P10 - PCA

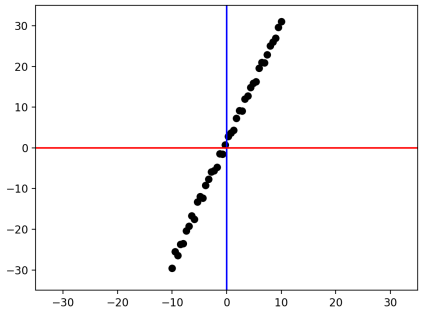
Luis Sa-Couto[1] and Andreas Wichert[2]

INESC-ID, Instituto Superior Tecnico, Universidade de Lisboa
{luis.sa.couto,andreas.wichert}@tecnico.ulisboa.pt

1 Principal Component Analysis (PCA)

Lower dimensional problems are generally easier to deal with than higher dimensional ones. Furthermore, it is often the case when data is represented with a needlessly large amount of features. For instance, think about images. A large number of pixels will correspond to a data collection with a large amount of features. However, images really live in a lower dimensional space since they capture a three-dimensional world. PCA is a technique that tries to discover the lower dimensional space where high dimensional points really live.

To exemplify the reasoning behind it, let us look at a group of two-dimensional points.



By looking at the data, we can see that although there are two features, the data seems to really vary accross a one dimensional diagonal line. The idea is to find that line and use it as our coordinates system.

The covariance matrix describes the variations of the data accross all dimensions. Through linear algebra, we can find the coordinate system that fits the data variation by computing the eigenvectors of the covariance matrix.

Eigenvalues and Eigenvectors An eigenvector of a matrix is a vector for which the transformation of that vector by that matrix yields the same vector multiplied by a scalar called the eigenvalue. In practice, this means the following:

$$\mathbf{Cu} = \lambda\mathbf{u} \iff (\mathbf{C} - \lambda\mathbf{I})\,\mathbf{u} = \mathbf{0}$$

Many eigenvectors exist with different scales. So, for the condition to met we must have an indetermined system with nontrivial solutions. This means that the determinant of the coefficients must equal zero. In fact, the determinant of the coefficients is called the characteristic polynomial:

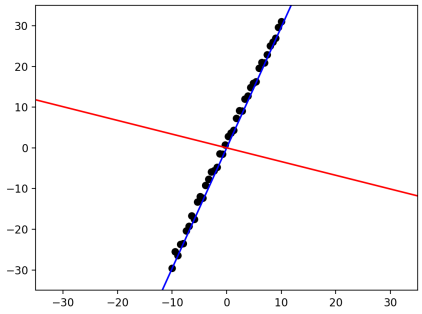
$$|\mathbf{C} - \lambda\mathbf{I}|$$

The eigenvalues λ are such that:

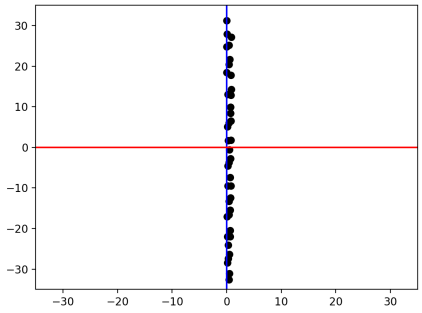
$$|\mathbf{C} - \lambda\mathbf{I}| = 0$$

Finding the list of eigenvalues we can find each corresponding eigenvector by choosing any vector that verifies $(\mathbf{C} - \lambda\mathbf{I})\,\mathbf{u} = \mathbf{0}$. However, it is usual to work with orthonormal eigenvectors, so, typically, one normalizes the chosen vector.

For our example, the eigenvectors define the two directions denoted by blue and red.



The K-L transform If we consider a coordinate system defined by the eigenvectors, we see in the figure below that mostly one feature varies while the other remains more or less constant.



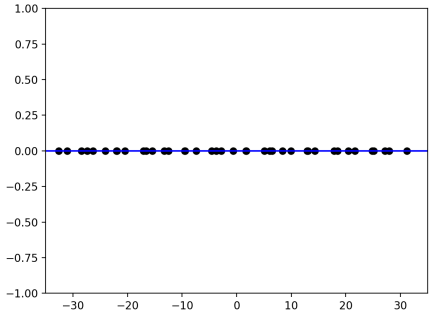
The matrix that rotates the coordinate system from the original one to the eigenvector defined one is called the K-L transform and is defined by placing the eigenvector on the columns:

$$\mathbf{U}_{KL} = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ | & | & & | \end{pmatrix}$$

So, to map a point $\mathbf{x}$ on the original coordinate system to the new one we do:

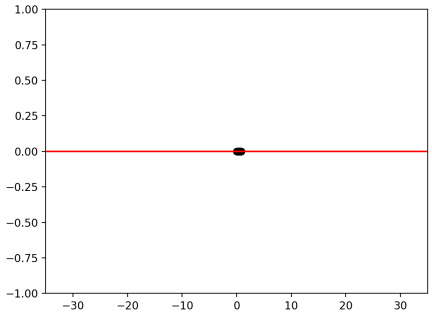
$$\mathbf{x}_{eig} = \mathbf{U}_{KL}^T \mathbf{x}$$

Mapping to a lower dimensional space Some dimensions are more relevant than others. The idea behind PCA is to discard the low importance dimensions. In our case, that would mean discard the red one and keep only the blue one:



As we can see, most variations is kept and few information was lost.

If we discard the most significant dimesion, we see that we lose all information since all points collapse:



With that said, the key idea here is to define importance. How to we know which eigendimensions to discard? The eigenvalues store this information. Larger eigenvalues mean more variation. So, one well known criterion, the Kaiser criterion, is to discard all dimensions with eigenvalue $\lambda_i < 1$.

So, to build the PCA transformation matrix, all we need to do is take the K-L matrix and remove the columns that correspond to eigenvectors we want to discard.

Exercise: Given the following training data:

$$\mathbf{x}^{(1)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{x}^{(2)} = \begin{pmatrix} 4 \\ 0 \end{pmatrix}, \mathbf{x}^{(3)} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \mathbf{x}^{(4)} = \begin{pmatrix} 6 \\ 3 \end{pmatrix}$$

a) Compute the K-L transformation.

Solution:
First, we compute the mean vector:

$$\mu = \frac{1}{4}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 4 \\ 0 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \end{pmatrix} + \begin{pmatrix} 6 \\ 3 \end{pmatrix}\right) = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$

Now, we can compute the covariance matrix. First, we compute the individual contributions from each point:

$$\mathbf{x}^{(1)} \rightarrow \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix}\right)\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix}\right)^T = \begin{pmatrix} -3 \\ -1 \end{pmatrix}\begin{pmatrix} -3 & -1 \end{pmatrix} = \begin{pmatrix} 9 & 3 \\ 3 & 1 \end{pmatrix}$$

$$\mathbf{x}^{(2)} \rightarrow \left(\begin{pmatrix} 4 \\ 0 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix}\right)\left(\begin{pmatrix} 4 \\ 0 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix}\right)^T = \begin{pmatrix} 1 \\ -1 \end{pmatrix}\begin{pmatrix} 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

$$\mathbf{x}^{(3)} \rightarrow \left(\begin{pmatrix} 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix}\right)\left(\begin{pmatrix} 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix}\right)^T = \begin{pmatrix} -1 \\ 0 \end{pmatrix}\begin{pmatrix} -1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

$$\mathbf{x}^{(4)} \rightarrow \left(\begin{pmatrix} 6 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix}\right)\left(\begin{pmatrix} 6 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix}\right)^T = \begin{pmatrix} 3 \\ 2 \end{pmatrix}\begin{pmatrix} 3 & 2 \end{pmatrix} = \begin{pmatrix} 9 & 6 \\ 6 & 4 \end{pmatrix}$$

Now, we can get the covariance matrix:

$$C = \frac{1}{4-1}\left(\begin{pmatrix} 9 & 3 \\ 3 & 1 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 9 & 6 \\ 6 & 4 \end{pmatrix}\right) = \frac{1}{3}\begin{pmatrix} 20 & 8 \\ 8 & 6 \end{pmatrix} = \begin{pmatrix} \frac{20}{3} & \frac{8}{3} \\ \frac{8}{3} & 2 \end{pmatrix}$$

To get the intended transformation we need to compute C's eigenvectors. We know that the eigenvalues λ of a matrix are the roots of the characteristic polynomial:

$$|\mathbf{C} - \lambda\mathbf{I}| = 0$$

Since we are working with two-dimensional data we know that there will be two eigenvalues. So, our equation becomes:

$$
\begin{aligned}
|\mathbf{C}-\lambda\mathbf{I}| &= \left|\begin{pmatrix} \frac{20}{3} & \frac{8}{3} \\ \frac{8}{3} & 2 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}\right| \\
&= \left|\begin{pmatrix} \frac{20}{3}-\lambda & \frac{8}{3}-0 \\ \frac{8}{3}-0 & 2-\lambda \end{pmatrix}\right| \\
&= \left(\frac{20}{3}-\lambda\right)(2-\lambda) - \left(\frac{8}{3}-0\right)\left(\frac{8}{3}-0\right) \\
&= \frac{40}{3} - \frac{20}{3}\lambda - 2\lambda + \lambda^2 - \frac{64}{9} \\
&= \lambda^2 - \frac{26}{3}\lambda + \frac{56}{9} \\
&= 0
\end{aligned}
$$

Solving the second degree equation, we get:

$$\lambda_1 = 0.79 \vee \lambda_2 = 7.88$$

Having the eigenvalues, we can get the eigenvectors. If we have that $\lambda_1 = 0.79$, then the corresponding eigenvector $u_1 = \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix}$ will verify the following:

$$\mathbf{C}\mathbf{u}_1 = \lambda_1\mathbf{u}_1 \iff (\mathbf{C}-\lambda_1\mathbf{I})\,\mathbf{u}_1 = 0$$

Which yields:

$$
\begin{gathered}
(\mathbf{C}-\lambda_1\mathbf{I})\,\mathbf{u}_1 = \mathbf{0} \\
\left[\begin{pmatrix} \frac{20}{3} & \frac{8}{3} \\ \frac{8}{3} & 2 \end{pmatrix} - \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_1 \end{pmatrix}\right]\begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
\begin{pmatrix} \frac{20}{3}-\lambda_1 & \frac{8}{3} \\ \frac{8}{3} & 2-\lambda_1 \end{pmatrix}\begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
\begin{pmatrix} \left(\frac{20}{3}-\lambda_1\right)u_{11} + \frac{8}{3}u_{12} \\ \frac{8}{3}u_{11} + (2-\lambda_1)\,u_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}
\end{gathered}
$$

So, the following condition must be met:

$$u_{12} = -\frac{3}{8}\left(\frac{20}{3}-\lambda_1\right)u_{11}$$

With that, a solution for the system will be:

$$\mathbf{u}_1 = \begin{pmatrix} u_{11} \\ -\frac{3}{8}\left(\frac{20}{3}-\lambda_1\right)u_{11} \end{pmatrix}$$

We can choose for instance $u_{11} = 1$ and get:

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ -\frac{3}{8}\left(\frac{20}{3}-0.79\right) \end{pmatrix} = \begin{pmatrix} 1 \\ -2.2 \end{pmatrix}$$

However, it is usual to work with normalized eigenvectors:

$$\mathbf{u}_1 = \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|_2} = \begin{pmatrix} 0.4138 \\ -0.9104 \end{pmatrix}$$

For $\lambda_2 = 7.88$, the corresponding eigenvector $\mathbf{u}_2 = \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix}$ will verify the following:

$$\mathbf{C}\mathbf{u}_2 = \lambda_2 \mathbf{u}_2 \Longleftrightarrow (\mathbf{C} - \lambda_2 \mathbf{I})\, \mathbf{u}_2 = 0$$

Which yields:

$$(\mathbf{C} - \lambda_2 \mathbf{I})\, \mathbf{u}_2 = \mathbf{0}$$

$$\left[\begin{pmatrix} \frac{20}{3} & \frac{8}{3} \\ \frac{8}{3} & 2 \end{pmatrix} - \begin{pmatrix} \lambda_2 & 0 \\ 0 & \lambda_2 \end{pmatrix} \right] \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} \frac{20}{3} - \lambda_2 & \frac{8}{3} \\ \frac{8}{3} & 2 - \lambda_2 \end{pmatrix} \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} \left(\frac{20}{3} - \lambda_2\right) u_{21} + \frac{8}{3} u_{22} \\ \frac{8}{3} u_{21} + (2 - \lambda_2)\, u_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Again, the following condition must be met:

$$u_{22} = -\frac{3}{8} \left(\frac{20}{3} - \lambda_2 \right) u_{21}$$

So, a solution for the system will be:

$$\mathbf{u}_2 = \begin{pmatrix} u_{21} \\ -\frac{3}{8} \left(\frac{20}{3} - \lambda_1 \right) u_{21} \end{pmatrix}$$

We can choose for instance $u_{21} = 1$ and get:

$$\mathbf{u}_2 = \begin{pmatrix} 1 \\ -\frac{3}{8} \left(\frac{20}{3} - 7.88 \right) \end{pmatrix} = \begin{pmatrix} 1 \\ 0.45 \end{pmatrix}$$

However, it is usual to work with normalized eigenvectors:

$$\mathbf{u}_2 = \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|_2} = \begin{pmatrix} 0.9119 \\ 0.4104 \end{pmatrix}$$

Having the eigenvectors, we can place them on the columns of a matrix to build the K-L tranformation:

$$U_{K-L} = \begin{pmatrix} 0.4138 & 0.9119 \\ -0.9104 & 0.4104 \end{pmatrix}$$

---

b) What is the rotation applied to go from the original coordinate system to the eigenvector coordinate system?

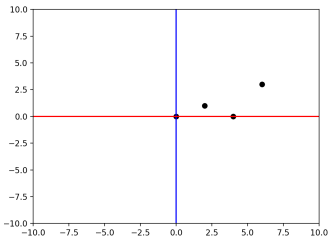
Solution:

$$\mathbf{U}\mathbf{e}_1 = \begin{pmatrix} 0.4138 & 0.9119 \\ -0.9104 & 0.4104 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \mathbf{u}_1$$

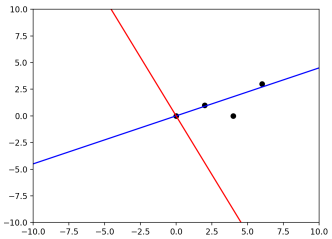
$$\mathbf{u}_1\mathbf{e}_1 = \left\| \begin{pmatrix} 0.4138 \\ -0.9104 \end{pmatrix} \right\|_2 \left\| \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\|_2 \cos\alpha = 0.4138$$

$$\alpha = \arccos(0.4138) = 1.1442 rad \simeq 65 deg$$

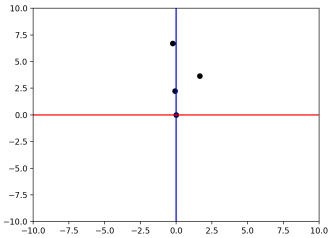
Let us graphically analyze what happened. The original points in the original coordinate system:



The eigenvectors are as follows:



Mapping the points to this new eigenspace through $\mathbf{x}_{eig} = \mathbf{U}^T\mathbf{x}$, we get:



Through this we can see that the 65 degree rotation was negative (i.e. counter-clockwise).

---

c) Which eigenvector is most significant?

Solution:
The most significant eigenvector is the largest one. In this case λ_2.

---

d) Can we apply the Kaiser criterion?

Solution:
Yes, since $\lambda_1 < 1$ we can apply the criterion and discard it.

---

e) Map the points onto the most significant dimension.

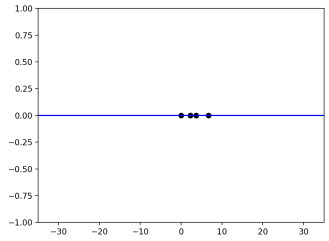
Solution:
We can map the points to the eigenspace through $\mathbf{x}_{eig} = \mathbf{U}^T\mathbf{x}$:

$$\begin{aligned}
\mathbf{x}_{eig}^{(1)} &= \begin{pmatrix} 0.4138 & 0.9119 \\ -0.9104 & 0.4104 \end{pmatrix}^T \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
\mathbf{x}_{eig}^{(2)} &= \begin{pmatrix} 0.4138 & 0.9119 \\ -0.9104 & 0.4104 \end{pmatrix}^T \begin{pmatrix} 4 \\ 0 \end{pmatrix} = \begin{pmatrix} 1.66 \\ 3.65 \end{pmatrix} \\
\mathbf{x}_{eig}^{(3)} &= \begin{pmatrix} 0.4138 & 0.9119 \\ -0.9104 & 0.4104 \end{pmatrix}^T \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.08 \\ 2.23 \end{pmatrix} \\
\mathbf{x}_{eig}^{(4)} &= \begin{pmatrix} 0.4138 & 0.9119 \\ -0.9104 & 0.4104 \end{pmatrix}^T \begin{pmatrix} 6 \\ 3 \end{pmatrix} = \begin{pmatrix} -0.25 \\ 6.70 \end{pmatrix}
\end{aligned}$$

PCA corresponds to keeping only the most significant dimensions. In this case, we discard the first dimension and get:

$$\begin{aligned}
\mathbf{x}_{PCA}^{(1)} &= \begin{pmatrix} 0 \end{pmatrix} \\
\mathbf{x}_{PCA}^{(2)} &= \begin{pmatrix} 3.65 \end{pmatrix} \\
\mathbf{x}_{PCA}^{(3)} &= \begin{pmatrix} 2.23 \end{pmatrix} \\
\mathbf{x}_{PCA}^{(4)} &= \begin{pmatrix} 6.70 \end{pmatrix}
\end{aligned}$$

Which yields:



---

2 Thinking Questions

When is an autoencoder equivalent to a PCA? What are the main differences?

Solution:

Linear units where the number of the hidden units corresponds to the number of the principal components (the eigenvectors with large eigenvalues) In PCA the eigenvalues indicate which dimensions are important. In a linear autoencoder the number of hidden units has to be determined by experiments

---