1 Mathematics of principal component analysis (PCA)

Assume we have some observations on p variables, $X_1, \dots X_p$. Main aim of principal component analysis is reducing the dimensionality of data, finding a subspace of dimension $q < p$ on which original data can be projected with a small error.

Mathematically, we ask the following problem. Given a $p-$dimensional random variable $X = (X_1, \dots X_p)$, consider the transformations

$$Y = m + P(X - m) \tag{1}$$

where $m \in \mathbb{R}^p$ and P is an orthogonal projection on a subspace of dimension $q < p$. We wish to find m and P such that $\mathbb{E}(\|X - Y\|^2)$ is minimal. If $p = 2$ and $q = 1$, the problem becomes finding a line r such that the square distance (measured as orthogonal distance) between the distribution of X and the line is minimal. We will assume that $\mathbb{E}(X) = \mu$ and the variance-covariance matrix $\mathrm{Cov}(X) = S$.

In order to show how to find m and P, I recall some properties that can be proved with simple computations:

1. if X is a p-dimensional random variable with $\mathbb{E}(X) = \mu$, then

$$\mathbb{E}(\|X - \mu\|^2) \leq \mathbb{E}(\|X - m\|^2) \qquad \forall\ m \in \mathbb{R}^p.$$

2. if X is a ($p-$dimensional) random variable with $\mathbb{E}(X) = \mu$ and $\mathrm{Cov}(X) = S$, and A is a matrix $k \times p$, then $\mathbb{E}(AX) = A\mu$ and $\mathrm{Cov}(AX) = ASA^t$.

3. A projection is an operator such that $P^2 = P$. The subspace on which it projects is $V = \mathrm{Im}(P)$. If P is a projection, also $I - P$ is a projection (the complementary one). P is an orthogonal projection if $P^t = P$. The orthogonal projection on a subspace V can be written as follows: let A a matrix $(p \times q)$ whose columns are an orthogonal base of V (hence $A^tA = I_q$, the q-dimensional identity matrix); then $P = AA^t$.

Let Y given by (1). Then

$$\begin{aligned} \mathbb{E}(\|X - Y\|^2) &= \mathbb{E}(\|X - m - P(X - m)\|^2) \\ &= \mathbb{E}(\|(I - P)X - (I - P)m\|^2) \geq \mathbb{E}(\|(I - P)X - (I - P)\mu\|^2 \end{aligned} \tag{2}$$

because of the property 1. Hence, whichever is the projection P, the optimal choice to minimize $\mathbb{E}(\|X - Y\|^2)$ is $m = \mu$ in (1).

From here onwards, we will take $m = \mu$.

Let us consider now

$$\begin{aligned} &\mathbb{E}(\|(I - P)(X - \mu)\|^2 = \textstyle\sum_{i=1}^p \mathbb{E}(((I - P)(X - \mu))_i^2) = \sum_{i=1}^p (\mathrm{Cov}((I - P)X)_{ii} \\ &\quad = \mathrm{tr}(\mathrm{Cov}((I - P)X)) = \mathrm{tr}((I - P)S(I - P)^t). \end{aligned} \tag{3}$$

Now use the property that, for all matrices A and B such that both AB and BA are well-defined, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. Then, from (3), we get

$$\begin{aligned} \mathbb{E}(\|(I - P)(X - \mu)\|^2 &= \mathrm{tr}((I - P)S(I - P)^t) = \mathrm{tr}((I - P)^2 S) \\ &= \mathrm{tr}((I - P)S) = \mathrm{tr}(S) - \mathrm{tr}(PS) \end{aligned} \tag{4}$$

using the assumption that P is an orthogonal projection.

Since $\mathrm{tr}(S)$ is a given number, minimizing $\mathbb{E}(\|X-Y\|^2)$ is equivalent to maximizing $\mathrm{tr}(PS)$.

For the sake of simplicity, we restrict to the case $q=1$ and so $P=vv^t$ where v is a vector of norm 1 ($v^tv=1$). Thanks to the spectral decomposition theorem (S is symmetric and positive semidefinite), we obtain $S=C^t\Lambda C$, where C is an orthogonal matrix ($C^tC=I$), while Λ is a diagonal matrix with elements $\lambda_1,\lambda_2,\ldots,\lambda_p$ cassumed to be ordered:

$$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0.$$

Then

$$\mathrm{tr}(PS)=\mathrm{tr}(vv^tC^t\Lambda C)=\sum_{i=1}^{p}\lambda_i(Cv)_i^2. \tag{5}$$

Since v has norm 1 and C is orthogonal, also Cv has norm 1, i.e. $\sum_{i=1}^{p}(Cv)_i^2=1$. It is then clear that, if $\lambda_1>\lambda_2$[1], the choice of the vector Cv that maximizes (5) is

$$Cv=\begin{pmatrix}1\\0\\\cdots\\0\end{pmatrix}\quad\text{so that}\quad v=C^t\begin{pmatrix}1\\0\\\cdots\\0\end{pmatrix}.$$

Moreover the value of the maximum is $\mathrm{tr}(PS)=\lambda_1$.

Let us observe that v is eigenvector of S relative to λ_1. In fact

$$Sv=C^t\Lambda CC^t\begin{pmatrix}1\\0\\\cdots\\0\end{pmatrix}=C^t\Lambda\begin{pmatrix}1\\0\\\cdots\\0\end{pmatrix}=\lambda_1C^t\begin{pmatrix}1\\0\\\cdots\\0\end{pmatrix}=\lambda_1v.$$

It has then been proved

Teorema 1 *Let Y be given by (1) with dim(Im(P)) = 1. Then*

$$\mathbb{E}(\|X-Y\|^2)\geq\mathbb{E}(\|X-\bar{Y}\|^2)=tr(S)-\lambda_1\quad with\quad\bar{Y}=\mu+vv^t(X-\mu) \tag{6}$$

where v is a (normalized) eigenvector relative to the eigenvalue λ_1.

Observe that, during the proof of the theorem, it has also been proved that, if $Z=v^tX$ (a one-dimensional random variable), the vector v with $\|v\|=1$ such that $V(Z)$ is maximal is the eigenvector of S relative to λ_1.

More generally, one has

Teorema 2 *Let Y given by (1) with dim(Im(P)) = q. Then the projection $\bar{P}$ that minimizes $\mathbb{E}(\|Y-X\|^2)$ under the rule (1) is the orthogonal projection with $Im(\bar{P})$ the subspace generated by the eigenvalues of S relative to $\lambda_1,\ldots,\lambda_q$. Then*

$$\mathbb{E}(\|X-\bar{Y}\|^2)=tr(S)-(\lambda_1+\cdots+\lambda_q)=\lambda_{q+1}+\cdots+\lambda_p.$$

We can then consider projections on subspaces of increasing dimension, every time adding a new eigenvector (orthogonal to the previous ones).

[1] if $\lambda_1=\lambda_2$, the minimum problem has not a unique solution.

2 Practical aspects

In reality, we will not have the covariance matrix S but only a finite number of points (in dimension p) $x_1, \ldots, x_n$. The computations will then be performed with sample mean and covariance.

Another problem is that the technique is very sensitive to the measure units of the variables. Unless there are scientific reasons to know that a certain scale is appropriate to the problem studied, it is often recommended to standardize the components $X_i,\ i = 1 \ldots p$ so that they have equal variance. This is equivalent to using the correlation matrix C among the variables $X_1, \ldots, X_p$) instead of the covariance matrix S.

Generally, one wants to be sure that few components are sufficient to approximate the variable X; there are several rules of thumb: one may choose a q such that $(\lambda_1 + \cdots + \lambda_q)/(\lambda_1 + \cdots + \lambda_p)$ is at least 80-90%; otherwise it may be checked that $\lambda_{q+1} \ll \lambda_q$; Kaiser's rule (used especially in factor analysis) suggests to stop when $\lambda_q > 1 > \lambda_{q+1}$.

After having selected the components, one may wish to plot the original observations in these new variables to check for some patterns. Another use (prominent especially in sociological and psychological research) is to interpret the first principal components (the eigenvectors of S) as being the main factors determining the observations; one then looks at how the principal components are written in the original variables, and tries to intepret the resulting "factors".

Some examples of its use in R can be seen in the scripts on the web.