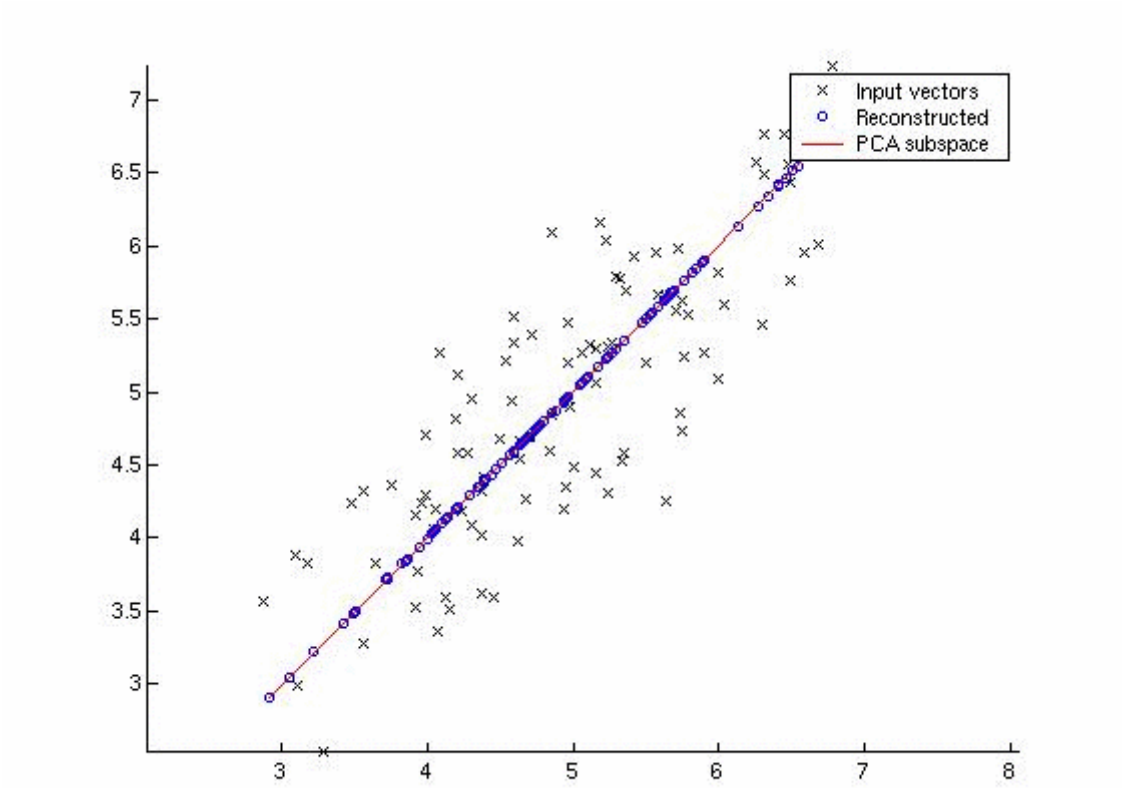
Principal Component Analysis

> Principal component analysis (PCA) is a dimensionality reduction technique. Given some dataset, its goal is to discover the "components" (explained below) that give the most information about the data.

The data we're dealing with might have a very large dimensionality, and some of those features might be useless. The idea, then is to choose the "most important" directions, and throw away the ones that don't matter as much. See, for example, the image below:



PCA is a linear projection from an n-dimensional space to a subspace. In the picture above, we project the data from 2 dimensions to 1, and the way to find the most important principal component is to simply look for the direction that maximizes the variance of the projected data.

But how do you find principal components?

First, we take the covariance matrix of the data Σ. Because Σ is square and symmetric, by the spectral theorem, we can find an *orthonormal basis* such that we can eigendecompose $\Sigma = Q\Lambda Q^\top$, with Λ being diagonal (note that because Q is orthonormal, we have $Q^{-1} = Q^\top$). That is, we *diagonalize* Σ by writing it in terms of its eigenvectors, which again we can do because it is symmetric.

After eigendecomposing Σ, we now have a diagonal matrix Λ written in an orthonormal eigenbasis of Σ. The correlation between points is now zero! It also becomes clear that (because eigenvalues are sorted in descending order by default) the maximum possible variance is achieved by projecting the data to the first coordinate axis (the first eigenvector). This is the first principal component.

Actually, it just so happens that the vectors that form eigenbasis of Σ are the directions with greatest variance! The goal of PCA is to find the direction(s) of maximal variance in the projection, i.e., it tries to minimize the reconstruction/projection error.

Reconstructing the Data

Let $\boldsymbol{X}$ be the $(n \times m)$ *centered* data matrix; that is, the data after subtracting the mean vector $\boldsymbol{\mu}$ from each row. We can then decompose the covariance matrix $(m \times m)$:

$$\Sigma = Q \Lambda Q^{-1}$$

Note that the eigenvectors are the columns of Q. Then, we choose the top k eigenvectors with the largest eigenvalues and collect them into the columns of a matrix - let's call it V, which is $(m \times k)$. Then, the PCA projections (or "scores"), in the lower dimensional space are given by:

$$\underbrace{\text{PCA Projections}}_{(n \times k)} = \underbrace{\boldsymbol{X}}_{(n \times m)} \cdot \underbrace{\boldsymbol{V}}_{(m \times k)}$$

Again, note that the n points now live in a k-dimensional subspace.

We can reconstruct the data (with possible information loss) by mapping it back to m dimensions with $\boldsymbol{V}^\top$. Finally, we add back the mean $\boldsymbol{\mu}$. The whole reconstruction looks like this:

$$\begin{aligned} \text{PCA Reconstruction} &= \text{PCA Projections} \cdot \text{Eigenvectors}^\top + \text{Mean} \\ &= \boldsymbol{X}\boldsymbol{V}\boldsymbol{V}^\top + \boldsymbol{\mu} \end{aligned}$$

We call $\boldsymbol{V}\boldsymbol{V}^\top$ a *projection* matrix. If all m eigenvectors are used, then there is no loss of information and $\boldsymbol{V}\boldsymbol{V}^\top$ is the identity matrix.

If PCA is done on the *correlation* matrix, and not the covariance matrix, the data is not only zero-centered by subtracting $\boldsymbol{\mu}$, but also scaled by dividing each column by its standard deviation σ. In this case, to reconstruct the data, we need to rescale the columns with σ and only then add back the mean vector.

How PCA turns from a geometric problem (with distances) to a linear algebra problem (with eigenvectors) Link

PCA is trying to find the direction such that the projection of the data on it has the highest possible variance. This direction is (by definition) called the first principal direction. We can formalize this as follows:

Given the covariance matrix Σ, we are looking for a vector $\boldsymbol{w}$ having unit length (why?), that is, $\|\boldsymbol{w}\| = 1$, such that $\boldsymbol{w}^\top \Sigma \boldsymbol{w}$ is maximal.

If $\boldsymbol{X}$ is the zero-centered (why?) data matrix, then the projection is given by $\boldsymbol{X}\boldsymbol{w}$ and its variance is:

$$\frac{1}{n-1}(\boldsymbol{X}\boldsymbol{w})^\top \cdot \boldsymbol{X}\boldsymbol{w} = \boldsymbol{w}^\top \cdot (\frac{1}{n-1}\boldsymbol{X}^\top \boldsymbol{X}) \cdot \boldsymbol{w} = \boldsymbol{w}^\top \Sigma \boldsymbol{w}$$

Also, an eigenvector of Σ is (by definition) any vector $\boldsymbol{v}$ such that $\boldsymbol{\Sigma v} = \boldsymbol{\lambda v}$.

It turns out that the first principal direction is given by the eigenvector with the largest eigenvalue. Proof:

Because of the spectral theorem, since Σ is symmetric, it is diagonal in its eigenvector basis. So we can choose an *orthogonal basis*, namely the one given by the eigenvectors, where Σ is diagonal and has eigenvalues λ_i on the diagonal. In that basis, $\boldsymbol{w}^\top \boldsymbol{\Sigma w}$ simplifies to $\sum \lambda_i w_i^2$. In other words, the variance is given by the weighted sum of the eigenvalues. To maximize this expression, we should simply take $\boldsymbol{w} = (1, 0, \ldots, 0)$, i.e., the first eigenvector, yielding variance λ_1. Indeed, deviating from this solution and "trading" parts of the largest eigenvalue for the parts of smaller ones will only lead to smaller overall variance. Also, note that we wrote $\boldsymbol{w}$ in terms of the new eigenbasis, that is, after the *change of basis/coordinates*; then, taking the $\boldsymbol{w}$ above means simply considering the first eigenvector.

Why do we need to zero-center the data?

If the data are not centered, then:

$$\frac{1}{n-1}(\boldsymbol{X}\boldsymbol{w})^\top \cdot \boldsymbol{X}\boldsymbol{w}$$

is not the variance. This is because the variance is, by definition, around the mean.

Why do the eigenvectors have to be unit length?

The reason why we constrain $\boldsymbol{w}$ to have unit length is that otherwise we can multiply it by any number and the expression above will increase by the square of that number - so the problem becomes ill-defined since the maximum of this expression is infinite. In fact, the variance of the projection of $\boldsymbol{w}$ is $\boldsymbol{w}^\top \Sigma \boldsymbol{w}$ only if $\boldsymbol{w}$ is unit length.

Questions on PCA Link

"Orthogonal" can mean two things:

- Variable axes as perpendicular axes (as in an orthogonal basis)
- Variables as *uncorrelated* by their data

Principal components are **orthogonal** (uncorrelated) variables. In general, orthogonality is not the same as uncorrelatedness, except in the case where at least one of the two random variables has an expected value of 0, which is the case in PCA.

If two variables are independent, then they are uncorrelated. However, not all uncorrelated variables are independent.

PCA is only sensitive to the scaling of the variables if we use the covariance matrix and not the correlation matrix. This is because covariance changes as the scale of the variables change, but correlation does not.

Note that we can only think in terms of the "variance explained by each eigenvalue" if we use the covariance matrix.

Correlated Variables vs. PCA Link

Should we remove highly correlated variables before doing PCA?
Well, probably.

If we think of what would happen if we pick one of the features of the data and create a bunch of near-copies of it, we will have a set of strongly correlated variables. Those will "load" onto several PCs (eigenvectors), differentially weighting them - and thus changing the directions of all eigenvectors too.

Say that we have two variables X and Y, which are very weakly correlated. PCA would identify two major PCs. If we add a variable Z which is an exact copy of Y, the points will stretch along the YZ directions, doubling their contribution to the variance. In this case, PCA would still identify two major PCs, but one would have twice the variance of the other.