Finite Sample Results on the convergence of PCA for spiked covariance models

Boaz Nadler

Department of CS and Applied Math
Weizmann Institute of Science

Spiked Covariance Model

Consider n independent observations $\{\boldsymbol{x}^\nu\}_{\nu=1}^n$ from the model

$$\boldsymbol{x} = \sum_{i=1}^{k} u_i \boldsymbol{v}_i + \sigma \boldsymbol{\xi}$$

where:

- $\boldsymbol{x} \in \mathbb{R}^p$
- u_i are independent random variables (components/latent variables) with $\mathbb{E}\{u_i\} = 0, \mathrm{Var}\{u_i\} = 1$.
- $\boldsymbol{v}_i$ are *fixed* orthogonal vectors in $\mathbb{R}^p$
- $\boldsymbol{\xi}$ is a noise vector, whose p components are i.i.d. $N(0,1)$ random variables, and σ is the level of noise.

(same model as in David Hoyle's lecture yesterday)

A Spiked Covariance Model

The population covariance matrix ($n \to \infty$) admits the form

$$C_\infty = \sum_{j=1}^{k} \boldsymbol{v}_i' \boldsymbol{v}_i + \sigma^2 I_p = \begin{pmatrix} \|\boldsymbol{v}_1\|^2 & 0 & & & \dots & 0 \\ & \ddots & & & & \\ 0 & & \|\boldsymbol{v}_k\|^2 & & & \\ \vdots & & & 0 & & \\ & & & & 0 & \\ 0 & & & & & 0 \end{pmatrix} + \sigma^2 I_p$$

This is a *spiked covariance model* with k large eigenvalues and an additional $p-k$ eigenvalues equal to σ^2.

Convergence in Spiked Covariance Model

Let $C_n = 1/n\boldsymbol{X}'\boldsymbol{X}$ denote the *sample* covariance matrix from n observations $\boldsymbol{x}^1, \dots, \boldsymbol{x}^n$.

Question: How close are the eigenvalues and eigenvectors of C_n to those of the population matrix C_∞ ?

Application: How good is PCA at detecting a characteristic signal inside noise.

Previous Work

1) **Classical Statistics:** the dimensionality p is fixed, noise level σ is fixed, number of samples $n \to \infty$ [Girshik 1936, Anderson 1963, etc]

2) **Statistical Mechanics/RMT:** both $p, n \to \infty$ with their ratio fixed $p/n = c$.

[In the statistics community: Bai and Silverstein, Baik and Silverstein, Johnstone, Paul, Onatski, Baik Ben Arous and Peche, El Karoui, ...

In the physics community: Hoyle & Rattray, Reimann & al, Biehl, Watson, ...]

3) **Stochastic Perturbation theory**: p,n fixed [GW Stewart 1990].

Previous Results - The Phase Transition

Limits of the leading Eigenvalue: As $p, n \to \infty$, p/n fixed

$$\lambda_{\text{PCA}} = \begin{cases} \sigma^2 \left(1 + \sqrt{\frac{p}{n}}\right)^2 & \text{if } n/p < \sigma^4/\|\boldsymbol{v}\|^4 \\ (\|\boldsymbol{v}\|^2 + \sigma^2)\left[1 + \frac{p}{n}\frac{\sigma^2}{\|\boldsymbol{v}\|^2}\right] & n/p \geq \sigma^4/\|\boldsymbol{v}\|^4 \end{cases} \tag{1}$$

Limits of the leading eigenvector:

$$R^2(p/n) = |\langle \boldsymbol{v}_{\text{PCA}}, \boldsymbol{v}\rangle|^2 = \begin{cases} 0 & \text{if } n/p < \sigma^4/\|\boldsymbol{v}\|^4 \\ \dfrac{\frac{n\|\boldsymbol{v}\|^4}{p\sigma^4} - 1}{\frac{n\|\boldsymbol{v}\|^4}{p\sigma^4} + \frac{\|\boldsymbol{v}\|^2}{\sigma^2}} & \text{if } n/p \geq \sigma^4/\|\boldsymbol{v}\|^4 \end{cases} \tag{2}$$

What happens for finite p, n ***?***
Asymptotica - Are we there yet ?

How close are the results for λ_{PCA} and $\boldsymbol{v}_{\mathrm{PCA}}$ to the limiting values as $p, n \to \infty$?

What is the origin of the phase transition ? What type of phase transition occurs for *finite* p, as a function of n ?

What is the explicit dependence of the eigenvalue and eigenvector on the noises ?

Our approach - Stochastic Matrix Eigenanalysis

We keep p, n ***fixed***, as well as the specific values of the latent variables u_j in the given training set.

This gives us a *finite* covariance matrix.

We will view σ - the noise level, as a *small* parameter and use

Matrix Perturbation Theory

Benefits

1) ***Probabilistic convergence:*** Under certain conditions, with probability $1-\varepsilon$, the eigenvector of sample PCA is "close" to that of population PCA, and we quantify this closeness.

2) ***Leading Order Perturbation analysis:*** We consider the limit $\sigma \to 0$. The leading eigenvectors and eigenvalues are *analytic* in σ, and we compute their Taylor expansion, with their explicit dependence on the noise terms - leading order mean values and variances.

3) ***A matrix analysis view of the phase transition.*** We present a matrix based explanation for a "phase transition type" behavior as a function of n for *finite* p and n.

Our Results

For simplicity, consider a training set of n samples $\{(\boldsymbol{x}^\nu, u^\nu, \boldsymbol{\xi}^\nu)\}_{\nu=1}^n$ from a 1-component model,

$$\boldsymbol{x} = u\boldsymbol{v} + \sigma\boldsymbol{\xi}$$

Define

$$s_u = \frac{1}{n}\sum_\nu (u^\nu)^2 \qquad \kappa^2 = s_u\|\boldsymbol{v}\|^2$$

$$\rho_j = \frac{1}{ns_u}\sum_\nu u^\nu \boldsymbol{\xi}_j^\nu \quad \beta_{ij} = \frac{1}{n}\sum_\nu \xi_i^\nu \xi_j^\nu$$

Remarks: $\rho_j \sim N(0, 1/n)$ are all i.i.d., $\beta_{ii} \sim \chi_n^2/n = O(1)$, $\beta_{ij} = O(1/\sqrt{n})$

The Covariance matrix

$$
\begin{aligned}
\frac{1}{n}X'X &= \begin{pmatrix} \kappa^2 & 0 & \dots & 0 \\ 0 & 0 & & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} + \sigma\kappa \begin{pmatrix} 2\rho_1 & \rho_2 & \dots & \rho_p \\ \rho_2 & 0 & & 0 \\ \vdots & & 0 & \vdots \\ \rho_p & 0 & & 0 \end{pmatrix} \\
& \quad + \sigma^2 \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,p} \\ \beta_{2,1} & \beta_{2,2} & & \vdots \\ & & & \\ \beta_{p,1} & & & \beta_{p,p} \end{pmatrix} \\
&= \mathcal{L}_0 \quad + \quad \sigma\mathcal{L}_1 \quad + \quad \sigma^2\mathcal{L}_2 \\
&= \text{signal} + \text{signal/noise interaction} + \text{noise}
\end{aligned}
$$

Perturbation Theory

EXAMPLE: Upon diagonalizing the minor of the noise matrix,

$$\begin{pmatrix} \kappa^2 + 2\kappa\sigma\rho_1 + \sigma^2\beta_{11} & \tilde{b}_{12} & \dots & \tilde{b}_{1p} \\ \tilde{b}_{12} & 0 & \dots & 0 \\ \tilde{b}_{13} & 0 & & 0 \\ \vdots & & \ddots & \vdots \\ \tilde{b}_{1p} & 0 & \dots & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \vdots & 0 & \lambda_3 & & 0 \\ 0 & & & \ddots & \\ 0 & & & & \lambda_p \end{pmatrix}$$

where $\tilde{b}_{ij} = \sigma\kappa\tilde{\rho}_j + \sigma^2\tilde{\beta}_{1j}$, and $\lambda_2 \geq \lambda_3 \geq \lambda_p$.

Determinant equation:

$$f(\lambda) = det(\lambda I - C_n) = \lambda - (\kappa^2 + 2\kappa\sigma\rho_1 + \sigma^2\beta_{11}) - \sum_{j=2}^{p} \frac{\tilde{b}_{1j}^2}{\lambda - \lambda_j} = 0$$

The Pulled Up Values

The eigenvalues of $1/n\mathbf{X}'\mathbf{X}$ are solutions of

$$f(\lambda) = \lambda - (\kappa^2 + 2\kappa\sigma\rho_1 + \sigma^2\beta_{11}) - \sum_{j=2}^{p} \frac{\tilde{b}_{1j}^2}{\lambda - \lambda_j} = 0$$

where $\tilde{b}_{ij} = \sigma\kappa\tilde{\rho}_j + \sigma^2\tilde{\beta}_{1j}$

Conditional on knowledge that the k-th direction has variance λ_k, $\tilde{b}_{ij} \sim N(0, \lambda_j\sigma^2(\kappa^2 + 2\sigma\kappa\rho_1 + \sigma^2\beta_{11})/n)$ and they are all independent. Therefore,

$$f(\lambda) = \lambda - (\kappa^2 + 2\kappa\sigma\rho_1 + \sigma^2\beta_{11}) - \frac{p-1}{n}(\kappa^2 + 2\kappa\sigma\rho_1 + \sigma^2\beta_{11})\mathbb{E}\frac{\lambda_j w_j^2}{\lambda - \lambda_j}$$

As $n, p \to \infty$, largest eigenvalue solves

$$f(\lambda) = \lambda - \|\boldsymbol{v}\|^2 - \sigma^2 = (\|\boldsymbol{v}\|^2 + \sigma^2)\sigma^2\frac{p}{n}\int \frac{x}{\lambda - x} f_{MP}(x)dx$$

which gives limiting eigenvalues, and location of phase transition.

Results I: Finite p, n eigenvalue and eigenvector

Theorem: Let $n < p$ be both finite and fixed. Using concentration of measure results from Szarek & Davidson, for $n < p$

$$\Pr\left\{\lambda_2 > \left(1 + \sqrt{\frac{p}{n}}\right)^2 + \frac{p}{n}\right\} \leq \varepsilon = \exp\left(-\frac{p}{2(1+\sqrt{5})^2}\right)$$

Assume

$$\kappa^2 \left(1 + \frac{2\sigma\rho_1}{\kappa} + \frac{\sigma^2}{\kappa^2}\beta_{11}\right) > \sigma^2 \left[\left(1 + \sqrt{\frac{p}{n}}\right)^2 + \frac{p}{n}\right]$$

Define

$$A = 1 + \frac{2\sigma\rho_1}{\kappa} + \frac{\sigma^2}{\kappa^2}\chi_1^2$$

then with probability at least $1 - \varepsilon$

$$\kappa^2 A + \sqrt{\sum_{j\geq 2} \tilde{b}_{1j}^2} \leq \lambda_{\text{PCA}} \leq \frac{\kappa^2 A + \sqrt{\kappa^4 A^2 + 4\sum \tilde{b}_{1j}^2}}{2}$$

Now $\sum_j \tilde{b}_{ij}^2 \approx \kappa^2 A \sigma^2 p/n$ so with probability at least $1 - \varepsilon'$

$$\kappa^2 A + \sigma\kappa\sqrt{A}\sqrt{p/n} \le \lambda_{\text{PCA}} \le \kappa^2 A + \frac{\sigma^2}{\kappa^2 A}\frac{p}{n}$$

Similar bounds on angle between $\boldsymbol{v}_{\text{PCA}}$ and $\boldsymbol{v}$,

$$\sin\theta_{\text{PCA}} \approx \frac{\sigma}{\kappa}\sqrt{\frac{p}{n}} + O(\sigma^2)\sqrt{\frac{p}{n}}$$

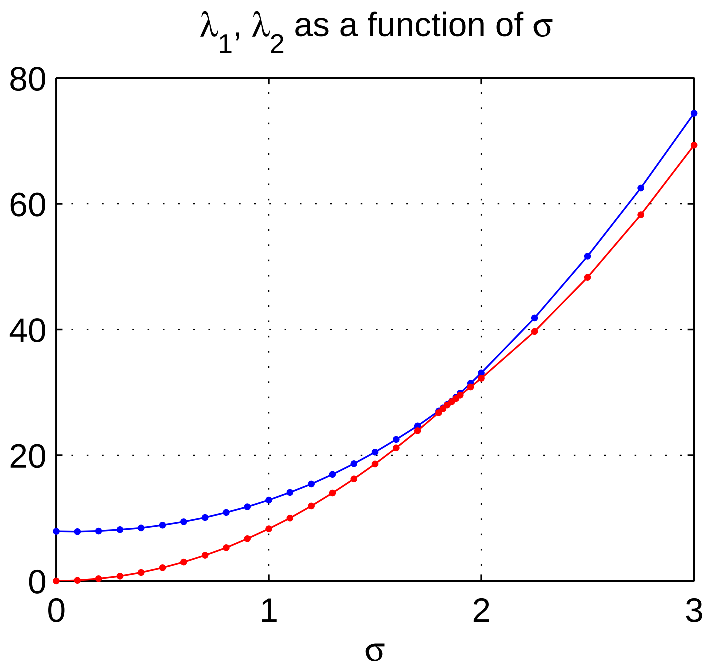
The Phase Transition

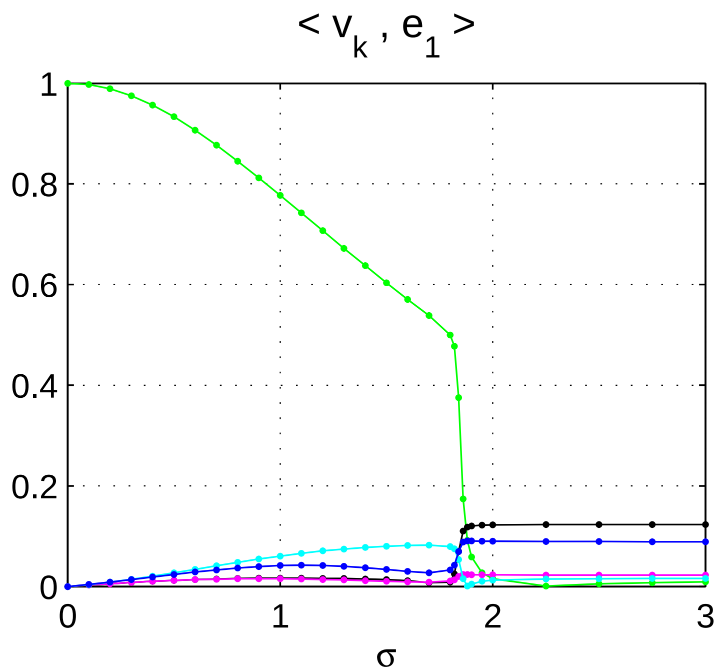
In the limit $n, p \to \infty$, p/n fixed,

$$R^2(p/n) = |\langle \boldsymbol{v}_{\mathrm{PCA}}, \boldsymbol{v} \rangle|^2 = \begin{cases} 0 & \text{if } n/p < \sigma^4/\|\boldsymbol{v}\|^4 \\ \dfrac{\frac{n\|\boldsymbol{v}\|^4}{p\sigma^4} - 1}{\frac{n\|\boldsymbol{v}\|^4}{p\sigma^4} + \frac{\|\boldsymbol{v}\|^2}{\sigma^2}} & \text{if } n/p \geq \sigma^4/\|\boldsymbol{v}\|^4 \end{cases} \tag{3}$$

Phase Transition for finite p as function of σ

First, a "thought experiment": Take training set $\{\boldsymbol{x}^\nu\}$ with finite p, n and start increasing σ. What should be the expected behavior of $R = \langle \boldsymbol{v}_{\mathrm{PCA}}, \boldsymbol{v} \rangle$ and of λ_{PCA} ?

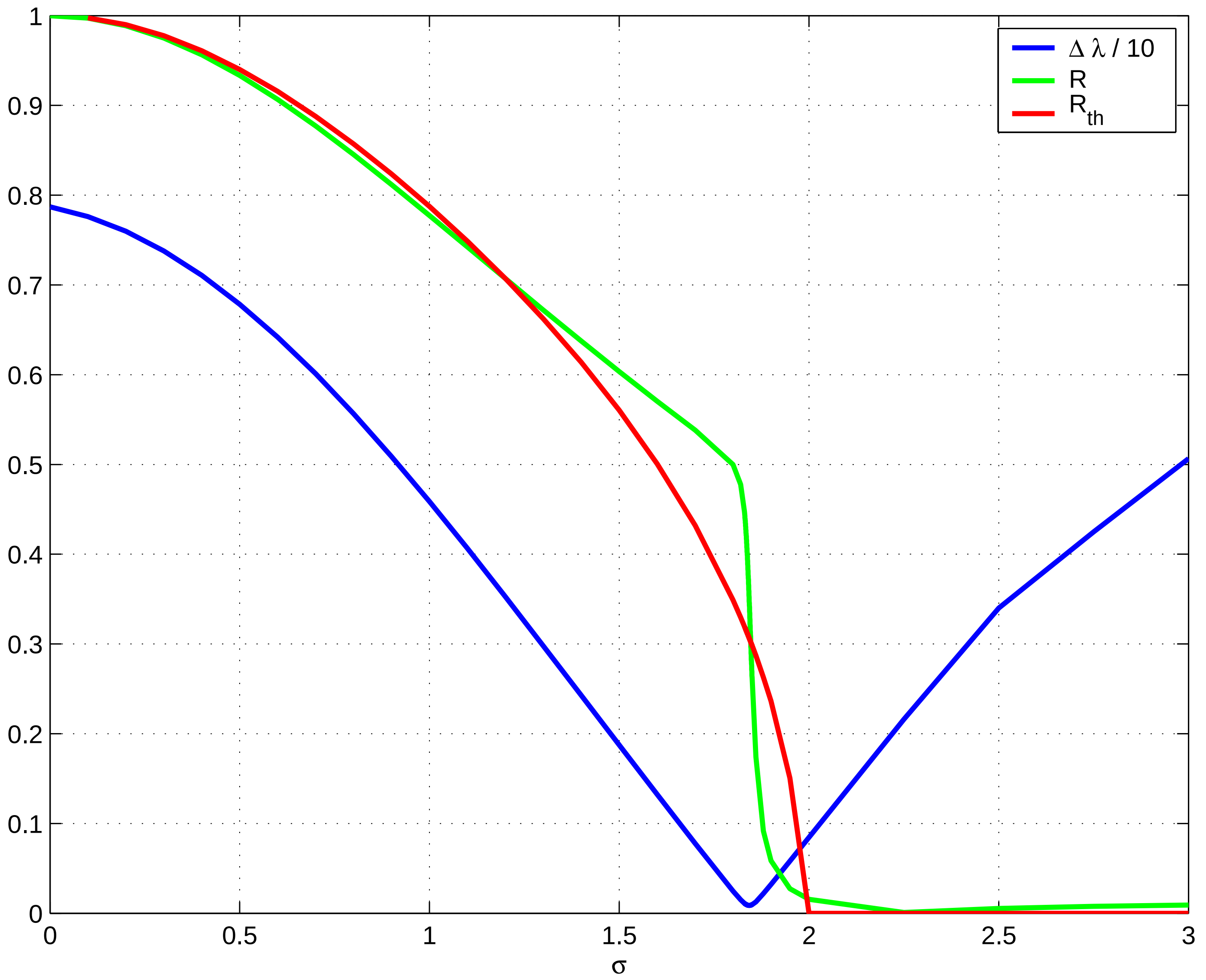




$\lambda \sim \kappa^2 + \sigma^2(1 + p/n)$ $\qquad$ $R \sim 1 - \sigma^2/\kappa^2 p/n$

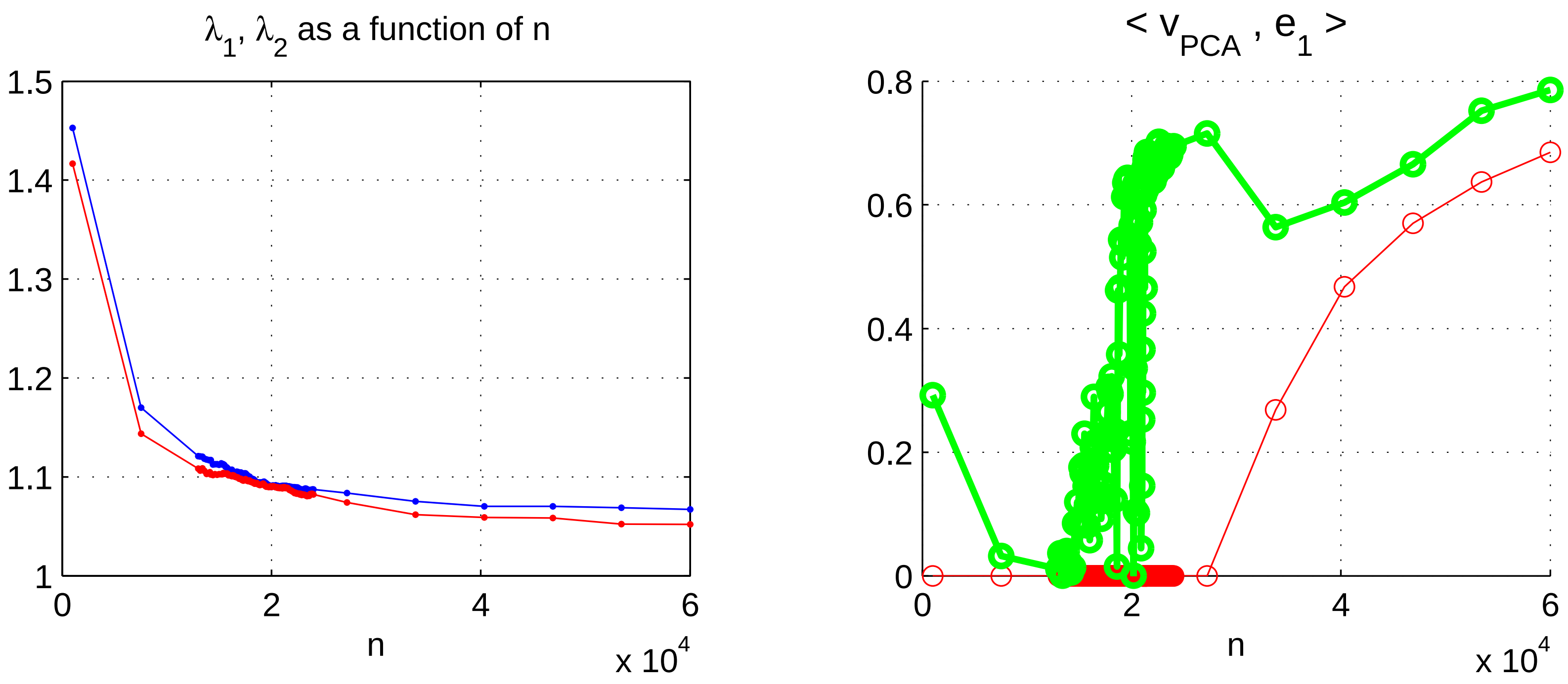
$$n = 50, p = 200, \kappa^2 = 7.87$$

Phase Transition as function of σ



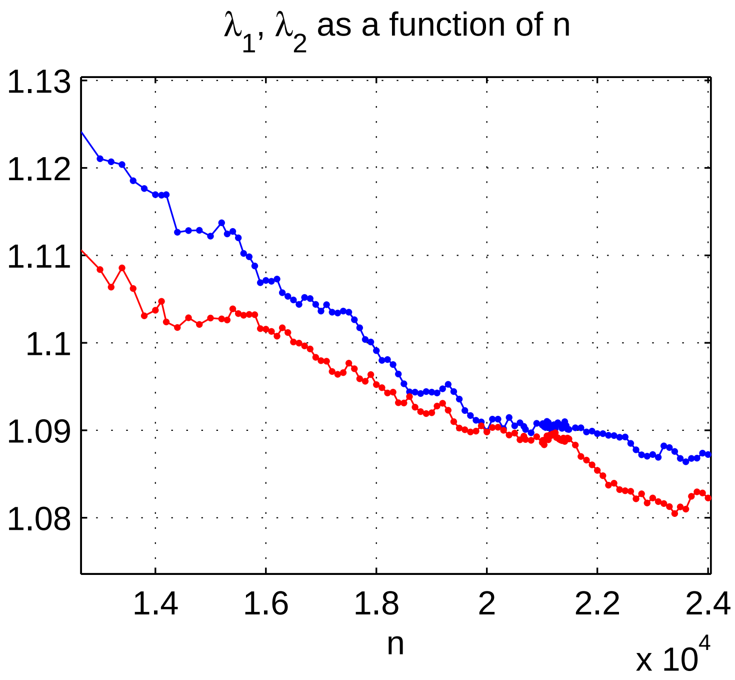
Phase Transition for finite p as function of n

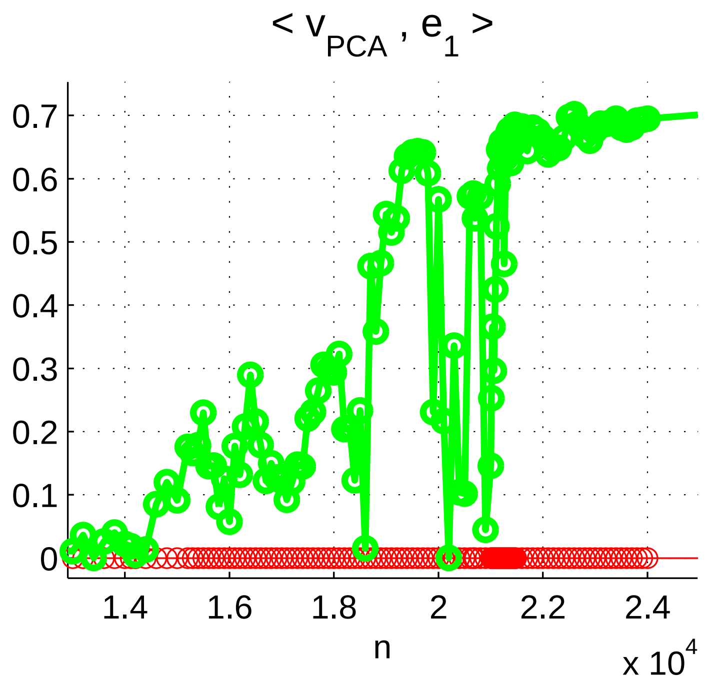
Now consider the behavior of λ_{PCA} and $R = \langle \boldsymbol{v}_{\text{PCA}}, \boldsymbol{v} \rangle$ as a function of the number of samples n:



$$p = 50, \|v\| = 1/5, \sigma = 1$$

Phase transition for finite p as function of n





$$p = 50, \|v\| = 1/5, \sigma = 1$$

Summary and some take home lessons

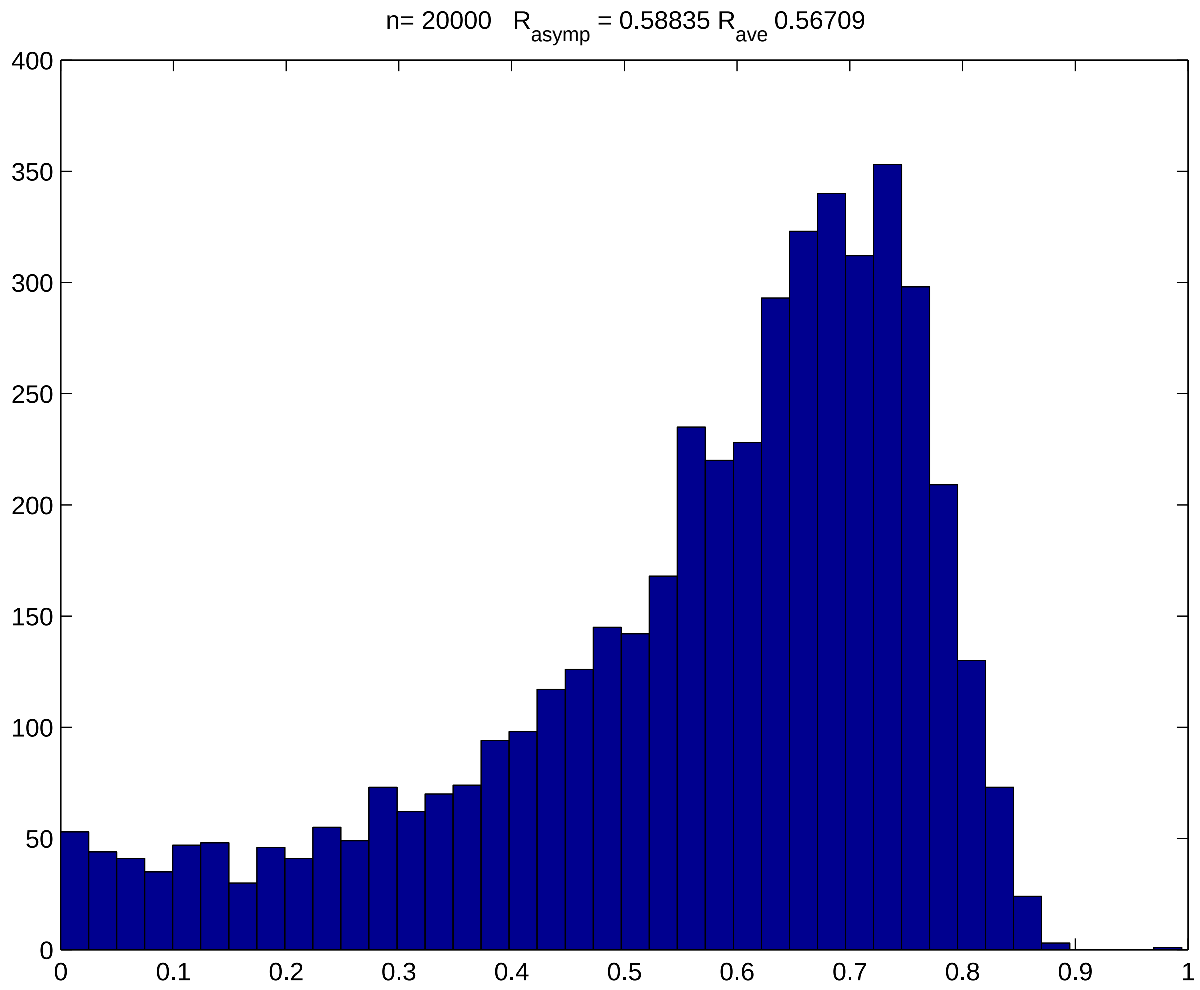
1. Asymptotic results are *not* the end of the story - finite p, n can have quite interesting behaviors.

2. Inference in high dimensional setting is qualitatively different from classical statistics, where error is $O(1/\sqrt{n})$.

3. If $n \ll p$ and noise is not negligible, PCA is not a good method to estimate the underlying principal vectors. Errors are $O(\frac{\sigma}{\|\boldsymbol{v}\|}\sqrt{\frac{p}{n}})$ and therefore important to perform "local"/wavelet type dimensional reduction prior to PCA. Other "global" methods for regression/classification such as partial least squares, classical least squares suffer from similar problems in the small n–large p case [Johnstone & Lu, Nadler & Coifman 05', Donoho & Buckheit 97', Raudys 80's]

The end :)

Asymptotica - not there yet !



Some open questions

Estimation / characterization of location of phase transition for finite p. Derivation of probabilities for $R > R_0$ as a function of p, n, σ ?

What is the "optimal" procedure to estimate $\boldsymbol{v}$ under various assumptions on the distribution of $\boldsymbol{v}$? - PCA is clearly not optimal

Can one derive bounds such as

$$\mathbb{E}\{R\} \leq 1 - f(p, n, \sigma)$$

for any possible procedure that tries to estimate $\boldsymbol{v}$, where averaging is over all possible noises and over an apriori assumed distribution of $\boldsymbol{v}$.

Taylor Expansion of leading eigenvalue and eigenvector

We now consider the dependence of λ_{PCA} and $\boldsymbol{v}_{\mathrm{PCA}}$ on the noise strength σ. Since these quantities are *analytic* in σ, we can expand them in a Taylor series:

$$\begin{aligned} \lambda_{\mathrm{PCA}} &= \lambda_0 + \sigma\lambda_1 + \sigma^2\lambda_2 + \ldots \\ \boldsymbol{v}_{\mathrm{PCA}} &= \boldsymbol{v}_0 + \sigma\boldsymbol{v}_1 + \sigma^2\boldsymbol{v}_2 + \ldots \end{aligned}$$

and insert into the eigenvalue equation

$$(\mathcal{L}_0 + \sigma\mathcal{L}_1 + \sigma^2\mathcal{L}_2)\boldsymbol{v} = \lambda\boldsymbol{v}$$

This gives the *explicit* leading order dependence as a function of the noise:

$$\lambda_{\text{PCA}} = \kappa^2 \left[1 + \frac{2\sigma\eta_1}{\kappa\sqrt{n}} + \frac{\sigma^2}{\kappa^2} \left(\frac{t_{p-1}}{n} + \beta_{11} \right) + O(\sigma^3) \right]$$

$$\boldsymbol{v}_{\text{PCA}} = \boldsymbol{e}_1 + \frac{\sigma}{\kappa\sqrt{n}} (0, \eta_2, \ldots, \eta_p)' + O(\sigma^2)$$

where η_i are all i.i.d. $N(0,1)$, $t_{p-1} \sim \chi^2_{p-1}$, and β_{11} has mean one.

Leading Order Mean and Variance

$$\mathbb{E}\{\lambda_{\text{PCA}}\} = \mathbb{E}\{u^2\}\|\boldsymbol{v}\|^2 + \sigma^2 \left(1 + \frac{p-1}{n}\right) + O(\sigma^4)$$

$$\text{Var}\{\lambda_{\text{PCA}}\} = \frac{1}{n}\mathbb{E}\{u^4\}\|\boldsymbol{v}\|^4 + \sigma^2 \frac{\mathbb{E}\{u^2\}\|\boldsymbol{v}\|^2}{n} + O(\sigma^4)$$

Leading order mean and variance

$$\mathbb{E}\{\sin\theta_{\text{PCA}}\} = \frac{\sigma}{\kappa\sqrt{n}} \frac{\sqrt{2}\Gamma\left(\frac{p-1}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{p-1}{2}\right)} + O(\sigma^2) \qquad (4)$$

The Γ functions arise from the average of the square root of a chi-squared variable. If $p \gg 1$, then

$$\mathbb{E}\{\sin\theta_{\text{PCA}}\} \approx \frac{\sigma}{\kappa}\sqrt{\frac{p-1}{n}}\left(1 - \frac{1}{4(p-1)} + O\left(\frac{1}{p^2}\right)\right) + O(\sigma^2)$$

and

$$\text{Var}\{\sin\theta_{\text{PCA}}\} \approx \frac{\sigma^2}{2\kappa^2 n}(1 + O(1/p)) + O(\sigma^2)$$