Statistics in Geophysics: Principal Component Analysis

Steffen Unkel

Department of Statistics
Ludwig-Maximilians-University Munich, Germany

Multivariate data

- Let $\mathbf{x} = (x_1, \ldots, x_p)^\top$ be a p-dimensional random vector with population mean $\boldsymbol{\mu}$ and population covariance matrix $\boldsymbol{\Sigma}$.
- Suppose that a sample of n realizations of $\mathbf{x}$ is available.
- These np measurements x_{ij} $(i = 1, \ldots, n; j = 1, \ldots, p)$ can be collected in a data matrix

$$\mathbf{X} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)})^\top = (\mathbf{x}_1, \ldots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$$

with $\mathbf{x}_{(i)}^\top = (x_{i1}, \ldots, x_{ip})$ being the i-th observation vector $(i = 1, \ldots, n)$ and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\top$ being the vector of the n measurements on the j-th variable $(j = 1, \ldots, p)$.

Preprocessing I

- It will be useful to preprocess $\mathbf{x}$ so that its components have commensurate means.
- This is done by centring $\mathbf{x}$, that is, $\mathbf{x} \leftarrow \mathbf{x} - \boldsymbol{\mu}$. For the transformed vector $\mathbf{x}$ it holds that $\mathrm{E}(\mathbf{x}) = \mathbf{0}_p$.
- In a sample setting, the centred data matrix in which all columns have zero mean can be computed as

$$\mathbf{X} \leftarrow \mathbf{C}_n \mathbf{X} \ ,$$

where $\mathbf{C}_n = (\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)$ is the centring matrix.

Preprocessing II

- Unless specified otherwise, it is always assumed in the sequel that both $\mathbf{x}$ and $\mathbf{X}$ are mean-centred.
- The sample covariance matrix of $\mathbf{X}$ is $\mathbf{S}_{\mathbf{X}} = \mathbf{X}^\top \mathbf{X}/(n-1)$.
- One can transform a mean-centred vector or mean-centred data further such that its variables have commensurate scales.

Preprocessing III

- Let $\mathbf{\Delta}$ be the $p \times p$ diagonal matrix whose elements on the main diagonal are the same as those of $\mathbf{\Sigma}$.

- The standardized random vector $\mathbf{z}$ with components having unit variance can be obtained as

$$\mathbf{z} = \mathbf{\Delta}^{-1/2}\mathbf{x} \ ,$$

where $\mathbf{\Delta}^{-1/2}$ is the diagonal matrix whose diagonal entries are the inverses of the square roots of those of $\mathbf{\Delta}$.

Preprocessing IV

- Let $\mathbf{D}$ denote the $p \times p$ diagonal matrix whose elements on the main diagonal are the same as those of $\mathbf{S}_{\mathbf{X}}$.

- A standardized data matrix $\mathbf{Z}$ with all its columns having variance equal to one can be computed as

$$\mathbf{Z} = \mathbf{X}\mathbf{D}^{-1/2} \ ,$$

  where $\mathbf{D}^{-1/2}$ is the diagonal matrix whose diagonal entries are the inverses of the square roots of those of $\mathbf{D}$.

- Thus, $\mathbf{Z}^{\top}\mathbf{Z}/(n-1)$ is the sample correlation matrix.

Preprocessing V

- A different form of scaling can be introduced such that the variables are normalized to have unit length.

- One can obtain such a normalized vector $\mathbf{z}$ as

$$\mathbf{z} = \frac{1}{\sqrt{n-1}} \mathbf{\Delta}^{-1/2} \mathbf{x} \ .$$

- In a sample analogue one finds $\mathbf{Z}$ as

$$\mathbf{Z} = \frac{1}{\sqrt{n-1}} \mathbf{X} \mathbf{D}^{-1/2} \ ,$$

  in which the columns have variance equal to $1/(n-1)$.

- Now $\mathbf{Z}^\top \mathbf{Z}$ is the matrix of observed correlations.

Eigendecomposition of the sample covariance matrix

- Let $\mathbf{S_X}$ be positive semi-definite with $\text{rank}(\mathbf{S_X}) = r$ ($r \leq p$).

- The eigenvalue decomposition (or spectral decomposition) of $\mathbf{S_X}$ can be written as

$$\mathbf{S_X} = \mathbf{E\Omega E}^\top = \sum_{i=1}^{r} \omega_i \mathbf{e}_i \mathbf{e}_i^\top \quad ,$$

where $\mathbf{\Omega} = \text{diag}(\omega_1, \ldots, \omega_r)$ is an $r \times r$ diagonal matrix containing the positive eigenvalues of $\mathbf{S_X}$, $\omega_1 \geq \cdots \geq \omega_r > 0$, on its main diagonal and $\mathbf{E} \in \mathbb{R}^{p \times r}$ is a column-wise orthonormal matrix whose columns $\mathbf{e}_1, \ldots, \mathbf{e}_r$ are the corresponding unit-norm eigenvectors of $\omega_1, \ldots, \omega_r$.

The aim of principal component analysis I

- Principal component analysis (PCA) provides a computationally efficient way of projecting the p-dimensional data cloud orthogonally onto a k-dimensional subspace.

- The aim of PCA is to derive k ($\ll p$) uncorrelated linear combinations of the p-dimensional observation vectors $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}$, called the sample principal components (PCs), which retain most of the total variation present in the data.

- This is achieved by taking those k components that successively have maximum variance.

The aim of principal component analysis II

- PCA looks for r vectors $\mathbf{e}_j \in \mathbb{R}^{p\times 1}$ $(j = 1, \ldots, r)$ which

$$\begin{array}{ll} \text{maximize} & \mathbf{e}_j^\top \mathbf{S_X} \mathbf{e}_j \\ \text{subject to} & \mathbf{e}_j^\top \mathbf{e}_j = 1 \quad \text{for } j = 1, \ldots, r \quad \text{and} \\ & \mathbf{e}_i^\top \mathbf{e}_j = 0 \quad \text{for } i = 1, \ldots, j-1 \quad (j \geq 2) \ . \end{array}$$

- It turns out that $\mathbf{y}_j = \mathbf{X}\mathbf{e}_j$ is the j-th sample PC with zero mean and variance ω_j, where $\mathbf{e}_j$ is an eigenvector of $\mathbf{S_X}$ corresponding to its j-th largest eigenvalue ω_j $(j = 1, \ldots, r)$.

- The total variance of the r PCs will equal the total variance of the original variables so that $\sum_{j=1}^{r} \omega_j = \text{trace}(\mathbf{S_X})$.

Singular value decomposition of the data matrix I

- The sample PCs can also be found using the singular value decomposition (SVD) of $\mathbf{X}$.

- Expressing $\mathbf{X}$ with rank r with $r \leq \min\{n, p\}$ by its SVD gives

$$\mathbf{X} = \mathbf{VDE}^\top = \sum_{j=1}^{r} \sigma_j \mathbf{v}_j \mathbf{e}_j^\top \ ,$$

  where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_r) \in \mathbb{R}^{n \times r}$ and $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_r) \in \mathbb{R}^{p \times r}$ are orthonormal matrices such that $\mathbf{V}^\top \mathbf{V} = \mathbf{E}^\top \mathbf{E} = \mathbf{I}_r$, and $\mathbf{D} \in \mathbb{R}^{r \times r}$ is a diagonal matrix with the singular values of $\mathbf{X}$ sorted in decreasing order, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$, on its main diagonal.

Singular value decomposition of the data matrix II

- The matrix $\mathbf{E}$ is composed of coefficients or loadings and the matrix of component scores $\mathbf{Y} \in \mathbb{R}^{n \times r}$ is given by $\mathbf{Y} = \mathbf{V}\mathbf{D}$.
- Since it holds that $\mathbf{E}^\top \mathbf{E} = \mathbf{I}_r$ and $\mathbf{Y}^\top \mathbf{Y}/(n-1) = \mathbf{D}^2/(n-1)$, the loadings are orthogonal and the sample PCs are uncorrelated.
- The variance of the j-th sample PC is $\sigma_j^2/(n-1)$ which is equal to the j-th largest eigenvalue, ω_j, of $\mathbf{S}_\mathbf{X}$ $(j = 1, \ldots, r)$.

Singular value decomposition of the data matrix III

- In practice, the leading k components with $k \ll r$ usually account for a substantial proportion

$$\frac{\omega_1 + \cdots + \omega_k}{\text{trace}(\mathbf{S_X})}$$

of the total variance in the data and the sum in the SVD of $\mathbf{X}$ is therefore truncated after the first k terms.

- If so, PCA comes down to finding a matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_k) \in \mathbb{R}^{n \times k}$ of component scores of the n samples on the k components and a matrix $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_k) \in \mathbb{R}^{p \times k}$ of coefficients whose k-th column is the vector of loadings for the k-th component.

Least squares property of the SVD

- PCA can be defined as the minimization of

$$||\mathbf{X} - \mathbf{Y}\mathbf{E}^\top||_F^2 \ ,$$

where $||\mathbf{B}||_F = \sqrt{\text{trace}(\mathbf{B}^\top \mathbf{B})}$ denotes the Frobenius norm of $\mathbf{B}$.

- When variables are measured on different scales or on a common scale with widely differing ranges, the data are often standardized prior to PCA.

- The sample PCs are then obtained from an eigenvalue decomposition of the sample correlation matrix. These components are not equal to those derived from $\mathbf{S_X}$.

Choosing the number of components I

(i) Retain the first k components which explain a large proportion of the total variation, say 70-80%.

(ii) If the correlation matrix is analyzed, retain only those components with eigenvalues greater than 1 (or 0.7).

(iii) Examine a scree plot. This is a plot of the eigenvalues versus the component number. The idea is to look for an "elbow" which corresponds to the point after which the eigenvalues decrease more slowly.

(iv) Consider whether the component has a sensible and useful interpretation.

Choosing the number of components II

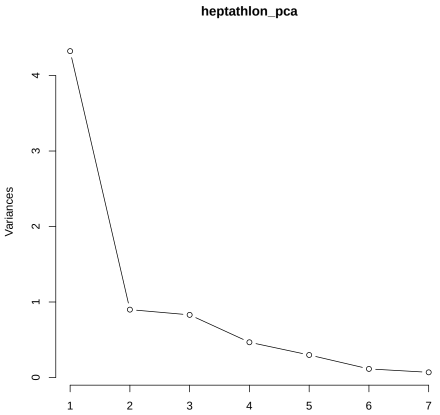


Figure: Scree diagram for the principal components of the Olympic heptathlon results.

Interpretation I

Correlations and covariances of variables and components

- The covariance of variable i with component j is given by

$$\mathrm{Cov}(x_i, y_j) = \omega_j e_{ji} \quad .$$

- The correlation of variable i with component j is therefore

$$r_{x_i,y_j} = \frac{\sqrt{\omega_j} e_{ji}}{s_i} \quad ,$$

  where s_i is the standard deviation of variable i.

- If the components are extracted from the correlation matrix, then

$$r_{x_i,y_j} = \sqrt{\omega_j} e_{ji} \quad .$$

Interpretation II

Rescaling principal components

- The coefficients $\mathbf{e}_j$ an be rescaled so that coefficients for the most important components are larger than those for less important components.
- These rescaled coefficients are calculated as

$$\mathbf{e}_j^* = \sqrt{\omega_j}\mathbf{e}_j \ ,$$

  for which ${\mathbf{e}_j^*}^\top \mathbf{e}_j^* = \omega_j$, rather than unity.
- When the correlation matrix is analyzed, this rescaling leads to coefficients that are the correlations between the components and the original variables.

Rotation I

- To enhance interpretation of the sample PCs, it is common in PCA to rotate the matrix of loadings by optimizing a certain "simplicity" criterion.
- The method of rotation emerged in Factor Analysis and was motivated both by solving the rotational indeterminacy problem and by facilitating the factors' interpretation.
- Rotation can be performed either in an orthogonal or an oblique (non-orthogonal) fashion.
- Several analytic orthogonal and oblique rotation criteria exist in the literature.

Rotation II

- To aid interpretation, all rotation criteria are designed to make the coefficients as simple as possible in some sense, with most loadings made to have values either ‘close to zero’ or ‘far from zero’, and with as few as possible of the coefficients taking intermediate values.
- After rotation, either one or both of the properties possessed by PCA, that is, orthogonality of the loadings and uncorrelatedness of the component scores, is lost.

PCA in the open-source software R

- Function `princomp()` in the **stats** package: Eigendecomposition of the covariance or correlation matrix. Alternative: use directly the function `eigen()`.

- Function `prcomp()` in the **stats** package: SVD of the (centered and possibly scaled) data matrix. Alternative: use directly the function `svd()`.

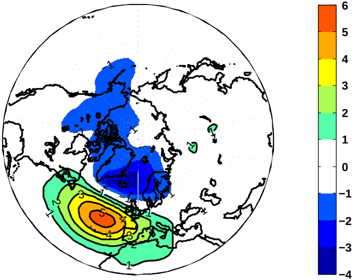
Description of the data

- For 41 cities in the United States the following seven variables were recorded:
  1. *SO2*: Sulphur dioxide content of air in micrograms per cubic meter
  2. *Temp*: Average annual temperature in degrees Fahrenheit
  3. *Manuf*: Number of manufacturing enterprises employing 20 or more workers
  4. *Pop*: Population size (1970 census) in thousands
  5. *Wind*: Average annual wind speed in miles per hour
  6. *Precip*: Average annual precipitation in inches
  7. *Days*: Average number of days with precipitation per year
- We shall examine how PCA can be used to explore various aspects of the data.
- Files: `chap3usair.dat` and `pcausair.R`

Description of the data

- Source: National Center for Environmental Prediction/National Center for Atmospheric Research.
- Winter monthly sea level pressures over the Northern Hemisphere north of 20°N.
- Gridded climate data with a 2.5°lat $\times$ 2.5°lon resolution ($p = 29 \times 144 = 4176$).
- Period: December 1948 to February 2006. Winter season is conventionally defined by December to February ($n = 174$).

Spatial patterns



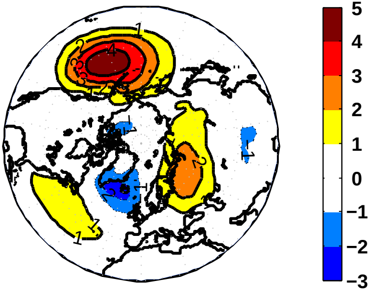


Figure: Spatial map representations of the two leading PCs for winter sea level pressure data (left: North Atlantic Oscillation, right: North Pacific Oscillation). The loadings have been multiplied by 100.