PCA : Principal Component Analysis - Intuitive explanation

Réda Arab

1 Goal

The goal of PCA is to transform our feature space of dimension p to a new feature space of dimension $k < p$.

It can be used for different purposes such as **visualization** ($k = 2, 3$), **compression** or **computational** (lower complexity).

Important use : having new ***decorrelated features***.

Principle : We project our data $x_1, x_2, ..., x_n \in \mathbb{R}^p$ into a new space of dimension $k < p$ such that the *total variance* (i.e the *spread* of our data) is maximized .

Notation : We write $X \in \mathbb{R}^{n*p}$ with n the size of the sample, p the number of features.

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} X_1 & X_2 & \dots & X_p \end{pmatrix}$$

In practice, we *center and reduce* our data before applying PCA (it simplifies the calculation and it prevents from having one feature which contains all the variance ; the importance of doing this transformation will be clear later).

$$x \to \frac{x - \hat{x}}{\sigma_x}$$

2 Preliminary Mathematics

2.1 SVD : Singular Value Decomposition

The reader may refer to the Wikipedia article (the schemes are interesting) `https://en.wikipedia.org/wiki/Singular_value_decomposition`

Statement : For X a real matrix of $\mathbb{R}^{n*p}$, it can be written as $X = UDV^T$ with U and V two orthogonal matrices of $\mathbb{R}^{n*n}$ and $\mathbb{R}^{p*p}$, and D a matrix with diagonal terms $D_{11} \geq D_{22} \geq \dots \geq D_{mm}$ with $m = min(n, p)$ and the other elements being null.

For example, if $n > p$, D will be of the form :

$$D = \begin{pmatrix} D_{11} & 0 & \dots & \dots \\ 0 & D_{22} & & \\ \vdots & & \ddots & \\ \vdots & & & D_{mm} \\ 0 & \dots & \dots & 0 \end{pmatrix}$$

Therefore :

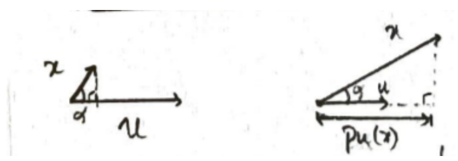
$$X^T X = (UDV^T)^T (UDV^T) = VD^T DV^T$$

$$\boxed{X^T X = V \Lambda V^T}$$

$\Lambda = D^T D$ contains **the eigenvalues** of $X^T X$ (singular values squared) which are real and positive or null.

V contains the **eigenvectors** which constitute an orthonormal basis of $\mathbb{R}^p$.

2.2 Scalar product and projection on R^2



$cos(\alpha) = \frac{p_u(x)}{||x||_2}$ where $p_u(x)$ the orthogonal projection of x on the line with u as direction vector.

Therefore $< x, u >= ||x||_2 \cdot ||u||_2 \cdot cos(\alpha) = ||u||_2 \cdot p_u(x)$.

If $||u||_2 = 1$, then $\boxed{< x, u >= p_u(x)}$

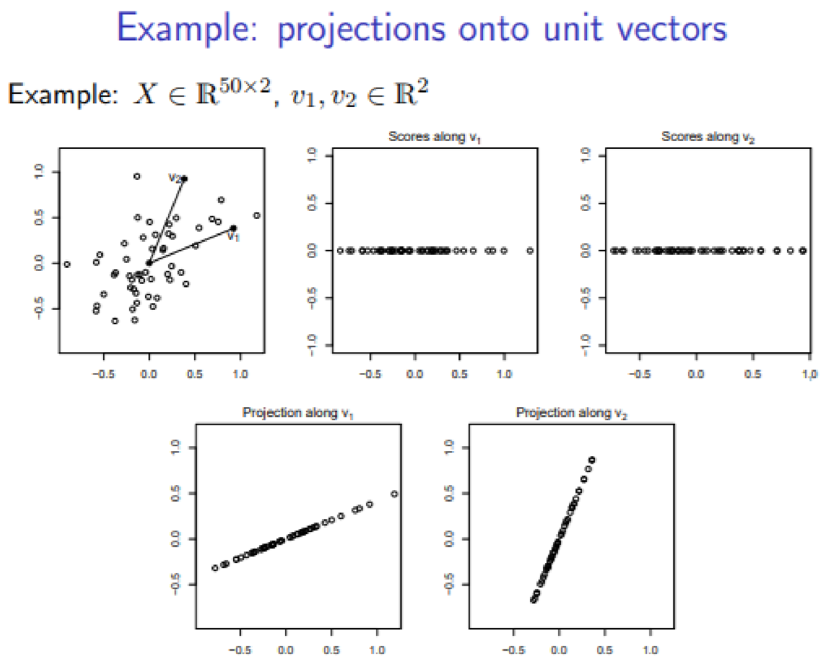


FIGURE 1 – From `https://www.stat.cmu.edu/~ryantibs/datamining/lectures/07-dim1.pdf`

2.3 Characterization of the orthogonal projection on a plane in $\mathbb{R}^n, n \geq 3$

Reminder : Given P a plane, $\Pi_P(x) = \underset{y\in P}{\text{argmin}}||y - x||_2^2$, where $\Pi_P(x)$ is the orthogonal projection of x on P.

Let us consider an orthonormal basis of P : (v_1, v_2). $\forall y \in P, y = \lambda_1 v_1 + \lambda_2 v_2$.

$||y - x||_2^2 = ||x||_2^2 - 2 < x, y > +||y||_2^2$

$||y - x||_2^2 = ||x||_2^2 - 2\lambda_1 < x, v_1 > -2\lambda_2 < x, v_2 > +\lambda_1^2 + \lambda_2^2$

Let us consider the function

$$\phi(\lambda_1, \lambda_2) = -2\lambda_1 < x, v_1 > -2\lambda_2 < x, v_2 > +\lambda_1^2 + \lambda_2^2$$

- $\frac{\partial\phi}{\partial\lambda_i} = -2 < x, v_i > +2\lambda_i \quad i = 1, 2$
- $\frac{\partial^2\phi}{\partial\lambda_i^2} = 2 \quad i = 1, 2$
- $\frac{\partial^2\phi}{\partial\lambda_1\partial\lambda_2} = 0$

Thereby $\frac{\partial^2\phi}{\partial\lambda^2} \succ 0$ (Hessian matrix positive definite) so the function is strictly convex. The minimum is reached for :

$$\frac{\partial\phi}{\partial\lambda} = 0 \quad \text{i.e}$$

$$\boxed{\lambda_i =< x, v_i >} \quad i = 1, 2$$

$$\boxed{\Pi_p(x) =< x, v_1 > v_1+ < x, v_2 > v_2}$$

Generalization

The previous formula can be generalized easily to a projection on a space E_k of dimension k by considering $(v_1, ..., v_k)$ an orthonormal basis of E_k. We get :

$$\boxed{\Pi_{E_k}(x) = \sum_{i=1}^{k} < x, v_i > v_i = \sum_{i=1}^{k} (x^T v_i) v_i}$$

3 PCA - Introduction

The **total variance** (*total sample variance*) is defined as :

$$\boxed{\frac{1}{n} \sum_{i=1}^{n} ||x_i - \hat{x}||_2^2 = \sum_{i=1}^{n} ||x_i||_2^2 = \sum_{i=1}^{p} ||X_i||_2^2}$$

as the datapoints are centered (i.e features). Note that the term $\frac{1}{n}$ is not important for the problem as we are looking for a maximum value.

$\rightarrow$ We are looking for a space of dimension k such that the projection of the x_i on this space will give us new datapoints (with new features associated) which keeps the total variance as large as possible.

As a result, we will have a new matrix Y such that :

$Y = \begin{pmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{pmatrix}$ with $y_i \in \mathbb{R}^k$ such that $\sum_{i=1}^{n} ||y_i||_2^2$ is "maximal" in the sense defined above.

Example

Let us consider an example when we project our datapoints from a p dimensional feature space to a space of dimension 1 (line).On the schemes, you can see from dimension 2 to 1.

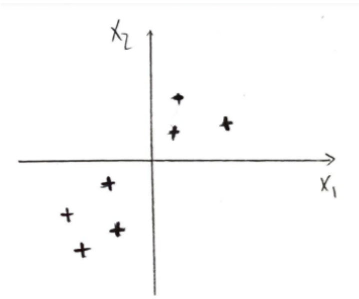


FIGURE 2 – Datapoints in 2D

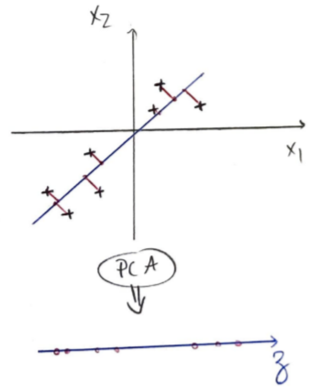


FIGURE 3 – New feature created with a large variance

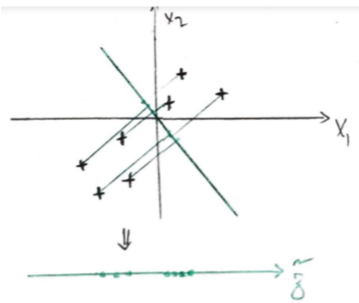


FIGURE 4 – New feature created with a low variance

Consider the case when we have datapoints with 2 features that we want to compress into one feature. So we are looking for a vector u (we can restrict our search to $||u||_2 = 1$ as we are looking for a direction) such that the projected datapoints on u ($x_i^T u, \quad i = 1, \ldots, n$) keeps a maximal total variance. .

i.e $Xu = \begin{pmatrix} x_1^T u \\ x_2^T u \\ \vdots \\ x_n^T u \end{pmatrix}$ and we know that $x_i^T u = p_u(x_i)$ for u with norm equals to 1.

Mathematically, we are looking for $u \in \mathbb{R}^2$ such that :

$\boxed{u = \underset{v \in \mathbb{R}^2, ||v||_2 = 1}{\text{argmax}} ||Xv||_2^2}$ (total variance maximized).

More generally, if we have a feature space of dimension p :

$$\boxed{u = \underset{v \in \mathbb{R}^p, ||v||_2 = 1}{\text{argmax}} ||Xv||_2^2}$$

We have : $||Xv||_2^2 = v^T X^T X v = v^T V D^T D V^T v$ with $||v||_2 = 1$.

Let us write $a = V^T v$. So $||a||_2 = ||V^T v||_2 = ||v||_2 = 1$ (as V orthogonal).

$\rightarrow ||Xv||_2^2 = a^T D^T D a = \sum_{i=1}^m a_i^2 D_{ii}^2 \leq D_{11}^2 \sum_{i=1}^m a_i^2 = D_{11}^2$

Therefore $a = (1, 0, \ldots, 0)^T$ maximizes $||Xv||_2^2$ i.e $\boxed{v = Va = V_1}$ which is **the eigenvector associated to** D_{11}^2. The direction which maximizes $||Xv||_2^2$ is V_1 and in this case $||XV_1||_2^2 = ||D_{11}U_1||_2^2 = D_{11}^2$

Remarks

- The projected datapoints $x_i^T v$ are centered :
  $\sum_{i=1}^n x_i^T v = (\sum_{i=1}^n x_i^T) v = 0$.
  So the total variance computed is $\sum_{i=1}^n (x_i^T v)^2 = ||Xv||_2^2$

- PCA and linear regression are different.

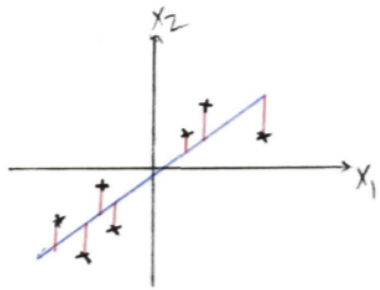


FIGURE 5 – Linear regression

The quantities we want to minimize in each case (the red lines in figure 3 and 5) are different.

4 Formulation of the problem

The problem can be summarized as :

$$\boxed{\underset{(v_1,v_2,\dots,v_k)\in\mathbb{R}^p}{\text{argmax}} \sum_{i=1}^{k}(||Xv_i||_2^2) \quad \text{such that} \quad v_i^T v_j = \delta_{i,j}}$$

Indeed, for a space E_k of dimension k and an orthonormal basis $(v_1, \dots, v_k)$:

$\Pi_{E_k}(x_i) = \sum_{i=1}^{k}(x_i^T v_l)v_l \quad$ and $\quad \sum_{i=1}^{n}||\Pi_{E_k}(x_i)||_2^2 = \sum_{i=1}^{n}\sum_{l=1}^{k}(x_i^T v_l)^2$

In this new space, we can write **the coordinates of the projected** x_i according to the orthonormal basis as :

$$Y = \begin{pmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{pmatrix} = \begin{pmatrix} x_1^T v_1 & x_1^T v_2 & \dots & x_1^T v_k \\ \vdots & \vdots & \dots & \vdots \\ x_n^T v_1 & x_n^T v_2 & \dots & x_n^T v_k \end{pmatrix} = \begin{pmatrix} Xv_1 & Xv_2 & \dots & Xv_k \end{pmatrix}$$

The total variance in this case is : $\sum_{i=1}^{n}\sum_{l=1}^{k}(x_i^T v_l)^2 = \sum_{i=1}^{k}||Xv_l||_2^2$

5 Problem resolution

We can solve the problem step by step, *direction by direction* (the problem is separable, it is a sum).

We can start with V_1 as a first new feature (*cf PCA -Introduction*) then, we are looking for a vector $v^{(2)}$ such that $||Xv^{(2)}||_2^2$ is maximized and $||v^{(2)}||_2 = 1$ and $(v^{(2)})^T V_1 = 0$.

We can easily show that $\boxed{v^{(2)} = V_2}$.

Recursively, we are looking a vector $v^{(j)}$ for $j = 2, \dots, k$, defined as below :

$$\boxed{\begin{array}{c} v^{(j)} \text{ maximise } \|Xv\|_2 \text{ over } v \in \mathbb{R}^p \text{ with the constraints} \\ \|v\|_2 = 1 \text{ and } (v^{(l)})^T v = 0 \text{ for all } l < j \end{array}}$$

We get $\boxed{(V_1, \dots, V_k)}$, **the eigenvectors** X^TX associated to the eigenvalues $D_{11}^2 \geq D_{22}^2 \geq \dots \geq D_{kk}^2 \geq 0$.

Therefore, the new datapoints are :

$$Y = \begin{pmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{pmatrix} = \begin{pmatrix} XV_1 & XV_2 & \dots & XV_k \end{pmatrix} = \begin{pmatrix} D_{11}U_1 & D_{22}U_2 & \dots & D_{kk}U_k \end{pmatrix}$$

The total variance of our new projected datapoints is :

$$\boxed{\sum_{l=1}^{k}||D_{ll}U_l||_2^2 = \sum_{l=1}^{k} D_{ll}^2}$$

The initial total variance, before the projection, is : $\boxed{\sum_{l=1}^{m} D_{ll}^2}$

Choice of k

We can define a way of choosing k (number of new features) as below.

$$\boxed{\min\{k \,|\, \frac{\sum_{l=1}^{k} D_{ll}^2}{\sum_{l=1}^{m} D_{ll}^2} \geq a\} \quad \text{with}\, a = 0.80, 0.90\, \text{or}\, 0.95\, \text{for example}}$$

Final remarks

— Our new datapoints are **centered** as shown in a previous section.

— The new features are **decorrelated** :

$$i \neq j, (XV_i)^T(XV_j) = V_i^T V \Lambda V^T V_j = 0.$$

— The new features are less interpretable, explainable.

— The general idea of this method of dimensionality reduction : **keeping the spread of our data as large as possible** .

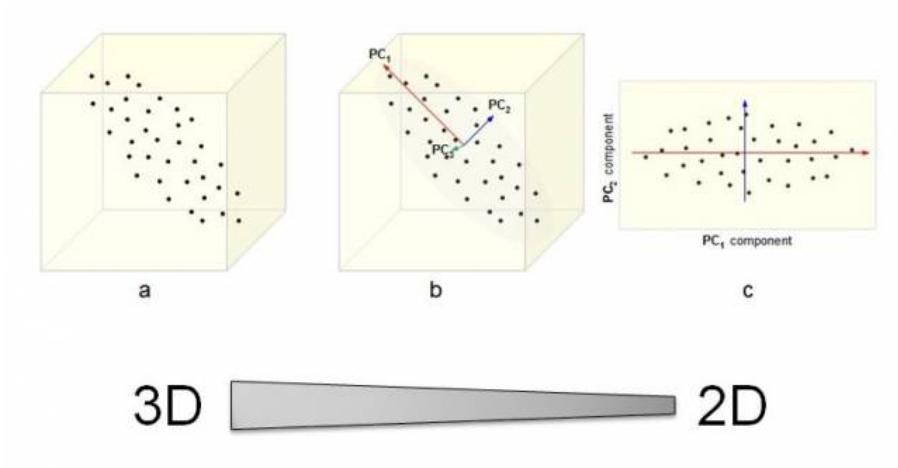


FIGURE 6 – From `https://www.davidzeleny.net/anadat-r/doku.php/en:pca`