Principal Component Analysis (PCA)

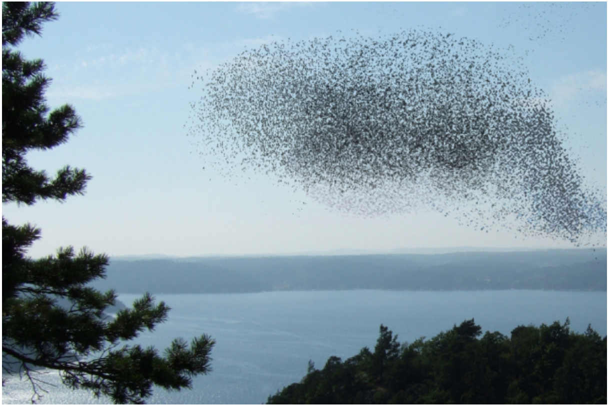
Nihan Acar-Denizli

11 April 2023

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Introduction to PCA

In the case of many variables, it is easier to describe the data using a few variables rather than all of the original variables.



Objectives

- to summarize data that composes of a large set of correlated variables with a smaller number of uncorrelated factors that explain most of the variability in the original data (dimension reduction)
- to cluster the individual observations by using the scores on the components (create indices)
- to represent data matrix in a low dimensional space (biplot)

Data Representation

PCA helps to represent data in a low dimensional space retaining as much as possible of the variation/information in the data set.

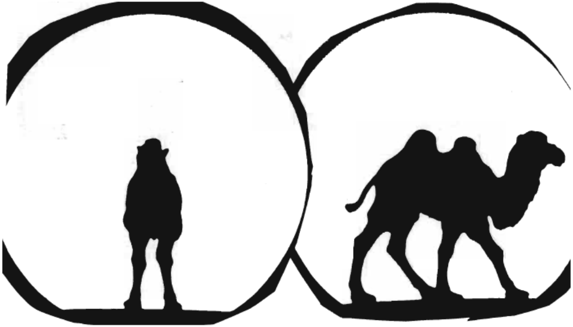


Figure: Camel or dromedary? (Illustration by J.P. Fénelon)

History of PCA

- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 6(2):559-572.

  Motivation: Finding lines and planes that best fit a set of points in p-dimensional space.

- Hotelling. H. (1933). Analysis of a complex of statistical variables into principal component analysis. *Journal of Educational Psychology*, 24:417-441.

  Motivation: Finding a smaller fundamental set of variables which determine the values of original p variables.

  For detailed information see, Joliffe, 2002.

Two Schools for Data Analysis

- French school of data analysis led by Jean Paul Benzecri. Based on projections and graphical displays, representation of both rows and variables are important (Husson et al., 2011)
- British school based on algebra and transformation of variables. Handles the case as an optimization problem with its constraints (Joliffe, 2002; James et al., 2007)

Steps of PCA

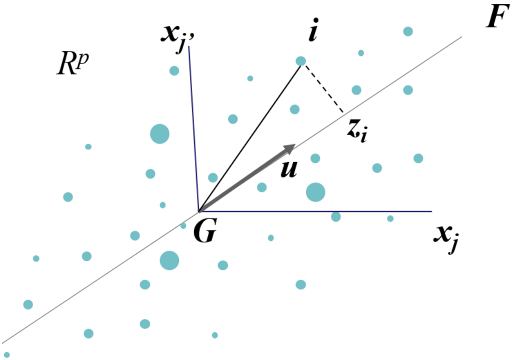
1. The correlations between variables are checked
2. The data matrix that composes of n number of observations and p number of variables is centered.
3. The covariance matrix $\mathbf{S}$ is computed.
4. The eigenvalues and eigenvectors of the covariance matrix are found.
5. m out of p components are chosen $(m < p)$.
6. The data is projected onto the eigenvectors.
7. Uncorrelated lower dimensional data is obtained.

Theory of PCA

Let $X_1, X_2, \ldots, X_p$ be a set of correlated variables.

PCA aims to find a new set of uncorrelated variables $Z_1, Z_2, \ldots, Z_p$ each of which is a linear combination of the p number of variables.

The new variables $Z_1, Z_2, \ldots, Z_p$ in order of importance (where Var$(Z_1) \geq$ Var$(Z_2) \geq \ldots \geq$ Var(Z_p)) are called principal components.



The Computation of Principal Components

The first PC is computed by,

$$Z_1 = u_{11}X_1 + u_{12}X_2 + \ldots + u_{1p}X_p,$$

where $u_{11}, u_{12}, \ldots, u_{1p}$ are called loadings and are the elements of the loading vector $u_1 = (u_{11}, u_{21}, \ldots, u_{p1})^t$ subject to the condition that

$$u_{11}^2 + u_{12}^2 + \ldots + u_{1p}^2 = 1.$$

The Computation of Principal Components

The loading vector u_1 defines a direction in the feature space along which the data vary the most.

If we project $X_1, X_2, \ldots, X_p$ onto this direction, the projected values are the principal component scores.

The Computation of Principal Components

The second PC is the linear combination of,

$$Z_2 = u_{21}X_1 + u_{22}X_2 + \ldots + u_{2p}X_p,$$

which is chosen to account for as much as possible of the remaining variation subject to two conditions:

1. Z_2 is uncorrelated with Z_1 (orthogonal to the first PC), $u_2 u_1 = 0$.
2. it is orthonormal $u_2 u_2 = 1$.

The Computation of Principal Components

The third PC is,

$$Z_3 = u_{31}X_1 + u_{32}X_2 + \ldots + u_{3p}X_p,$$

where

- $u_3 u_3 = u_{31}^2 + u_{32}^2 + \ldots + u_{3p}^2 = 1.$
- Z_3 uncorrelated with Z_1 and Z_2 ($u_3 u_1 = 0$ and $u_3 u_2 = 0$).

The Computation of Principal Components

Similarly the jth component can be written in the form of the linear combination of,

$$Z_j = u_{j1}X_1 + u_{j2}X_2 + \ldots + u_{jp}X_p,$$

subject to the conditions

- orthonormality $(u_j u_j = 1)$,
- orthogonality $(u_j u_i = 0 \text{ for } i < j)$.

The Computation of Principal Components

The principal component loading vectors and the eigenvalues are obtained from spectral decomposition of the covariance matrix:

$$\mathbf{S} = \mathbf{U}\mathbf{D}_\lambda\mathbf{U}'.$$

In matrix notation, the principal component scores are found by,

$$\mathbf{Z}_{n\times p} = \mathbf{X}^c_{n\times p}\mathbf{U}_{p\times p}$$

where $\mathbf{X}^c$ is centered data matrix.
Then the eigenvalues are the diagonal elements of $\mathbf{D}_\lambda$ which is computed by,

$$\begin{aligned}\frac{1}{n-1}\mathbf{Z}'\mathbf{Z} =& \frac{1}{n-1}(\mathbf{X}^c\mathbf{U})'\mathbf{X}^c\mathbf{U} = \frac{1}{n-1}\mathbf{U}'\mathbf{X}^{c\prime}\mathbf{X}^c\mathbf{U}\\ =& \mathbf{U}'\mathbf{S}\mathbf{U} = \mathbf{U}'\mathbf{U}\mathbf{D}_\lambda\mathbf{U}'\mathbf{U} = \mathbf{D}_\lambda\end{aligned}$$

The Computation of Principal Components

An alternative way to compute PCA is singular value decomposition of centered data matrix $\mathbf{X}^c$,

$$\mathbf{X}^c = \mathbf{VDU}'$$

The principal components are found from:

$$\mathbf{Z} = \mathbf{X}^c\mathbf{U} = \mathbf{VD}$$

Then the squared singular values relate to the variance of the principal components:

$$\frac{1}{n-1}\mathbf{Z}'\mathbf{Z} = \frac{1}{n-1}(\mathbf{VD})'\mathbf{VD} = \frac{1}{n-1}\mathbf{D}^2 = \mathbf{D}_\lambda$$

Eigenvalues and Eigenvectors

The variances of principal components are the eigenvalues of the variance-covariance matrix $\mathbf{S}$.

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0,$$

where λ_i corresponds to the ith principal component ($i = 1, \ldots, p$). $u_{i1}, u_{i2}, \ldots, u_{ip}$ are the elements of corresponding eigenvector.

The total variance of principal components is equal to the sum of the variances of original variables $s_1^2, s_2^2, \ldots, s_p^2$,

$$\sum_{j=1}^{p} \lambda_j = s_1^2 + s_2^2 + \ldots + s_p^2,$$

which is equivalent to,

$$\sum_{j=1}^{p} \lambda_j = \text{trace}(\mathbf{S}).$$

Scaling Variables

If variables $X_1, X_2, \ldots, X_p$ are standardized variables (have zero mean), then variance covariance matrix $\mathbf{S}$ becomes a correlation matrix $\mathbf{R}$.

In this case the sum of the eigenvalues will be equal to the number of variables p.

Standardization is necessary when the units of measurements of the observed variables differ.

How to decide number of components

- Percentage of explained variation (80% or more (Manly,2004))
- Size of the eigenvalue (according to Kaiser's rule the components with eigenvalues greater than 1)
- The scree plot (all the components up to the point where the bend occurs)

The Proportion of Explained Variation

The jth principal component accounts for a proportion of P_j of the total variation of the original data,

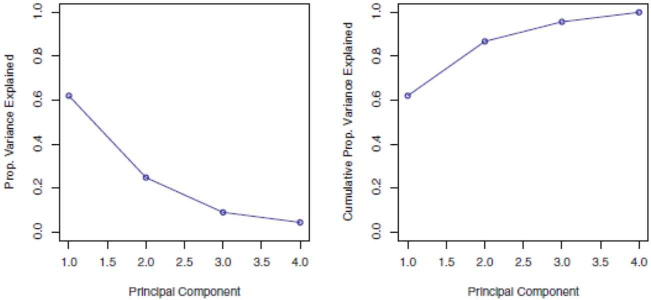
$$P_j = \frac{\lambda_j}{\text{trace}(\mathbf{S})}.$$

The first m principal components account for a proportion of,

$$p^m = \frac{\sum_{j=1}^{m} \lambda_j}{\text{trace}(\mathbf{S})}$$

Proportions of explained variation are shown on a graph called "scree plot" and it is used to decide optimum number of components.

The Scree Plot



Three Dimensions to two dimensions

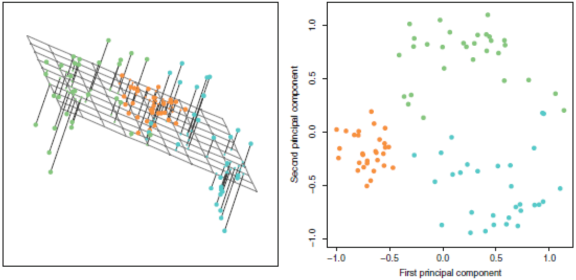


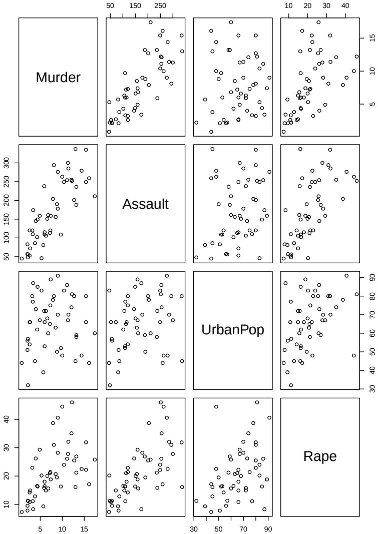
Figure: **Left:** The first two principal component directions span the plane that best fits the data. It minimizes the sum of squared distances from each point to the plane. **Right:** The first two principal component score vectors give the coordinates of the projection of the 90 observations onto the plane. The variance in the plane is maximized.

Example: USArrests Data Set

Consider the USArrests data set in HSAUR package in R.

- For 50 states the number of arrests per 100,000 residents for each of three crimes: Assault, Murder, Rape and the percent of the population in each state living in urban areas (UrbanPop) is recorded.

Correlations Among Variables



Correlations Among Variables

```
> crime<-USArrests
> summary(crime)
     Murder          Assault         UrbanPop          Rape      
 Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30  
 1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07  
 Median : 7.250   Median :159.0   Median :66.00   Median :20.10  
 Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23  
 3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18  
 Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00  
> cor(crime)
             Murder   Assault   UrbanPop      Rape
Murder   1.00000000 0.8018733 0.06957262 0.5635788
Assault  0.80187331 1.0000000 0.25887170 0.6652412
UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412
Rape     0.56357883 0.6652412 0.41134124 1.0000000
```

Principal Component Loadings

The principal component score vectors have length $n = 50$, and the principal component loading vectors have length $p = 4$.

	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

Eigenvalues

```
> summary(crimepc)
Importance of components:
                          Comp.1      Comp.2      Comp.3       Comp.4
Standard deviation     82.8908472 14.06956001 6.424204055 2.4578367034
Proportion of Variance  0.9655342  0.02781734 0.005799535 0.0008489079
Cumulative Proportion   0.9655342  0.99335156 0.999151092 1.0000000000
> crimepc2<-princomp(crime,cor=TRUE)
> summary(crimepc2)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation     1.5748783 0.9948694 0.5971291 0.41644938
Proportion of Variance 0.6200604 0.2474413 0.0891408 0.04335752
Cumulative Proportion  0.6200604 0.8675017 0.9566425 1.00000000
> vz<-(summary(crimepc)$sdev)^2
> vz
     Comp.1      Comp.2      Comp.3      Comp.4
6870.892554  197.952519   41.270398    6.040961
> vz<-(summary(crimepc2)$sdev)^2
> vz
   Comp.1    Comp.2    Comp.3    Comp.4
2.4802416 0.9897652 0.3565632 0.1734301
```

Biplot for USArrests Data Set

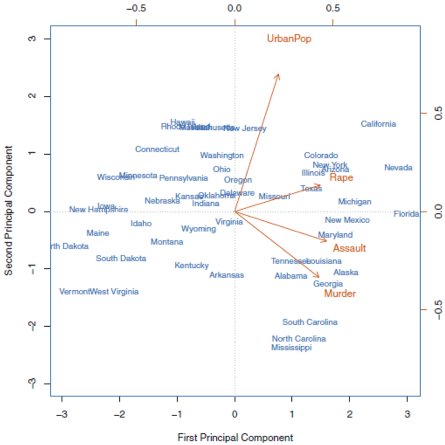


Figure: The orange arrows indicate the first two principal component loading vectors. The blue state names represent the scores for the first two principal components.

Interpretation of Biplot

- The first loading vector places approximately equal weight on Assault, Murder, and Rape, but less weight on UrbanPop. This component corresponds to the crimes.
- The second loading vector places most of its weight on UrbanPop and much less weight on the other three features. Hence, this component roughly corresponds to the level of urbanization of the state.
- The crime-related variables (Murder, Assault, and Rape) are located close to each other, and that the UrbanPop variable is far from the other three. This indicates that the crime-related variables are correlated with each other.

Assumptions

1. Normality: All variables and their linear combinations should be normally distributed.
2. Linearity: The relationships among pairs of variables should be linear.
3. Absence of outliers among individuals: Outliers on individuals could have more influence on the factor solution than the other cases.
4. Absence of outliers among variables:The variables that are unrelated to other variables in the data set effect the factor results (A variable with a low correlation with other variables and the important factors is an outlier among variables.).
5. Factorability: A factorable data set should include several sizeable correlations (the correlations should exceed 0.30).

Assumptions

- Bartlett's (1959) test of sphericity is used to test the hypothesis that the correlations in a correlation matrix are zero.

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right) ln|R| \approx \chi^2_{p(p-1)/2}$$

- Kaiser's (1970,1974) measure (Kaiser-Meyer-Olkin (KMO) test) check sampling adequacy. KMO-value is required to be 0.6 and above for a good PCA or Factor Analysis.

$$\text{KMO} = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} a_{ij}^2}$$

where r_{ij} are the elements of correlation matrix and a_{ij} are the partial correlations.

Introduction to Factor Analysis

Factor Analysis is introduced by Spearman (1904). Factor analysis is based on the model:

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \ldots + a_{ik}F_k + e_i$$

where
X_i : ith standardized variable with mean zero and unit variance
$a_{i1}, a_{i2}, \ldots, a_{ik}$: factor loadings for the ith variable
$F_i,\ i = 1, \ldots, k$: uncorrelated factors with mean zero and unit variance
e_i :error term which is uncorrelated with the common factors and has zero mean.

Introduction to Factor Analysis

Variance of X_i is equivalent to,

$$\mathsf{Var}(X_i) = 1 = a_{i1}^2\mathsf{Var}(F_1) + a_{i2}^2\mathsf{Var}(F_2) + \ldots + a_{ik}^2\mathsf{Var}(F_k) + \mathsf{Var}(e_i)$$

Then, $a_{i1}^2 + a_{i2}^2 + \ldots + a_{ik}^2$ is called the communality of (the part of variance X_i that is related to the common factors) X_i.

Factor Analysis by Using Principal Components

Let $Z_1, Z_2, \ldots, Z_p$ be the principal components of p variables,

$$
\begin{aligned}
Z_1 &= b_{11}X_1 + b_{12}X_2 + \ldots + b_{1p}X_p \\
Z_2 &= b_{21}X_1 + b_{22}X_2 + \ldots + b_{2p}X_p \\
&\vdots \\
Z_p &= b_{p1}X_1 + b_{p2}X_2 + \ldots + b_{pp}X_p
\end{aligned}
$$

where b_{ij} are the eigenvectors of the correlation matrix.

Factor Analysis by Using Principal Components

Since the transformation from X values to Z values are orthogonal, the inverse relationship is written as:

$$
\begin{aligned}
X_1 &= b_{11}Z_1 + b_{21}Z_2 + \ldots + b_{p1}Z_p \\
X_2 &= b_{12}Z_1 + b_{22}Z_2 + \ldots + b_{p2}Z_p \\
&\vdots \\
X_p &= b_{1p}Z_1 + b_{2p}Z_2 + \ldots + b_{pp}Z_p
\end{aligned}
$$

Factor Analysis by Using Principal Components

If m components are selected then the relationship becomes:

$$
\begin{aligned}
X_1 &= b_{11}Z_1 + b_{21}Z_2 + \ldots + b_{m1}Z_m + e_1 \\
X_2 &= b_{12}Z_1 + b_{22}Z_2 + \ldots + b_{m2}Z_m + e_2 \\
&\vdots \\
X_p &= b_{1p}Z_1 + b_{2p}Z_2 + \ldots + b_{mp}Z_m + e_p
\end{aligned}
$$

where $e_i,\ i = 1, \ldots, m$ are linear combinations of principal components from Z_{m+1} to Z_p.

Factor Analysis by Using Principal Components

Principal components can be converted to factors by scaling so as to have unit variance $F_i = Z_i/\sqrt{\lambda_i}$.
The equations then become:

$$
\begin{aligned}
X_1 &= \sqrt{\lambda_1} b_{11} F_1 + \sqrt{\lambda_2} b_{21} F_2 + \ldots + \sqrt{\lambda_m} b_{m1} F_m + e_1 \\
X_2 &= \sqrt{\lambda_1} b_{12} F_1 + \sqrt{\lambda_2} b_{22} F_2 + \ldots + \sqrt{\lambda_m} b_{m2} F_m + e_2 \\
&\vdots \\
X_p &= \sqrt{\lambda_1} b_{1p} F_1 + \sqrt{\lambda_2} b_{2p} F_2 + \ldots + \sqrt{\lambda_m} b_{mp} F_m + e_p
\end{aligned}
$$

Factor Analysis by Using Principal Components

Then the factor model is written as,

$$\begin{aligned}
X_1 &= a_{11}F_1 + a_{12}F_2 + \ldots + a_{1m}F_m + e_1 \\
X_2 &= a_{21}F_1 + a_{22}F_2 + \ldots + a_{2m}F_m + e_2 \\
&\vdots \\
X_p &= a_{p1}F_1 + a_{p2}F_2 + \ldots + a_{pm}F_m + e_p
\end{aligned}$$

where $a_{ij} = \sqrt{\lambda_j} b_{ji}$.

Comparison of PCA and Factor Analysis

Principal Component Analysis (PCA)

- It is not based on any particular model.
- The aim is to explain the as much of the variation of the variables as possible.
- In PCA the components are composites of observed variables.
- PCA tries to reduce number of variables by creating principal components based on original variables.

Factor Analysis (FA

- It is based on a model.
- The aim is to explain correlations among variables.
- In FA, the original variables are defined as the linear combination of factors.
- FA tries to identify latent variables to explain the original data.

Rotation

Rotation is used to improve the interpretability and scientific utility of the solution.

Some rotation methods are:

- Orthogonal Rotation
  - Varimax: Minimize complexity of factors.
  - Quartimax: Minimize complexity of variables.
  - Equamax: simplfy both variables and factors(compromise between varimax and quartimax).
- Oblique Rotation
  - Direct oblimin
  - Direct quartimax

For detailed information, see Tabachnik and Fidell (2013).

Some Real Life Applications of PCA and Factor Analysis

- In Banking to profile customers based on profiles.
- For image processing such as face recognition.
- In psychology to understand psychological scales.

REFERENCES

- Manly, B.F.J (1989). Multivariate statistical methods: a primer. 3rd edition. Chapman and Hall, London.
- Johnson and Wichern (2002). Applied Multivariate Statistical Analysis, 5th edition, Prentice Hall.
- Peña, D. (2002). Análisis de datos multivariantes, McGraw Hill.
- Joliffe, I.T.(2002). Principal Component Analysis, 2nd Edition, Springer.
- Tabachnik, B.G. and Fidell, L.S. (2013). Using Multivariate Statistics, Pearson.
- Grafelman, J. (2021). Lecture Notes of Data Analysis (Bachelor in Data Science,FIB,UPC).

THANK YOU FOR YOUR ATTENTION!