Principal component analysis

Julien Calbert

December 5, 2023

1 Theoretical reminder

- The eigenvalues are given in decreasing order

  $$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$$

  note that we can have $\lambda_i = \lambda_j$ for some i, j.
  Given an Hermitian matrix $A \in \mathbb{C}^{n\times n}$, we denote by $R_A(x)$, the *Rayleigh quotient* of a nonzero vector $x \in \mathbb{C}^n$

  $$R_A(x) \coloneqq \frac{x^\top A x}{x^\top x}, \qquad x \neq 0.$$

  According to the course notes, we have the following lemma and theorem,

  Lemma 1 (3.25)**.** Let $\mathcal{S}_j \subseteq \mathbb{C}^n$ be a subspace of dimension j. Then, it holds that

  $$\min_{x\neq 0 \in \mathcal{S}_j} R_A(x) \leq \lambda_j, \qquad \max_{x\neq 0\in\mathcal{S}_j} R_A(x) \geq \lambda_{n-j+1}.$$

  Theorem 1 (Courant-Fisher)**.**

  $$\max_{\mathcal{S}_j} \min_{x\neq 0\in\mathcal{S}_j} R_A(x) = \lambda_j, \qquad \min_{\mathcal{S}_j} \max_{x\neq 0\in\mathcal{S}_j} R_A(x) = \lambda_{n-j+1}.$$

  We can prove the following theorem (which is not in the lecture note):

  Theorem 2. Given an Hermitian matrix $A \in \mathbb{C}^{n\times n}$, we define

  $$\mathcal{S}_{n-j} = \{x \in \mathbb{C}^n \mid (x|x_i) = 0, i = 1, \ldots j\} = \operatorname{span}\langle x_1, \ldots, x_j\rangle^\perp$$

  where x_j is an eigenvector of A associated with λ_j and $(x_i|x_k) = 0\ \forall i = 1, \ldots, j,\ k = 1, \ldots, j, i \neq k$. So $\mathcal{S}_{n-j}$ is the subspace of the vectors orthogonal to the eigenvectors associated to the j largest eigenvalues. We have

  $$\max_{x\neq 0\in\mathcal{S}_{n-j}} R_A(x) = \lambda_{j+1} \quad \text{and} \quad x_{j+1} = \operatorname*{argmax}_{x\neq 0\in\mathcal{S}_{n-j}} R_A(x)$$

  where $x_{j+1} \in \mathcal{S}_{n-j}$ is an eigenvector associated with λ_{j+1}: $Ax_{j+1} = \lambda_{j+1}x_{j+1}$.

Proof. Since the matrix A is symmetric, we have $m_g(A) = m_a(A)$, and we know that there exists an orthogonal basis of eigenvectors: $x_1, \ldots, x_n$. Since $\mathcal{S}_{n-j} = \text{span}\langle x_1, \ldots, x_j\rangle^\perp = \text{span}\langle x_{j+1}, \ldots, x_n\rangle$ we have

$$\dim(\mathcal{S}_{n-j}) = n - j.$$

Therefore using lemma (1), and since the stationary points of the Rayleigh quotient $R_A(x)$ are exactly the eigenvectors of A (and that $R_A(x_k) = \lambda_k$), we have

$$\max_{x \neq 0 \in \mathcal{S}_{n-j}} R_A(x) = \lambda_{n-(n-j)+1} = \lambda_{j+1}$$

with $R_A(x_{j+1}) = \lambda_{j+1}$ and $x_{j+1} \in \mathcal{S}_{n-j}$. □

- Given $A \in \mathbb{C}^{m \times n}$, we can compute the SVD of A:

$$A = U\Sigma V^*$$

with $UU^* = U^*U = I_m$, $VV^* = V^*V = I_n$ and $\Sigma \in \mathbb{R}^{m \times n}$ a "diagonal" matrix. The columns of U are called the left singular vectors of A. The columns of V are called the right singular vectors of A. The diagonal entries of A are the singular values of A.

If A is an Hermitian matrix ($A = A^*$), therefore we have

  - $AA^* = U\Sigma V^* V \Sigma^* U^* = U\Sigma^2 U^*$ (which is the eigenvalue decomposition of AA^*). We have $\sigma_i(A)^2 = \lambda_i(AA^*)$ and the left singular vectors of A are the eigenvectors of AA^*.
  - $A^*A = V\Sigma^* U^* U \Sigma V^* = V\Sigma^2 V^*$ (which is the eigenvalue decomposition of A^*A). We have $\sigma_i(A)^2 = \lambda_i(A^*A)$ and the right singular vectors of A are the eigenvectors of A^*A.

- Best low rank approximation in Frobenius norm

Theorem 3. Given a matrix $A \in \mathbb{C}^{m \times n}$ of rank r and $l \leq r$, we have

$$\begin{aligned} \min_{B \in \mathbb{R}^{m \times n}} \quad & \|A - B\|_F^2 = \sum_{i=l+1}^{r} \sigma_i^2(A) \\ \text{s.t.} \quad & \text{rank}(B) \leq l \end{aligned} \tag{1}$$

And the solution is given the truncated SVD of $A = U\Sigma V^*$: $B = \sum_{i=1}^{l} \sigma_i(A) u_i v_i^*$.

2 PCA

Given $X \in \mathbb{R}^{m \times n}$, find $P \in \mathbb{R}^{m \times m}$ **orthogonal** such that the data in the new coordinate system

$$Y = PX$$

1. minimize (linear) redundancy
2. maximize the variance ("information")

The **rows** of P are the new direction of the new coordinate system. Indeed, let $p_1, p_2, \dots, p_m$ be the new directions. I want that p_i becomes e_i after my change of variables, i.e., I want that

$$p_i = P^{-1} e_i \Leftrightarrow p_i = P^\top e_i$$

since P is orthogonal (or see the justification in the document).

It is assumed that the data are centered.
Here are 3 equivalent formulations of the PCA.

2.1

PCA computes the new orthonormal basis with the following **greedy** algorithm.

Algorithm (2)

1. select a normalized direction $p_1 \in \mathbb{R}^m$ along with the variance in X is maximized.
2. For $k = 2, \dots, m$:
   search for a direction $p_k \in \mathbb{R}^m$ among the direction orthogonal to $p_1, \dots, p_{k-1}$, along with the variance is maximized.

The variance of the point $x_i \in \mathbb{R}^m$ along the direction $p \in \mathbb{R}^m$ ($\|p\|_2 = 1$) is $(x_i|p)^2$ (it is the squared norm of the projection of x_i onto $\text{span}\langle p\rangle$: $\text{VAR}(x_i) = \|(x_i|p)p\|_2^2 = (x_i|p)^2\|p\|_2^2 = (x_i|p)^2$.

$$\begin{aligned}
p_1 &= \underset{\|p\|_2=1}{\text{argmax}}\ \frac{1}{n}\sum_{i=1}^{n}(x_i|p)^2 = \underset{\|p\|_2=1}{\text{argmax}}\ \frac{1}{n}\|X^\top p\|_2^2 = \underset{\|p\|_2=1}{\text{argmax}}\ p^\top\left(\frac{XX^\top}{n}\right)p \\
&= \underset{p\neq 0}{\text{argmax}}\ \frac{p^\top C_X p}{p^\top p} = \underset{\|p\|_2=1}{\text{argmax}}\ R_{C_X}(p) \\
&= \frac{w_1}{\|w_1\|_2}.
\end{aligned}$$

where w_1 is an eigenvector of $C_X = \frac{XX^\top}{n}$ associated with the largest eigenvalue λ_1 of C_X. And the value is $R_{C_X}(p_1) = R_{C_X}(w_1) = \lambda_1$.

The other directions are obtained by successively applying the following procedure for $k = 2, \dots, m$

$$\begin{aligned}
p_k &= \underset{\|p\|_2=1}{\text{argmax}}\quad \frac{1}{n}\sum_{i=1}^{n}(x_i|p)^2 \\
&\qquad \text{s.t.}\quad (p|p_i) = 0 \quad i = 1, \dots, k-1 \\
&= \underset{p\in\mathcal{S}_{n-k+1}}{\text{argmax}}\ R_{C_X}(p) \\
&= \frac{w_k}{\|w_k\|_2}.
\end{aligned}$$

Where $\mathcal{S}_{n-k+1} = \text{span}\langle p_1, \dots, p_{k-1}\rangle^\perp$ ($\dim(\mathcal{S}_{n-k+1}) = n-(k-1) = n-k+1$). Therefore using theorem (1), the solution is given by w_k which is an eigenvector of $C_X = \frac{XX^\top}{n}$ associated with the k^{th} largest eigenvalue λ_k of C_X. And the value is $R_{C_X}(p_k) = R_{C_X}(w_k) = \lambda_k$.

Therefore the matrix $P = \begin{pmatrix} p_1 & p_2 & \dots & p_m \end{pmatrix}^\top$ is equal to the orthonormal matrix composed of the eigenvectors of C_X transpose. And the variance of the new data Y along the direction $p_1, \dots, p_m$ is equal to $\sum_{i=1}^m \lambda_i(C_X)$. So

- PCA projects the data along the directions where the data varies the most.
- These directions are determined by the eigenvectors of the covariance matrix corresponding to the largest eigenvalues.
- The magnitude of the eigenvalues corresponds to the variance of the data along the eigenvector directions.

2.2 PCA with eigenvalue decomposition

We search for an orthogonal matrix $P \in \mathbb{R}^{m\times m}$ such that $C_Y = \frac{YY^\top}{n}$ is diagonalized. Note: there are several way of diagonalizing a matrix, PCA will chose the "simplest" way.

So, $C_Y = \frac{1}{n}YY^\top = P\frac{1}{n}XX^\top P^\top = PC_XP^\top$.

We compute the eigenvector decomposition of C_X

$$C_X = EDE^\top.$$

Therefore, if we choose $P = E^\top$, we have

$$C_Y = E^\top EDE^\top E = D.$$

The entries D_{ii} which are the variance of the dataset Y (or X) along the new directions P are the eigenvalues of C_X: $[C_Y]_{ii} = \lambda_i(C_X)$. Therefore PCA computes the transformation P that diagonalize the empirical sample covariance matrix. The matrix P obtained is the same that we obtained with algorithm 2.

2.3 PCA with singular value decomposition

We compute the SVD of $Z = \frac{1}{\sqrt{n}}\ X^\top = U\Sigma V^\top$, $U \in \mathbb{R}^{n\times n}, V \in \mathbb{R}^{m\times m}, \Sigma \in \mathbb{R}^{n\times m}$. We have $C_X = \frac{XX^\top}{n} = V\Sigma^2V^\top$.
If we apply the following orthogonal transformation $Y = PX$ with $P = V^\top$, we obtain

$$C_Y = \frac{YY^\top}{n} = \frac{1}{n}V^\top XX^\top V = V^\top V\Sigma^2V^\top V = \Sigma^2.$$

Therefore the matrix P computed by PCA with algorithm (2) is $P = E^\top = V^\top$, i.e., the rows of P are

- the directions of the new bases
- the eigenvectors of C_X

- the right singular vectors of $\frac{1}{\sqrt{n}}X^\top$.

The diagonal entries of C_Y are

- the variance of X along the directions of the new basis (p_i)
- the eigenvalues of C_X
- the squared singular values of $\frac{1}{\sqrt{n}}X^\top$.

3 Dimensionality reduction

We keep only $l \leq m$ features in the new coordinate system. We will prove that PCA minimizes the loss of "information" (variance) when we back-project to the original coordinate system.

Let $P_l = \begin{pmatrix} p_1 & p_2 & \dots & p_l \end{pmatrix}^\top \in \mathbb{R}^{l \times m}$. Let

$$Y_l = P_l X \in \mathbb{R}^{l \times n}$$

be the data in the new coordinate system where we have kept only the l-features. Let

$$\hat{X} = P_l^\top Y_l = P_l^\top P_l X,$$

be the reconstruction of the data from their reduction in the new coordinate system.
In the particular case where $l = m$, we have $\hat{X} = P^\top P X = X$. Otherwise we have $P_l P_l^\top = I_l$ but $P_l^\top P_l \neq I_m$.

Given $\frac{1}{\sqrt{n}}X^\top = U\Sigma V^\top$. We partition the matrices $V = [V_l\ \tilde{V}_l]$ with $V_l \in \mathbb{R}^{m \times l}$ the l first columns of V, $U = [U_l\ \tilde{U}_l]$ and $\Sigma = \begin{pmatrix} \Sigma_l & 0 \\ 0 & \tilde{\Sigma}_l \end{pmatrix}$.

We have $P = V^\top$ and $P_l = V_l^\top$

$$\begin{aligned} \hat{X} &= P_l^\top P_l X = V_l V_l^\top X = V_l V_l^\top \sqrt{n} V \Sigma U^\top = V_l [I_l\ 0] \Sigma U^\top \sqrt{n} = V_l \begin{pmatrix} \Sigma_l & 0 \end{pmatrix} U^\top \sqrt{n} = \sqrt{n} V_l \Sigma_l U_l^\top \\ &= l\text{-truncated SVD of } X \end{aligned}$$

and

$$\begin{aligned} C_{\hat{X}} &= \frac{\hat{X}\hat{X}^\top}{n} = V_l \Sigma_l U_l^\top U_l \Sigma_l V_l^\top = V_l \Sigma_l^2 V_l^\top \\ &= l\text{-truncated SVD of } C_X \end{aligned}$$

Both matrices $\hat{X}$ and $C_{\hat{X}}$ are of rank l. Therefore using Theorem (3), we have that PCA minimize the error of reconstruction

$$\begin{aligned} \min_{\hat{X} \in \mathbb{R}^{m \times n}} \quad & \|X - \hat{X}\|_F^2 = \sum_{i=l+1}^{r} \sigma_i^2(X) = \sum_{i=l+1}^{r} \lambda_i(XX^\top) = n \sum_{i=l+1}^{r} \lambda_i(C_X) \\ \text{s.t.} \quad & \text{rank}(\hat{X}) \leq l \end{aligned} \tag{3}$$

(the solution of this problem is the matrix reconstructed from PCA).
Therefore, the mean squared error of reconstruction is

$$E = \frac{1}{n}\|X - \hat{X}\|_F^2 = \frac{1}{n}\sum_{i=1}^{n}\|x_i - \hat{x}_i\|_2^2 = \sum_{i=l+1}^{r}\lambda_i(C_X).$$

We also have

$$\begin{aligned} \min_{C_X \in \mathbb{R}^{m\times n}} \quad & \|C_X - C_{\hat{X}}\|_F^2 = \sum_{i=l+1}^{r}\sigma_i^2(C_X) \\ \text{s.t.} \quad & \operatorname{rank}(C_{\hat{X}}) \leq l \end{aligned} \tag{4}$$

(the solution of this problem is the covariance matrix of the reconstructed data obtained with PCA).

PCA is therefore the (linear) dimensionality reduction that minimizes the reconstruction error in the Frobenius norm.

3.1 Inconvenient

- The covariance matrix represents only second order statistics among the vector values.
- Since the new variables (features) are linear combinations of the original variables, it is usually difficult to interpret their meaning.
- (see paper)