# Principle component analysis (PCA)

## Notation and preliminaries
- Data matrix $X$ ................................................................. 3
- Summary statistics .......................................................... 4
- Useful R-commands ...................................................... 5
- A small warning ........................................................... 6
- Measures of multivariate spread ........................................ 7

## Introduction to PCA
- Introduction ................................................................. 9
- Definition of PCA .......................................................... 10
- Examples ............................................................... 11
- Possible uses of PCA .................................................... 12
- Possible uses of PCA .................................................... 13
- Possible uses of PCA .................................................... 14

## Now a bit more math (MKB, Ch 8)
- Brief review of linear algebra ............................................. 16
- Spectral decomposition theorem .......................................... 17
- Formal definition of PCA - population case ......................... 18
- Example in 2-dimensions .................................................. 19
- Properties ................................................................. 20
- Properties ................................................................. 21
- Properties ................................................................. 22
- Corollaries ................................................................. 23
- Sample version ............................................................. 24
- Properties of sample version ............................................ 25

## Some practical issues
- Two R-implementations ................................................. 27
- Using correlation matrix instead of $\Sigma$ .......................... 28
- Scaling of the variables ................................................. 29
- Robust versions of PCA .................................................. 30
- How many principal components to use? .......................... 31
- Effect of ignoring some components on individual variables ... 32
- Some limitations of PCA ................................................. 33
Data matrix $X$

- Data matrix $X$ has $n$ rows (cases) and $p$ columns (variables). Sometimes written as $X(n \times p)$.
- Rows of $X$ are denoted by $x'_1, \ldots, x'_n$. Note that $x_i$ is the $i$th row of $X$ written as a column.
- Columns of $X$ are denoted by $x^{(1)}, \ldots, x^{(p)}$.
- So we have:
  $$X = \begin{bmatrix}
    x'_1 \\
    x'_2 \\
    \vdots \\
    x'_n
  \end{bmatrix} = [x_1', \ldots, x_n']' = [x^{(1)}, \ldots, x^{(p)}],$$
  where $x'_i = [x_{i1}, \ldots, x_{ip}]$, and $x^{(j)}_i = [x_{ij}, \ldots, x_{nj}]$.
- Usually, $x_1, \ldots, x_n$ are a random sample, and $x^{(1)}, \ldots, x^{(p)}$ are not.

Summary statistics

- Sample mean of $j$th variable: $\bar{x}_j = \frac{1}{n} \sum_{r=1}^{n} x_{rj}$
- Sample variance of $j$th variable: $s_j = s_j^2 = \frac{1}{n} \sum_{r=1}^{n} (x_{rj} - \bar{x}_j)^2$
- Sample covariance between $i$th and $j$th variable: $s_{ij} = \frac{1}{n} \sum_{r=1}^{n} (x_{rj} - \bar{x}_j)(x_{ri} - \bar{x}_i)$.
- Sample correlation coefficient between $i$th and $j$th variable: $r_{ij} = s_{ij} / (s_i s_j)$. The correlation coefficient is invariant under changes of scale and origin. Note that $|r_{ij}| \leq 1$.
- Sample mean vector: $\bar{x} = [\bar{x}_1, \ldots, \bar{x}_p]'$. Can be viewed as the center of gravity of the sample points in $\mathbb{R}^p$.
- Sample covariance matrix $S = (s_{ij})$ (size $p \times p$). $S$ is an unbiased estimate of the population covariance matrix $\Sigma$ if we use $1/(n - 1)$ instead of $1/n$.
- Sample correlation matrix $R = (r_{ij})$ (size $p \times p$).

Useful R-commands

- For sample mean vector: `apply(data, 2, mean)`
- For sample covariance matrix: `cov(data)`
- For sample correlation matrix: `cor(data)`
- `cov2cor` scales a covariance matrix into a correlation matrix efficiently
A small warning
- The multivariate summary measures we saw so far are simple generalizations of their 1-dimensional counterparts:
  - sample mean ⇒ sample mean vector
  - sample variance ⇒ sample covariance matrix
- This does not hold always. For example, the generalization of the sample median to the sample median vector is not a good measure of the center of the data.
  - Exercise: construct a data example where the component wise median lies on the convex hull of the data points.

Measures of multivariate spread
- The sample covariance matrix $S$ measures the spread of the data points about the mean.
- Sometimes it is convenient to have a single number that indicates the spread. Two common measures:
  - The generalized variance, $|S|$ (determinant of $S$)
  - The total variation, $\text{tr}(S)$ (sum of diagonal elements of $S$)
For both measures, a large value indicates a large amount of spread.

Introduction to PCA

Introduction
- Technique quite old: Pearson (1901) and Hotelling (1933), but still one of the most used multivariate techniques today
- Main idea:
  - Start with variables $X_1, \ldots, X_p$
  - Find a rotation of these variables, say $Y_1, \ldots, Y_p$ (called principal components), so that:
    - $Y_1, \ldots, Y_p$ are uncorrelated. Idea: they measure different dimensions of the data.
    - $\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \ldots \geq \text{Var}(Y_p)$. Idea: $Y_1$ is most important, then $Y_2$, etc.
Definition of PCA

- Given \( X = (X_1, \ldots, X_p)' \)
- We call \( a'X \) a standard linear combination (SLC) if \( \sum a_i^2 = 1 \)
- Find the SLC \( a'_{(1)} = (a_{11}, \ldots, a_{p1}) \) so that \( Y_1 = a'_{(1)}X \) has maximal variance
- Find the SLC \( a'_{(2)} = (a_{12}, \ldots, a_{p2}) \) so that \( Y_2 = a'_{(2)}X \) has maximal variance, subject to the constraint that \( Y_2 \) is uncorrelated to \( Y_1 \).
- Find the SLC \( a'_{(3)} = (a_{13}, \ldots, a_{p3}) \) so that \( Y_3 = a'_{(3)}X \) has maximal variance, subject to the constraint that \( Y_3 \) is uncorrelated to \( Y_1 \) and \( Y_2 \)
- Etc...

Examples

- See R-code for example in 2-dimensions
- See R-code for example using Bumpus' sparrow data

Possible uses of PCA

- Interest in first principal component:
  - Example: How to combine the scores on 5 different examinations to a total score? One could simply take the average. But it may be better to use the first principal component. Since the first principal component maximizes the variance, it spreads out the scores as much as possible.
  - Example: How to combine different cost factors into a cost of living index? Use first principal component. (Same rationale as above.)

Possible uses of PCA

- Interest in 2nd - pth principal components:
  - When all measurements are positively correlated, the first principal component is often some kind of average of the measurements (e.g., size of birds, severity index of psychiatric symptoms)
  - Then the other principal components give important information about the remaining pattern (e.g., shape of birds, pattern of psychiatric symptoms)
### Possible uses of PCA

- Interest in first few principal components:
  - Dimension reduction: summarize the data with a smaller number of variables, losing as little information as possible.
  - Can be used for graphical representations of the data.
- Use PCA as input for regression analysis:
  - Highly correlated explanatory variables are problematic in regression analysis.
  - One can replace them by their principal components, which are uncorrelated by definition.

### Brief review of linear algebra

- $\lambda_1$ is an eigenvalue of a matrix $A$ if $Ax = \lambda_1 x$ for some $x \neq 0$.
- The vector $x$ is called the corresponding eigenvector.
- If $x$ is scaled so that $\|x\| = 1$, then $x$ is called a standardized eigenvector.
- An orthogonal matrix $Q$ is a square matrix whose transpose is its inverse: $QQ' = Q'Q = I$.
- A $p \times p$ symmetric matrix $Q$ is positive definite if $x^TQx > 0$ for all $0 \neq x \in \mathbb{R}^p$.
- For any symmetric matrix $A$, tr($A$) equals the sum of its eigenvalues.
- For any symmetric matrix $A$, $|A|$ equals the product of its eigenvalues.

### Spectral decomposition theorem

- Any symmetric matrix $A(p \times p)$ can be written as

  $A = \Gamma \Lambda \Gamma'$

  where $\Lambda$ is a diagonal matrix of eigenvalues of $A$ and $\Gamma$ is an orthogonal matrix whose columns consist of the corresponding standardized eigenvectors.
Formal definition of PCA - population case

■ We first consider the population case: Let $X \in \mathbb{R}^p$ be a random vector with mean $\mu$ and covariance matrix $\Sigma$ (note that we don’t make any assumptions about the distribution of $X$).
■ Then the principal component transformation is the transformation

$$X \rightarrow Y = \Gamma'(X - \mu)$$

where $\Gamma$ is the orthogonal matrix consisting of the standardized eigenvectors corresponding to the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ of $\Sigma$. Thus, $\Sigma = \Gamma \Lambda \Gamma'$, or equivalently, $\Gamma' \Sigma \Gamma = \Lambda$.

Example in 2-dimensions

■ See board

Properties

■ Theorem 8.2.1 of MKB:
  ◆ $E(Y_i) = 0$
  ◆ $\text{Var}(Y_i) = \lambda_i$
  ◆ $\text{Cov}(Y_i, Y_j) = 0$ if $i \neq j$
  ◆ $\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \cdots \geq \text{Var}(Y_p)$
  ◆ $\sum_{i=1}^{p} \text{Var}(Y_i) = \text{tr}(\Sigma)$
  ◆ $\prod_{i=1}^{p} \text{Var}(Y_i) = |\Sigma|$

See proof on board.

Properties

■ Theorem 8.2.2 of MKB:
  ◆ No SLC of $X$ has a larger variance than $\lambda_1$, the variance of the first principal component.

See proof on board.
Properties

- Theorem 8.2.3 of MKB:
  - If $\alpha = a'X$ is a SLC of $X$ which is uncorrelated with the first $k$ principal components of $X$, then the variance of $\alpha$ is maximized when $\alpha$ is the $(k+1)$th principal component of $X$.

  See proof on board.

Corollaries

- The sum of the first $k$ eigenvalues divided by the sum of all the eigenvalues,
  $$(\lambda_1 + \cdots + \lambda_k)/(\lambda_1 + \cdots + \lambda_p),$$
  represents the proportion of total variation explained by the first $k$ principal components.

- If the covariance matrix of $X$ has rank $r < p$, then the last $p-r$ eigenvalues of $\Sigma$ are zero, and the total variation can be entirely explained by the first $r$ principal components. See 2-dimensional example on board.

Sample version

- $X = (x_1, \ldots, x_n)'$ is a $(n \times p)$ data matrix.
- $a$ is a standardized vector: $a'a = 1$
- Then $Xa$ gives $n$ observations on a new variable defined as a weighted sum of the columns of $X$. The variance of this new variable is $a'Sa$, where $S$ is the sample covariance matrix of $X$.
- Which $a$ gives the largest variance?
- In the population case, we had $Y_1 \in \mathbb{R}$ and $X \in \mathbb{R}^p$: $Y_1 = \gamma'(1)(X - \mu) = (X - \mu)'\gamma(1)$
- In the sample case things are analogous: $y(1) = (X - \bar{x})'g(1)$, where $1 \in \mathbb{R}^n$ is a vector of ones, and $g(1)$ is the standardized eigenvector corresponding to the largest eigenvalue of $S$.
- In general: $Y = [y(1), \ldots, y(p)] = (X - \bar{x}')G$
  (where $Y$ is $n \times p$, $X$ is $n \times p$, and $G$ is $p \times p$)

Properties of sample version

- Columns of $Y$ have mean zero
- Columns of $Y$ are uncorrelated
- The sample variance of $y(i)$ equals $l_i$, $i = 1, \ldots, p$, where $l_1 \geq \cdots \geq l_p$ are the eigenvalues of $S$.

See proof on board.
Some practical issues

**Two R-implementations**

- There are two R-implementations of PCA: `prcomp()` and `princomp()`.
- `prcomp` is numerically more stable, and therefore preferred.
- `princomp` has a few more options.

**Using correlation matrix instead of Σ**

- If we first standardize $X$ to have mean zero and variance 1, then the covariance matrix of the standardized $X$ equals the correlation matrix of $X$. See board.
- Hence, working with the correlation matrix instead of Σ is equivalent to working with standardized variables (see also the options `scale = T/F` and `center = T/F` in `prcomp()`).

**Scaling of the variables**

- PCA is not scale invariant.
- Example:
  - Suppose we have measurements in kg and meters.
  - We want to have principal components expressed in grams and hectometers.
  - Option 1: multiply measurements in kg by 1000, multiply measurements in meters by 1/100, and then apply PCA.
  - Option 2: apply PCA on original measurements, and then re-scale to the appropriate units.
  - These two options will generally give different results!
- Reason: eigenvectors are not scale invariant.
- See example in R-code.

**Robust versions of PCA**

- PCA is sensitive to outliers, since it is based on the sample covariance matrix $S$ which is sensitive to outliers.
- Robust versions of PCA can be obtained by using a robust version of $S$, see option `covmat` in the function `princomp()`.
How many principal components to use?

■ When the goal of PCA was dimension reduction, how many principal components should we use?
■ Various rules of thumb:
  ◆ Take the smallest number of components needed to explain, say 80% or 90% of the total variance
  ◆ Exclude principal components whose eigenvalues are less than average (i.e., less than one if the correlation matrix was used) - tends to include few components
  ◆ Use screeplot (see R-code) - tends to include more components

Effect of ignoring some components on individual variables

■ Let’s consider the correlation between the original variables \(X\) and the principal components \(Y\).
■ Let \(\rho_{ij}\) be the correlation between \(X_i\) and \(Y_j\).
■ \(\rho_{ij}^2\) can be interpreted as the amount of variation of \(X_i\) that is explained by \(Y_j\) (compare to \(R^2\) in linear regression).
■ Sample version: look at \(r_{ij}^2\).
■ See R-code.
■ Effect of throwing away some principal components is not the same for all variables.

Some limitations of PCA

■ The directions with largest variance are assumed to be of most interest.
■ We only consider orthogonal transformations (rotations) of the original variables. (Kernel PCA is an extension of PCA that allows non-linear mappings).
■ PCA is based only on the mean vector and the covariance matrix of the data. Some distributions (e.g. multivariate normal) are completely characterized by this, but others are not.
■ Dimension reduction can only be achieved if the original variables were correlated. If the original variables were uncorrelated, PCA does nothing, except for ordering them according to their variance. See 2-dimensional example on board.
■ PCA is not scale invariant.