1 Principal components: concepts and calculation

We would like to explain the variance-covariance structure of a set of variables by a few linear combinations of these variables.

Let \( \vec{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \) be a random vector with population mean \( \vec{\mu} \) and population covariance matrix \( \Sigma \). Denote the spectral decomposition of \( \Sigma \) as

\[ \Sigma = \lambda_1 \vec{v}_1 \vec{v}_1^\top + \ldots + \lambda_p \vec{v}_p \vec{v}_p^\top, \]

where \( \lambda_1 \geq \ldots \geq \lambda_p > 0 \). Each eigenvector is represented as

\[ \vec{v}_k = [v_{k1}, \ldots, v_{kp}]^\top. \]

1.1 The first principal component

Consider a linear combination of the variates by \( \vec{a} = [a_1 \ldots a_p]^\top \):

\[ Y_1 = \vec{a}^\top \vec{X} = a_1 X_1 + a_2 X_2 + \ldots + a_p X_p. \]

In order to explain the variance-covariance of \( \vec{X} \) as much as possible, we want to maximize the variance of \( Y_1 \). At the same time, in order to fix the scale, we impose the constraint \( \|\vec{a}\| = 1 \). Then the first principal component for \( \vec{X} \) is defined by the following optimization:

\[
\begin{aligned}
\max & \quad \text{Var}(Y_1) = \text{Var}(\vec{a}^\top \vec{X}) = \vec{a}^\top \Sigma \vec{a} \\
\text{s.t.} & \quad \|\vec{a}\|^2 = 1
\end{aligned}
\]

The Lagrangian for the above optimization is

\[ f(\vec{a}; \lambda) = \vec{a}^\top \Sigma \vec{a} - \lambda(\vec{a}^\top \vec{a} - 1). \]

By setting the gradient to be equal to zero, we get

\[ \nabla_{\vec{a}} f(\vec{a}; \lambda) = 2\Sigma \vec{a} - 2\lambda \vec{a} = \vec{0}, \]

that is

\[ \Sigma \vec{a} = \lambda \vec{a}, \]
This implies that ⃗a must be a unit eigenvector, and λ is the corresponding eigenvalue. Notice what we aim to maximize is
\[ \text{Var}(Y_1) = \text{Var}(⃗a^\top ⃗X) = ⃗a^\top Σ⃗a = λ⃗a^\top ⃗a = λ. \]
Since λ₁ ≥ λ₂ ≥ ... ≥ λₚ > 0, in order to maximize Var(Y₁), we must have
\[
\begin{cases}
⃗a = ⃗v_1 \\
λ = λ_1.
\end{cases}
\]
To sum up, we have the following result:

**Proposition 1.** The first principal component is
\[ Y_1 = ⃗v_1^\top ⃗X = v_{11}X_1 + v_{12}X_2 + \ldots + v_{1p}X_p, \]
where ⃗v_1 is the eigenvector corresponding to the leading eigenvalue λ₁. Moreover, Var(Y₁) = λ₁.

**1.2 The second principal component**

In order to define the second principal component, we also look for some linear combination of the variates
\[ Y_2 = a^\top ⃗X = a_1X_1 + a_2X_2 + \ldots + a_pX_p, \]
such that its variance is as large as possible. However, here we have two constraints:
First, we need to impose ||⃗a||² = 1 in order to fix the scale; Second, we require that Y₂ explains the variance-covariance of ⃗X that has not been explained by Y₁, which amounts to
\[ \text{cov}(Y_2, Y_1) = 0. \]
Notice that
\[ \text{cov}(Y_2, Y_1) = \text{cov}(⃗a^\top ⃗X, ⃗v_1^\top ⃗X) = ⃗a^\top \text{Cov}(⃗X)⃗v_1 = ⃗a^\top Σ⃗v_1 = λ_1⃗a^\top ⃗v_1, \]
so the constraint cov(Y₂, Y₁) = 0 is equivalent to ⃗a^\top ⃗v_1 = 0. Consequently, the second principal component is defined through the following optimization
\[
\begin{align*}
\max & \quad \text{Var}(Y_2) = \text{Var}(⃗a^\top ⃗X) = ⃗a^\top Σ⃗a \\
\text{s.t.} & \quad ⃗a^\top ⃗a = 1, \\
& \quad ⃗a^\top ⃗v_1 = 0.
\end{align*}
\]
The resulting Lagrangian is thus
\[ f(⃗a; λ, γ) = ⃗a^\top Σ⃗a - λ(∥⃗a∥² - 1) - γ⃗a^\top ⃗v_1. \]
Again, by setting the gradient to be equal to zero, we get
\[ \nabla_⃗a f(⃗a; λ, γ) = 2Σ⃗a - 2λ⃗a - γ⃗v_1 = 0. \]
Taking the inner product of both sides with ⃗v₁, we have
\[ ⃗v_1^\top (2Σ⃗a - 2λ⃗a - γ⃗v_1) = 0. \]
The first term
\[ \vec{v}_1^\top \Sigma \vec{a} = \vec{a}^\top \Sigma \vec{v}_1 = \vec{a}^\top (\lambda_1 \vec{v}_1) = \lambda_1 \vec{a}^\top \vec{v}_1 = 0. \]
By the constraint, the second term is \( \vec{v}_1^\top \vec{a} = 0 \). As a result, \( \gamma \|\vec{v}_1\|^2 = \gamma = 0 \). Then the stationary condition is still in the form of
\[ \Sigma \vec{a} = \lambda \vec{a}. \]
This implies that \( \vec{a} \) must be a unit eigenvector while \( \lambda \) is the corresponding eigenvalue. Note that we actually want to maximize
\[ \text{Var}(Y_2) = \vec{a}^\top \Sigma \vec{a} = \lambda \|\vec{a}\|^2 = \lambda. \]
Since \( \vec{a}^\top \vec{v}_1 = 0 \), the best we can get is
\[
\begin{align*}
\vec{a} &= \vec{v}_2 \\
\text{Var}(Y_2) &= \lambda_2.
\end{align*}
\]
To sum up, we have the following result:

**Proposition 2.** The second principal component is
\[ Y_2 = \vec{v}_2^\top \vec{X} = v_{21}X_1 + v_{22}X_2 + \ldots v_{2p}X_p, \]
where \( \vec{v}_2 \) is the eigenvector corresponding to the second largest eigenvalue \( \lambda_2 \). Moreover, \( \text{Var}(Y_2) = \lambda_2 \).

### 1.3 General concepts on principal components

In general, the \( k \)-th principal component can be defined iteratively through the following procedure. Given the existing principal components \( Y_1, \ldots, Y_{k-1} \), we look for some linear combination of the variates
\[ Y_k = \vec{a}^\top \vec{X} = a_1X_1 + a_2X_2 + \ldots + a_pX_p, \]
such that its variance is as large as possible. Still, we need to impose \( \|\vec{a}\|^2 = 1 \) in order to fix the scale, and require that \( Y_k \) to explain the variance-covariance of \( \vec{X} \) that has not been explained by \( Y_1, \ldots, Y_{k-1} \), which amounts to
\[ \text{cov}(Y_k, Y_1) = \text{cov}(Y_k, Y_2) = \ldots = \text{cov}(Y_k, Y_{k-1}) = 0. \]
Consequently, the \( k \)-th principal component is defined through the following optimization
\[
\begin{align*}
\text{max} & \quad \text{Var}(Y_k) = \text{Var}(\vec{a}^\top \vec{X}) = \vec{a}^\top \Sigma \vec{a} \\
\text{s.t.} & \quad \vec{a}^\top \vec{a} = 1, \\
& \quad \vec{a}^\top \vec{v}_1 = 0, \\
& \quad \vdots \\
& \quad \vec{a}^\top \vec{v}_{k-1} = 0.
\end{align*}
\]
In general, we have the following result
Proposition 3. The $k$-th principal component is

$$Y_k = \bar{v}_k^\top X = v_{k1}X_1 + v_{k2}X_2 + \ldots v_{kp}X_p,$$

where $\bar{v}_k$ is the eigenvector corresponding to the $k$-th largest eigenvalue $\lambda_k$. Moreover, $\text{Var}(Y_k) = \lambda_k$.

The coefficients $v_{k1}, \ldots, v_{kp}$ are referred to as loadings on the random variables $X_1, \ldots, X_p$ for the $k$-th principal component $Y_k$.

1.4 Verification of covariance structures of the PCs

Denote

$$V = \begin{bmatrix} \bar{v}_1 & \ldots & \bar{v}_p \end{bmatrix} = \begin{bmatrix} v_{11} & v_{21} & \ldots & v_{p1} \\ v_{12} & v_{22} & \ldots & v_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1p} & v_{2p} & \ldots & v_{pp} \end{bmatrix},$$

where the row and column indices require attention. Recall that $\Sigma = V\Lambda V^\top$, where

$$\Lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_p \end{bmatrix}.$$

The random vector of population principal components can thus be written as

$$\bar{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix} = \begin{bmatrix} \bar{v}_1^\top \\ \vdots \\ \bar{v}_p^\top \end{bmatrix} X = \bar{v}_1^\top X = \bar{v}_2^\top X = \ldots = \bar{v}_p^\top X = V^\top \bar{X}.$$

The linear relationship $\bar{Y} = V^\top \bar{X}$ gives

$$\text{Cov}(\bar{Y}) = V^\top \text{Cov}(\bar{X})V = V^\top V\Lambda V^\top V = \Lambda,$$

which implies

$$\text{Var}(Y_k) = \lambda_k, \text{ for } k = 1, \ldots, p.$$

and

$$\text{Cov}(Y_j, Y_k) = 0 \text{ for } j \neq k.$$

1.5 Standardization

In certain applications, it is common to standardize the original variates $X_1, \ldots, X_p$ into

$$Z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}, \quad Z_2 = \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}}, \ldots, \quad Z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}.$$
Then it is straightforward to get

\[
\text{Cov}(\bar{Z}) = \begin{bmatrix}
\rho_{11} & \rho_{12} & \cdots & \rho_{1p} \\
\rho_{21} & \rho_{22} & \cdots & \rho_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{p1} & \rho_{p2} & \cdots & \rho_{pp}
\end{bmatrix} = \text{Corr}(\bar{X})
\]

where

\[
\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}
\]

If we still represent the spectral decomposition for the covariance of the standardized variables as

\[
\text{Cov}(\bar{Z}) = \lambda_1 \bar{v}_1 \bar{v}_1^\top + \cdots + \lambda_p \bar{v}_p \bar{v}_p^\top
\]

with \(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0\). The principal components of \(Z_1, \ldots, Z_p\) are

\[
Y_k = \bar{v}_k^\top \bar{Z} = v_{k1} Z_1 + \ldots v_{kp} Z_p.
\]

2 Basic Principal Component Analysis

2.1 Contribution of variables to the determination of PCs

One standard method to compare the contributions of different variables to the determination of a particular PC is through the formula:

\[
Y_k = \bar{v}_k^\top \bar{X} = v_{k1} X_1 + v_{k2} X_2 + \cdots v_{kp} X_p.
\]

In other words, we compare the contributions of \(X_1, \ldots, X_p\) to the determination of \(Y_k\) based on the loadings \(v_{k1}, \ldots, v_{kp}\).

Here we introduce the second method: Compare the contributions of \(X_1, \ldots, X_p\) to the determination of \(Y_k\) based on the correlations \(\text{Corr}(X_1, Y_k), \ldots, \text{Corr}(X_p, Y_k)\). Recall that we have \(\bar{Y} = V^\top \bar{X}\), where \(V\) is defined in (1.1). Then,

\[
\text{Cov}(\bar{Y}, \bar{X}) = \text{Cov}(V^\top \bar{X}, \bar{X}) = V^\top \text{Cov}(\bar{X}) = V^\top \Sigma = V^\top \Lambda V^\top = \Lambda V^\top
\]

The covariance between the \(k\)-th principal component and the \(j\)-th variate is

\[
\text{Cov}(Y_k, X_j) = \lambda_k v_{kj}.
\]
We further have
\[ \text{Corr}(Y_k, X_j) = \frac{\text{Cov}(Y_k, X_j)}{\sqrt{\text{Var}(Y_k)\text{Var}(X_j)}} = \frac{\lambda_k v_{kj}}{\sqrt{\lambda_k \sigma_{jj}}} = v_{kj} \sqrt{\frac{\lambda_k}{\sigma_{jj}}}. \]

This gives the second method to compare the contributions of \( X_j \)'s to the determination of \( Y_k \) through the correlation coefficients \( v_{kj} \sqrt{\frac{\lambda_k}{\sigma_{jj}}} \) for \( j = 1, \ldots, p \).

When the variables are standardized from \( X_j \) to \( Z_j \), we have and
\[ \text{Corr}(Y_k, Z_j) = v_{kj} \sqrt{\frac{\lambda_k}{\rho_{jj}}} = v_{kj} \sqrt{\frac{\lambda_k}{\sigma_{jj}}}. \]

This implies that for a fixed \( k \), the loadings and correlation coefficients between the \( k \)-th PC \( Y_k \) and \( Z_1, \ldots, Z_p \) are proportional. Therefore, there is no difference in comparing of the contributions of variables to the determination of \( Y_k \) based on either loadings or correlations.

### 2.2 Selecting the number of PCs

Recall that the spectral decomposition of the population covariance is
\[ \Sigma = \lambda_1 \vec{v}_1 \vec{v}_1^\top + \lambda_2 \vec{v}_2 \vec{v}_2^\top + \ldots + \lambda_p \vec{v}_p \vec{v}_p^\top. \]

Denote
\[ \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{bmatrix} = V \Lambda V^\top. \]

The trace formula gives
\[ \text{trace}(\Sigma) = \text{trace}(V \Lambda V^\top) = \text{trace}(\Lambda V^\top V) = \text{trace}(\Lambda), \]
which is equivalent to
\[ \sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \lambda_1 + \lambda_2 + \ldots + \lambda_p. \]

Since \( \text{Var}(Y_k) = \lambda_k \) for \( k = 1, \ldots, p \) and \( \text{Var}(X_j) = \sigma_{jj} \), the trace formula gives
\[ \text{Var}(X_1) + \ldots + \text{Var}(X_p) = \text{Var}(Y_1) + \ldots + \text{Var}(Y_p) \]

**Definition 4.** The proportion of the total variance due to the first \( k \) principal components is defined as
\[ \frac{\text{Var}(Y_1) + \ldots + \text{Var}(Y_k)}{\text{Var}(Y_1) + \ldots + \text{Var}(Y_p)} = \frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_p} = \frac{\lambda_1 + \ldots + \lambda_k}{\sigma_{11} + \ldots + \sigma_{pp}}. \]

Example: If
\[ \frac{\lambda_1 + \lambda_2 + \lambda_3}{\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp}} > 90\%, \]
then we can replace \( X_1, \ldots, X_p \) with \( Y_1, Y_2 \) and \( Y_3 \) without much loss of information.
3 Sample PCA

3.1 Summary of results

Let \( \mathbf{x}_1, \ldots, \mathbf{x}_n \) be a sample with sample mean \( \bar{\mathbf{x}} \) and sample covariance \( \mathbf{S} \). By considering the spectral decomposition of the sample covariance

\[
\mathbf{S} = \hat{\lambda}_1 \mathbf{u}_1 \mathbf{u}_1^\top + \ldots + \hat{\lambda}_p \mathbf{u}_p \mathbf{u}_p^\top,
\]

where \( \mathbf{u}_k = [u_{k1}, \ldots, u_{kp}]^\top \), we have the following results about sample PCs in the analogy to population PCs:

- The \( k \)-th sample PC is defined as
  \[
  \hat{Y}_k = u_{k1} X_1 + u_{k2} X_2 + \ldots + u_{kp} X_p.
  \]
  The coefficients \( u_{k1}, \ldots, u_{kp} \) are referred to as loadings for the \( k \)-th sample principal component \( \hat{Y}_k \). In particular, the \( i \)-th observation of the \( k \)-th sample principal component as
  \[
  \hat{y}_{ik} = \mathbf{u}_k^\top \mathbf{x}_i = u_{k1} x_{i1} + u_{k2} x_{i2} + \ldots + u_{kp} x_{ip}.
  \]
- The sample variance of \( \hat{Y}_k \) is \( \hat{\lambda}_k \), and for \( k \neq j \), the sample covariance between \( \hat{Y}_k \) and \( \hat{Y}_j \) is 0.
- The sample correlation between \( Y_k \) and \( X_j \) is
  \[
  u_{kj} \sqrt{\frac{\hat{\lambda}_k}{s_{jj}}}.
  \]
- The total sample covariances is
  \[
  s_{11} + s_{22} + \ldots + s_{pp} = \hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_p,
  \]
  and the proportion of the total variance due to the first \( k \) sample principal components:
  \[
  \frac{\hat{\lambda}_1 + \ldots + \hat{\lambda}_k}{\hat{\lambda}_1 + \ldots + \hat{\lambda}_p} = \frac{\hat{\lambda}_1 + \ldots + \hat{\lambda}_k}{s_{11} + \ldots + s_{pp}}.
  \]

3.2 Reduction of number of columns in the dataset

Consider the spectral decomposition of the sample covariance in the matrix form:

\[
\mathbf{S} = \hat{\lambda}_1 \mathbf{u}_1 \mathbf{u}_1^\top + \ldots + \hat{\lambda}_p \mathbf{u}_p \mathbf{u}_p^\top = \mathbf{U} \hat{\mathbf{\Lambda}} \mathbf{U}^\top,
\]

where

\[
\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p \geq 0,
\]

\[
\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p] = \begin{bmatrix}
  u_{11} & u_{21} & \ldots & u_{p1} \\
  u_{12} & u_{22} & \ldots & u_{p2} \\
  \vdots & \vdots & \ddots & \vdots \\
  u_{1p} & u_{2p} & \ldots & u_{pp}
\end{bmatrix},
\]
and

\[ \hat{\Lambda} = \begin{bmatrix} \hat{\lambda}_1 \\ \vdots \\ \hat{\lambda}_k \end{bmatrix} \]

Then the \( i \)-th observation of all sample principal components is

\[ \hat{\mathbf{y}}_i = \begin{bmatrix} \hat{y}_{i1} \\ \vdots \\ \hat{y}_{ip} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{y}}_1 \\ \hat{\mathbf{y}}_2 \\ \vdots \\ \hat{\mathbf{y}}_n \end{bmatrix} = \begin{bmatrix} u_{11} x_{i1} + u_{12} x_{i2} + \ldots + u_{1p} x_{ip} \\ u_{21} x_{i1} + u_{22} x_{i2} + \ldots + u_{2p} x_{ip} \\ \vdots \\ u_{p1} x_{i1} + u_{p2} x_{i2} + \ldots + u_{pp} x_{ip} \end{bmatrix} = U^\top \hat{x}_i. \]

Then, the data matrix of sample principal components is

\[ \hat{\mathbf{Y}} := \begin{bmatrix} \hat{\mathbf{y}}_1^\top \\ \hat{\mathbf{y}}_2^\top \\ \vdots \\ \hat{\mathbf{y}}_n^\top \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{x}}_1^\top U \\ \hat{\mathbf{x}}_2^\top U \\ \vdots \\ \hat{\mathbf{x}}_n^\top U \end{bmatrix} = U = XU = \begin{bmatrix} \mathbf{u}_1, \ldots, \mathbf{u}_p \end{bmatrix}. \]

In particular, if we only keep the observations of the first two sample PCs, we get

\[ \begin{bmatrix} \hat{y}_{11} & \hat{y}_{12} \\ \hat{y}_{21} & \hat{y}_{22} \\ \vdots & \vdots \\ \hat{y}_{n1} & \hat{y}_{n2} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ u_{12} & u_{22} \\ \vdots & \vdots \\ u_{1p} & u_{2p} \end{bmatrix}. \]

In practice, one is interested in plotting the PC scores of the observations for \( \hat{\mathbf{Y}}_1 \) and \( \hat{\mathbf{Y}}_2 \), i.e., the scatter plot of

\[ \begin{bmatrix} \hat{y}_{11} \\ \hat{y}_{12} \\ \vdots \\ \hat{y}_{n1} \end{bmatrix}, \begin{bmatrix} \hat{y}_{21} \\ \hat{y}_{22} \\ \vdots \\ \hat{y}_{n2} \end{bmatrix}. \]

Meanwhile, each variable \( X_j \) should be also presented in the plot as the vector of loadings

\[ \begin{bmatrix} u_{1j} \\ u_{2j} \end{bmatrix}, \]

which will be helpful for the interpretation of \( \hat{\mathbf{Y}}_1 \) and \( \hat{\mathbf{Y}}_2 \).

### 4 Data analysis and interpretation

See, e.g., Example 8.5 on page 451.