Tutorial 5 : PCA

1 Theory - Principal Component Analysis

A set of vectors is given: $\{x_k\}_{k=1}^n,\ x_k \in \mathbb{R}^d$, with a large dimension $d \gg 1$.

Objective

Represent these vectors (or their information) by vectors of a smaller dimension $m < d$.

Assumption: $\hat{\mu} = \frac{1}{n}\sum_{k=1}^n x_k = 0$. Otherwise, we pre-process the samples: $x_k \leftarrow x_k - \hat{\mu}$.

Define $\hat{\Sigma}_n \triangleq \frac{1}{n}\sum_{k=1}^n x_k x_k^T$ as the empirical covariance of the set.

The empirical variance in the direction of the unit vector $w \in \mathbb{R}^d$ is defined by:

$$\frac{1}{n}\sum_{k=1}^n \left(w^T x_k\right)^2 = w^T \left(\frac{1}{n}\sum_{k=1}^n x_k x_k^T\right) w = w^T \hat{\Sigma}_n w \triangleq S_n(w).$$

Recall that the projection of x_k onto w is given by $\langle w, x_k \rangle = w^T x_k$.

First principal component of $\{x_k\}_{k=1}^n$ - The unit vector $w_1 \in \mathbb{R}^d$ which maximizes $S_n(w)$.

$\mathbf{m^{th}}$ principal component of $\{x_k\}_{k=1}^n$ - The vector $w_m \in \mathbb{R}^d$ which maximizes $S_n(w)$, among all unit vectors that are orthogonal to $w_1, w_2, ..., w_{m-1}$.

<u>**Claim**</u> (Proven in Class)

The principal components are given by $w_m = v_m$ where v_m is the eigenvector of $\hat{\Sigma}_n$ that corresponds to the eigenvalue λ_m $(\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n)$.

PCA - Summary:

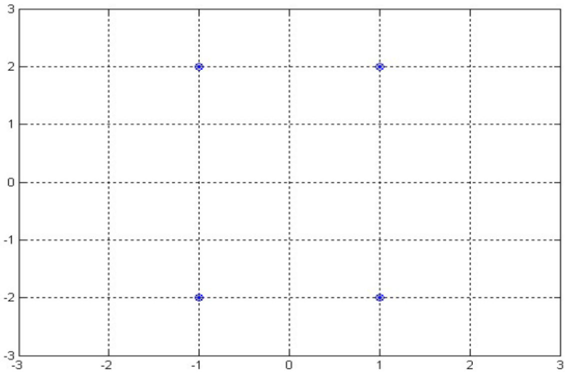
1. Centering - $x_k \leftarrow x_k - \hat{\mu}$ where $\hat{\mu} = \frac{1}{n}\sum_{k=1}^n x_k$ is the empirical average.
2. Compute the empirical covariance matrix - $\hat{\Sigma}_n = \frac{1}{n}\sum_{k=1}^n x_k x_k^T$.
3. Perform eigen-decomposition:
$$\hat{\Sigma}_n = V \Lambda V^T,$$
where $V = \{v_1, v_2, ..., v_n\}$ is a matrix whose columns are the eigenvectors and Λ is a diagonal matrix with the eigenvalues on its diagonal.
4. Choose the m principal components to be the eigenvectors corresponding to the m largest eigenvalues.

2 Practice

Question 1

Consider the classification problem with an input space $X = \mathbb{R}^2$ and an output space $Y = \{+1, -1\}$. A training set $\{x_k, y_k\}_{k=1}^4$ is given

$$\begin{array}{cccc} x_1 = (-1,2)^T & x_2 = (1,2)^T & x_3 = (-1,-2)^T & x_4 = (1,-2)^T \\ y_1 = +1 & y_2 = +1 & y_3 = -1 & y_4 = -1 \end{array}$$

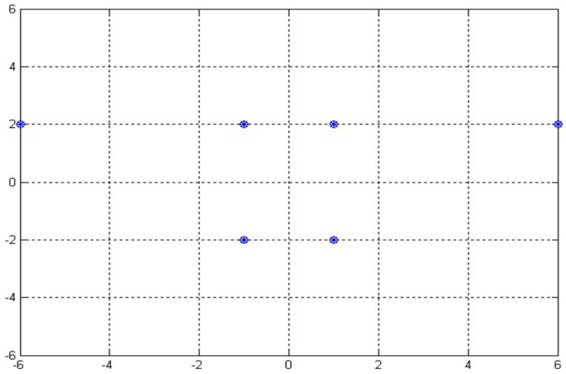


(a) Calculate the empirical covariance $\hat{\Sigma}_4$ of the given set of examples. What is the first principal component of the set?

(b) Suggest a linear classifier which classifies without errors the given set of examples, based on the projection of the examples on the first principal component.

(c) Now, assume that two examples are added to the set:

$$\begin{array}{cc} x_5 = (6,2)^T & x_6 = (-6,2)^T \\ y_5 = +1 & y_6 = +1 \end{array}$$



Calculate the empirical variance $\hat{\Sigma}_6$ of the new given set of examples.
What is the first principal component?

(d) In this case, can the training set be classified without error by using a linear classifier, based on the projection of the examples on the first principal component only?

Solution

(a) First we compute the empirical mean

$$\hat{\mu} = \frac{1}{4}\sum_{k=1}^{4} = \frac{1}{n}(-1+1-1+1, 2+2-2-2)^T = (0,0)^T.$$

Hence, the samples are centered. Next, we compute the empirical covariance

$$\begin{aligned}\hat{\Sigma}_4 &= \frac{1}{4}\left[(-1,2)^T(-1,2) + (1,2)^T(1,2) + (-1,-2)^T(-1,-2) + (1,-2)^T(1,-2)\right] \\ &= \frac{1}{4}\left[\begin{bmatrix}1 & -2\\ -2 & 4\end{bmatrix} + \begin{bmatrix}1 & 2\\ 2 & 4\end{bmatrix} + \begin{bmatrix}1 & 2\\ 2 & 4\end{bmatrix} + \begin{bmatrix}1 & -2\\ -2 & 4\end{bmatrix}\right] = \frac{1}{4}\begin{bmatrix}4 & 0\\ 0 & 16\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 4\end{bmatrix}\end{aligned}$$

The covariance matrix is diagonal, hence, it is easy to that the first principal component is $v_1 = (0,1)^T$, i.e., the second coordinate (as can be seen from the drawing).

(b) The classifier $y_k = sign(x_k^T v_1)$ classifies without errors the given set of examples. One may think that the principles components are related to the labels but this is not necessarily true.

(c) First we compute a empirical mean

$$\hat{\mu} = \frac{1}{6}\sum_{k=1}^{6} = \frac{1}{6}(0, 2+2)^T = (0, \frac{2}{3})^T$$

Therefore we need to center the samples

$$\begin{aligned}\tilde{x}_1 &= (-1, \frac{4}{3})^T,\ \tilde{x}_2 = (1, \frac{4}{3})^T,\ \tilde{x}_3 = (-1, -\frac{8}{3})^T, \\ \tilde{x}_4 &= (1, -\frac{8}{3})^T,\ \tilde{x}_5 = (6, \frac{4}{3})^T,\ \tilde{x}_6 = (-6, \frac{4}{3})^T.\end{aligned}$$

Then, the empirical covariance is

$$\begin{aligned}\hat{\Sigma}_6 &= \frac{1}{6}\left[(-1,\frac{4}{3})^T(-1,\frac{4}{3}) + (1,\frac{4}{3})^T(1,\frac{4}{3}) + (-1,-\frac{8}{3})^T(-1,-\frac{8}{3}) + (1,-\frac{8}{3})^T(1,-\frac{8}{3}) + (6,\frac{4}{3})^T(6,\frac{4}{3}) + (-6,\frac{4}{3})^T(-6,\frac{4}{3})\right] \\ &= \begin{bmatrix}12\frac{2}{3} & 0\\ 0 & 3\frac{5}{9}\end{bmatrix}.\end{aligned}$$

Hence, now the first principle component is $v_1 = (1,0)^T$.

(d) Lets compute the projection of samples x_2 and x_4 onto $v_1 = (1,0)^T$:

$$\begin{aligned}x_2^T v_1 &= (1,2)(1,0)^T = 1, \\ x_4^T v_1 &= (1,-2)(1,0)^T = 1.\end{aligned}$$

Both samples have the same projection but their labels are different. Therefore, there is no linear classifier based on v_1 which can classify the training set without errors. We can conclude that in general there in no relation between dimensionality reduction using PCA and the labels. This is expected since we do not consider the labels when performing PCA.

Question 2

Consider a given set of examples $\{x_k\}_{k=1}^n,\ x_k \in \mathbb{R}^d$. A linear transformation, represented by a unitary matrix $A \in \mathbb{R}^{d\times d}$, is applied to the samples:

$$\tilde{x}_k = Ax_k \in \mathbb{R}^d,\ k = 1, 2, ..., n.$$

(a) What are the principal components of $\{\tilde{x}_k\}_{k=1}^n$? Compute the projection of the set onto the principal component. Explain the results.

(b) Now, following the linear transformation, noise is added to each example:

$$\tilde{x}_k = Ax_k + \epsilon_k \in \mathbb{R}^d,\ k = 1, 2, ..., N,$$

where

$$\frac{1}{n}\sum_{k=1}^n \epsilon_k = 0,\ \frac{1}{n}\sum_{k=1}^n x_k\epsilon_k^T = 0,\ \frac{1}{n}\sum_{k=1}^n \epsilon_k\epsilon_k^T = \lambda I\ (\lambda > 0).$$

Repeat (a) and (b) for this case.

Solution

(a) Lets compute the covariance matrix

$$\widetilde{\Sigma}_n = \frac{1}{n}\sum_{k=1}^n \tilde{x}_k\tilde{x}_k^T = \sum_{k=1}^n Ax_kx_k^TA^T = A\left(\sum_{k=1}^n x_kx_k^T\right)A^T = A\hat{\Sigma}_nA^T.$$

Using spectral decomposition we can write:

$$\hat{\Sigma}_n = V\Lambda V^T,$$

where the columns of V are the principal components of $\{x_k\}_{k=1}^n$. Therefore,

$$\widetilde{\Sigma}_n = AV\Lambda V^TA^T \equiv \tilde{V}\Lambda\tilde{V}^T$$

where the columns of $\tilde{V} = AV$ are the principal components of $\{\tilde{x}_k\}_{k=1}^n$. Hence, we can conclude that applying a linear transformation on the samples results in applying the same transformation on the principal components.
The projection of sample x_k on the mth principal components is given by

$$\langle\tilde{x}_k, \tilde{v}_m\rangle = \tilde{x}_k^T\tilde{v}_m = (Ax_k)^T(Av_m) = x_k^TA^TAv_m \underset{A^TA=I}{=} x_k^Tv_m = \langle x_k, v_m\rangle.$$

This results implies that applying an orthogonal linear transformation on the samples does not change the projections (representations). Note this is true also for $A \in \mathbb{R}^{s\times d}$ with $s \geq d$ as long as $A^TA = I \in \mathbb{R}^{d\times d}$.

(b) The empirical mean in this case is

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^n \tilde{x}_k = \frac{1}{n}\sum_{k=1}^n Mx_k + \frac{1}{n}\sum_{k=1}^n \epsilon_k = M\underbrace{\frac{1}{n}\sum_{k=1}^n x_k}_{=0} + \underbrace{\frac{1}{n}\sum_{k=1}^n \epsilon_k}_{=0} = 0.$$

The empirical covariance matrix is given by

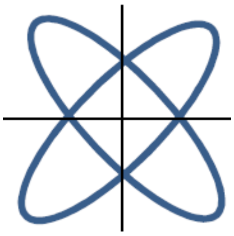
$$\begin{aligned}
\widetilde{\Sigma}_n = \frac{1}{n}\sum_{k=1}^n \tilde{x}_k\tilde{x}_k^T &= \frac{1}{n}\sum_{k=1}^n Ax_kx_k^TA^T + \frac{2}{n}\sum_{k=1}^n Ax_k\epsilon_k^T + \frac{1}{n}\sum_{k=1}^n \epsilon_k\epsilon_k^T \\
&= A\hat{\Sigma}_nA^T + 2A\left(\underbrace{\frac{1}{n}\sum_{k=1}^n x_k\epsilon_k^T}_{=0}\right) + \underbrace{\frac{1}{n}\sum_{k=1}^n \epsilon_k\epsilon_k^T}_{=\lambda I} \\
&= A\hat{\Sigma}_nA^T + \lambda I \\
&= AV\Lambda(AV)^T + \lambda I \quad \text{(spectral decomposition)} \\
&= AV\Lambda(AV)^T + AV(\lambda I)(AV)^T \quad (A \text{ is unitary} \Rightarrow AV \text{ is unitary}) \\
&= AV(\Lambda + \lambda I)(AV)^T.
\end{aligned}$$

Therefore, the principal components (the eigenvectors) are modified by a linear transformation (as before) while the eigenvalues are increased by a constant λ. Since the same value is added to all the eigenvalues, their order remains, thus, also the order of the principal components given by $\{MV_i\}_{i=1}^d$. This implies that PCA is insensitive to the addition of iid noise. However, when the noise is with different variance at each coordinate the order of the principal components might change and it should be taken into consideration.

Question 3 (Self-Reading)

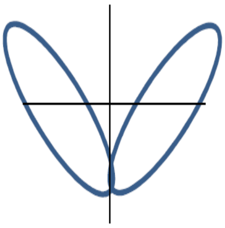
Consider a two dimensional random vector X, which with probability 0.5 has a Gaussian distribution $\mathcal{N}(\mu_1, \Sigma_1)$, and with probability 0.5 has a Gaussian distribution $\mathcal{N}(\mu_2, \Sigma_2)$ where

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ \Sigma_1 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \ \mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ \Sigma_2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$



(a) Find the principal components v_1, v_2 of a set $\{x_k\}_{k=1}^n$ of samples of X for $n \to \infty$. What is the variance at an arbitrary direction defined by a unit vector $\hat{w}$?

(b) Repeat (a) for $\mu_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mu_2 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$ (Σ_1, Σ_2 remain unchanged). Would you expect to get the same result for the variance at direction $\hat{w}$? (No need for explicit computation).



Solution

(a) According the central limit theorem, the empirical mean and covariance will converge to the true mean and covariance for $n \to \infty$. Therefore, we need to compute the mean $\mu = E[X]$ and the covariance $\Sigma = E\big[(X-\mu)(X-\mu)^T\big]$. To that end, we define a random variable Z where $Z = 1$ if the first Gaussian $\mathcal{N}(\mu_1, \Sigma_1)$ is chosen and $Z = 2$ if the second Gaussian $\mathcal{N}(\mu_2, \Sigma_2)$ is chosen. By smoothing theorem:

$$\mu = E[X] = E_Z\left[E_{X|Z}[X|Z]\right] = \frac{1}{2}\begin{bmatrix} 0 \\ 0 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$\Sigma = E[XX^T] = \frac{1}{2}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = 2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Hence, the principal components are $v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. The variance at direction $\hat{w}$ is

$$\sigma_{\hat{W}}^2 = \hat{w}^T \Sigma \hat{w} = 2\hat{w}^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \hat{w} = 2||\hat{w}||_2^2 = 2.$$

Hence, the variance does not dependent on the direction and is a constant for all directions. This is expected since the distribution of X is symmetric (see the first drawing) and the variance at directions v_1 and v_2 are the same, thus, it will remain the same at each direction that is a (normalized) linear combination of v_1 and v_2.

(b) We compute the new mean

$$\tilde{\mu} = E[X] = E_Z\left[E_{X|Z}[X|Z]\right] = \frac{1}{2}\begin{bmatrix}1\\0\end{bmatrix} + \frac{1}{2}\begin{bmatrix}-1\\0\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}.$$

To compute the new covariance matrix we the following relation for a random vector Y with mean μ_Y:

$$\Sigma_Y = E[YY^T] - \mu_Y\mu_Y^T \quad \Rightarrow \quad E[YY^T] = \Sigma_Y + \mu_Y\mu_Y^T.$$

Hence,

$$\begin{aligned}\tilde{\Sigma} &= E\big[(X-\tilde{\mu})(X-\tilde{\mu})^T\big] = E[XX^T] = \frac{1}{2}E[XX^T|Z=1] + \frac{1}{2}E[XX^T|z=2]\\ &= \frac{1}{2}(\Sigma_1 + \mu_1\mu_1^T) + \frac{1}{2}(\Sigma_2 + \mu_2\mu_2^T)\\ &= \frac{1}{2}\left(\begin{bmatrix}2\,1\\1\,2\end{bmatrix} + \begin{bmatrix}1\,0\\0\,0\end{bmatrix} + \begin{bmatrix}2\ -1\\-1\,2\end{bmatrix} + \begin{bmatrix}1\,0\\0\,0\end{bmatrix}\right) = \begin{bmatrix}3\,0\\0\,2\end{bmatrix}\end{aligned}$$

We can see from the above that the principal components remain the same, however, the variance at each direction is different (no symmetry), hence, the variance at a certain direction $\hat{w}$ will depend on the direction.