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

June 2018
Part 1: Classical Asymptotic Statistics
\[ x_1, \ldots, x_n \in \mathbb{R}^p \text{ assumed i.i.d. from r.v. } X. \]

**Sample Mean:**

\[ \bar{x} = \frac{1}{n} \sum_i x_i \]
\( \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p \) assumed i.i.d. from r.v. \( X \).

**Sample Mean:**

\[
\bar{x} = \frac{1}{n} \sum_i x_i
\]

**Sample Covariance Matrix:**

\[
S_n = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T
\]
\( \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p \) assumed i.i.d. from r.v. \( X \).

**Sample Mean:**

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

**Sample Covariance Matrix:**

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

**Eigendecomposition / Principal Component Analysis**

\[
S_n = \sum_j \ell_j \hat{\mathbf{v}}_j \hat{\mathbf{v}}_j^T
\]
Reminder: CLT, if $x_i$ all i.i.d. from r.v. $X$ with $\mathbb{E}[X] = \mu$ and $\text{Var}[X] = \sigma^2 < \infty$, then as $n \to \infty$

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

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, $n \gg p$, $\hat{x}$ and $S_n$ are accurate estimators of $\mu$ and $\Sigma$. 
Furthermore, as for eigendecomposition,

\[ \ell_j \rightarrow \lambda_j \] and for eigenvalues with multiplicity one \( \hat{v}_j \rightarrow v_j \)

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

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

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

Also,

\[
\hat{v}_j = v_j + O_P \left( \frac{1}{\sqrt{n}} \right)
\]
Asymptotic Eigenvalue Distribution

**Example:** Single signal in noise

\[ \Sigma = \lambda_1 e_1 e_1^T + \sigma^2 I_m \]
Asymptotic Eigenvalue Distribution

Example: Single signal in noise

$$\Sigma = \lambda_1 e_1 e_1^T + \sigma^2 I_m$$

$m = 6, n = 150, \lambda_1 = 10$

Empirical $\mathcal{N}(0,1)$
Eigenvector Asymptotics

Example:
Signal strength $\lambda$ in noise variance $\sigma^2$.

$$\lambda_1 = \lambda + \sigma^2, \quad \lambda_j = \sigma^2.$$  
Without loss of generality, assume $h = e_1$. 
Eigenvector Asymptotics

Example:
Signal strength \( \lambda \) in noise variance \( \sigma^2 \).

\[
\lambda_1 = \lambda + \sigma^2, \quad \lambda_j = \sigma^2.
\]

Without loss of generality, assume \( h = e_1 \).

Asymptotically,

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

Hence

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

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

\[ m = 5, n = 120, \lambda_1 = 10 \]

Empirical \( \chi^2_{m-1} \)

\[ n \cdot \frac{\chi^2}{(\lambda + 1) \cdot (1/R - 1)} \]
Part II: What happens when dimension $p$ is large when $p$ and $n$ are comparable, or even $p \gg n$?
Example: Consider $x_1, \ldots, x_n$ all i.i.d. from $\mathcal{N}(0, I_p)$. Namely, $\Sigma = I_p$, 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!
Example: Consider $x_1, \ldots, x_n$ all i.i.d. from $\mathcal{N}(0, I_p)$. Namely, $\Sigma = I_p$, all its $p$ eigenvalues are equal $\lambda_j = 1$.

How do eigenvalues of $S_n$ look like when $p, n$ are comparable?
**Example:** Consider $x_1, \ldots, x_n$ all i.i.d. from $\mathcal{N}(0, I_p)$. Namely, $\Sigma = I_p$, 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

iter: 5000 m = 25 n = 5000 Nbins= 64
Simulation: Eigenvalue Spread

iter: 5000 m = 25 n = 1000 Nbins= 64
Simulation: Eigenvalue Spread

iter: 5000  m = 25  n = 500  Nbins= 64
Let $\{\ell_i\}_{i=1}^m$ be the eigenvalues of a random symmetric matrix $H$.

*Empirical Spectral Distribution Function:*

$$F_m(t) = \frac{1}{m} \# \{ \ell_i \leq t \}$$
The Quarter-Circle Law

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 $\ell_i$ be sample eigenvalues of $S_n$, then

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

where $a = (1 - \sqrt{c})^2$, $b = (1 + \sqrt{c})^2$
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 $\ell_i$ be sample eigenvalues of $S_n$, then

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

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
Consider model whereby

\[ \Sigma = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k, 0, \ldots, 0) + \sigma^2 I_m \]

*Spiked covariance* with \( k \) spikes.

Observe \( n \) vectors \( x_i \in \{\mathbb{R}, \mathbb{C}\}^m \) from this model.

**Question:** What happens to largest sample eigenvalues and eigenvectors as \( n, m \to \infty \), with \( k, \lambda_j \) 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 $j = 1, \ldots, k$,

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

Phenomena known as *retarded learning* in statistical physics.
Theorem: As $m, n \to \infty$ with $m/n \to c$, 

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

In statistical physics:

[Hoyle and Rattray, Reimann & al, Biehl, Watson]
Phase Transition / Eigenvectors

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

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

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

\[ H = \frac{1}{n} X'X = \begin{pmatrix} z_1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix} + \sigma \begin{pmatrix} 0 & b_2 & \ldots & b_m \\ b_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ b_m & 0 & \ldots & 0 \end{pmatrix} + \sigma^2 \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & z_{2,2} & \ldots & z_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & z_{m,2} & \ldots & z_{m,m} \end{pmatrix} \]

\[ = A_0 + \sigma A_1 + \sigma^2 A_2 \]

\[ = \text{signal} + \text{signal/noise interaction} + \text{noise} \]
Proof of Phase Transition: Single Spike

**Trick:** Diagonalize noise part:

\[ H = \frac{1}{n} X'X = \begin{pmatrix} z_1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix} + \sigma \begin{pmatrix} 0 & \tilde{b}_2 & \ldots & \tilde{b}_m \\ \tilde{b}_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{b}_m & 0 & \ldots & 0 \end{pmatrix} + \sigma^2 \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & \mu_2 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mu_m \end{pmatrix} \]
Proof of Phase Transition: Single Spike

**Trick:** Diagonalize noise part:

\[ H = \frac{1}{n} X'X = \begin{pmatrix} z_1 & 0 & \ldots & 0 \\ 0 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix} + \sigma \begin{pmatrix} 0 & \tilde{b}_2 & \ldots & \tilde{b}_m \\ \tilde{b}_2 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \tilde{b}_m & 0 & 0 & 0 \end{pmatrix} + \sigma^2 \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & \mu_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & \mu_m \end{pmatrix} \]

**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 $m/n \rightarrow c$,

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

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

$$b_j \rightarrow \mathcal{N}(0, z_1 \mu_j / n)$$

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 $\sigma$

First, a "thought experiment": Take clean signal data $\{x'\}$ with finite $m, n$, add noise and start increasing $\sigma$. What should be the expected behavior of $|\langle v_{\text{PCA}}, v \rangle|$ and of $\ell_1$?

\[
\begin{align*}
\lambda & \sim \kappa^2 + \sigma^2(1 + m/n) \\
n & = 50, \ m = 200, \ \kappa^2 = 7.87
\end{align*}
\]

$R \sim 1 - \sigma^2/\kappa^2 m/n$
Phase Transition as function of $\sigma$

![Graph showing phase transition as a function of $\sigma$]
Part III:

Can we do better in high dimensions with additional information?

*Sparsity* of covariance or of principal components