Machine Learning for Data Science (CS 4786)

Lecture 2 & 3: Principal Component Analysis

The text in black outlines high level ideas. The text in blue provides simple mathematical details to "derive" or get to the algorithm or method. The text in red are mathematical details for those who are interested.

1 Dimensionality Reduction and Linear Projection

We are provided with data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ where each $\mathbf{x}_t \in \mathbb{R}^d$ is a d-dimensional vector. The goal in dimensionality reduction is to compress these points into vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^K$ where K is smaller than d.

In this lecture we will consider dimensionality reduction through linear transformations, meaning, the low dimensional representation $\mathbf{y}_t$ for each datapoint $\mathbf{x}_t$ is obtained by setting

$$\mathbf{y}_t = W^\top \mathbf{x}_t$$

where W defines the linear transformation given by a $d \times k$ matrix Notice that one can represent the matrix W as

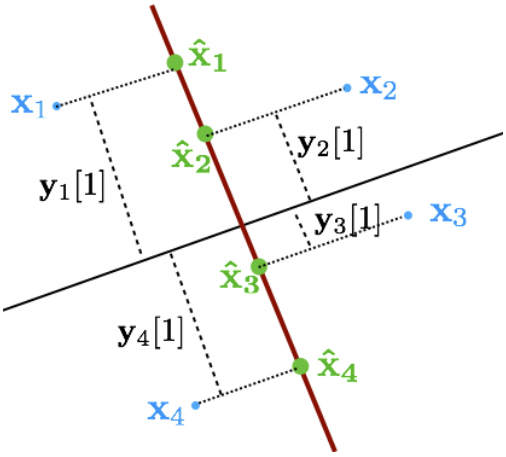
$$W = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$$

where $\mathbf{w}_1, \ldots, \mathbf{w}_K$ are each d-dimensional vectors.

2 PCA to One Dimension ($K = 1$)

For the case when $K = 1$, $\mathbf{y}_t[1] = \mathbf{w}_1^\top \mathbf{x}_t$. Now note that arbitrary scaling of $\mathbf{w}_1$ to say $\alpha \mathbf{w}_1$ for some $\alpha \in \mathbb{R}$ simply leads to $\mathbf{y}_t[1]$'s being scaled by α. Such scaling does not really affect our projections however. Hence without loss of generality we can cimply set $\|\mathbf{w}_1\|_2 = 1$, that is find a vector of unit norm.

The figure below illustrates the basic idea when we consider the projection down to only one dimension. The points in blue are the original $\mathbf{x}_1, \ldots, \mathbf{x}_n$. $\mathbf{w}_1$ the first direction of projection is illustrated in the figure by the red line. The points in green represent the reconstructions $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n$. The one dimensional representation $\mathbf{y}_1[1], \ldots, \mathbf{y}_n[1]$ are illustrated by the lengths in the figure.

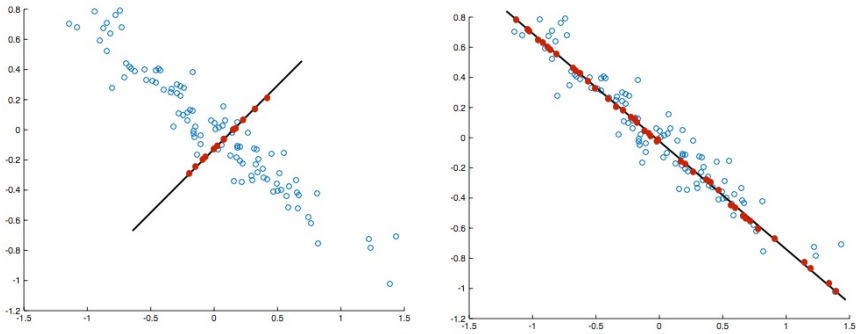


Now the main question at hand boils down to, ***How do we pick the right linear transformation*** W ***so a to retain as much information about the original data points as possible.***

2.1 Maximizing Spread (variance)

Basic idea: Pick the directions along which data is maximally spread (or variance is high).

As an example in the illustration below we would like to pick the second option as the spread if the points across the chosen direction is larger in the second figure.



How do we formalize this (for $K = 1$)?
Let us first consider the first direction to pick $\mathbf{w}_1$. We want to pick the direction long which variance of $\mathbf{y}_1[1], \ldots, \mathbf{y}_n[1]$ is largest. That is we want to pick the direction $\mathbf{w}_1$ (unit norm vector) that maximize the sample variance in the projected direction which is given by:

$$\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{y}_t[1] - \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t[1]\right)^2 = \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}_1^\top \mathbf{x}_t - \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}_1^\top \mathbf{x}_t\right)^2$$

Thus the solution $\mathbf{w}_1$ is given by,

$$\mathbf{w}_1 = \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2^2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top \mathbf{x}_t\right)^2$$

Let $\mu = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t$. Now let us simplify the above expression further,

$$\begin{aligned}
\mathbf{w}_1 &= \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \mathbf{x}_t - \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top \mathbf{x}_t\right)^2 \\
&= \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top \left(\mathbf{x}_t - \frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t\right)\right)^2 \\
&= \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{w}^\top (\mathbf{x}_t - \mu)\right)^2 \\
&= \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \frac{1}{n}\sum_{t=1}^{n}\mathbf{w}^\top (\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top \mathbf{w} \\
&= \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top \left(\frac{1}{n}\sum_{t=1}^{n}(\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)^\top\right)\mathbf{w} \\
&= \arg \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top \Sigma \mathbf{w}
\end{aligned}$$

Where Σ is sample the covariance matrix.

The first direction we pick will be the the unit vector $\mathbf{w}_1$ that maximizes $\mathbf{w}_1^\top \Sigma \mathbf{w}_1$

(Roughly speaking) Whenever we want to maximize (or minimize) a function subject to a constraint, we can use the idea of Lagrange multipliers. What the result says is that there exists $\lambda_1 \in \mathbb{R}$ such that the solution to $\mathbf{w}_1$ can be alternatively written down as :

$$\mathbf{w}_1 = \arg \max_{\mathbf{w}\in\mathbb{R}^d} \mathbf{w}^\top \Sigma \mathbf{w} - \lambda \|\mathbf{w}\|_2^2$$

To optimize the above we simply take derivative and equate to 0. This gives us

$$\Sigma \mathbf{w}_1 - \lambda \mathbf{w}_1 = 0$$

The direction $\mathbf{w}_1$ we obtain by maximizing the variance in the direction is some unit vector that satisfies

$$\Sigma \mathbf{w}_1 = \lambda \mathbf{w}_1$$

But this is exactly the definition of an eigen vector of matrix Σ. In Dutch the word "eigen" means self or own. Eigen vector of a matrix multiplied to the matrix results in a vector that is the just a scaled version of the eigenvector itself.

So we see that $\mathbf{w}_1$ is an eigenvector of Σ. The next question is, which eigen vector to choose. To this end note that we want to maximize $\mathbf{w}_1^\top \Sigma \mathbf{w}_1$ and we just saw that $\Sigma \mathbf{w}_1 = \lambda \mathbf{w}_1$. Hence

$$\mathbf{w}_1^\top \Sigma \mathbf{w}_1 = \lambda \|\mathbf{w}_1\|_2^2 = \lambda$$

Since we want to maximize the above quantity it stands to reason that we pick $\mathbf{w}_1$ to be the eigen vector corresponding to the largest eigen value of Σ.

3 What about $K > 1$?

For simplicity, for this section we shall assume that $\frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t = 0$ because if not we can simply center the points by subtracting the mean from each one.

A simple fact from linear algebra is that, if we consider any orthonormal basis of $\mathbb{R}^d$ given by $\mathbf{w}_1, \ldots, \mathbf{w}_d \in \mathbb{R}^d$, then any vector in $\mathbb{R}^d$ can be represented as a linear combination of the d basis. Recall that orthonormal vectors are vectors $\mathbf{w}_1, \ldots, \mathbf{w}_d$ such that each vector is of unit length, that is

$$\forall i \leq d, \qquad \|\mathbf{w}_i\|_2^2 = \sum_{j=1}^{d} \mathbf{w}_i[j]^2 = 1$$

and the vectors are orthogonal to each other, that is

$$\forall i, j \text{ s.t. } i \neq j, \quad \mathbf{w}_i^\top \mathbf{w}_j = 0$$

The key idea we are going to use to produce the K dimensional representation of the n points is that we shall first find orthonormal basis $\mathbf{w}_1, \ldots, \mathbf{w}_d$ in which to represent each point $\mathbf{x}_t$. But however we shall pick only K of the d basis $\mathbf{w}_1, \ldots, \mathbf{w}_d$ and approximate the data points in the this chosen K dimensional subspace spanned by $\mathbf{w}_1, \ldots, \mathbf{w}_K$. Thus the matrix W will be got by considering only these K basis.

Since we can write any vector in d-dimension as a linear combination of the orthonormal basis, let us write each

$$\mathbf{x}_t = \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j \tag{1}$$

where for each $\mathbf{x}_t$, $\mathbf{y}_t[j]$ represents the coefficient on the jth basis $\mathbf{w}_j$. Now the K dimensional representation of the point $\mathbf{x}_t$ is given by the K numbers $\mathbf{y}_t[1], \ldots, \mathbf{y}_t[K]$. What this means is that we can view the data point $\mathbf{x}_t$ being approximated by the reconstruction

$$\hat{\mathbf{x}}_t = \sum_{j=1}^{K} \mathbf{y}_t[j]\mathbf{w}_j \tag{2}$$

Now further note that, since $\mathbf{x}_t = \sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j$ and since $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are orthonormal, if we consider $W = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$,

$$W^\top \mathbf{x}_t = W^\top \left(\sum_{j=1}^{d} \mathbf{y}_t[j]\mathbf{w}_j \right) = \begin{bmatrix} \mathbf{y}_t[1] \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{y}_t[K] \end{bmatrix} \tag{3}$$

which coincides with our definition of linear transformation in the first section.

3.1 View I: Maximize Total Variance

Basic idea: Pick the orthogonal directions along which data is maximally spread in each of the coordinates.

How do we formalize this? We want to maximize the sum of the variances of $\mathbf{y}_t$'s. That is, we want to maximize

$$\sum_{j=1}^{K} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{y}_t[j] - \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_t[j] \right)^2 = \sum_{j=1}^{K} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{w}_j^\top \mathbf{x}_t - \frac{1}{n} \sum_{t=1}^{n} \mathbf{w}_j^\top \mathbf{x}_t \right)^2$$

That is we want to solve the optimization problem:

$$\begin{aligned}(\mathbf{w}_1, \ldots, \mathbf{w}_K) &= \underset{\text{orthonormal } W}{\operatorname{argmax}} \sum_{j=1}^{K} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{w}_j^\top \mathbf{x}_t - \frac{1}{n} \sum_{t=1}^{n} \mathbf{w}^\top \mathbf{x}_t \right)^2 \\ &= \underset{\text{orthonormal } W}{\operatorname{argmax}} \sum_{j=1}^{K} \mathbf{w}_j^\top \Sigma \mathbf{w}_j\end{aligned}$$

To solve the above, first we note that we can solve $\mathbf{w}_1$ just as the $K = 1$ case which yields the top eigen vector as solution for $\mathbf{w}_1$. Next we can solve $\mathbf{w}_2$, subject to it being orthogonal to $\mathbf{w}_1$ and this exactly yields the second largest eigenvector. Next $\mathbf{w}_3$ is the third largest and so forth. Thus the solution we get is the first K largest eigen vectors.

3.2 View II: Minimizing Reconstruction Error

Basic idea: Another way of thinking about which orthonormal basis to choose is to pick the basis such that the reconstruction error is minimized. That is pick the orthonormal basis $\mathbf{w}_1, \ldots, \mathbf{w}_K$ such that

$$\frac{1}{n} \sum_{t=1}^{n} \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_2^2$$

is minimized. (See the figure on page 2).

Let us simplify the above term,

$$\begin{aligned}\frac{1}{n} \sum_{t=1}^{n} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 &= \frac{1}{n} \sum_{t=1}^{n} \left\| \sum_{j=1}^{K} \mathbf{y}_t[j] \mathbf{w}_j - \mathbf{x}_t \right\|_2^2 \\ &= \frac{1}{n} \sum_{t=1}^{n} \left\| \sum_{j=1}^{K} \mathbf{y}_t[j] \mathbf{w}_j - \sum_{j=1}^{d} \mathbf{y}_t[j] \mathbf{w}_j \right\|_2^2 \qquad \text{(using the fact that } \mathbf{y}_t[j] = \mathbf{w}_j^\top \mathbf{x}_t \text{ from Eq. 3)} \\ &= \frac{1}{n} \sum_{t=1}^{n} \left\| \sum_{j=K+1}^{d} \mathbf{y}_t[j] \mathbf{w}_j \right\|_2^2 \\ &= \frac{1}{n} \sum_{t=1}^{n} \left\| \sum_{j=K+1}^{d} (\mathbf{w}_j^\top (\mathbf{x}_t - \mu)) \mathbf{w}_j \right\|_2^2\end{aligned}$$

from the above to the next equation is not hard but just takes a bit of staring. Note that for any vector $\mathbf{v}$, $\|\mathbf{v}\|_2^2 = \mathbf{v}^\top \mathbf{v}$. Expanding the above and noticing that $\mathbf{w}_j$'s are orthonormal yields the below. It's ok if this step seems hard, just take it as given.

$$
\begin{aligned}
&= \frac{1}{n} \sum_{t=1}^{n} \sum_{j=K+1}^{d} \left(\mathbf{w}_j^\top \mathbf{x}_t \right)^2 \\
&= \frac{1}{n} \sum_{t=1}^{n} \sum_{j=K+1}^{d} \mathbf{w}_j^\top \mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}_j \\
&= \sum_{j=K+1}^{d} \mathbf{w}_j^\top \left(\frac{1}{n} \sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t^\top \right) \mathbf{w}_j
\end{aligned}
$$

since $\mathbf{x}_t$'s are centered,

$$
= \sum_{j=K+1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j
$$

The orthonormal basis $\mathbf{w}_1, \ldots, \mathbf{w}_d$ that we shall pick are the ones that minimize the reconstruction error and are hence the orthonormal basis that minimize, $\sum_{j=K+1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j$.

We again use the Lagrangian multipliers to rewrite the constrained minimization problem (with the unit norm constraints) into an unconstrained minimization problem. Specifically we see that there exists $\lambda_1, \ldots, \lambda_d$ such that the orthonormal basis $\mathbf{w}_1, \ldots, \mathbf{w}_d$ are the ones that minimize,

$$
\sum_{j=K+1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j - \sum_{j=1}^{d} \lambda_j \|\mathbf{w}_j\|_2^2
$$

Taking derivative and equating to 0 we find that for any index $K+1 \leq j \leq d$,

$$
\Sigma \mathbf{w}_j - \lambda_j \mathbf{w}_j = 0
$$

Thus again we find that the $\mathbf{w}$'s are eigen vectors.

To minimize the reconstruction error we simply pick the eigen basis of Σ and retain K of them while throwing away the remaining. Now since each $\mathbf{w}_j$ is an eigen vector we have that

$$
\Sigma \mathbf{w}_j = \lambda_j \mathbf{w}_j
$$

where λ_j is the corresponding eigen value. Now the question remains, which Eigen vector to keep and which to throw away. To this end recall that we want to minimize

$$
\sum_{j=K+1}^{d} \mathbf{w}_j^\top \Sigma \mathbf{w}_j = \sum_{j=K+1}^{d} \lambda_j \mathbf{w}_j^\top \mathbf{w}_j = \sum_{j=K+1}^{d} \lambda_j
$$

Thus it stands to reason that to minimize the above we pick $\mathbf{w}_{K+1}, \ldots, \mathbf{w}_d$ to be the eigenvectors with the $d-K$ smallest eigenvalues.

4 PCA Algorithm

What both the views tell us is that the matrix W we shall use for PCA is got by taking the K eigenvectors corresponding to the top K eigen values. As for the PCA algorithm, one way to implement it is to first compute the covariance matrix given the data. This can be done by calculating the mean vector and then covariance matrix as

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t \qquad \Sigma = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{x}_t - \boldsymbol{\mu})(\mathbf{x}_t - \boldsymbol{\mu})^\top$$

Next we perform eigen decomposition of the matrix and take the top K eigen vectors and set

$$W = \begin{bmatrix} \mathbf{w}_1^\top \\ \cdot \\ \cdot \\ \mathbf{w}_K^\top \end{bmatrix} = \text{eigs}(\Sigma, K)$$

4.1 Projection to Lower Dimension

Projection is simply given by

$$\mathbf{y}_t = W(\mathbf{x}_t - \boldsymbol{\mu})$$

The above is the version where we center $\mathbf{x}_t$'s which is represented by the fact that $\boldsymbol{\mu}$ is subtracted from each $\mathbf{x}_t$.

4.2 Reconstruction

Reconstruction of the data points based on low dimensional representation is given by

$$\hat{\mathbf{x}}_t = W^\top \mathbf{y}_t + \boldsymbol{\mu}$$