How close are $\hat{\mu}_n$ and S_n to the population mean and variance

Boaz Nadler

Department of Computer Science and Applied Mathematics
The Weizmann Institute of Science

Dec. 2021

Part 1: Classical Asymptotic Statistics

Reminder

$\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ assumed i.i.d. from r.v. X.

Sample Mean:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_i \mathbf{x}_i$$

Reminder

$\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ assumed i.i.d. from r.v. X.

Sample Mean:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_i \mathbf{x}_i$$

Sample Covariance Matrix:

$$S_n = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{x})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

Reminder

$\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ assumed i.i.d. from r.v. X.

Sample Mean:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_i \mathbf{x}_i$$

Sample Covariance Matrix:

$$S_n = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{x})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

Eigendecomposition / Principal Component Analysis

$$S_n = \sum_j \ell_j \hat{\mathbf{v}}_j \hat{\mathbf{v}}_j^T$$

Classical Asymptotics, p fixed, $n \to \infty$

Reminder: CLT, if x_i all i.i.d. from r.v. $X \in \mathbb{R}^p$ with $\mathbb{E}[X] = \mu$ and $Var[X_i] = \sigma^2 < \infty$, then as $n \to \infty$

$$\sqrt{n}(\bar{\mathbf{x}}_i - \mu_i) \sim \mathcal{N}(0, \sigma^2)$$

Similarly, if X has finite fourth moment, element-wise,

$$(S_n)_{ij} - \Sigma_{ij} = O_P\left(\frac{1}{\sqrt{n}}\right)$$

Main Point: If p fixed, $n \gg p$, $\bar{\mathbf{x}}$ and S_n are accurate estimators of $\boldsymbol{\mu}$ and Σ.

Classical Asymptotics, p fixed, $n \to \infty$

Furthermore, as for eigendecomposition,

$$\ell_j \to \lambda_j \text{ and for eigenvalues with multiplicity one } \hat{\mathbf{v}}_j \to \mathbf{v}_j$$

Theorem: For eigenvalue λ_i of multiplicity one, under mild assumptions on $\mathbf{x}$, as $n \to \infty$, $\ell_i \sim \mathcal{N}(\mu, \sigma^2)$ where

$$\mu = \mathbb{E}[\ell_i] = \lambda_i + \frac{1}{n} \sum_j \frac{\lambda_i \lambda_j}{\lambda_i - \lambda_j} + o\left(\frac{1}{n}\right)$$

$$\sigma^2 = Var[\ell_i] = \frac{2}{n\beta} \lambda_i^2 + o\left(\frac{1}{n}\right)$$

Also,

$$\hat{\mathbf{v}}_j = \mathbf{v}_j + O_P\left(\frac{1}{\sqrt{n}}\right)$$

Asymptotic Eigenvalue Distribution

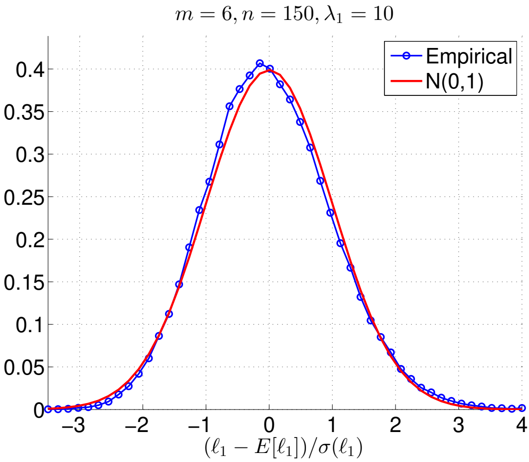
Example: Single signal in noise

$$\Sigma = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T + \sigma^2 I_m$$

Asymptotic Eigenvalue Distribution

Example: Single signal in noise

$$\Sigma = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T + \sigma^2 I_m$$



Eigenvector Asymptotics

Example:

Signal strength λ in noise variance σ^2.

$$\lambda_1 = \lambda + \sigma^2,\ \lambda_j = \sigma^2.$$

Without loss of generality, assume $\mathbf{h} = \mathbf{e}_1$.

Eigenvector Asymptotics

Example:

Signal strength λ in noise variance σ^2.

$$\lambda_1 = \lambda + \sigma^2,\ \lambda_j = \sigma^2.$$

Without loss of generality, assume $\mathbf{h} = \mathbf{e}_1$.
Asymptotically,

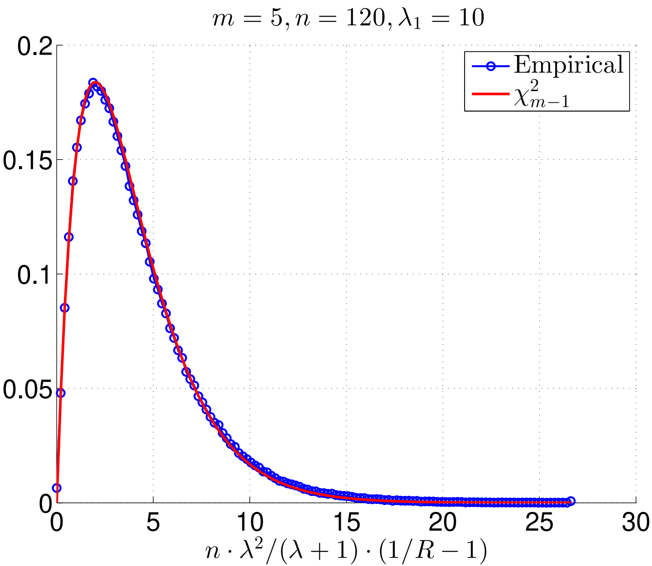
$$\hat{\mathbf{v}}_1 = (1, 0, \ldots, 0) + \frac{\sigma}{\sqrt{n}}\sqrt{\frac{\lambda + \sigma^2}{\lambda^2}}(0, \xi_2, \ldots, \xi_m)$$

Hence

$$R = \left|\langle \frac{\hat{\mathbf{v}}_1}{\|\hat{\mathbf{v}}_1\|}, \mathbf{h}\rangle\right|^2 \approx \frac{1}{1 + \frac{\sigma^2}{n}\frac{\lambda+\sigma^2}{\lambda^2}\chi^2_{m-1}} \approx \frac{1}{1 + \frac{p-1}{n}\frac{\sigma^2}{\lambda}}$$

If $n \gg p$ and $\lambda \gg \sigma^2$, good overlap between first sample and population principal components, $R \approx 1$.

Simulation Example: Eigenvector Spread



Part II: What happens when dimension p is large

when p and n are comparable, or even $p \gg n$?

Modern Asymptotics: Random Matrix Theory

Example: Consider $\mathbf{x}_1, \ldots, \mathbf{x}_n$ all i.i.d. from $\mathcal{N}(0, \mathbf{I}_p)$.
Namely, $\Sigma = \mathbf{I}_p$, all its p eigenvalues are equal $\lambda_j = 1$.

Modern Asymptotics: Random Matrix Theory

Example: Consider $\mathbf{x}_1, \ldots, \mathbf{x}_n$ all i.i.d. from $\mathcal{N}(0, \mathbf{I}_p)$.
Namely, $\Sigma = \mathbf{I}_p$, all its p eigenvalues are equal $\lambda_j = 1$.

How do eigenvalues of S_n look like when p, n are comparable ?

Modern Asymptotics: Random Matrix Theory

Example: Consider $\mathbf{x}_1, \ldots, \mathbf{x}_n$ all i.i.d. from $\mathcal{N}(0, \mathbf{I}_p)$.
Namely, $\Sigma = \mathbf{I}_p$, all its p eigenvalues are equal $\lambda_j = 1$.

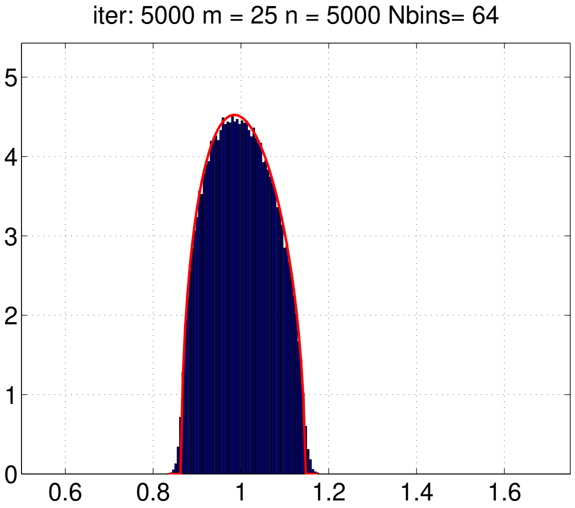
How do eigenvalues of S_n look like when p, n are comparable ?

Note: If $p > n$ then S_n not even invertible !
It has $p - n - 1$ eigenvalues exactly equal to zero !

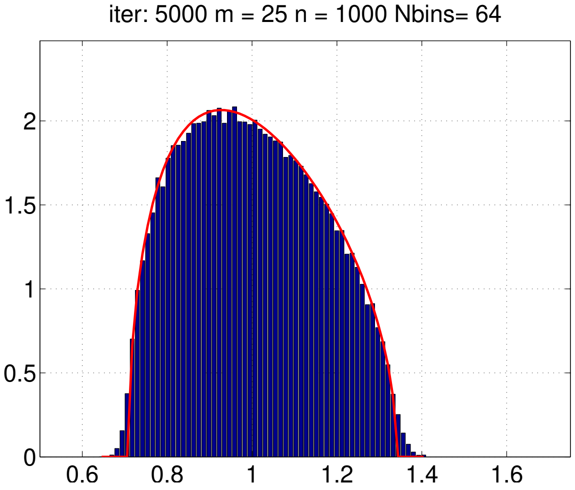
Numerical Illustration

```
X = randn(m,n);
S = 1/n X X' ;
L = eig(S) ;
histL = hist(L,x);
```

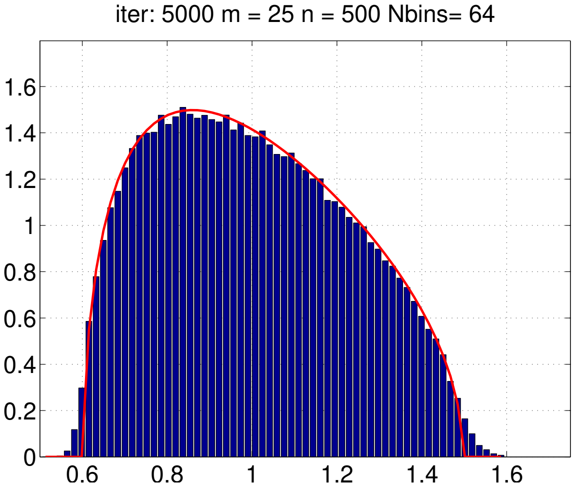
Simulation: Eigenvalue Spread



Simulation: Eigenvalue Spread



Simulation: Eigenvalue Spread



Spread of Sample Eigenvalues

Let $\{\ell_i\}_{i=1}^m$ be the eigenvalues of a random symmetric matrix H.

Empirical Spectral Distribution Function:

$$F_m(t) = \frac{1}{m}\#\{\ell_i \le t\}$$

The Quarter-Circle Law

[Marchenko & Pastur, 1967]

Let S_n be sample covariance of n Gaussian observations from $\mathcal{N}(0, I_p)$.

Theorem: For $\Sigma = I$, as $p, n \to \infty$ with $p/n \to c$, $(c < 1)$ let ℓ_i be sample eigenvalues of S_n, then

$$\boxed{f_{MP}(t) = \frac{1}{2\pi ct}\sqrt{(b-t)(t-a)} \quad t \in [a, b]}$$

where $a = (1-\sqrt{c})^2$, $b = (1+\sqrt{c})^2$

The Quarter-Circle Law

[Marchenko & Pastur, 1967]

Let S_n be sample covariance of n Gaussian observations from $\mathcal{N}(0, I_p)$.

Theorem: For $\Sigma = I$, as $p, n \to \infty$ with $p/n \to c$, $(c < 1)$ let ℓ_i be sample eigenvalues of S_n, then

$$\boxed{f_{MP}(t) = \frac{1}{2\pi ct}\sqrt{(b-t)(t-a)} \quad t \in [a, b]}$$

where $a = (1-\sqrt{c})^2$, $b = (1+\sqrt{c})^2$

If $c > 1$, then $a = 0$, and there are $p - n - 1$ sample eigenvalues exactly at zero.

Now consider data of the form signal+noise

Spiked Covariance Models

Consider model whereby

$$\Sigma = diag(\lambda_1, \lambda_2, \ldots, \lambda_k, 0, \ldots, 0) + \sigma^2 I_m$$

Spiked covariance with k spikes.

Observe n vectors $\mathbf{x}_i \in \{\mathbb{R}, \mathbb{C}\}^m$ from this model.

Question: What happens to largest sample eigenvalues and eigenvectors as $n, m \to \infty$, with k, λ_j fixed ?

Phase Transition

[complex case, Ben-Arous, Baik, Peche]
[real case, Baik and Silverstein]

Theorem: For spike model with k spikes, as $n, m \to \infty$ with $m/n \to c$, for $j = 1, \ldots, k$,

$$\ell_j \to \begin{cases} (\lambda_j + \sigma^2)\left(1 + \frac{m-k}{n}\frac{\sigma^2}{\lambda_j}\right) & \lambda_j > \sigma^2\sqrt{m/n} \\ \sigma^2(1+\sqrt{m/n})^2 & \lambda_j < \sigma^2\sqrt{m/n} \end{cases}$$

Phenomena known as *retarded learning* in statistical physics.

Phase Transition / Eigenvectors

[D. Paul 07', Nadler 08']

Theorem: As $m, n \to \infty$ with $m/n \to c$,

$$R^2(m/n) = |\langle \mathbf{v}_{\text{PCA}}, \mathbf{v} \rangle|^2 = \begin{cases} 0 & \text{if } \lambda < \sigma^2\sqrt{m/n} \\ \dfrac{\frac{\lambda^2}{c\sigma^4} - 1}{\frac{\lambda^2}{c\sigma^4} + \frac{\lambda}{\sigma^2}} & \text{if } \lambda > \sigma^2\sqrt{m/n} \end{cases}$$

In statistical physics:

[Hoyle and Rattray, Reimann & al, Biehl, Watson]

Phase Transition / Eigenvectors

[D. Paul 07', Nadler 08']

Theorem: As $m, n \to \infty$ with $m/n \to c$,

$$R^2(m/n) = |\langle \mathbf{v}_{\text{PCA}}, \mathbf{v} \rangle|^2 = \begin{cases} 0 & \text{if } \lambda < \sigma^2 \sqrt{m/n} \\ \dfrac{\frac{\lambda^2}{c\sigma^4} - 1}{\frac{\lambda^2}{c\sigma^4} + \frac{\lambda}{\sigma^2}} & \text{if } \lambda > \sigma^2 \sqrt{m/n} \end{cases}$$

In statistical physics:

[Hoyle and Rattray, Reimann & al, Biehl, Watson]

Asymptotic $\sqrt{n}$-Gaussian fluctuations for both eigenvalue and eigenvector [Paul, 07]

$$\sqrt{n}(\ell_1 - \mathbb{E}[\ell_1]) \sim \mathcal{N}(0, \sigma^2(\lambda_1))$$

Proof of Phase Transition: Single Spike

$$H = \frac{1}{n}X'X = \begin{pmatrix} z_1 & 0 & \dots & 0 \\ 0 & 0 & & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} + \sigma \begin{pmatrix} 0 & b_2 & \dots & b_m \\ b_2 & 0 & & 0 \\ \vdots & & 0 & \vdots \\ b_m & 0 & & 0 \end{pmatrix}$$

$$+\sigma^2 \left(\begin{array}{c|ccc} 0 & 0 & \dots & 0 \\ \hline 0 & z_{2,2} & & z_{2m} \\ & \vdots & \ddots & \vdots \\ 0 & z_{m,2} & & z_{m,m} \end{array}\right)$$

$$= \quad A_0 \quad + \quad \sigma A_1 \quad + \quad \sigma^2 A_2$$

= *signal* + *signal/noise interaction* + noise

Proof of Phase Transition: Single Spike

Trick: Diagonalize noise part:

$$H = \frac{1}{n}X'X = \begin{pmatrix} z_1 & 0 & \dots & 0 \\ 0 & 0 & & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} + \sigma \begin{pmatrix} 0 & \tilde{b}_2 & \dots & \tilde{b}_m \\ \tilde{b}_2 & 0 & & 0 \\ \vdots & & 0 & \vdots \\ \tilde{b}_m & 0 & & 0 \end{pmatrix}$$

$$+\sigma^2 \left(\begin{array}{c|ccc} 0 & 0 & \dots & 0 \\ \hline 0 & \mu_2 & & 0 \\ & & \ddots & \\ 0 & 0 & & \mu_m \end{array}\right)$$

Proof of Phase Transition: Single Spike

Trick: Diagonalize noise part:

$$H = \frac{1}{n}X'X = \begin{pmatrix} z_1 & 0 & \dots & 0 \\ 0 & 0 & & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} + \sigma \begin{pmatrix} 0 & \tilde{b}_2 & \dots & \tilde{b}_m \\ \tilde{b}_2 & 0 & & 0 \\ \vdots & & 0 & \vdots \\ \tilde{b}_m & 0 & & 0 \end{pmatrix}$$

$$+\sigma^2 \left(\begin{array}{c|ccc} 0 & 0 & \dots & 0 \\ \hline 0 & \mu_2 & & 0 \\ & & \ddots & \\ 0 & 0 & & \mu_m \end{array}\right)$$

Arrowhead Matrix: Its eigenvalues are roots of *secular equation*

$$\ell - z_1 = \sum_j \frac{\tilde{b}_j^2}{\ell - \mu_j}$$

Proof of Phase Transition

$(m-1) \times (m-1)$ matrix Z is of pure noise $\rightarrow \mu_2, \ldots, \mu_m$ are eigenvalues of $W_{m-1}(n, \sigma^2 I)$.

As $m, n \rightarrow \infty$ with $m/n \rightarrow c$,

$$z_1 \rightarrow (\lambda + \sigma^2)$$

$$\mu_2, \ldots, \mu_p \rightarrow \text{Marchenko Pastur density}$$

$$\tilde{b}_j \rightarrow \mathcal{N}(0, z_1 \mu_j / n)$$

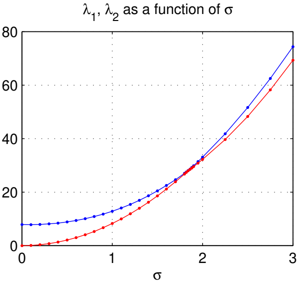
$$\text{sum converges to integral}$$

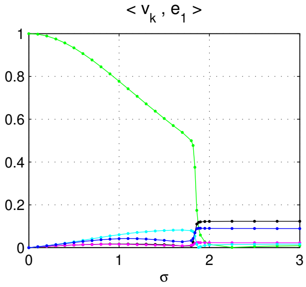
$$\ell - (\lambda + \sigma^2) = c \int (\lambda + \sigma)^2 \frac{\mu}{\ell - \mu} f_{MP}(\mu) d\mu$$

Integral can be computed explicitly, gives quadratic equation. Its solution gives the phase transition formula.

Phase Transition for finite m as function of σ

First, a "thought experiment": Take clean signal data $\{\mathbf{x}^\nu\}$ with finite m, n, add noise and start increasing σ. What should be the expected behavior of $|\langle \mathbf{v}_{\mathrm{PCA}}, \mathbf{v}\rangle|$ and of ℓ_1 ?

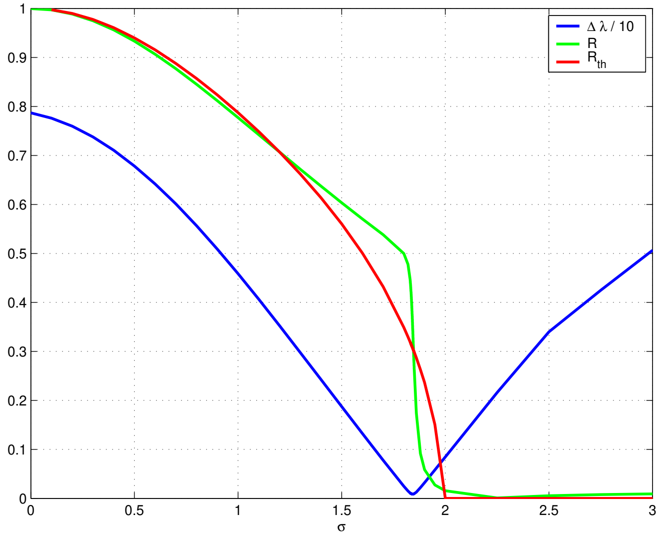




$$\lambda \sim \kappa^2 + \sigma^2(1 + m/n) \qquad\qquad R \sim 1 - \sigma^2/\kappa^2 m/n$$

$$n = 50, m = 200, \kappa^2 = 7.87$$

Phase Transition as function of σ



Part III:

Can we do better in high dimensions
with additional information ?
Sparsity of covariance or of principal components