Data Learning 1 : Principal Component Analysis

Principal Component Analysis is one of the most popular tool of multifactorial analysis. It consists in replacing the given (correlated) variables of a data set by a smaller number of uncorrelated (hopefully meaningfull) one that captures as much as possible of the variability of the datas. We consider a data set consisting of p variables that have been measured on n individuals, each of which been of weight w_i (usually we will simply choose $w_i = 1/n$). The data set can be wiewed as a $n \times p$ table, the *data matrix* $X = (x_i^j)_{1 \leq i \leq n, 1 \leq j \leq p}$. Let X' denote the transpose of X which is a $p \times n$ matrix.

The space of individuals : The data set can be consider as a set of n points in a p dimensional vector space E, each point representing an individual x_i with coordinates $x_i' = (x_i^1, x_i^2, \ldots, x_i^p)$. In order to measure the dissimilarities between them, assume that the space has an euclidien structure given by a symetric $p \times p$ matrix M :

$$< x_i, x_l >= x_i' M x_l \quad \text{and} \quad \|x_i\|_M = x_i' M x_i.$$

Notice that in the space of individuals, each axis represents one variable and a change of basis consists in replacing initial variables by new synthetic ones obtained as linear combination of the initial ones.

Inertia and centroïd of a cloud :

Inertia is a name taken from the *moment of inertia* in mechanics. A physical object has a center of gravity or *centroïd* and every particle of the object has a certain mass m and a certain distance d from the centroïd. The moment of inertia of the object is the quantity md^2 summed over all the particles that constitute the object. This concept has an analogy for the cloud of individual points in the space of individuals. The *centroïd* of the cloud, $\overline{x}$, is the *average individual* whose j^{th} coordinate is defined by $\overline{x}_j = \sum_{i=1}^n w_i x_i^j$ and the *inertia* of the cloud is defined by

$$\mathcal{I} = \sum_{i=1}^n w_i \|x_i - \overline{x}\|_M^2.$$

This quantity measures the dispersion of the cloud from the point $\overline{x}$. It can be easily generalized to define the inertia with respect to any point of the space (and not only with respect to its centroïd $\overline{x}$). But it can be shown that the centroïd is precisely the point that minimizes the inertia of the cloud with respect to a point (indeed, this can be verified for one dimensional spaces and can be extended to n-dimensional spaces using Pythagoras formula).

Let us say that each individual point *contributes* to the inertia of the whole cloud because one term of the sum is a certain percentage of the total sum. With the same idea, let us define the *contribution of one direction to the inertia* as the percentage of the total inertia that remain after an M-orthogonal projection of the cloud on the one dimensional subspace corresponding to that direction.

The space of variables : The data set can also be considered as a set of p points in an $n-$dimensional euclidien space F, each point representing one variable x^j. If all the variables are centered variables ($\overline{x}_j = 0$ for all j) and if we chose the metric given by the matrix $D = diag(w_1, \ldots, w_n)$, then we have a statistical interpretation of the euclidien structure of F. Indeed the norme of one variable and the cosine of the angle of two variables are nothing else than the variance and the correlation. Thus standardized variables, that is centered and of variance 1, are points of the unitary sphere (of dimension $n-1$) in F.

First factorial axis (or first PC) : In order to perform the *Principal Component Analysis* of the data set, consider first the space of individuals. We are looking on an origine O in this space E and a vector $u \in E$ such that the M-orthogonal projection of the cloud of individuals on the line L defined by 0 and u loses as little information as possible. First we can choose a unitary u, i.e. $\|u\|_M = 1$. Then let choose L in order to minimize

$$\sum_{i=0}^n w_i \|x_i - z_i\|_M^2$$

where z_i denote the orthogonal projection of individual x_i on L. As this quantity, which is the moment of inertia of the cloud with respect to L, is precisely the inertia of the projection of the cloud on the subspace of E M-orthogonal to L, the property of the centroïd recalled before shows that O must be the centroïd. Thus, assume that the centroïd of the cloud is the origine of the space (or assume that X has been replaced by $X - 1_n \overline{x}'$, where 1_n is the individual point with all coordinates equal to 1). Now, by Pithagoras' formula, one has

$$\|x_i - z_i\|_M^2 = \|x_i\|_M^2 - \|z_i\|_M^2$$

and thus, as the inertia $\sum_{i=0}^{n} w_i \|x_i\|_M^2$ does not depend on L, L is as well the line that maximizes $\sum_{i=0}^{n} w_i \|z_i\|_M^2$, that is the line that contributes the most to the inertia. Using the definition of the vector z_i, it easy to show that its lengths is $< x_i, u >= x_i' M u$ and thus

$$\text{Max}_u \sum_{i=0}^{n} w_i \|z_i\|_M^2 = \text{Max}_u \{(u'MX'DXMu), \;\; u'Mu = 1\}. \tag{1}$$

Let A denote the symetric $p \times p$ matrix $A := X'DXM$. Notice that this matrix is equal to the covariance matrix of the variables when $M = Id_n$ is the trivial metric. It can be shown that the vector u that maximize $u'MAu$ under the constrain $u'Mu = 1$ is precisely the eigenvector u_1 of the matrix A associated with the largest eigenvalue λ_1.

Other factorial axis : The first factorial axis is by definition the one dimensional subspace that contributes the most to the inertia of the set of individuals. In order to determine the two dimensional space with the same property, it suffice to project the set of individuals on the M-orthogonal to u_1 subspace, that is a $(p-1)$ dimensional subspace, and to solve the same problem as before in this subspace. Its solution will be the unitary eigenvector u_2 of A associated with the second largest eigenvalue λ_2. And so on. As the matrix A is by definition a non negative symetric matrix, it admits a basis of M-orthonormal eigenvectors $(u_1, u_2, \ldots, u_p)$ corresponding to positive and decreasing eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. The axis defined by the p eigenvectors of this basis are called *factorial axis* or *principal components*.

It is easy to see using (1) that each eigenvalue λ_j is equal to the inertia of the projection of the set of individuals on the j^{th} factorial axis and that the sum of all eigenvalues $\sum_{j=1}^{p} \lambda_j$, which is the trace of A, is nothing else than the total inertia of the cloud.

Principal components in the space of variables : Instead of constructing a new basis in the space of individuals it is also possible to compute a new basis $v_1, v_2, \ldots, v_n$ in the space of variables. Using the same reasoning as before, one can show that the v_i will be a solution of

$$\text{Max}_v \{(v'DBv), \;\; v'Dv = 1\}$$

where B is the non negative and symetric matrix $B := XMX'D$. Thus they will be eigenvectors of B. But the non zero eigenvalues of A and B been the same, the formula between the $u_1, u_2, \ldots, u_p$ and the $v_1, v_2, \ldots, v_n$ is simply, for eigenvectors associated with non zero eigenvalues, $v_i = \frac{1}{\sqrt{\lambda_i}} X u_i$ if the v_i have been chosen orthonormal as well. **Graphic representations and measure of quality :** The computation of the principal components (PC) allows to represent the set of n individuals through a projection in the *first factorial plan* that is the two dimensional subspace generated by the two first PC. Usually the projection of the original variables are also plotted, as vectors and not as points, in this plan and sometime this suggests a possible meaning of the new axis in term of the original ones. When the data set contains more individuals than the n that have not been used to compute the PC, they can also be plotted on this plan (as projections) and interpreted in terms of the new variables.

The measure of the quality of the representation of the data set by the j^{th} first PC is given by the percentage of the total inertia contained in the corresponding space. For $j = 1, \ldots, p$, the percentage contained in the j first dimensions is given by

$$\frac{\sum_{1 \leq i \leq j} \lambda_i}{\sum_{1 \leq i \leq n} \lambda_i}.$$

Another plot that is usually done when a PC analysis is performed is the so called *scree diagram* that shows the decreasing sequence of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. If one is interested to determine how many dimensions are needed to represent well the data set, then the answer is easier if the scree shows a corner, the two or three first eigenvalues been much larger then all the others ; indeed, this means that two or three dimensions will be enought to take into account almost all the variability of the data set.

The representation of the data set in the plan of the two first PC in the space of variables is called *the correlation disk*. Endeed as the variables are usually contained in a sphere their projections on this plan belongs to the disk of radius 1 and their lengths are all the closer to one as they are located in the space close to this plan and all the closer to zero as they are located close to the orthogonal subspace. But the most important thing to look at in the correlation disk is the angle between the variables because the cosine of them represent the correlation between them.