Lecture 08: Finite Sample PCA I

Nikola Zlatanov

Innopolis University

Advanced Statistics

27-th of March to 3-th of April, 2023

Motivation

- To do PCA of data vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N \in \mathbb{R}^d$, coming from some distribution $f_{\boldsymbol{X}}(\boldsymbol{x})$, we needed to have the covariance matrix $\boldsymbol{\Sigma}$.
- But what happens in cases when we do not know the covariance matrix $\boldsymbol{\Sigma}$. Can we still do PCA then as well?
- Specifically, assume we have data vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$, coming from some unknown distribution $f_{\boldsymbol{X}}(\boldsymbol{x})$ and unknown covariance matrix $\boldsymbol{\Sigma}$.
- How to do PCA then? And what will be the error we make compared to the case when we knew the covariance matrix?

Finite Sample PCA

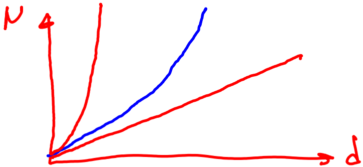
- Let us have a sample of N i.i.d. data vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_N \in \mathbb{R}^d$, coming from some unknown distribution $f_{\boldsymbol{X}}(\boldsymbol{x})$ and unknown covariance matrix $\boldsymbol{\Sigma}$.
- Let us estimate the unknown covariance matrix $\boldsymbol{\Sigma}$ as $\hat{\boldsymbol{\Sigma}}_N$ using the N samples as

$$\hat{\boldsymbol{\Sigma}}_N = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^T, \tag{1}$$

- Then, let's do PCA, using the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}_N$, on the data vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_N$.
- Hopefully, the results using the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}_N$ will be close to doing the PCA, using the actual covariance matrix $\boldsymbol{\Sigma}$, on the data vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_N$.

Finite Sample PCA

- An immediate question in when using the estimated covariance matrix $\hat{\mathbf{\Sigma}}_N$ to do PCA is the following.
- How large should the sample size N be in order for the two PCA results to be close?
- Since each of the samples $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_N$ is of dimension d, $\mathbf{\Sigma}$ is a $d \times d$ matrix.
- Then, for sure, the accuracy of the PCA using the estimated covariance matrix $\hat{\mathbf{\Sigma}}_N$ will depend on how N depends on d.
- How will N depend on d? Maybe for a good accuracy we need $N = O(d)$, or maybe $N = O(d^2)$ or maybe $\underline{N = O(e^d)}$



$d = 10^6$

$N = \exp(10^6) \sim$

$N = 10^{12}$

$\boxed{N = 10^6}$

Finite Sample PCA

- We know that the PCA result using the covariance matrix $\boldsymbol{\Sigma}$ is a function of its eigenvalues, $\lambda_i(\boldsymbol{\Sigma})$, and eigenvectors, $\boldsymbol{v}_i(\boldsymbol{\Sigma})$.
- Hence, if $\lambda_i(\hat{\boldsymbol{\Sigma}}_N)$ and $v_i(\hat{\boldsymbol{\Sigma}}_N)$ are the eigenvalues and eigenvectors of the estimated covariance matrix, $\hat{\boldsymbol{\Sigma}}_N$, then for the two PCA results to be close, the following has to hold

$$\begin{gathered} \lambda_i(\boldsymbol{\Sigma}) \approx \lambda_i(\hat{\boldsymbol{\Sigma}}_N) \\ \text{AND} \\ \boldsymbol{v}_i(\boldsymbol{\Sigma}) \approx \boldsymbol{v}_i(\hat{\boldsymbol{\Sigma}}_N), \end{gathered} \tag{2}$$

  i.e., the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}_N$ are very close to each other.
- How large should the sample size N be in order for the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}_N$ to be very close to each other?

Covariance Estimation Problem

Step 1:

- Let us measure how far $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}_N$ are from each other using some measurement/metric.
- For matrices, so far, we learned that we can use the Frobenius norm or the operator norm.
- Let's try the operator norm:

$$||\hat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}||_{\text{op}} = \sigma_1(\hat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}), \tag{3}$$

  where $\sigma_1\left(\hat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}\right)$ is the largest singular value of the matrix $\hat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}$

- Now, we have

$$\begin{aligned}\sigma_1(\hat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}) &= \left|\lambda_1(\hat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma})\right| = \max_{\boldsymbol{v},||\boldsymbol{v}||_2=1} \left|\boldsymbol{v}^T(\hat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma})\boldsymbol{v}\right| \\ &= \max_{\boldsymbol{v},||\boldsymbol{v}||_2=1} \left|\boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_N\boldsymbol{v} - \boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}\right| \end{aligned} \tag{4}$$

Distance Between True and Sampled Covariance Matrice

- The first element in (4) is

$$\boldsymbol{v}^T\hat{\boldsymbol{\Sigma}}_N\boldsymbol{v} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{v}^T\boldsymbol{X}_i\boldsymbol{X}_i^T\boldsymbol{v} = \frac{1}{N}\sum_{i=1}^{N}Y_iY_i = \frac{1}{N}\sum_{i=1}^{N}Y_i^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}\langle\boldsymbol{v},\boldsymbol{X}_i\rangle^2, \tag{5}$$

  where $Y_i = \boldsymbol{v}^T\boldsymbol{X}_i = \boldsymbol{X}_i^T\boldsymbol{v} = \langle\boldsymbol{v},\boldsymbol{X}_i\rangle$.

- The second element in (4) is

$$\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v} = \boldsymbol{v}^T E[\boldsymbol{X}\boldsymbol{X}^T]\boldsymbol{v} = E[\boldsymbol{v}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{v}] = E[YY] = E[Y^2]$$
$$= E[\langle\boldsymbol{v},\boldsymbol{X}\rangle^2] \tag{6}$$

  where $Y = \boldsymbol{v}^T\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{v} = \langle\boldsymbol{v},\boldsymbol{X}\rangle$.

Distance Between True and Sampled Covariance Matrice

- Inserting (5) and (6) into (4), and then (4) into (3), we obtain

$$||\hat{\mathbf{\Sigma}}_N - \mathbf{\Sigma}||_{\text{op}} = \max_{\boldsymbol{v},||\boldsymbol{v}||_2=1} \left| \frac{1}{N} \sum_{i=1}^{N} \langle \boldsymbol{v}, \boldsymbol{X}_i \rangle^2 - E\left[\langle \boldsymbol{v}, \boldsymbol{X} \rangle^2\right] \right|, \tag{7}$$

which is a random variable (RV).

- Let us define $Z(\boldsymbol{v})$ as

$$Z(\boldsymbol{v}) = \frac{1}{N} \sum_{i=1}^{N} \langle \boldsymbol{v}, \boldsymbol{X}_i \rangle^2 - E\left[\langle \boldsymbol{v}, \boldsymbol{X} \rangle^2\right], \text{ s.t. } ||\boldsymbol{v}||_2 = 1. \tag{8}$$

Then

$$||\hat{\mathbf{\Sigma}}_N - \mathbf{\Sigma}||_{\text{op}} = \max_{\boldsymbol{v},||\boldsymbol{v}||_2=1} |Z(\boldsymbol{v})| \tag{9}$$

Covariance Estimation Theorem

- Thm (Covariance Estimation Theorem): Let $\boldsymbol{X}_i \in \mathbb{R}^d$ be N sample vectors independently and identically drawn from the Gaussian distribution $\mathcal{N}(0, \boldsymbol{\Sigma})$. Let each element of $\boldsymbol{X}_i$ be with variance σ^2. Let $\delta \geq 0$. Then, the following holds

$$\begin{aligned}
&\Pr\left\{ ||\hat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}||_{\text{op}} \leq \delta \right\} \\
&\geq 1 - 2\exp\left(d\sqrt{\frac{2}{\epsilon}} \right) \exp\left(-N \frac{(1-\epsilon)\delta}{2\sigma^2} \min\left\{ \frac{(1-\epsilon)\delta}{2\sigma^2}, 1 \right\} \right) \qquad (10)
\end{aligned}$$

- What this Thm says is that the distance between $\hat{\boldsymbol{\Sigma}}_N$ and $\boldsymbol{\Sigma}$ can be made smaller than δ, with probability that goes to one as N increases, as long as N satisfies

$$N \frac{(1-\epsilon)\delta}{2\sigma^2} \min\left\{ \frac{(1-\epsilon)\delta}{2\sigma^2}, 1 \right\} > d\sqrt{\frac{2}{\epsilon}} \qquad (11)$$

Covariance Estimation Theorem

- What the Covariance Estimation Theorem tells us is that $N = o(d)$.
- In other words, the number of vector samples N we need to accurately estimate the covariance matrix $\boldsymbol{\Sigma}$ using

$$\hat{\boldsymbol{\Sigma}}_N = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^T \tag{12}$$

  depends linearly on the vectors $\boldsymbol{X}_i$ length d.
- This is very good news since there is no curse-of-dimensionality in this problem.
- Imagine if it was $N = o(e^d)$. Then, it would have been impossible to estimate $\hat{\boldsymbol{\Sigma}}_N$ using (12) since for $d = 100$ we would need at least e^{100} samples, which is more than the number of atoms in the observable universe.

Distance Between Finite Sample PCA and True PCA

- So far we only proved that we can make $||\hat{\mathbf{\Sigma}}_N - \mathbf{\Sigma}||_{\text{op}}$ vary small by increasing N, given that condition (11) holds.
- But PCA does not depend on $||\hat{\mathbf{\Sigma}}_N - \mathbf{\Sigma}||_{\text{op}}$ explicitly, and instead depends on the distances between the eigenvalues and eigenvectors of $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}_N$.
- In fact, for the PCA results using $\hat{\mathbf{\Sigma}}_N$ to be very close to the PCA results using $\mathbf{\Sigma}$, we needed the following to hold for all dimensions i over which we do the PCA (what happens in the other dimensions is irrelevant)

$$\lambda_i(\mathbf{\Sigma}) \approx \lambda_i(\hat{\mathbf{\Sigma}}_N) \quad \text{AND} \quad \boldsymbol{v}_i(\mathbf{\Sigma}) \approx \boldsymbol{v}_i(\hat{\mathbf{\Sigma}}_N), \tag{18}$$

  i.e., the eigenvalues and eigenvectors of $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}_N$ are very close to each other in all dimensions i over which we do the PCA.
- In the following, we will show that by making $||\hat{\mathbf{\Sigma}}_N - \mathbf{\Sigma}||_{\text{op}}$ vary small, we will implicitly achieve (18).

Distance Between Eigenvalues

- Thm (Weyl's Inequality): $\forall$ $d \times d$ symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, the following holds

$$\max_i |\lambda_i(\boldsymbol{A}) - \lambda_i(\boldsymbol{B})| \leq ||\boldsymbol{A} - \boldsymbol{B}||_{\text{op}} \tag{19}$$

- We will not prove it!
- What this theorem says is that the largest distance between the eigenvalues of any two symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is upper bounded by the operator norm of $\boldsymbol{A} - \boldsymbol{B}$
- Hence, if we make $||\hat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}||_{\text{op}}$ vary small, we implicitly make the distance between the eigenvalues of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}_N$ also very small. i.e., the following holds

$$\lambda_i(\boldsymbol{\Sigma}) \approx \lambda_i(\hat{\boldsymbol{\Sigma}}_N) \tag{20}$$

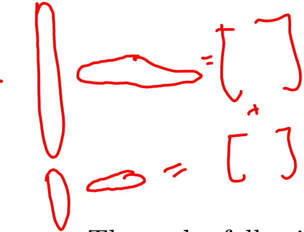
- How about the eigenvectors. Now, they are more tricky, but we can still do it as shown in the following.

Distance Between Eigenvectors

- Thm (Davis-Kahan Inequality): For any symmetric matrices $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}_N$ with eigenvectors $\boldsymbol{v}_1(\boldsymbol{\Sigma}), \boldsymbol{v}_2(\boldsymbol{\Sigma}), ..., \boldsymbol{v}_k(\boldsymbol{\Sigma})$ and $\boldsymbol{v}_1(\hat{\boldsymbol{\Sigma}}_N), \boldsymbol{v}_2(\hat{\boldsymbol{\Sigma}}_N), ..., \boldsymbol{v}_k(\hat{\boldsymbol{\Sigma}}_N)$, respectively, let $\boldsymbol{P}_{\boldsymbol{\Sigma}}$ and $\boldsymbol{P}_{\hat{\boldsymbol{\Sigma}}_N}$ be defined as

$d >> \underline{k}$

1, 2, 3



$$\boldsymbol{P}_{\boldsymbol{\Sigma}} = \sum_{i=1}^{k} \boldsymbol{v}_i(\boldsymbol{\Sigma})\boldsymbol{v}_i^T(\boldsymbol{\Sigma}) \tag{21}$$

$$\boldsymbol{P}_{\hat{\boldsymbol{\Sigma}}_N} = \sum_{i=1}^{k} \boldsymbol{v}_i(\hat{\boldsymbol{\Sigma}}_N)\boldsymbol{v}_i^T(\hat{\boldsymbol{\Sigma}}_N). \tag{22}$$

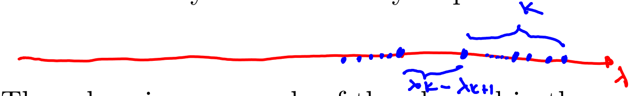
Then, the following holds

$$||\boldsymbol{P}_{\hat{\boldsymbol{\Sigma}}_N} - \boldsymbol{P}_{\boldsymbol{\Sigma}}||_{\text{op}} \leq \frac{||\hat{\boldsymbol{\Sigma}}_N - \boldsymbol{\Sigma}||_{\text{op}}}{\lambda_k(\boldsymbol{\Sigma}) - \lambda_{k+1}(\boldsymbol{\Sigma}_{\cancel{N}})} \tag{23}$$

- We will not prove it!

Distance Between Eigenvectors

- What the Davis-Kahan Inequality Thm says is that by making $||\hat{\mathbf{\Sigma}}_N - \mathbf{\Sigma}||_{\text{op}}$ vary small, we implicitly make the first k eigenvectors of $\hat{\mathbf{\Sigma}}_N$ to be very close to the first k eigenvectors of $\mathbf{\Sigma}_N$, as long as there is a 'spectral gap' between the k-th and the $(k+1)$-th eigenvalue of $\mathbf{\Sigma}_N$. $d >> k$
- 'Spectral gap' between the k-th and the $(k+1)$-th eigenvalue of $\mathbf{\Sigma}_N$, means that they are as far away as possible. See this Figure:



- This Thm also gives us a rule of thumb used in the current PCA algos: Do PCA dimension reduction up to the point when we observe a spectral gap. Anything beyond the spectral gap consider as noise and throw it away.
- In the next lecture, we will arrive at the same concept, noise that needs to be thrown away, via a different method which will reveal us even more insights into random matrices.