Principal Components Analysis

Tsung-Chi Cheng

Department of Statistics
National Chengchi University
Taipei 11605, Taiwan

E-mail: chengt@nccu.edu.tw

Notations

- Define the $n \times p$ data matrix

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

  where x_{ij} is the element in the ith row and jth column for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, p$.

- The data in the rth row (observation) of $\boldsymbol{X}$ are denoted as

$$\boldsymbol{x}_r^T = (x_{r1}\ x_{r2}\ \cdots\ x_{rp}) \text{ or } \boldsymbol{x}_r = \begin{pmatrix} x_{r1} \\ x_{r2} \\ \vdots \\ x_{rp} \end{pmatrix}$$

Mean vector and covariance matrix

- Mean vector of $\boldsymbol{X}$

$$\boldsymbol{\mu} = E(\boldsymbol{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

- Covariance matrix

$$\begin{aligned} \boldsymbol{\Sigma} = Var(\boldsymbol{X}) &= E(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T \\ &= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \end{aligned}$$

where

$$\begin{aligned} \sigma_{ii} &= Var(X_i) = E(X_i - \mu_i)^2, \ i = 1, 2, \cdots, p \\ \sigma_{ij} &= Cov(X_i, X_j) = E(X_i - \mu_i)(X_j - \mu_j) \ i \neq j = 1, 2, \cdots, p \end{aligned}$$

Correlation matrix

- Correlation matrix

$$\boldsymbol{R} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}$$

where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

Eigenvalues and eigenvectors

- $\lambda_1, \lambda_2, \cdots, \lambda_p$ denote the eigenvalues of $\mathbf{\Sigma}$ or $\boldsymbol{R}$
- $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_p$ are the corresponding orthogonal eigenvectors

$$\begin{aligned}
\boldsymbol{a}_1^T &= c(a_{11} \quad a_{12} \quad \cdots \quad a_{1p}) \\
\boldsymbol{a}_2^T &= c(a_{21} \quad a_{22} \quad \cdots \quad a_{2p}) \\
&\vdots \\
\boldsymbol{a}_p^T &= c(a_{p1} \quad a_{p2} \quad \cdots \quad a_{pp})
\end{aligned}$$

- eigen() in R

Principal Components Analysis (PCA)

- PCA is a statistical method that explains the correlation structure explained by the correlated number of p variables with the uncorrelated number of m variables which the linear combinations of the original variables provide ($p > m$).
- Eigenvalues and eigenvectors of the covariance or correlation matrices are used to find the linear combinations of the p variables in the data matrix, $\boldsymbol{X}$.
- The aim of PCA is to find out new independent measures that represent different linear combinations.
- Principal components may be useful for regression analysis when
  - There are too many explanatory variables relative to the number of observations
  - The explanatory variables are highly correlated

- One of the primary benefits of using PCA is that the directions of greatest variability give the most information about the configuration of the data in multidimensional space.
- The first PC has the greatest variance and extracts the largest amount of information from the data.
- The second component is orthogonal to the first one and has the greatest variance, in that the subspace is orthogonal to the first component; and, it extracts the greatest information in that subspace, and so on.
- The PCs also minimize the sum of the squared deviations of the residuals from the projection into linear subspaces of dimensions 1, 2, etc.
- The first PC gives a line such that the projections of the data onto this line have the smallest sum of squared deviations among all possible lines.
- The first two PCs define a plane that minimizes the sum of the squared deviations of the residuals, and so on.

Finding the sample principal components

- The first principal component of the observations is that linear combination of the original variables whose sample variance is greatest amongst all possible such linear combinations.
- The second principal component is defined as that linear combination of the original variables that accounts for a maximal proportion of the remaining variance subject to being uncorrelated with the first principal component.
- Subsequent components are defined similarly.
- The question now arises as to how the coefficients specifying the linear combinations of the original variables defining each component are found.

PC 1

- The first PC of the observations, $\boldsymbol{y}_1$, is the linear combination

$$\begin{aligned}\boldsymbol{y}_1 &= \boldsymbol{a}_1^T \boldsymbol{X} \\ &= a_{11}\boldsymbol{x}_1 + a_{12}\boldsymbol{x}_2 + \cdots + a_{1p}\boldsymbol{x}_p\end{aligned}$$

  whose sample variance is greatest among all such linear combinations.
- Because the variance of $\boldsymbol{y}_1$ could be increased without limit simply by increasing the coefficients $\boldsymbol{a}_1$, a restriction must be placed on these coefficients.
  - A sensible constraint is to require that the sum of squares of the coefficients should take the value one, although other constraints are possible and any multiple of the vector $\boldsymbol{a}_1$ produces basically the same component.
  - To choose the elements of the vector $\boldsymbol{a}_1$ so as to maximise the variance of $\boldsymbol{y}_1$ subject to the sum of squares constraint, which can be written $\boldsymbol{a}_1^T \boldsymbol{a}_1 = 1$
  - The sample variance of $\boldsymbol{y}_1$ that is a linear function of the $\boldsymbol{X}$ variables is given by $\boldsymbol{a}_1^T \boldsymbol{S} \boldsymbol{a}_1 = 1$.

Remarks

- To maximise a function of several variables subject to one or more constraints, the method of *Lagrange multipliers* is used.
- We simply state that the Lagrange multiplier approach leads to the solution that $\boldsymbol{a}_1$ is the *eigenvector* or *characteristic vector* of the sample covariance matrix, $\boldsymbol{S}$, corresponding to this matrix's largest *eigenvalue* or *characteristic root*.
- The eigenvalues $\boldsymbol{\lambda}$ and eigenvectors $\boldsymbol{\Gamma}$ of a $p \times p$ matrix $\boldsymbol{A}$ are such that $\boldsymbol{A\Gamma} = \boldsymbol{\lambda\Gamma}$.

PC 2

- The second principal component, $\boldsymbol{y}_2$, is defined to be the linear combination

$$\begin{aligned} \boldsymbol{y}_2 &= \boldsymbol{a}_2^T \boldsymbol{X} \\ &= a_{21}\boldsymbol{x}_1 + a_{22}\boldsymbol{x}_2 + \cdots + a_{2p}\boldsymbol{x}_p \end{aligned}$$

  has the greatest variance subject to the following two conditions:

$$\boldsymbol{a}_2^T \boldsymbol{a}_2 = 1$$
$$\boldsymbol{a}_1^T \boldsymbol{a}_2 = 0$$

- The second condition ensures that $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are uncorrelated.
  - the sample correlation is zero

PC j

- Similarly, the jth PC is that linear combination $\boldsymbol{y}_j$, is the linear combination

  $$\begin{aligned} \boldsymbol{y}_j &= \boldsymbol{a}_j^T \boldsymbol{X} \\ &= a_{j1}\boldsymbol{x}_1 + a_{j2}\boldsymbol{x}_2 + \cdots + a_{jp}\boldsymbol{x}_p \end{aligned}$$

  that has the greatest sample variance subject to the conditions, for $i < j$

  $$\begin{aligned} \boldsymbol{a}_j^T \boldsymbol{a}_j = 1 \\ \boldsymbol{a}_i^T \boldsymbol{a}_j = 0 \end{aligned}$$

- Application of the Lagrange multiplier technique demonstrates that the vector of coefficients defining the jth PC, $\boldsymbol{a}_j$, is the eigenvector of $\boldsymbol{S}$ associated with its jth largest eigenvalue.

PCA

- If the p eigenvalues of $\boldsymbol{S}$ are denoted by $\lambda_1, \lambda_2, \cdots, \lambda_p$, then by requiring that $\boldsymbol{a}_i^T \boldsymbol{a}_i = 1$ be shown that the variance of the ith principal component is given by λ_i.
- The total variance of the p principal components will equal the total variance of the original variables so that

$$\begin{aligned} \sum_{i=1}^{p} \lambda_i &= s_1^2 + s_2^2 + \cdots + s_p^2 \\ &= trace(\boldsymbol{S}) \end{aligned}$$

  where s_i^2 is the sample variance of X_i.

PCA

- Consequently, the jth principal component accounts for a proportion P_j of the total variation of the original data, where

$$P_j = \lambda_j / trace(\boldsymbol{S})$$

- The first m principal components, where $m < p$ account for a proportion $P^{(m)}$ of the total variation in the original data, where

$$P^{(m)} = \sum_{j=1}^{m} \lambda_j \bigg/ trace(\boldsymbol{S})$$

Remarks I

- In geometrical terms, it is easy to show that the first PC defines the line of best fit (in the sense of minimising residuals orthogonal to the line) to the p-dimensional observations in the sample.
- These observations may therefore be represented in one dimension by taking their projection onto this line
  - that is, finding their first principal component score
- If the observations happen to be collinear in p dimensions, this representation would account completely for the variation in the data and the sample covariance matrix would have only one non-zero eigenvalue.
- In practise, of course, such collinearity is extremely unlikely, and an improved representation would be given by projecting the p-dimensional observations onto the space of the best fit, this being defined by the first two PCs.
  - Similarly, the first m components give the best fit in m dimensions.

Remarks I

- If the observations fit exactly into a space of m dimensions, it would be indicated by the presence of $p - m$ zero eigenvalues of the covariance matrix.
- This would imply the presence of $p - m$ linear relationships between the variables.
- Such constraints are sometimes referred to as structural relationships.
- In practise, in the vast majority of applications of principal components analysis, all the eigenvalues of the covariance matrix will be non-zero.

Choosing the number of components

- Retain just enough components to explain some specified large percentage of the total variation of the original variables.
  - Values between 70% and 90% are usually suggested.
- Exclude those principal components whose eigenvalues are less than the average, $\sum_{j=1}^{p} \lambda_j / p$.
- When the components are extracted from the correlation matrix, $trace(\boldsymbol{R}) = p$, and the average variance is therefore one, so applying the rule in the previous bullet point, components with eigenvalues less than one are excluded.
- Scree diagram: the plot of the λ_j against j
- Modified scree diagram: the plot of the $\log(\lambda_j)$ against j

S or R?

- The structure of the PCs derived from the sample covariance matrix, $\boldsymbol{S}$, will depend upon the essentially arbitrary choice of units of measurement.
- PCs should only be extracted from $\boldsymbol{S}$ when all the original variables have roughly the same scale
- But this is rare in practise and consequently, in practise, PCs are extracted from the correlation matrix of the variables, $\boldsymbol{R}$.
- Extracting the components as the eigenvectors of $\boldsymbol{R}$ is equivalent to calculating the PCs from the original variables after each has been standardised to have unit variance.
- There is rarely any simple correspondence between the components derived from $\boldsymbol{S}$ and those derived from $\boldsymbol{R}$.
- Choosing to work with $\boldsymbol{R}$ rather than with $\boldsymbol{S}$ involves a definite but possibly arbitrary decision to make variables "equally important".

Concluding remarks

- PCA is a multivariate technique with the central aim of reducing the dimensionality of a multivariate data set while accounting for as much of the original variation as possible present in the data set.
- This aim is achieved by transforming to a new set of variables, the *principal components*, that are linear combinations of the original variables, which are uncorrelated and are ordered so that the first few of them account for most of the variation in all the original variables.
- princomp() in R