Unsupervised Learning

Tommaso Proietti
University of Rome Tor Vergata

Principal Components Analysis

Principal Components Analysis

Objective: given a set of p measurements on N individuals, we aim at determining $r \leq p$ orthogonal (uncorrelated) variables, called principal components, defined as linear combinations of the original ones.
The PCs are uncorrelated and have decreasing variance.

- Synthesis: information dimensionality reduction.
- Interpretation: express the original data in terms of a reduced number of underlying variables (factors).
- Score the individual profiles, with a summary score.
- Obtain multivariate displays (scatterplot) of the units in two or three dimensions.

The first component is designed to capture as much of the variability in the data as possible, and the succeeding components in turn extract as much of residual variability as possible.

Let $\mathbf{x}_i$ be a vector containing p measurements for unit i, $i = 1, \ldots, N$. We assume that the measurements are centred. The mean vector and the covariance matrix are

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i = \mathbf{0}, \quad \mathbf{S} = \frac{1}{N}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'.$$

Given a vector $\mathbf{a}$, $||\mathbf{a}|| = 1$, we denote by $\mathbf{x}_i^*$ the orthogonal projection of $\mathbf{x}_i$ along $\mathbf{a}$, obtained by a contraction or an expansion of $\mathbf{a}$,

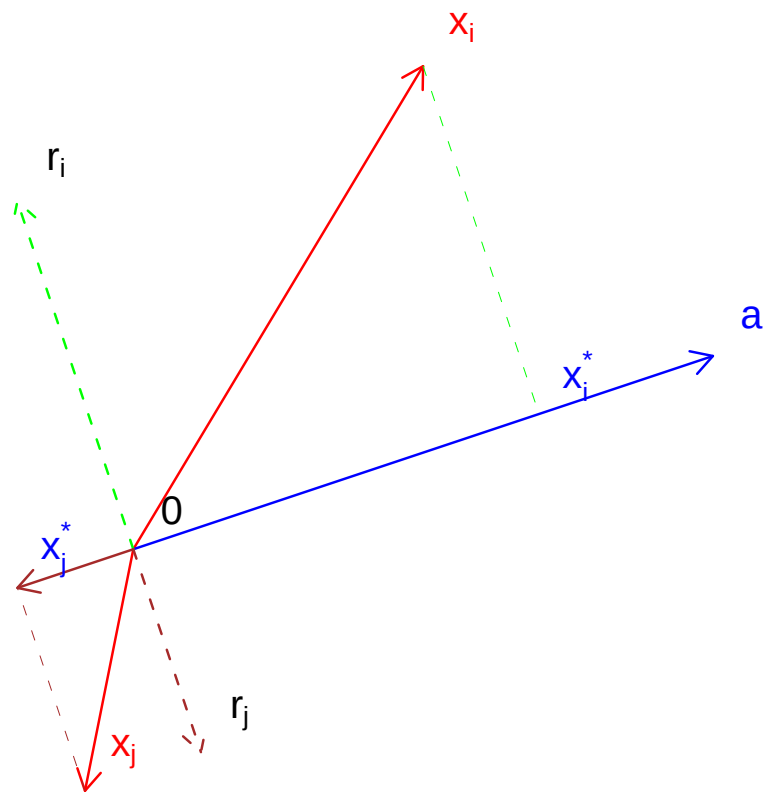
$$\mathbf{x}_i^* = z_i\mathbf{a}.$$

By the parallelogram rule we can find a vector $\mathbf{r}_i$, orthogonal to $\mathbf{x}_i^*$ ($\mathbf{r}_i'\mathbf{x}_i^* = 0$) such that:

$$\mathbf{x}_i = \mathbf{x}_i^* + \mathbf{r}_i, \tag{1}$$

which implies (Pythagora's thm)

$$||\mathbf{x}_i||^2 = ||\mathbf{x}_i^*||^2 + ||\mathbf{r}_i||^2 = {z_i}^2 + ||\mathbf{r}_i||^2.$$


x_i
r_i
a
x_i^*
0
x_j^*
r_j
x_j

The scalar z_i is the coordinate of unit i in the subspace generated by $\mathbf{a}$. It is obtained by the scalar product:

$$z_i = \frac{\mathbf{x}_i'\mathbf{a}}{||\mathbf{a}||^2} = \mathbf{x}_i'\mathbf{a}, \quad i = 1, \ldots, N.$$

The new variable z is centred around zero:

$$\bar{z} = \frac{1}{N}\sum z_i = \left(\frac{1}{N}\sum \mathbf{x}_i'\right)\mathbf{a} = 0$$

and has variance

$$\frac{1}{N}\sum_{i=1}^{N} {z_i}^2 = \frac{1}{N}\sum_{i=1}^{N} \mathbf{a}'\mathbf{x}_i\mathbf{x}_i'\mathbf{a} = \mathbf{a}'\mathbf{S}\mathbf{a}.$$

We now aim at choosing $\mathbf{a}$ (the subspace of dimension 1) such that the information loss from considering the projection $\mathbf{x}_i^*$ (and thus z_i), en lieu of the original $\mathbf{x}_i$, is a minimum:

$$\begin{aligned} \min\left\{\sum_{i=1}^n ||\mathbf{r}_i||^2\right\} &= \min\left\{\sum_{i=1}^n ||\mathbf{x}_i||^2 - \sum_{i=1}^n ||\mathbf{x}_i^*||^2\right\} \\ &\equiv \max\left\{\sum_{i=1}^N {z_i}^2\right\} = \max\left\{\mathbf{a}'\mathbf{S}\mathbf{a}\right\}. \end{aligned}$$

Solving the constrained optimization problem

$$\max\{\mathbf{a}'\mathbf{S}\mathbf{a}\} \qquad \text{s.v. } \mathbf{a}'\mathbf{a} = 1, \tag{2}$$

amounts to maximising the Lagrangian

$$\phi(\mathbf{a}, \lambda) = \mathbf{a}'\mathbf{S}\mathbf{a} - \lambda(\mathbf{a}'\mathbf{a} - 1).$$

Differentiating w.r.t $\mathbf{a}$ leads to the system $\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$, with $\mathbf{a}'\mathbf{a} = 1$.
The solution is to choose $\mathbf{a}$ as the eigenvector of the covariance matrix $\mathbf{S}$ corresponding to the largest eigenvalue, λ_1. Denote this by $\mathbf{a}_1, \mathbf{a}_1'\mathbf{a}_1 = 1$.

The vector $\mathbf{z}_1 = \mathbf{X}\mathbf{a}_1$ is know as the first principal component. It has zero mean and variance λ_1.
We can imagine extracting a second component, defined by the linear combination with weights (aka loadings) $\mathbf{a}_2$, $\mathbf{z}_2 = \mathbf{X}\mathbf{a}_2$, orthogonal to (uncorrelated with) the first, with maximal variance. The solution is provided by the eigenvector corresponding to the second eigenvalue, λ_2. The latter is also the variance of the second principal component.
The remaining p.c.'s are determined sequentially according to the same logic. Letting $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_p]$, and $\mathbf{S} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}'$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, PCA yields p orthogonal variables from p linear combinations

$$\mathbf{Z} = \mathbf{X}\mathbf{A}$$

The covariance matrix of the new variables is

$$\frac{1}{N}\mathbf{Z}'\mathbf{Z} = \mathbf{\Lambda}.$$

Geometrically, we aim at representing the matrix $\mathbf{X}$ in an orthogonal subspace (an hyperplane) with $r \leq p$ dimensions. The coordinates of the points along the projection subspace are given by the $N \times r$ matrix $\mathbf{Z}$: $\mathbf{Z} = \mathbf{X}\mathbf{A}_r$, and are obtained as linear combinations of the original measurements ($\mathbf{z}_k = \mathbf{X}\mathbf{a}_k, k = 1, 2, \ldots, r$).
The coefficients of the linear combinations are contained in the $p \times r$ matrix $\mathbf{A}_r$,

$$\mathbf{A}_r = [\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_r], \quad \mathbf{A}_r'\mathbf{A}_r = \mathbf{I}_r$$

and are called *loadings*, as they provide the weight assigned to the original variables used for constructing the PCs.

Standardisation of the variables and Mahalanobis distance

When the variables are standardized,

$$\mathbf{X} \to \mathbf{U}, \quad x_{ik} \longrightarrow u_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}$$

PCA is obtained from the spectral decomposition of the correlation matrix $\mathbf{R} = \frac{1}{N}\mathbf{U}'\mathbf{U}$.
An important result is the following: the Mahalanobis distance is equivalent to the Euclidean distance computed on the standardised PCs.

$$_M d_{ij} = \sqrt{\sum_{k=1}^{p} \frac{(z_{ik} - z_{jk})^2}{\lambda_k}}$$

Example: Hatco dataset

The dataset consists of $N = 100$ observations on $p = 7$ variables relating to the the perception of HATCO on seven attributes identified as the most influential in the choice of suppliers.
A survey of purchasing managers of firms buying from HATCO, a fictional industrial supplier, was conducted.
Each of these variables was measured on a graphic rating scale, ranging from zero to ten.

1. X1 Delivery speed
2. X2 Price level
3. X3 Price flexibility
4. X4 Manufacturer's image
5. X5 Service
6. X6 Salesforce's image
7. X7 Product quality

Correlation matrix

```
> R = cor(X)
     X1   X2   X3   X4   X5   X6   X7
X1 1.00 0.93 0.88 0.57 0.71 0.67 0.93
X2 0.93 1.00 0.84 0.54 0.75 0.47 0.94
X3 0.88 0.84 1.00 0.70 0.64 0.64 0.85
X4 0.57 0.54 0.70 1.00 0.59 0.15 0.41
X5 0.71 0.75 0.64 0.59 1.00 0.39 0.57
X6 0.67 0.47 0.64 0.15 0.39 1.00 0.57
X7 0.93 0.94 0.85 0.41 0.57 0.57 1.00
```

Eigenvalues and eigenvectors

```
> eigen(R, symmetric=TRUE)
$values
[1] 5.035 0.934 0.498 0.421 0.081 0.020 0.011
$vectors
      [,1]    [,2]    [,3]   [,4]     [,5]  [,6]   [,7]
[1,] -0.43  0.1118  0.0755 -0.042  0.63249  0.34  0.528
[2,] -0.42 -0.0293  0.4425  0.011 -0.00012 -0.79  0.099
[3,] -0.42 -0.0092 -0.2042 -0.325 -0.70103  0.16  0.399
[4,] -0.29 -0.6684 -0.4515 -0.303  0.26101 -0.11 -0.300
[5,] -0.35 -0.2949 -0.0059  0.847 -0.17426  0.20 -0.072
[6,] -0.29  0.6424 -0.6038  0.154  0.08696 -0.24 -0.228
[7,] -0.41  0.2004  0.4340 -0.246 -0.04958  0.37 -0.636
```

The best rank 1 approximation to $\mathbf{R}$ is provided by $R_1 = \lambda_1 \mathbf{a}_1 \mathbf{a}_1' =$

```
print(R1, digits = 2)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,] 0.95 0.92 0.92 0.64 0.76 0.63 0.89
[2,] 0.92 0.89 0.89 0.62 0.74 0.61 0.86
[3,] 0.92 0.89 0.89 0.62 0.74 0.61 0.86
[4,] 0.64 0.62 0.62 0.44 0.52 0.43 0.60
[5,] 0.76 0.74 0.74 0.52 0.61 0.51 0.72
[6,] 0.63 0.61 0.61 0.43 0.51 0.42 0.59
[7,] 0.89 0.86 0.86 0.60 0.72 0.59 0.84
```

The best rank 2 approximation is $R_2 = \lambda_1 \mathbf{a}_1 \mathbf{a}_1' + \lambda_2 \mathbf{a}_2 \mathbf{a}_2$,

```
print(R2, digits = 2)
     [,1] [,2] [,3]  [,4] [,5]  [,6] [,7]
[1,] 0.96 0.91 0.92 0.573 0.73 0.698 0.91
[2,] 0.91 0.89 0.89 0.641 0.75 0.594 0.86
[3,] 0.92 0.89 0.89 0.630 0.74 0.607 0.86
[4,] 0.57 0.64 0.63 0.853 0.70 0.028 0.48
[5,] 0.73 0.75 0.74 0.701 0.69 0.331 0.66
[6,] 0.70 0.59 0.61 0.028 0.33 0.806 0.71
[7,] 0.91 0.86 0.86 0.479 0.66 0.713 0.87
```

Principal components analysis

```
> pca = princomp(Z)
> print(loadings(pca),digits=2)

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
X1 -0.43   0.11                0.63   0.34   0.53
X2 -0.42          0.44                -0.79
X3 -0.42         -0.20  -0.32  -0.70   0.16   0.40
X4 -0.29  -0.67  -0.45  -0.30   0.26  -0.11  -0.30
X5 -0.35  -0.29          0.85  -0.17   0.20
X6 -0.29   0.64  -0.60   0.15         -0.24  -0.23
X7 -0.41   0.20   0.43  -0.25          0.37  -0.64
```

```
> summary(pca)
Importance of components:
                          Comp.1    Comp.2     Comp.3
Standard deviation     2.2212397 0.9564757 0.69854231
Proportion of Variance 0.7192283 0.1333594 0.07113139
Cumulative Proportion  0.7192283 0.8525877 0.92371910
                           Comp.4     Comp.5      Comp.6
Standard deviation     0.64251115 0.28181487 0.141187184
Proportion of Variance 0.06017793 0.01157720 0.002905805
Cumulative Proportion  0.98389702 0.99547423 0.998380033
                            Comp.7
Standard deviation     0.105418078
Proportion of Variance 0.001619967
Cumulative Proportion  1.000000000
```

Selection of the number of components

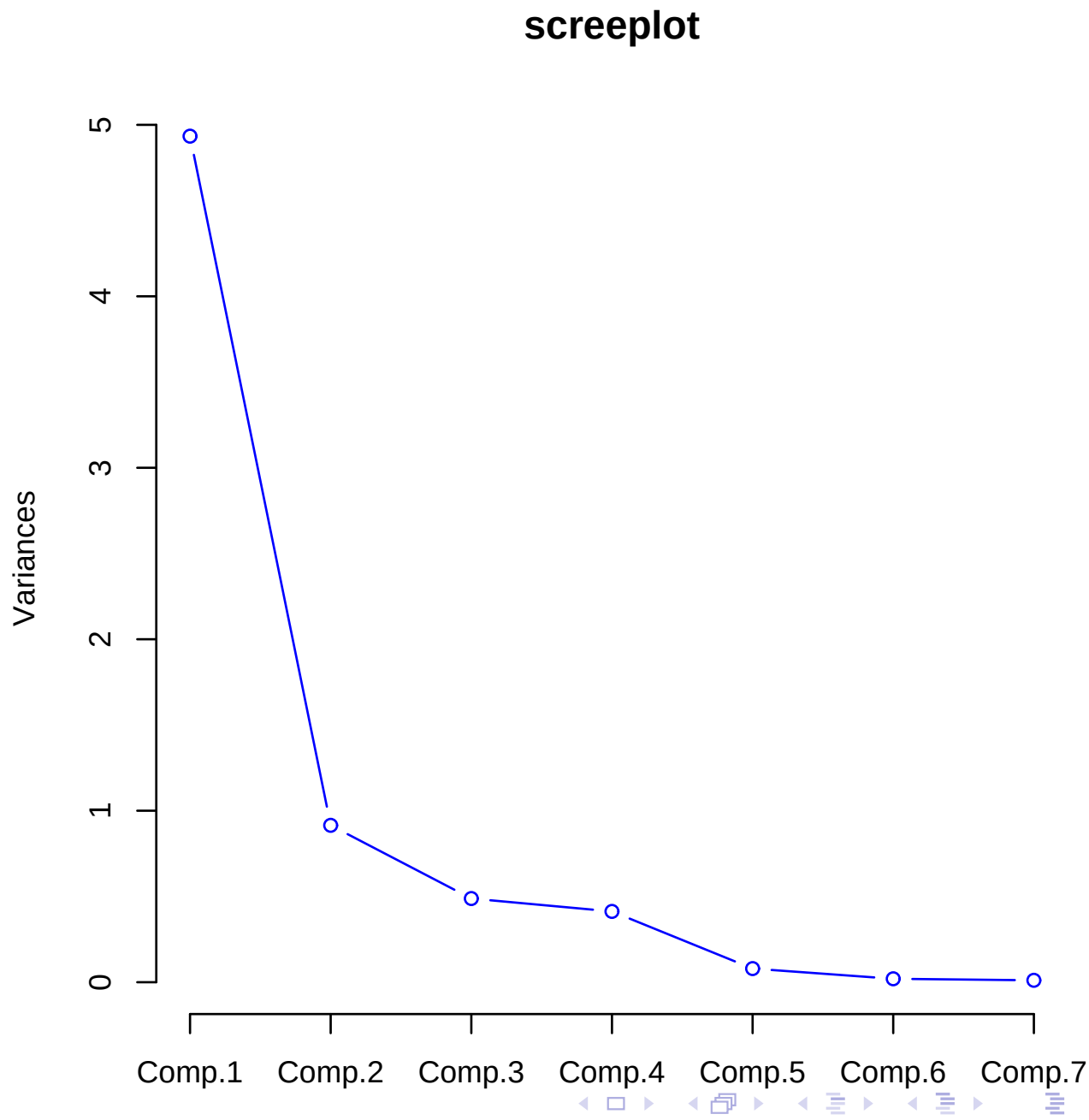
- Consider the share of the total variance absorbed by the first r components, Q_r,

$$Q_r = \frac{\sum_{h=1}^{r} \lambda_h}{\sum_{h=1}^{p} \lambda_h}$$

  Select r so that $Q_r \geq q$, where q is a large number.
- Kaiser criterion: compute the average eigenvalue

$$\bar{\lambda} = \frac{1}{p} \sum_{h=1}^{p} \lambda_h.$$

  Select the first r components for which $\lambda_h > \bar{\lambda}$. Note: if the variables are standardised, $\bar{\lambda} = 1$.
- Locate the elbow in the screeplot (plot of the λ_r vs r).


screeplot
Variances
0
1
2
3
4
5
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6
Comp.7

Appendix I: Spectral (eigen-) decomposition

Let $\mathbf{S}$ be a $p \times p$ **symmetric** matrix, i.e. $\mathbf{S} = \mathbf{S}'$.
Consider the problem of determining a scalar λ and a vector $\mathbf{a}$ that satisfy the linear equations system

$$\mathbf{Sa} = \lambda\mathbf{a}$$

under the normalization constraint: $\mathbf{a}'\mathbf{a} = 1$ ($\mathbf{a}$ has unit norm).
The homogenous system $(\mathbf{S} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0}$ has nontrivial solutions if $|\mathbf{S} - \lambda\mathbf{I}| = 0$. This condition yields an homogeneous equation of order p in λ with p solutions, denoted $\lambda_h, h = 1, \ldots, p$, which are known as eigenvalues. We order them in non-increasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ (≥ 0 if $\mathbf{S}$ is s.p.d.).

For each eigenvalue we can determine p corresponding eigenvectors, $\mathbf{a}_h, h = 1, \ldots, p$, so that $\mathbf{S}\mathbf{a}_h = \lambda \mathbf{a}_h$.
A key property is that eigenvectors corresponding to distinct eigenvectors are orthogonal: $\mathbf{a}_h' \mathbf{a}_k = 0$ for $h \neq k$.
Concatenating the eigenvectors into the matrix

$$\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_p]$$

we have $\mathbf{A}'\mathbf{A} = \mathbf{I}$ e $\mathbf{A}\mathbf{A}' = \mathbf{I}$, i.e. $\mathbf{A}$ is an orthogonal matrix ($\mathbf{A}^{-1} = \mathbf{A}'$).
Letting $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$, we have the following decomposition (spectral decomposition of $\mathbf{S}$):

$$\mathbf{S} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}' = \sum_{h=1}^{p} \lambda_h \mathbf{a}_h \mathbf{a}_h'$$

Some uses of this decomposition:

$$tr(\mathbf{S}) = \sum_{h=1}^{p} \lambda_h$$

$$\mathbf{S}^p = \mathbf{A}\mathbf{\Lambda}^p\mathbf{A}'$$

for p real (e.g. the square root of a matrix is $\mathbf{A}\mathbf{\Lambda}^{1/2}\mathbf{A}'$ where $\mathbf{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})$.
Note: these results hold only for symmetric matrices. In general λ_h need not be real and $\mathbf{a}_h$ are not orthogonal so that the decomposition is $\mathbf{S} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^{-1}$.

Example

$$\mathbf{S} = \begin{bmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix}.$$

Eigenvalues: $\lambda_1 = 1.8$ e $\lambda_2 = 0.2$. The eigenvector corresponding to λ_1 is

$$\mathbf{a}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.71 \\ 0.71 \end{bmatrix}$$

The second eigenvector, corresponding to $\lambda_2 = .2$ is:

$$\mathbf{a}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} 0.71 \\ -0.71 \end{bmatrix}$$

The spectral decomposition of $\mathbf{S}$ is

$$\mathbf{S} = 1.8\frac{1}{2}\mathbf{i}\mathbf{i}' + 0.2\frac{1}{2} \begin{bmatrix} 1 \\ -1 \end{bmatrix} [1 \quad -1];$$

The first addend is an approximation of rank 1 of $\mathbf{S}$.