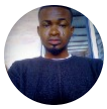
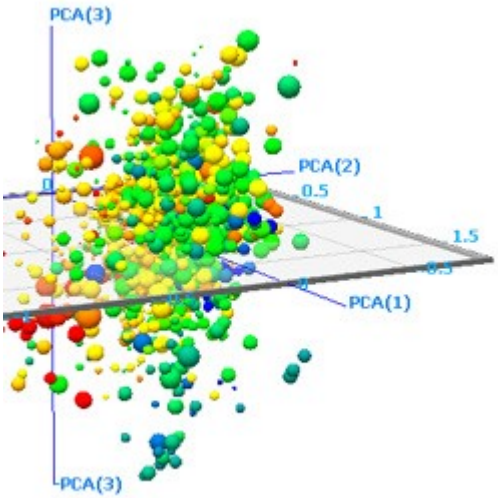


Lukmon AYINLA Follow

Apr 21 · 8 min read ★

The Mathematics and Intuitions of Principal Component Analysis (PCA) Using Truncated Singular Value Decomposition (SVD)



As data scientists or Machine learning experts, we are faced with tonnes of columns of data to extract insight from, among these features are redundant ones, in more fancier mathematical term—co-linear features. The numerous columns of features without prior treatment leads to curse of dimensionality which in turn leads to over fitting.

To ameliorate this curse of dimensionality, principal component analysis (PCA for short) which is one of many ways to address this, is employed using truncated Singular Value Decomposition (SVD).

Principal Component Analysis starts to make sense when the number of measured variables are more than three (3) where visualization of the cloud of the data point is difficult and it is near impossible to get insight from.

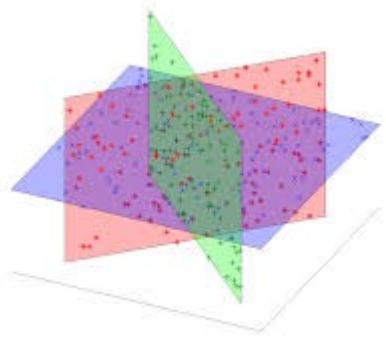
First: Let's try to grasp the goal of Principal Component Analysis. The following should put in mind:

1. Is the number of samples (m) much less than the number of measured variables(d)—i.e number of columns in a data frame m<< d
2. The goal of PCA is to get a subspace (i.e a lower dimensions) and projecting the cloud of data points to the subspace without losing information embedded in the original data with higher dimensions, so are looking for vector that maximizes the variation of the data.

Assuming the original data is in d-dimensional space, the goal is to arrive at a new data points in p dimension where p <<< d (for worst case scenario is p = d). That is the **intuition**.

The **intuition** can be represented mathematically as thus;

$$x \in \mathbb{R}^d \quad \Rightarrow \quad y \in \mathbb{R}^p \qquad \text{where } p << d$$



2D subspaces in a 3D space.

I would like to divide this analysis into the following;

1. Understanding the concept of projection.
2. Principal Component Analysis itself.
3. PCA via a truncated Singular Value Decomposition.

Understanding the concept of projection.

As discussed earlier the main goal of PCA is to arrive at a subspace i.e. lower dimensional space without losing much information.

Intuition: Before principal component analysis is carried out on a data set, the data must be centered by subtracting the mean of each measured variables from data points corresponding to the mean axis, so that the summation of all data points is zero.

Centered Data

Mathematics:

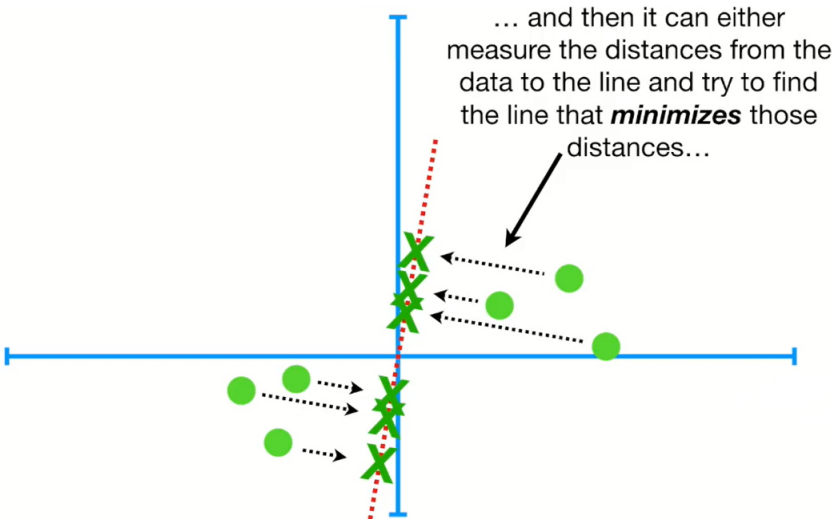
$$X = [x^1, ..., x^d] \forall x^i \epsilon \mathbb{R}^d$$

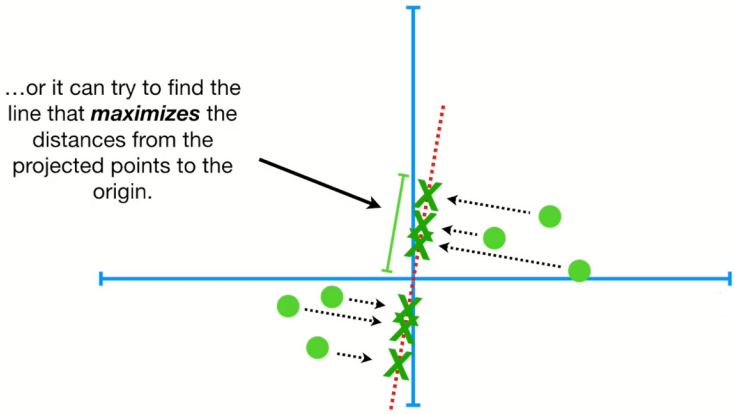
$$\overline{x} = \frac{1}{m-1} \sum_{m-1}^{i} x_i$$

$$X_{centered} = X^i - \bar{x}^i \ \ where \ i = 1, ..., d$$

The center shifted data point does not change their positions relative to each other.

Intuition: After the data is centered on the origin, we start drawing a random line given that it has to pass through the origin and fitting the line as much as possible on the data points. So, how does PCA know that a line is of the best fit ? This is done by projecting the data points on the line and checking the line that best minimizes the distances between the data points. and the line or the line the best maximizes the distances between the point of the projected data on the line and the origin. This forms an ***optimization problem.***





Mathematics: The length of the projected data points from the origin.

$$||\widetilde{x}|| = x^T.\varphi \quad let\, \varphi \; = candidate\; axis$$

$$\|\varphi\|_2 = 1$$

As said earlier, this results into an optimization problem considering the fitted line that maximizes the length of the projected data points from the origin.

$$J(\varphi) = \frac{1}{2m} \sum_{k=0}^{m-1} (\widehat{x_k}^T . \varphi)^2$$

The lengths of the projections are squared to avoid cancellation of positive values by the negative values.

Let's express J(phi) in terms of sample covariance of the data.

Intuition: A quick one, ***Covariance*** is the joint variation between two random variables from their corresponding expected value or mean.

Mathematics: Where covariance, C is expressed as

$$C = \frac{1}{m} X^T X$$

J(phi) can be written in terms of covariance as:

$$J(\varphi) = \frac{1}{2} \varphi^T . C . \varphi$$

Principal Component Analysis itself.

Intuition: The purpose of PCA is to find that candidate axis with the maximum of sum of squared distances between the projected point of the fitted line and the origin, J(phi).

Mathematics:

$$\varphi^* = argmax J(\varphi)$$

Let's recall the objective function to be optimized and the span of the candidate axis.

$$J(\varphi) = \frac{1}{2}\varphi^T.C.\varphi$$

$$\|\varphi\|_2 = 1$$

Squaring both of the above equation.

$$\|\varphi\|^2 = 1$$

Rewriting the above equation which forms the constraint of the maximization of the objective function.

$$\varphi^T\varphi = 1$$

There is a mathematical trick to achieve this which is known as the ***Lagrange Multiplier***.

$$\widehat{J}(\varphi, \lambda) = J(\varphi) + \frac{\lambda}{2}(1 - \varphi^T\varphi)$$

Recall;

$$J(\varphi) = \frac{1}{2}\varphi^T.C.\varphi$$

So,

$$\widehat{J}(\varphi, \lambda) = \frac{1}{2}\varphi^T C\varphi + \frac{\lambda}{2}(1 - \varphi^T\varphi)$$

Finding the derivative of J with respect to phi and lambda respectively

$$\left.\begin{array}{l} \bigtriangledown_{\varphi}\widehat{J} = C\varphi - \lambda\varphi \\ \bigtriangledown_{\lambda}\widehat{J} = \frac{1}{2}(1 - \varphi^{T}\varphi) \end{array}\right\} = 0$$

Equating the derivative of the Lagrange Multiplier to zero forms an Eigen problem.

$$\left.\begin{array}{l} C\varphi^{*} = \lambda^{*}\varphi^{*} \\ \varphi^{*T}\varphi^{*} = 1 \end{array}\right\} = Eigen\ Problem$$

So, we are on the search for eigen pairs

$$(\varphi^{*}, \lambda^{*})$$

We are close to getting the principal components of the subspace of the original data, and there arise an eigen problem where the only know variable is formula above is the covariance matrix. This takes us to the final stage of solving PCA.

PCA via truncated Singular Value Decomposition (SVD).

> *Singular value decomposition (SVD) is known as a Swiss Army Knife of Linear Algebra*

Intuition: And what we want, is to solve the eigen problem that came up in Principal Components Analysis (PCA). There is a fact from the concept of Linear Algebra that every matrix X has a singular value decomposition. This implies that we can factorize or decompose X into three (3) matrices which are:

1. Left singular vector, U.
2. Diagonal matrix, SIGMA.
3. Right singular vector, V.

Mathematics:

$$X = U \sum V^T$$

The dimension of X is (m x d), usually m greater than or equal to d, the dimension of U is (s x m), the dimension of sigma is (s x s) while the dimension of V transpose is (s x d) where;

$$s = min(m, d)$$

The left singular vector, U is orthogonal which implies that:

$$U^T U = I_s$$

where I is an Identity matrix with dimension (s x s).

$$\sum = \begin{bmatrix} \delta_0 & 0 & 0 \\ . & \delta_1 & . \\ 0 & 0 & \delta_{s-1} \end{bmatrix}$$

sigma is the singular values. where

$$\delta_0 \geq \delta_1 \geq ... \geq \delta_{s-1} \geq 0$$

Starting with the covariance matrix.

$$C = \frac{1}{m} X^T X$$

Recall;

$$X = U \sum V^T$$

Substitute X in the covariance matrix.

$$C = \frac{1}{m}(U \sum V^T)^T (U \sum V^T)$$

$$C = \frac{1}{m}(V \sum^T U^T)(U \sum V^T)$$

Recall;

$$U^T U = I_s$$

We will be left with,

$$C = \frac{1}{m} V \sum^T \sum V^T$$

Rearrange;

$$CV = V \frac{1}{m} (\sum)^2$$

$$V \equiv [v_0, \cdots, v_{s-1}]$$

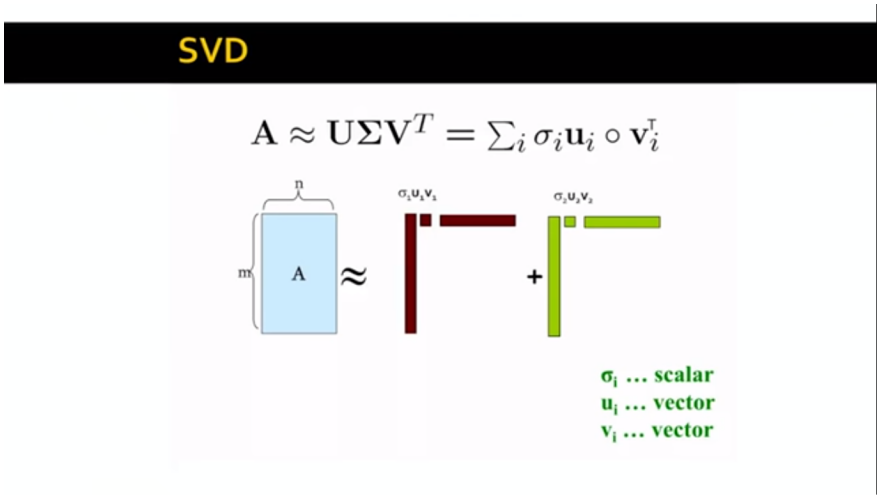
$$\sum \equiv diag(\delta_0, \cdots, \delta_{s-1})$$

At elemental level,

$$\begin{aligned} CV_0 &= \tfrac{\delta_0^2}{m} v_0 \\ &\cdot \\ &\cdot \\ &\cdot \\ CV_{s-1} &= \tfrac{\delta_{s-1}^2}{m} v_{s-1} \end{aligned}$$

$$(\frac{\delta_k^2}{m}, v_k) \; are \; eigen \; pairs$$

So, let's start viewing SVD as this



$$X = U \sum V^T \approx \sum_{k=0}^{s-1} U_k \delta_k v_k^T$$

By convention, the singular values occur in descending order.

$$U \sum V^T \approx \sum_{k=0}^{r-1} u_k \delta_k v_k^T$$

We can look at X as a running sum of different matrices. So, let's approximate the original singular value decomposition (SVD) using the first r-term where r < or = s.

$$U \sum V^T \approx U_r \sum_r V_r^T$$

In summary:

1. Centralized the data X
2. Compute the below using r-truncated SVD

$$X = U_r \sum_r V_r^T$$

3. Let the right singular vector Vr be the new axis that is, the principal component, such that:

$$compress(X) \approx X \cdot V_r^T$$

Let's have a chit chat in comment below if you find anything wrong in this article, or anything you feel is missing, I am willing and ready to learn.

You can follow me on Medium and also on Twitter **@LukmonAyinla2** for more update