Mathematical foundations of Principal Component Analysis

Hardik Prabhu

October 2020

1 Introduction

Dimensionality reduction is the transformation of data from a higher dimensional space to low-dimensional space, such that the information loss is minimum. Dimensionality reduction can be used for noise reduction, data visualization, cluster analysis, or as an intermediate step to facilitate other analyses. Principal Component analysis(PCA) is one such technique.

In PCA, we reduce n-dimensional data to a lower q-dimensional representation, by projecting the data along q orthonormal vectors(principal components) such that the average squared Euclidean distance (L2 norm) between the original data points and the projected points is minimum.

Equivalently, we will find that it is the same as trying to find a subspace of q-dimensions such that the variance of projected data along the vectors of the orthonormal basis consisting of the principal components is maximized.
While deriving, we will come across the assumption of the data being centered. It's a crucial assumption and if not, then the data should be centered first by subtracting the mean.

In the end, we will conclude by proving that the principal components are the eigenvectors of the sample covariance matrix, and the eigenvalues are the variance in the directions corresponding to the principal components.

2 Linear Algebraic Preliminaries

Let V be any finite vector space, with inner product $\langle\, ,\rangle$.

Definition: For any subset of vectors, $S \subseteq V$, the orthogonal subspace of S, $S^{\perp}$ is defined as,

$$S^{\perp} = \{x \in V : \langle x, y\rangle = 0, \forall y \in S\}$$

.

Theorem 1: For any subspace W of V, V is the direct sum of W and $W^{\perp}$.

$$W \oplus W^{\perp} = V$$

Proof: Let $x \in W \cap W^{\perp}$. Then, by definition of orthogonal subspace, $\langle x, x\rangle = 0$. It implies that $x = 0_v$, the zero vector. Therefore $W \cap W^{\perp} = \{0_v\}$. **(R1)**
Let $x \in V$, β be an orthonormal basis of W.

$$\beta = \{v_1, v_2, v_3..v_m\}$$

such that $\langle v_i, v_i\rangle = 1$ and for $i \neq j$, $\langle v_i, v_j\rangle = 0$.

Let $w \in W$ such that,

$$w = \sum_{i=1}^{m} \langle x, v_i\rangle v_i$$

$\Rightarrow$ for any $v_i \in \beta$, $\langle x - w, v_i\rangle = 0$
$\Rightarrow x - w \in W^{\perp}$.
$\Rightarrow x = w + (x - w) \in W \cup W^{\perp}$
Therefore, $V = W \cup W^{\perp}$ **(R2)**

Combining the results **(R1)** and **(R2)**, we prove the theorem.

Note: We assumed the existence of orthonormal basis for finite vector space W. It is a true statement. We can apply the gram schmidt process to derive an orthonormal basis from any arbitrary basis.

Corollary 1: For a vector $v \in V$, and a subspace W of V. The vector v has 2 components, one in the subspace W and, another orthogonal to W. The component in W is also known as the orthogonal projection of v in W.

Definition: We define the distance between 2 vectors, x and y, in an inner product space V as,

$$dist(x, y) = \parallel x - y \parallel = \sqrt{\langle x - y, x - y \rangle}$$

Corollary 2: The **"closest vector"** to v in W is it's orthogonal projection in W.

Proof: We know that, for vectors x, y such that $\langle x, y \rangle = 0$, x and y are orthogonal to each other. Then $\parallel x + y \parallel^2 = \parallel x \parallel^2 + \parallel y \parallel^2$ **(R3)**

Let v ∈ V, and W be a subspace of V. By applying **Corollary 1**,

$$v = v_W + v_{W^\perp}$$

Let $w \in W$, then $(dist(v, w))^2$ is given as,

$$\parallel v - w \parallel^2 = \parallel v_W + v_{W^\perp} - w \parallel^2$$

By applying **(R3)**, we get,

$$\parallel v - w \parallel^2 = \parallel v_W - w \parallel^2 + \parallel v_{W^\perp} \parallel^2$$

$\Rightarrow$

$$\parallel v - w \parallel^2 \geq 0 + \parallel v_{W^\perp} \parallel^2$$

$\Rightarrow$

$$\parallel v - w \parallel^2 \geq \parallel v - v_W \parallel^2$$

$\Rightarrow$

$$\parallel v - w \parallel \geq \parallel v - v_W \parallel$$

Hence, we have proved that the closest point indeed is the orthogonal projection in W.

Interpretation in Euclidean Space

In the n-dimensional real space R^n with standard inner product, $\langle x, y \rangle = y^\top x$. Considering the euclidean distance (L2 norm), the above corollaries hold true. Let us look at a simple illustration.

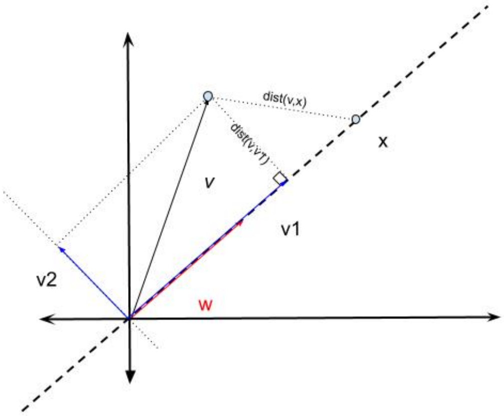

Figure 1: Projecting 2-D point on a 1-D line

Consider a vector $v \in R^2$. Let w be an arbitrary unit vector in R^2, $\parallel w \parallel = 1$. Then the line passing through origin in the direction of w, a 1-D subspace, is given as $L = \{x \in R^2 : x = \alpha w, \alpha \in R\}$.
Then, $v = v1 + v2$, such that $v1 \in L$ and $v2 \in L^{\perp}$.
The point v1 is closet to v in L, $dist(v, x) \geq dist(v, v1), \forall x \in L$.
The vector $v1$ is given as $v1 = ww^{\top}v$. The magnitiude of the projected vector is given by $\langle v, w \rangle$.

In general, the orthogonal projection of a vector $v \in R^n$ in m-dimensional subspace W is given as,

$$v_w = \sum_{i=1}^{m} w_i w_i^{\top} v$$

where $\beta = \{w_1, w_2 ... w_m\}$, is the orthonormal basis of W. It can be derived directly from the proof of Theorem 1.

3 Derivation of PCA

Before we consider reduction from higher n-dimension space to lower q-dimension space, we look at a one-dimensional projection. That is, we have n-dimensional feature vectors, and we want to project them on to a line through the origin. We are trying to find a best fit line to represent all sample data points. Let the line be in the direction of an unit vector w.
The error of the fit is given by the euclidean distance between the projection and the original data point. That is, if x is a data point in R^n, then the error of projecting is given as $\parallel x - ww^{\top}x \parallel$.

Consider our data as m samples of n-dimensional feature space (R^n). Let x_i denote the ith sample.
Then the mean square error of the projection is given as,

$$MSE = \frac{1}{m} \sum_{i=1}^{m} \parallel x_i - ww^{\top}x_i \parallel^2$$

After simplification,

$$MSE = \frac{1}{m} \sum_{i=1}^{m} \langle\, x_i, x_i \rangle - \langle\, x_i, w \rangle^2$$

Our objective is to find such w that minimizes the MSE. This is equivalent to maximising $\frac{1}{m}\sum_{i=1}^{m}\langle x_i, w\rangle^2$. We can see that $\frac{1}{m}\sum_{i=1}^{m}\langle x_i, w\rangle^2$ is the square of sample mean of random variable $z_i = \langle x_i, w\rangle$.
z_i is the representation of x_i in the 1-D space.

If we have a centered data, that is, the mean, $\frac{1}{m}\sum_{i=1}^{m} x_i = 0$
Then,

$$\frac{1}{m}\sum_{i=1}^{m}\langle x_i, w\rangle = \frac{1}{m}\sum_{i=1}^{m} w^\top x_i$$

$$= w^\top(\frac{1}{m}\sum_{i=1}^{m} x_i) = 0$$

Note: The data being centered is a very important assumption. If not, the data should be centred by subtracting the samples with mean, that is, $x_i := x_i - \mu$.

we now move ahead with the assumption, that our data is centered.
$\Rightarrow$

$$\frac{1}{m}\sum_{i=1}^{m}\langle x_i, w\rangle^2 = \frac{m-1}{m}var(z)$$

Therefore, in order to minimize the MSE, we have to maximize the sample variance of z. Which is the variance along the line onto which our data is projected. This is a strong assertion, we have to see if it generalizes to projecting the data to more than one principal component.

In general, we are trying to project the n-dimensional data into a lower dimensional subspace of dimension q, such that MSE is minimum. We have shown before that for any data point, the closest representation of the data point in subspace of lower dimension is its orthogonal projection onto it. The question is which is the right subspace.
Let the orthonormal basis of the subspace of dimension p, $p < n$, be $\beta = \{w_1, w_2 ... w_p\}$. The basis vectors are known as the principal components. The projected data can be represented by the p-dimensional coordinates given by $z_i = [z_{i1}, z_{i2}, ...zip]^\top$, where $z_{ij} = \langle x_i, w_j\rangle$, which is the projection along direction w_j.
In order to find the right subspace, we have to solve the following optimization problem associated with it:

Minimize: $\frac{1}{m}\sum_{i=1}^{m} \| x_i - \sum_{j}^{p} w_j w_j^\top x_i \|^2$

Constraint: $\beta = \{w_1, w_2 ... w_p\}$ is an orthonormal basis.

If we expand the expression, a lot of cross product terms will cancel out. Recall the fact that the square of norm of sum of orthogonal vectors, is the sum of squares of norms of the two vectors. **(R3)**
Therefore,

$$\| x_i \|^2 = \| x_i - \sum_{j}^{p} w_j w_j^\top x_i + \sum_{j}^{p} w_j w_j^\top x_i \|^2$$

$$= \| x_i - \sum_{j}^{p} w_j w_j^\top x_i \|^2 + \| \sum_{j}^{p} w_j w_j^\top x_i \|^2$$

Therefore, the MSE is given by,

$$\frac{1}{m}\sum_{i=1}^{m} \| x_i \|^2 - \frac{1}{m}\sum_{i=1}^{m} \| \sum_{j}^{p} w_j w_j^\top x_i \|^2$$

We can see that, minimizing MSE is same as maximizing $\frac{1}{m}\sum_{i=1}^{m} \| \sum_{j}^{p} w_j w_j^\top x_i \|^2$
By applying **(R3)**,

$$\| \sum_{j}^{p} w_j w_j^\top x_i \|^2 = \| \sum_{j}^{p} \langle x_i, w_j\rangle w_j \|^2$$

$$= \sum_{j}^{p} \langle x_i, w_j\rangle^2 \| w_j \|^2 = \sum_{j}^{p}\langle x_i, w_j\rangle^2$$

Therefore, we have to maximize the expression $\frac{1}{m}\sum_j^p \sum_{i=1}^m \langle x_i, w_j \rangle^2$.

This expression is similar to the case of projecting to a line. Therefore by trying to maximize the sum for each component j, we are left with maximizing the sum of the sample variances of the projections on each of the components.

Maximize:

$$\frac{1}{m}\sum_j^p \sum_{i=1}^m \langle x_i, w_j \rangle^2 = \frac{m-1}{m}\sum_j^p var(z_j)$$

The only constraint is the fact that all w_j's form an othonormal basis.

If we could maximize the sum of variances without violating the constraint then we essentially have a solution for our optimization problem.

Before going any further we will look at some statistical concepts which are required.

4 Key Statistical Concepts

Definition: The expected value of a random matrix $\mathbf{X}$ is given as a matrix such that the entries of $E(\mathbf{X})$ are the expected values respective entries of $\mathbf{X}$

$$E(\mathbf{X})_{ij} = E(\mathbf{X}_{ij})$$

The linearity of the expectation operator easily follows from this definition, which is stated below.

Theorem 2: Let W_1 and W_2 be m x n matrices of random variables, let A_1 and A_2 be k x m matrices of constants, and let B be an n x l matrix of constants. Then

$$E[A_1 W_1 B + A_2 W_2] = A_1 E[W_1] B + A_2 E[W_2]$$

Definition: Let $X = [X_1, X_2..X_n]^\top$ be n x 1 vector. It's covariance matrix is defined as

$$Cov(X) = E[(X - E(X))(X - E(X))^\top$$

We can see that $Cov(X)_{ij} = Cov(X_i, X_j)$ (Covariance matrix is a symmetric matrix)

Upon further algebraic manipulation, along with linearity of expectation, we get alternative expression for $Cov(X)$ as

$$Cov(X) = E(XX^\top) - E(X)E(X)^\top$$

Theorem 3: Let A be an m x n matrix of constants. Then

$$Cov(AX) = ACov(X)A^\top$$

The proof is straightforward. We can get this result by using the alternative expression of covariance and the linearity of expectation.

Theorem 4: $Cov(X)$ is a positive semi-definite matrix.

Proof: Consider an arbitrary vector $Z \in R^\top$. Then we know that the variance of $Z^\top X$ is non-negative. Therefore,

$$0 \leq var(Z^\top X) = Cov(Z^\top X) = Z^\top Cov(X) Z$$

We already know that $Cov(X)$ is symmetric. Therefore, $Cov(X)$ is positive semi-definite.

Note: For sample covariance matrix, we replace the covairances with sample covariances of respective entries. We can similarly show that sample covariance matrix is a positive semi-definite matrix.

5 Maximizing the sum of variances

Let $\mathbf{X}$ denote the entire data. The rows of $\mathbf{X}$ are the n-dimenisonal data points, which are denoted as x_i.

$$\mathbf{X} = \left[\begin{array}{c} x_1^\top \\ x_2^\top \\ \vdots \\ x_m^\top \end{array} \right]$$

Since we have assumed the data to be centered, sample covariance matrix is given as

$$C = \frac{1}{m-1}\mathbf{X}^\top\mathbf{X}$$

Consider the expression $\sum_i^m \langle x_i, w_j\rangle^2$

$$\sum_i^m \langle x_i, w_j\rangle^2 = \sum_i^m w_j^\top x_i x_i^\top w_j$$

$$= w_j^\top \sum_i^m x_i x_i^\top w_j$$

$$= w_j^\top \mathbf{X}^\top \mathbf{X} w_j$$

$$= (m-1) w_j^\top C w_j$$

Therefore the sample variance along each principal component j,

$$var(z_j) = w_j^\top C w_j$$

Returning back to the optimization problem.

The Lagrange function is given as

$$L(w_1..wp, \lambda) = \sum_{j=1}^p w_j^\top C w_j - \lambda_j(w_j^\top w_j - 1)$$

with $w_j^\top w_j = 1$ for all j.

By taking the gradient with respect to w_j and equating to zero we get

$$2Cw_j - 2\lambda_j w_j = 0$$

$\Rightarrow$

$$Cw_j = \lambda_j w_j$$

Thus, desired vector w_j is an eigenvector of the covariance matrix C.
Also, $L = \sum_{j=1}^p \lambda_j$.

Since C is a symmetric matrix, it has a singular value decomposition same as the eigenvalue decompostion.

That is,

$$C = \Lambda\Sigma\Lambda^\top$$

Σ is the diagonal matrix consisting of all the eigenvalues of C arranged in decreasing order from top to bottom. The columns of Λ forms an orthonormal basis of R^n.
Thus, if we select the first p eigenvalues, and the first p columns of Λ, we get the p principal components and the variances of projections along them (eigenvalues).