c\left(\left(\sum_{i=1}^{p} \mu_i^2 \right) \wedge p \right) \quad c \in (1/2, 1)$$

- Soft is better than JS for sparse signal O(k log p) < O(p)

*Risk Comparison Between Soft- and Hard-Thresholding [Johnstone, GE]

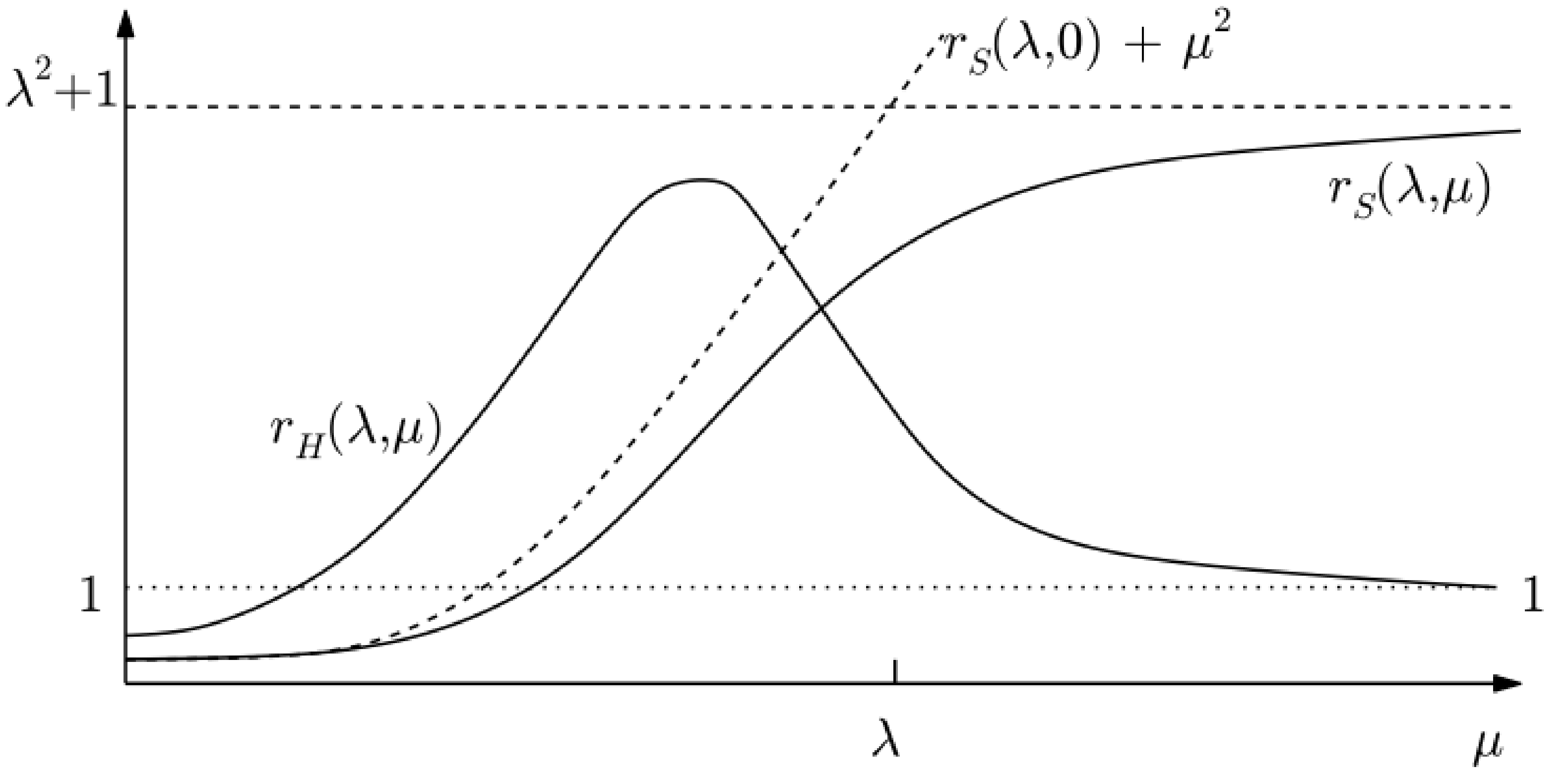


Figure 8.1 Schematic diagram of risk functions of soft and hard thresholding. Dashed lines indicate upper bounds for soft thresholding of Lemma 8.3.

Random Matrix Theory and PCA

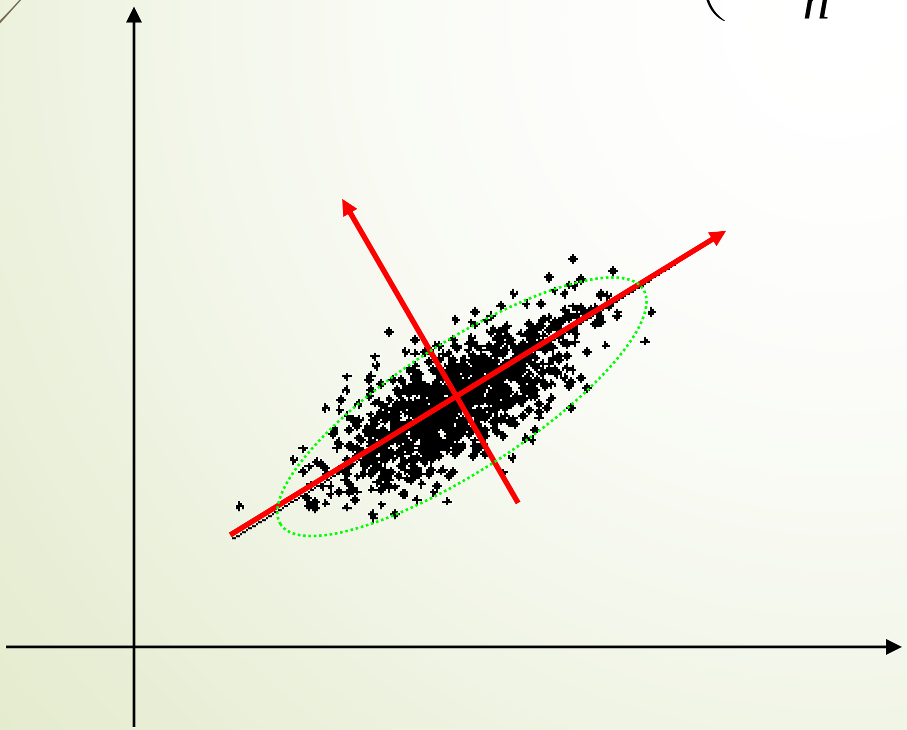
Principal Component Analysis (PCA)

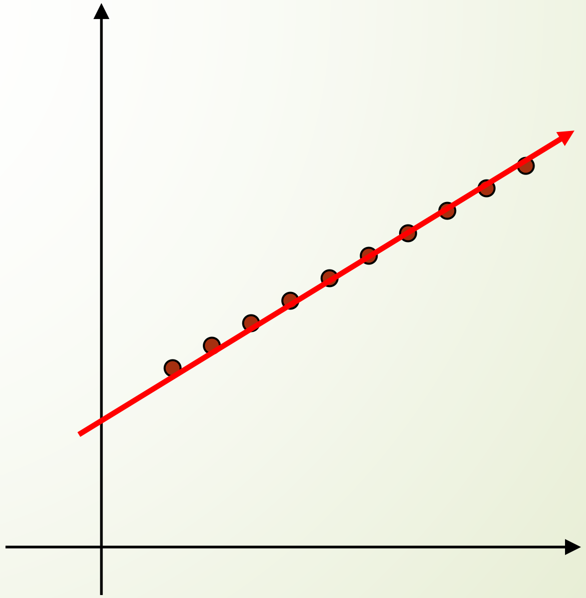
- Principal Component Analysis (PCA)

$$X_{n\times p} = [X_1 \quad X_2 \quad \dots \quad X_p]$$

$$\Sigma_{ij} = [\mathrm{cov}(X_i, X_j)] = E[(X_i - \mu_i)(X_j - \mu_j)]$$

Eigen-decomposition of $\hat{\Sigma} = X^T \left(I - \frac{1}{n} ee^T \right)^2 X \to \Sigma$ for fixed p and n->∞





One Dimensional Manifold

PCA may go wrong if p ≥ n

- For $p/n = \gamma > 0$, assume p-dimensional X_i

$$X_i = \sum_{\nu=1}^{M} \sqrt{\lambda_\nu} u_{\nu i} \theta_\nu + \sigma Z_i, \quad u_{\nu i} \approx N(0,1), \quad Z_i \approx N_p(0, I_p),$$

where θ_ν are orthonormal, then PCA is inconsistent by Random Matrix Theory

$$\langle \hat{\theta}_\nu, \theta_\nu \rangle \to \begin{cases} 0 & \lambda_\nu \in [0, \sqrt{\gamma}] \\ \dfrac{1 - \gamma / \lambda_\nu^2}{1 + \gamma / \lambda_\nu} & \lambda_\nu > \sqrt{\gamma} \end{cases}, \quad \sigma = 1$$

- Phase transition:
  - Below the threshold, estimation is orthogonal to the truth
  - Above the threshold, the angle decreases as eigenvalue grows, but always biased

Johnstone (2006) High Dimensional Statistical Inference and Random Matrices, arxiv.org/abs/math/0611589v1

E.g. Rank-1 model

$$Y = X + \varepsilon,$$

where signal lies in an one-dimensional subspace $X = \alpha u$ with $\alpha \sim \mathcal{N}(0, \sigma_X^2)$ and noise $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I_p})$ is i.i.d. Gaussian. For multi-rank models, please see [KN08]. Therefore $Y \sim \mathcal{N}(0, \Sigma)$ where

$$\Sigma = \sigma_X^2 uu' + \sigma_\varepsilon^2 I_p.$$

- Can we recover signal direction (u) by PCA?

Noise has a spectrum: Marcenko-Pastur Distribution

$X_i \sim \mathcal{N}(0, I_p)$

- A. When p fixed and $n \to \infty$, the classical Law of Large Numbers tells us

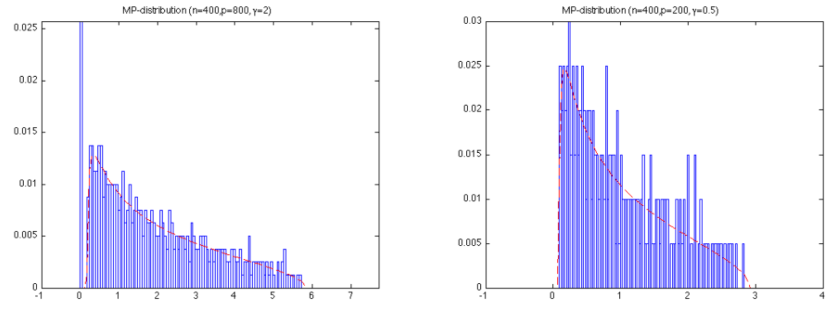
$$\widehat{\Sigma}_n = \frac{1}{n} X X' \to \mathrm{I_p}. \tag{20}$$

- B. But when $\frac{p}{n} \to \gamma \neq 0$, the distribution of the eigenvalues of $\widehat{\Sigma}_n$ follows

if $\gamma \leq 1$,

$$\mu^{MP}(t) = \begin{cases} 0 & t \notin [a, b] \\ \frac{\sqrt{(b-t)(t-a)}}{2\pi\gamma t} dt & t \in [a, b] \end{cases}$$

and has an additional point mass $1 - 1/\gamma$ at the origin if $\gamma > 1$. Note that $a = (1-\sqrt{\gamma})^2, b = (1+\sqrt{\gamma})^2$.



(a) (b)

A Phase Transition in PCA

$$p/n \to \gamma$$

Define the signal-noise ratio $SNR = R = \frac{\sigma_X^2}{\sigma_\varepsilon^2}$, where for simplicity $\sigma_\varepsilon^2 = 1$

$$\lambda_{max}(\widehat{\Sigma}_n) \to \begin{cases} (1+\sqrt{\gamma})^2 = b, & \sigma_X^2 \le \sqrt{\gamma} \\ (1+\sigma_X^2)(1+\frac{\gamma}{\sigma_X^2}), & \sigma_X^2 > \sqrt{\gamma} \end{cases}$$

$$|\langle u, v_{max}\rangle|^2 \to \begin{cases} 0 & \sigma_X^2 \le \sqrt{\gamma} \\ \frac{1-\frac{\gamma}{\sigma_X^4}}{1+\frac{\gamma}{\sigma_X^2}}, & \sigma_X^2 > \sqrt{\gamma} \end{cases}$$

Sparsity plays a central role

4.5. Further Comments. When $\frac{log(p)}{n} \to 0$, we need to add more restrictions on $\widehat{\Sigma}_n$ in order to estimate it faithfully. There are typically three kinds of restrictions.

- Σ sparse
- Σ^{-1} sparse, also called–Precision Matrix
- banded structures (e.g. Toeplitz) on Σ or Σ^{-1}

Sparsity triggers a dawn of the Science for High Dimensional Data Analysis

- When p>>n, PCA may work with additional requirements
  - Σ is sparse or fast decay
  - Σ is low rank
  - Σ^{-1} is sparse
  - Various extensions…
- Geometry and topology begin to enter this Odyssey: data concentrate on
  - low-dimensional subspaces, manifolds …