

PRINCIPAL COMPONENT ANALYSIS

1. Introduction

Principal component analysis (PCA) is concerned with explaining the variance-covariance structure of the data through a few *linear combinations* of the original variables. Its general objectives are:

- data reduction
- interpretation.

Data reduction. Although the original data set contains p variables, often much of the variability can be accounted for by a smaller number (m) of principal components. When there is (almost) as much information in the m components as there is in the original p variables, the original data set consisting of n observations on p variables can be reduced to one consisting of n observations on m principal components.

Interpretation. A PCA can show relationships that were not previously suspected, and it allows interpretations that would not ordinarily result.

2. Construction of population principal components

Let the random vector $X = [X_1, X_2, \dots, X_p]^T$ have the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$.

Consider the linear combinations

$$\begin{aligned}
 Y_1 &= l_1^T X = l_{11}X_1 + l_{21}X_2 + \dots + l_{p1}X_p \\
 Y_2 &= l_2^T X = l_{12}X_1 + l_{22}X_2 + \dots + l_{p2}X_p \\
 &\vdots \\
 Y_h &= l_h^T X = l_{1h}X_1 + l_{2h}X_2 + \dots + l_{ph}X_p \\
 &\vdots \\
 Y_p &= l_p^T X = l_{1p}X_1 + l_{2p}X_2 + \dots + l_{pp}X_p
 \end{aligned} \tag{1}$$

The variances and covariances of the linear combinations are:

$$Var(Y_h) = l_h^T \Sigma l_h \quad h = 1, 2, \dots, p \tag{2}$$

$$Cov(Y_k, Y_h) = l_h^T \Sigma l_k \quad h, k = 1, 2, \dots, p \tag{3}$$

Basic idea: The principal components are *uncorrelated* linear combinations Y_1, Y_2, \dots, Y_p whose variances in (2) are as large as possible.

The first principal component is the linear combination with maximum variance. It maximizes $Var(Y_1) = l_1^T \Sigma l_1$ under the constraint $l_1^T l_1 = 1$.

We define

- $Y_1 =$ first PC $=$ linear combination $l_1^T X$ that maximizes $Var(l_1^T X)$
 subject to $l_1^T l_1 = 1$.
 $Y_2 =$ second PC $=$ linear combination $l_2^T X$ that maximizes $Var(l_2^T X)$
 subject to $l_2^T l_2 = 1$ and $Cov(l_1^T X, l_2^T X) = 0$.
 \vdots
 $Y_h =$ h th PC $=$ linear combination $l_h^T X$ that maximizes $Var(l_h^T X)$
 subject to $l_h^T l_h = 1$ and $Cov(l_h^T X, l_k^T X) = 0$ for $k < h$.
 \vdots
 $Y_p =$ p th PC $=$ linear combination $l_p^T X$ that maximizes $Var(l_p^T X)$
 subject to $l_p^T l_p = 1$ and $Cov(l_p^T X, l_k^T X) = 0$ for $k < p$.

Result 1. Let B be a positive definite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ and associated normalized eigenvectors e_1, e_2, \dots, e_p . Then

$$\max_{x \neq 0} \frac{x^T B x}{x^T x} = \lambda_1 \quad (\text{attained when } x = e_1)$$

$$\max_{x \perp e_1, \dots, e_k} \frac{x^T B x}{x^T x} = \lambda_{k+1} \quad (\text{attained when } x = e_{k+1}, k = 1, 2, \dots, p-1)$$

Proof.

Let P be the orthogonal matrix whose columns are the eigenvectors e_1, e_2, \dots, e_p and Λ be the diagonal matrix with eigenvalues $\lambda_1, \dots, \lambda_p$ along the main diagonal. Let $B^{1/2} = P\Lambda^{1/2}P^T$ and $y = P^T x$.

Consequently, $x \neq 0$ implies $y \neq 0$. Thus,

$$\begin{aligned} \frac{x^\tau Bx}{x^\tau x} &= \frac{x^\tau B^{1/2} B^{1/2} x}{x^\tau P P^\tau x} = \frac{x^\tau P \Lambda^{1/2} P^\tau P \Lambda^{1/2} P^\tau x}{y^\tau y} = \frac{y^\tau \Lambda y}{y^\tau y} \\ &= \frac{\sum_{j=1}^p \lambda_j y_j^2}{\sum_{j=1}^p y_j^2} \leq \lambda_1 \frac{\sum_{j=1}^p y_j^2}{\sum_{j=1}^p y_j^2} = \lambda_1 \end{aligned}$$

Setting $x = e_1$ gives

$$y = P^\tau e_1 = [1, 0, \dots, 0]^\tau$$

For this choice of x , we get $\frac{y^\tau \Lambda y}{y^\tau y} = \lambda_1$, or

$$\frac{e_1^\tau B e_1}{e_1^\tau e_1} = \lambda_1$$

A similar argument produces the second part of the result.

Now, $x = Py = y_1 e_1 + y_2 e_2 + \dots + y_p e_p$, so $x \perp e_1, \dots, e_k$ implies

$$0 = e_j^\tau x = y_1 e_j^\tau e_1 + y_2 e_j^\tau e_2 + \dots + y_p e_j^\tau e_p = y_j, \quad j \leq k$$

Therefore, for x perpendicular to the first k eigenvectors e_j , the left-hand side of the inequality becomes

$$\frac{x^\tau Bx}{x^\tau x} = \frac{\sum_{j=k+1}^p \lambda_j y_j^2}{\sum_{j=k+1}^p y_j^2}$$

Taking $y_{k+1} = 1, y_{k+2} = \dots = y_p = 0$ gives the asserted maximum.

□

Result 2. Let Σ be the covariance matrix associated with the random vector $X = [X_1, X_2, \dots, X_p]^T$. Let Σ have the eigenvalue-eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ and $\{e_1, e_2, \dots, e_p\}$ is orthonormal. (If some λ_h are equal, the choices of the corresponding coefficient vectors e_h and Y_h are not unique.) Denote the coordinates of e_h as $e_h = [e_{1h}, e_{2h}, \dots, e_{ph}]^T$. The h th *principal component* is then given by

$$Y_h = e_h^T X = e_{1h}X_1 + e_{2h}X_2 + \dots + e_{ph}X_p \quad h = 1, 2, \dots, p \quad (4)$$

i.e. put $l_1 = e_1, \dots, l_p = e_p$. With these choices,

$$\begin{aligned} \text{Var}(Y_h) &= e_h^T \Sigma e_h = \lambda_h & h = 1, 2, \dots, p \\ \text{Cov}(Y_k, Y_h) &= e_h^T \Sigma e_k = 0 & h \neq k \end{aligned} \quad (5)$$

Proof. From result 1 it follows that

$$\begin{aligned} \max_{l \neq 0} \frac{l^T \Sigma l}{l^T l} &= \lambda_1 = \frac{e_1^T \Sigma e_1}{e_1^T e_1} = e_1^T \Sigma e_1 = \max_{l^T l=1} l^T \Sigma l = \text{Var}(Y_1) \\ \max_{l \perp e_1, e_2, \dots, e_k} \frac{l^T \Sigma l}{l^T l} &= \lambda_{k+1} = \frac{e_{k+1}^T \Sigma e_{k+1}}{e_{k+1}^T e_{k+1}} = e_{k+1}^T \Sigma e_{k+1} \\ &= \max_{\substack{l \perp e_1, e_2, \dots, e_k \\ l^T l=1}} l^T \Sigma l = \text{Var}(Y_{k+1}) \end{aligned}$$

For any two eigenvectors e_h and e_k with $h \neq k$ we have $e_h^T e_k = 0$.

So we conclude

$$\text{Cov}(Y_h, Y_k) = e_h^T \Sigma e_k = e_h^T \lambda_k e_k = 0$$

for any $h \neq k$. □

Result 2 has some important corollaries. Note that the total variance of a distribution is defined as the trace of the covariance matrix, $tr(\Sigma)$.

Result 3. The total variance equals

$$\begin{aligned} tr(\Sigma) &= \sigma_1^2 + \sigma_2^2 + \dots + \sigma_p^2 = \sum_{j=1}^p Var(X_j) \\ &= \lambda_1 + \lambda_2 + \dots + \lambda_p = \sum_{h=1}^p Var(Y_h) \end{aligned}$$

Proof.

The spectral decomposition of Σ is $\Sigma = \sum_{j=1}^p \lambda_j e_j e_j^\tau = P \Lambda P^\tau$ where Λ is the diagonal matrix of eigenvalues and $P = [e_1, \dots, e_p]$ so $P P^\tau = \mathbf{I} = P^\tau P$. Therefore

$$tr(\Sigma) = tr(P \Lambda P^\tau) = tr(\Lambda P P^\tau) = tr(\Lambda) = \lambda_1 + \lambda_2 + \dots + \lambda_p$$

and the proof is complete. □

The proportion of the total variance due to the h th principal component is therefore equal to:

$$\frac{\lambda_h}{\lambda_1 + \lambda_2 + \dots + \lambda_p} \quad h = 1, 2, \dots, p$$

If most of the total population variance can be attributed to the first one, two or three components, then these components can “replace” the original p variables without much loss of information. There are several criteria to select this number of principal components, e.g.

1. 80% or 90% of the total variance
2. $\lambda_j > 0.7 \text{ave}_{j=1}^p \lambda_j$
3. plot λ_j versus index j

Consider the coefficient vector $e_h = [e_{1h}, \dots, e_{jh}, \dots, e_{ph}]^T$. The magnitude of e_{jh} measures the importance of the j th variable to the h th principal component, irrespective of the other variables. In particular, e_{jh} is proportional to the correlation coefficient between X_j and Y_h .

Result 4. If $Y_1 = e_1^\tau X, Y_2 = e_2^\tau X, \dots, Y_p = e_p^\tau X$ are the principal components obtained from the covariance matrix Σ , then

$$Cor(X_j, Y_h) = \frac{e_{jh}\sqrt{\lambda_h}}{\sigma_j} \quad j, h = 1, 2, \dots, p$$

are the correlation coefficients between the variables X_j and the components Y_h .

Proof.

Set $l_j = [0, \dots, 0, 1, 0, \dots, 0]^\tau$ then

$$X_j = l_j^\tau X$$

$$Cov(X_j, Y_h) = Cov(l_j^\tau X, e_h^\tau X) = l_j^\tau \Sigma e_h$$

Since

$$\Sigma e_h = \lambda_h e_h$$

it follows that

$$Cov(X_j, Y_h) = l_j^\tau \lambda_h e_h = \lambda_h e_{jh}.$$

Then since

$$Var(Y_h) = \lambda_h \quad \text{and} \quad Var(X_j) = \sigma_j^2$$

it follows that

$$Cor(X_j, Y_h) = \frac{Cov(X_j, Y_h)}{\sqrt{Var(Y_h)}\sqrt{Var(X_j)}} = \frac{e_{jh}\sqrt{\lambda_h}}{\sigma_j} \quad j, k = 1, 2, \dots, p$$

□

Example 1. Suppose the random variables X_1, X_2 , and X_3 have the covariance matrix

$$\begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

It may be verified that the eigenvalue-eigenvector pairs are

$$\begin{aligned} \lambda_1 &= 5.83, & e_1^T &= [0.383, -0.924, 0] \\ \lambda_2 &= 2.00, & e_2^T &= [0, 0, 1] \\ \lambda_3 &= 0.17, & e_3^T &= [0.924, 0.383, 0] \end{aligned}$$

The principal components become

$$\begin{aligned} Y_1 &= e_1^T X = 0.383X_1 - 0.924X_2 \\ Y_2 &= e_2^T X = X_3 \\ Y_3 &= e_3^T X = 0.924X_1 + 0.383X_2 \end{aligned}$$

The variance of the first principal component and the covariance between the first and the second component are

$$\begin{aligned} \text{Var}(Y_1) &= \text{Var}(0.383X_1 - 0.924X_2) \\ &= 5.83 \\ &= \lambda_1 \\ \text{Cov}(Y_1, Y_2) &= \text{Cov}(0.383X_1 - 0.924X_2, X_3) \\ &= 0 \end{aligned}$$

It is clear that

$$\begin{aligned} \sigma_1^2 + \sigma_2^2 + \sigma_3^2 &= 1 + 5 + 2 \\ &= \lambda_1 + \lambda_2 + \lambda_3 \\ &= 5.83 + 2.00 + 0.17 \end{aligned}$$

The first two components account for a proportion $(5.83 + 2)/8 = 0.98$ of the population variance. In this case the components Y_1 and Y_2 could replace the three original variables with little loss of information.

Finally, using Result 4, we have that

$$\text{Cor}(X_1, Y_1) = 0.925$$

$$\text{Cor}(X_2, Y_1) = -0.998$$

$$\text{Cor}(X_1, Y_2) = 0$$

$$\text{Cor}(X_2, Y_2) = 0$$

$$\text{Cor}(X_3, Y_2) = 1$$

We conclude that X_1 and X_2 are about equally important to the first principal component.

3. Geometrical interpretation

Geometrically, these linear combinations represent the selection of a new coordinate system obtained by orthogonally transforming the original system, with e_1, e_2, \dots, e_p as the new coordinate axes. The new axes represent the directions with maximum variability.

Result 5. Consider the p dimensional ellipsoid $X^T \Sigma^{-1} X = c^2$.
The principal components define the axes of the ellipsoid.

Proof. We already know that

- If Σ is positive definite then Σ^{-1} exists, and

$$\Sigma e = \lambda e \quad \text{implies} \quad \Sigma^{-1} e = (1/\lambda) e$$

and Σ^{-1} is positive definite as well.

- The spectral decomposition of Σ^{-1} is $\Sigma^{-1} = \frac{1}{\lambda_1} e_1 e_1^T + \dots + \frac{1}{\lambda_p} e_p e_p^T$

With the use of the spectral decomposition of Σ^{-1} we get

$$c^2 = X^T \Sigma^{-1} X = \frac{1}{\lambda_1} (e_1^T X)^2 + \frac{1}{\lambda_2} (e_2^T X)^2 + \dots + \frac{1}{\lambda_p} (e_p^T X)^2$$

where $e_1^T X, e_2^T X, \dots, e_p^T X$ are the principal components of X .

Setting

$$Y_1 = e_1^T X, Y_2 = e_2^T X, \dots, Y_p = e_p^T X$$

we have

$$c^2 = \frac{1}{\lambda_1} Y_1^2 + \frac{1}{\lambda_2} Y_2^2 + \dots + \frac{1}{\lambda_p} Y_p^2$$

This equation defines an ellipsoid in a coordinate system with axes lying in the directions of e_1, e_2, \dots, e_p respectively, and with half-length $c\sqrt{\lambda_h}$ in the direction e_h . \square

Result 6. Suppose X is distributed as $N_p(\mu, \Sigma)$. Then the principal components are the axes of the ellipsoids with constant density.

Proof. The density function of X is

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^\tau \Sigma^{-1}(x-\mu)}$$

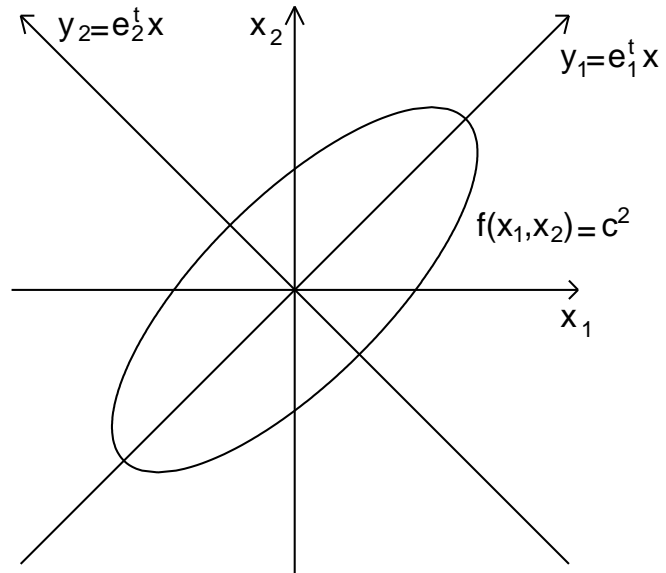
The density of X is constant if

$$\frac{1}{2}(x - \mu)^\tau \Sigma^{-1}(x - \mu) = c^2$$

We can set $\mu = 0$. This can be done because the normal random vector X can always be translated to the normal random vector $W = X - \mu$ with $E(W) = 0$ and $Cov(X) = Cov(W)$.

With the help of the previous result, the proof is complete. \square

Example: $p = 2$, $\mu = 0$, and $Cor(X_1, X_2) = 0.75$.



4. Principal components obtained from standardized variables

Consider the standardized variables

$$\begin{aligned} Z_1 &= \frac{(X_1 - \mu_1)}{\sigma_1} \\ Z_2 &= \frac{(X_2 - \mu_2)}{\sigma_2} \\ &\vdots \\ Z_p &= \frac{(X_p - \mu_p)}{\sigma_p} \end{aligned}$$

In matrix notation,

$$Z = (V^{\frac{1}{2}})^{-1}(X - \mu)$$

with the diagonal standard deviation matrix

$$V^{\frac{1}{2}} = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p \end{bmatrix}$$

We conclude that

$$E(Z) = 0 \quad \text{and} \quad Cov(Z) = (V^{\frac{1}{2}})^{-1}\Sigma(V^{\frac{1}{2}})^{-1} = Cor(X) = \rho$$

The principal components of Z can be obtained from the eigenvectors of the correlation matrix ρ of X .

Result 7. The h th principal component of the standardized variables $Z = [Z_1, Z_2, \dots, Z_p]^T$, with $Cov(Z) = \rho$, and $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ the eigenvalue-eigenvector pairs for ρ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, is given by

$$Y_h = e_h^T Z = e_h^T (V^{\frac{1}{2}})^{-1} (X - \mu), \quad h = 1, 2, \dots, p$$

Moreover,

$$\sum_{h=1}^p Var(Y_h) = \sum_{j=1}^p Var(Z_j) = p$$

and

$$Cor(Z_j, Y_h) = e_{jh} \sqrt{\lambda_h}, \quad j, h = 1, 2, \dots, p$$

Proof.

This follows immediately from results 2, 3, and 4. □

The proportion of total variance explained by the h th principal component of Z is therefore equal to

$$\frac{\lambda_h}{p} \quad h = 1, 2, \dots, p.$$

Example 2. Consider the covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$$

and the resulting correlation matrix

$$\rho = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$$

The eigenvalue-eigenvector pairs from Σ are

$$\begin{aligned}\lambda_1 &= 100.16, & e_1 &= [0.040, 0.999]^\tau \\ \lambda_2 &= 0.84, & e_2 &= [0.999, -0.4]^\tau\end{aligned}$$

The eigenvalue-eigenvector pairs from ρ are

$$\begin{aligned}\lambda_1 &= 1.4, & e_1 &= [0.707, 0.707]^\tau \\ \lambda_2 &= 0.6, & e_2 &= [0.707, -0.707]^\tau\end{aligned}$$

The principal components are

$$\Sigma : \left\{ \begin{array}{l} Y_1 = 0.040X_1 + 0.999X_2 \\ Y_2 = 0.999X_1 - 0.040X_2 \end{array} \right\}$$

and

$$\rho : \left\{ \begin{array}{l} Y_1 = 0.707Z_1 + 0.707Z_2 = 0.707(X_1 - \mu_1) + 0.0707(X_2 - \mu_2) \\ Y_2 = 0.707Z_1 - 0.707Z_2 = 0.707(X_1 - \mu_1) - 0.0707(X_2 - \mu_2) \end{array} \right\}$$

We see that X_2 completely dominates the first principal component of Σ .

This first principal component explains a proportion

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{100.16}{101} = 0.992$$

of the total population variance.

In contrast, the variables Z_1 and Z_2 contribute equally to the principal components of ρ .

$$\begin{aligned}Cor(Z_1, Y_1) &= e_{11}\sqrt{\lambda_1} = 0.707\sqrt{1.4} = 0.837 \\ Cor(Z_2, Y_1) &= e_{21}\sqrt{\lambda_1} = 0.707\sqrt{1.4} = 0.837\end{aligned}$$

In this case, the first principal component explains a proportion

$$\frac{\lambda_1}{p} = 0.7$$

of the total standardized population variance.

We conclude that the relative importance of the variables is affected by the standardization.

Conclusion. The principal components of Σ differ from those of ρ . The variables are often standardized when they have different units or widely different scales.

5. PCA for covariance matrices with special structures

5.1. Diagonal Σ

Suppose that

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p^2 \end{bmatrix}$$

Setting $e_h = [0, \dots, 0, 1, 0, \dots, 0]^\tau$, with 1 in the h th position, we observe that

$$\Sigma e_h = \sigma_h^2 e_h$$

and we conclude that (σ_h^2, e_h) is an eigenvalue-eigenvector pair. The set of principal components is just the original set of uncorrelated random variables, but possibly in a different order (i.e., ranked by decreasing σ_j^2).

Conclusion. It is not necessary to use principal components when the covariance matrix is diagonal.

If X is distributed as $N_p(\mu, \Sigma)$ with a diagonal matrix Σ , the contours of constant density are ellipsoids whose axes are parallel to the original coordinate axes. When we standardize the variables we find

$$\rho = I_p \quad \text{and} \quad \rho e_h = 1 e_h$$

hence the eigenvalue 1 has multiplicity p and $e_h = [0, \dots, 0, 1, 0, \dots, 0]$, $h = 1, 2, \dots, p$ are the eigenvectors. The multivariate normal ellipsoids of constant density then become spheres.

5.2. Equicorrelated Σ

Let Σ be of the form

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \dots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \dots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \dots & \sigma^2 \end{bmatrix}$$

The resulting correlation matrix

$$\rho = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}$$

implies that the variables X_1, X_2, \dots, X_p are equally correlated. The p eigenvalues of ρ are then

$$\lambda_1 = 1 + (p - 1)\rho$$

$$\lambda_2 = \lambda_3 = \dots = \lambda_p = 1 - \rho$$

with associated eigenvectors

$$e_1^\tau = \left[\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \dots, \frac{1}{\sqrt{p}} \right]$$

$$e_2^\tau = \left[\frac{1}{\sqrt{1*2}}, \frac{-1}{\sqrt{1*2}}, 0, \dots, 0 \right]$$

$$e_3^\tau = \left[\frac{1}{\sqrt{2*3}}, \frac{1}{\sqrt{2*3}}, \frac{-2}{\sqrt{2*3}}, 0, \dots, 0 \right]$$

$$e_i^\tau = \left[\frac{1}{\sqrt{(i-1)*i}}, \dots, \frac{1}{\sqrt{(i-1)*i}}, \frac{-(i-1)}{\sqrt{(i-1)*i}}, 0, \dots, 0 \right]$$

$$e_p^\tau = \left[\frac{1}{\sqrt{(p-1)*p}}, \dots, \frac{1}{\sqrt{(p-1)*p}}, \frac{-(p-1)}{\sqrt{(p-1)*p}} \right]$$

(which correspond to Helmert's orthogonal transformation).

The first principal component

$$Y_1 = e_1^\tau X = \frac{1}{\sqrt{p}} \sum_{j=1}^p X_j$$

is then proportional to the sum of the p original variables.

This principal component explains a proportion

$$\frac{\lambda_1}{p} = \rho + \frac{1 - \rho}{p}$$

of the total population variability. (We see that $\frac{\lambda_1}{p} \approx \rho$ for ρ close to 1 or p large.)

If the standardized variables Z_1, Z_2, \dots, Z_p have a multivariate normal distribution with an equicorrelated covariance matrix, the ellipsoids of constant density are “cigar-shaped” with the major axis proportional to the first principal component $Y_1 = \frac{1}{\sqrt{p}}[1, 1, \dots, 1]X$, that is, parallel to the main diagonal. The minor axes occur in spherically symmetric directions perpendicular to the major axis.

Remark. The relation with Helmert’s transformation stems from the following. Let X_1, \dots, X_n be i.i.d. according to $N_1(0, 1)$ and put $V = [X_1 - \bar{X}, \dots, X_n - \bar{X}]$. Then we have for each $i = 1, \dots, n$ that

$$\begin{aligned} \text{var}[V_i] &= E[(X_i - \bar{X})(X_i - \bar{X})] \\ &= E[X_i^2] - 2E[X_i \bar{X}] + E[\bar{X}^2] \\ &= 1 - \frac{2}{n}E[X_i(X_1 + \dots + X_n)] + \frac{1}{n} \\ &= 1 - \frac{2}{n} + \frac{1}{n} = 1 - \frac{1}{n} \end{aligned}$$

because $\bar{X} \sim N_1(0, \frac{1}{n})$.

For any $i \neq h$ we find

$$\begin{aligned} \text{Cov}(V_i, V_h) &= E[(X_i - \bar{X})(X_h - \bar{X})] \\ &= E[X_i X_j] - E[X_i \bar{X}] - E[X_h \bar{X}] + E[\bar{X}^2] \\ &= 0 - \frac{1}{n} - \frac{1}{n} + \frac{1}{n} = -\frac{1}{n} \end{aligned}$$

Therefore the n by n covariance matrix $\Sigma = \text{Cov}[V]$ is equicorrelated, with

$$\begin{aligned} \sigma^2 &= 1 - \frac{1}{n} > 0 \\ \rho &= -\frac{1}{n} / (1 - \frac{1}{n}) = -1/(n - 1) < 0 \end{aligned}$$

The total variance is thus $n - 1$ instead of n .

6. Sample principal components

Let $x_i = [x_{i1}, \dots, x_{ij}, \dots, x_{ip}]^\tau$. Assume the data x_1, x_2, \dots, x_n represent n independent observations from some elliptic p -dimensional population with mean vector μ and covariance matrix Σ . These data yield the sample mean vector \bar{x} , the sample covariance matrix S , and the sample correlation matrix R .

We know that the n values of any linear combination

$$l_1^\tau x_i = l_{11}x_{i1} + l_{21}x_{i2} + \dots + l_{p1}x_{ip} \quad i = 1, 2, \dots, n$$

have sample mean $l_1^\tau \bar{x}$ and sample variance $l_1^\tau S l_1$. The pairs of values $(l_1^\tau x_i, l_2^\tau x_i)$ have sample covariance $l_1^\tau S l_2$.

We obtain the following results concerning sample principal components.

Result 8. If S is the $p \times p$ sample covariance matrix with eigenvalue-eigenvector pairs $(\hat{\lambda}_1, \hat{e}_1), (\hat{\lambda}_2, \hat{e}_2), \dots, (\hat{\lambda}_p, \hat{e}_p)$, the h th sample principal component is given by

$$\hat{y}_h = \hat{e}_h^\tau x = \hat{e}_{1h}x_1 + \hat{e}_{2h}x_2 + \dots + \hat{e}_{ph}x_p \quad h = 1, 2, \dots, p$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p \geq 0$ and x is any observation on the variables X_1, X_2, \dots, X_p . Also

$$\begin{aligned} \text{var}_{i=1}^n(\hat{y}_{ih}) &= \hat{\lambda}_h & \text{for } h = 1, 2, \dots, p \\ \text{cov}_{i=1}^n(\hat{y}_{ih}, \hat{y}_{ik}) &= 0 & \text{for } h \neq k \end{aligned}$$

Result 9.

$$\text{Total sample variance} = \text{tr}(S) = \sum_{j=1}^p s_{jj} = \hat{\lambda}_1 + \hat{\lambda}_2 + \dots + \hat{\lambda}_p = \sum_{h=1}^p \hat{\lambda}_h$$

and

$$r_{x_j, \hat{y}_h} = \frac{\hat{e}_{jh} \sqrt{\hat{\lambda}_h}}{\sqrt{s_{jj}}} \quad j, h = 1, 2, \dots, p$$

The sample principal components of the standardized observations are given by Result 8, with the matrix R instead of S . Again the principal components of S and R are not the same.

Comment. The observations x_i are often “centered” by subtracting \bar{x} . This has no effect on the sample covariance matrix S and gives the h th principal component

$$\hat{y}_h = \hat{e}_h^\tau (x - \bar{x}) \quad \text{for } h = 1, 2, \dots, p$$

for any observation vector x . If we consider the values of the h th component

$$\hat{y}_{ih} = \hat{e}_h^\tau (x_i - \bar{x}) \quad \text{for } h = 1, 2, \dots, p$$

then

$$\text{ave}_{i=1}^n(\hat{y}_{ih}) = \bar{\hat{y}}_h = \frac{1}{n} \sum_{i=1}^n \hat{e}_h^\tau (x_i - \bar{x}) = \frac{1}{n} \hat{e}_h^\tau \left(\sum_{i=1}^n (x_i - \bar{x}) \right) = 0$$

The sample mean of each principal component is thus zero.

Example 3. We have a study of size and shape relationships for turtles which measures carapace length, width and height. In studies of size-and-shape relationships, one often uses a logarithmic transformation. The natural logarithms of the measurements of 24 male turtles have sample mean vector

$\bar{x}^T = [4.725, 4.478, 3.703]$ and covariance matrix

$$S = 10^{-3} \begin{bmatrix} 11.072 & 8.019 & 8.160 \\ 8.019 & 6.417 & 6.005 \\ 8.160 & 6.005 & 6.773 \end{bmatrix}$$

A principal component analysis in S-PLUS yields the following result.

Standard deviations:

Comp. 1	Comp. 2	Comp. 3
0.1494402	0.02394526	0.01856994

The loadings are:

	Comp. 1	Comp. 2	Comp. 3
V1	0.683	-0.159	0.713
V2	0.510	-0.594	-0.622
V3	0.523	0.788	-0.324

The Cumulative percentage of total variance is:

Comp. 1	Comp. 2	Comp. 3
0.9605077	0.9851684	1

The first principal component has an interesting subject-matter interpretation. Since

$$\begin{aligned} \hat{y}_1 &= 0.683 \ln(\text{length}) + 0.510 \ln(\text{width}) + 0.523 \ln(\text{height}) \\ &= \ln[(\text{length})^{0.683}(\text{width})^{0.510}(\text{height})^{0.523}] \end{aligned}$$

the first PC may be viewed as a multiple of $\ln(\text{volume})$.

Example 4. The weekly rates of return for five stocks listed on the New York Stock Exchange were determined for the period January 1975 through December 1976. Let x_1, x_2, \dots, x_5 denote observed weekly rates of return for the five stocks. Then

$$\bar{x} = [0.0054, 0.0048, 0.0057, 0.0063, 0.0037]^T$$

and

$$R = \begin{bmatrix} 1.000 & 0.577 & 0.509 & 0.387 & 0.462 \\ 0.577 & 1.000 & 0.599 & 0.389 & 0.322 \\ 0.509 & 0.599 & 1.000 & 0.436 & 0.426 \\ 0.387 & 0.389 & 0.436 & 1.000 & 0.523 \\ 0.462 & 0.322 & 0.426 & 0.523 & 1.000 \end{bmatrix}$$

The eigenvalues and corresponding normalized eigenvectors of R are:

$$\begin{aligned} \hat{\lambda}_1 &= 2.857 & \hat{e}_1 &= [0.464, 0.457, 0.470, 0.421, 0.421]^T \\ \hat{\lambda}_2 &= 0.809 & \hat{e}_2 &= [0.240, 0.509, 0.260, -0.526, -0.582]^T \\ \hat{\lambda}_3 &= 0.540 & \hat{e}_3 &= [-0.612, 0.178, 0.335, 0.541, -0.435]^T \\ \hat{\lambda}_4 &= 0.452 & \hat{e}_4 &= [0.387, 0.206, -0.662, 0.472, -0.382]^T \\ \hat{\lambda}_5 &= 0.343 & \hat{e}_5 &= [-0.451, 0.676, -0.400, -0.176, 0.385]^T \end{aligned}$$

Using the standardized variables, we obtain the first two sample principal components

$$\begin{aligned} \hat{y}_1 &= \hat{e}_1^T \mathbf{z} = 0.464z_1 + 0.457z_2 + 0.470z_3 + 0.421z_4 + 0.421z_5 \\ \hat{y}_2 &= \hat{e}_2^T \mathbf{z} = 0.240z_1 + 0.509z_2 + 0.260z_3 - 0.526z_4 - 0.582z_5 \end{aligned}$$

These components account for

$$\left(\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p} \right) \times 100\% = 73\%$$

of the total sample variance. The first component is an equally weighted sum, or “index”, of the five stocks. This component might be called a market

component. The second component represents a contrast between the first three stocks (which were chemical stocks) and the last two stocks (oil stocks). It might be called an industry component.

Example 5. The body weight (in grams) for $n = 150$ female mice were obtained immediately after birth of their first four litters. The sample mean vector and sample correlation matrix were

$$\bar{x} = [39.88, 45.08, 48.11, 49.95]^T$$

$$R = \begin{bmatrix} 1.000 & 0.7501 & 0.6329 & 0.6363 \\ 0.7501 & 1.000 & 0.6925 & 0.7386 \\ 0.6329 & 0.6925 & 1.000 & 0.6625 \\ 0.6363 & 0.7386 & 0.6625 & 1.000 \end{bmatrix}$$

The eigenvalues of this matrix are

$$\hat{\lambda}_1 = 3.058, \quad \hat{\lambda}_2 = 0.382, \quad \hat{\lambda}_3 = 0.342, \quad \text{and} \quad \hat{\lambda}_4 = 0.217$$

and $e_1 = [0.493, 0.522, 0.487, 0.497]^T$.

The first eigenvalue is nearly equal to $1 + (p - 1)\bar{r} = 3.056$. The remaining eigenvalues are small and about equal. Thus there is some evidence that the corresponding population correlation matrix ρ may be of the “equal-correlation” form. This will be explored further in Example 7.

The first principal component accounts for $100(\hat{\lambda}_1/p)\% = 76\%$ of the total variance. The average post-birth weights increase over time. The variation in weights is fairly well explained by the first principal component with nearly equal coefficients.

Comment. An unusually small value for the last eigenvalue can indicate an unnoticed linear dependency in the data set.

Consider a situation where x_1 , x_2 , and x_3 are subtest scores and the total score x_4 is the sum $x_1 + x_2 + x_3$. Although the linear combination

$$[1, 1, 1, -1]x = x_1 + x_2 + x_3 - x_4$$

is always zero, rounding error in the computation of eigenvalues may lead to a small nonzero value.

Thus eigenvalues very close to zero are important!

7. Inference

The eigenvalues and eigenvectors of the covariance (correlation) matrix are the essence of a principal component analysis. The total variance can be “explained” with fewer than p dimensions when the first few eigenvalues are much larger than the rest.

In practice, the quality of the principal component approximation is defined on the basis of the eigenvalue-eigenvector pairs $(\hat{\lambda}_h, \hat{e}_h)$ extracted from S or R . Because of sampling variation, these eigenvalues and eigenvectors will differ from the population matrices Σ and ρ . Best to have large n !

7.1. Inference about eigenvalues

We assume that the observations X_1, X_2, \dots, X_n are a random sample from a normal population and that the (unknown) eigenvalues of Σ are **distinct** and positive, so that $\lambda_1 > \lambda_2 > \dots > \lambda_p > 0$.

Result 10.

Put $\hat{\lambda} = [\hat{\lambda}_1, \dots, \hat{\lambda}_p]^\tau$ and let Λ be the diagonal matrix of eigenvalues $\lambda_1, \dots, \lambda_p$ of Σ . Then

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow[n \rightarrow \infty]{\mathcal{D}} N_p(0, 2\Lambda^2)$$

This result implies that, for large n , the $\hat{\lambda}_h$ are independent and they have an approximate $N(\lambda_h, 2\lambda_h^2/n)$ distribution. Using the normal distribution, $P[|\hat{\lambda}_h - \lambda_h| \leq \Phi^{-1}(1 - \frac{\alpha}{2})\lambda_h\sqrt{(2/n)}] = 1 - \alpha$. A large-sample $100(1 - \alpha)\%$

confidence interval for λ_h is thus provided by

$$\frac{\hat{\lambda}_h}{1 + \Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{2/n}} \leq \lambda_h \leq \frac{\hat{\lambda}_h}{1 - \Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{2/n}} \quad (7.1)$$

Example 6. We construct a 95% confidence interval for λ_1 (the variance of the first population principal component) for the stock price data of Example 4. Assume the stock rates of return represent independent drawings from an $N_5(\mu, \Sigma)$ population, where Σ is positive definite with distinct eigenvalues $\lambda_1 > \lambda_2 > \dots > \lambda_5 > 0$. Since $n = 100$ is large, we use (7.1) with $h = 1$ to construct a 95% confidence interval for λ_1 .

We can calculate that

$$\hat{\lambda}_1 = 0.0036 \quad \Phi^{-1}(0.975) = 1.96$$

Therefore, with 95% confidence,

$$0.0028 \leq \lambda_1 \leq 0.005$$

7.2. Testing for Equicorrelation

When $\text{Cor}(X_j, X_k) = \rho$ for all $j \neq k$ the eigenvalues of Σ are not all distinct, hence the previous results do not apply. Let

$$H_0 : \rho \in \left\{ \rho_0 = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \dots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix} ; -1 \leq \rho \leq 1 \right\}$$

and

$$H_1 : \rho \text{ is not of this form.}$$

The test procedure assumes that n is large, and requires the quantities

$$\bar{r}_k = \frac{1}{p-1} \sum_{\substack{j=1 \\ j \neq k}}^p r_{jk} \quad k = 1, 2, \dots, p$$

$$\bar{r} = \frac{2}{p(p-1)} \sum_{j < k} r_{jk}$$

$$\hat{\gamma} = \frac{(p-1)^2 [1 - (1 - \bar{r})^2]}{p - (p-2)(1 - \bar{r})^2}$$

where \bar{r}_k is the average of the off-diagonal elements in the k th column of R , and \bar{r} is the overall average of the off-diagonal elements.

The test has the form: Reject H_0 in favor of H_1 iff

$$T = \frac{(n-1)}{(1-\bar{r})^2} \left[\sum_{j < k} (r_{jk} - \bar{r})^2 - \hat{\gamma} \sum_{k=1}^p (\bar{r}_k - \bar{r})^2 \right] > \chi_{(p+1)(p-2)/2, 1-\alpha}^2$$

Example 7. We shall use the correlation matrix of Example 5 about mice ($n=150$) to illustrate this test. Here $p = 4$, so

$$H_0 : \rho \in \left\{ \rho_0 = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix} ; -1 \leq \rho \leq 1 \right\}$$

$$H_1 : \rho \notin H_0$$

We obtain that

$$\begin{aligned} \bar{r}_1 &= 0.6731 & \bar{r}_2 &= 0.7271 \\ \bar{r}_3 &= 0.6626 & \bar{r}_4 &= 0.6791 \\ \bar{r} &= 0.6855 \end{aligned}$$

and

$$\begin{aligned} \sum_{j < k} \sum (r_{jk} - \bar{r})^2 &= 0.01277 \\ \sum_{k=1}^4 (\bar{r}_k - \bar{r})^2 &= 0.00245 \\ \hat{\gamma} &= 2.1329 \end{aligned}$$

$$T = 11.4$$

Since $(p + 1)(p - 2)/2 = 5$, the 5% critical value for the test of equicorrelation is $\chi_{5,0.95}^2 = 11.07$. The value of our test statistic is approximately equal to the large-sample 5% critical point. The evidence against H_0 is thus not overwhelming.

We saw that the smallest eigenvalues $\hat{\lambda}_2, \hat{\lambda}_3$, and $\hat{\lambda}_4$ are slightly different. With the large sample size in this problem ($n=150$), small differences from the equicorrelation structure show up as statistically significant.

8. Graphing the principal components

Plots of the principal components can

- reveal suspect observations
- provide checks on the assumption of normality

The last principal components can help to detect suspect observations. Each observation x_i can be expressed as a linear combination

$$\begin{aligned} x_i &= (x_i^\top \hat{e}_1) \hat{e}_1 + (x_i^\top \hat{e}_2) \hat{e}_2 + \dots + (x_i^\top \hat{e}_p) \hat{e}_p \\ &= \hat{y}_{i1} \hat{e}_1 + \hat{y}_{i2} \hat{e}_2 + \dots + \hat{y}_{ip} \hat{e}_p \end{aligned}$$

of the complete set of eigenvectors $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_p$ of S . The magnitudes of the last principal components determine how well the first few components fit the observations, because

$$\hat{y}_{i1} \hat{e}_1 + \hat{y}_{i2} \hat{e}_2 + \dots + \hat{y}_{i,q-1} \hat{e}_{q-1}$$

differs from x_i by

$$\hat{y}_{iq} \hat{e}_q + \dots + \hat{y}_{ip} \hat{e}_p$$

whose squared length is

$$\hat{y}_{iq}^2 + \dots + \hat{y}_{ip}^2.$$

Suspect observations will often be such that at least one of the coordinates $\hat{y}_{iq}, \dots, \hat{y}_{ip}$ contributing to this squared length will be large.

Method

1. To help check the normal assumption, construct scatter diagrams for pairs of the first few principal components. Also make Q-Q plots of the sample values generated by each principal component.

2. Construct scatter diagrams and Q-Q plots of the last few principal components to identify suspect observations.

Example 8. If we consider the male turtle data discussed in Example 3, the three sample principal components are

$$\hat{y}_1 = 0.683(x_1 - 4.725) + 0.510(x_2 - 4.478) + 0.523(x_3 - 3.703)$$

$$\hat{y}_2 = -0.159(x_1 - 4.725) - 0.594(x_2 - 4.478) + 0.788(x_3 - 3.703)$$

$$\hat{y}_3 = -0.713(x_1 - 4.725) + 0.622(x_2 - 4.478) + 0.324(x_3 - 3.703)$$

where $x_1 = \ln(\text{length})$, $x_2 = \ln(\text{width})$, and $x_3 = \ln(\text{height})$. Figure 1 shows the Q-Q plot of \hat{y}_2 and Figure 2 shows the scatterplot of (\hat{y}_1, \hat{y}_2) . The observation of the first turtle lies in the upper right corner of the Q-Q plot and in the upper left corner of the scatterplot. This point is suspect, and therefore it should be checked for recording errors. Apart from the first turtle, the scatterplot appears to be reasonably elliptical.

Remark. We can also use principal components to check some assumptions of a multivariate multiple regression model. Consider the:

Residual vector = (observation vector) - (vector of predicted values)

$$r_i = y_i - z_i^\tau \hat{\beta} \quad i = 1, 2, \dots, n$$

We can compute the principal components derived from the covariance matrix of the residuals of a multivariate regression:

$$\frac{\sum_{i=1}^n (r_i - \bar{r})(r_i - \bar{r})^\tau}{n - p}$$

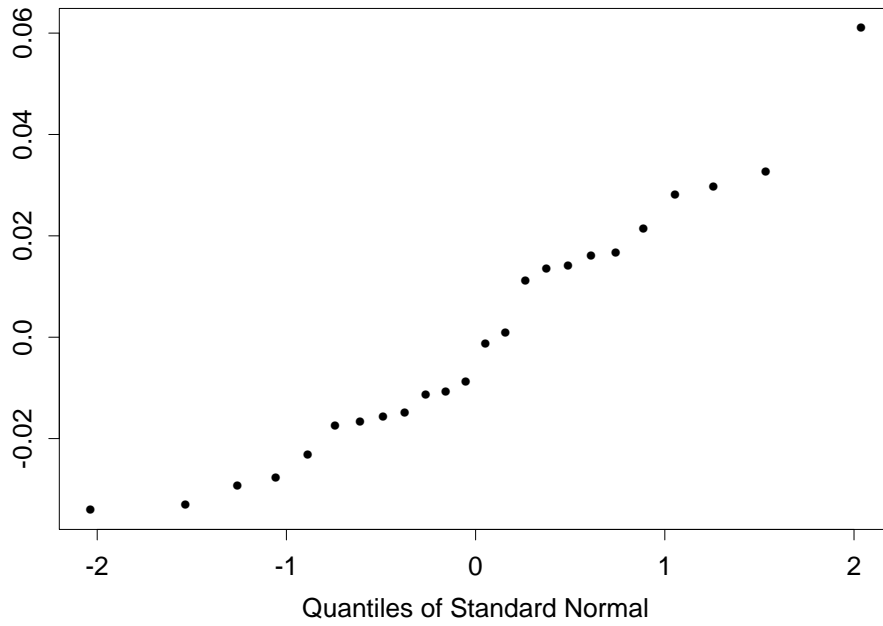


Figure 1.1. Q-Q plot for the second PC of the male turtle data.

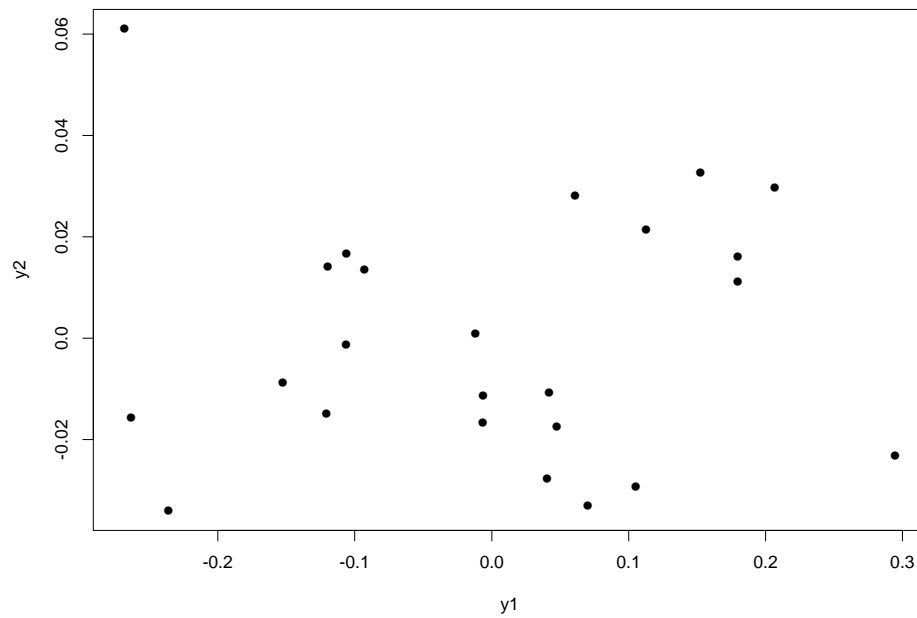


Figure 1.2. Scatterplot of the principal components \hat{y}_1 and \hat{y}_2 (male turtle data).

in the same way as for a random sample. Be aware that there are linear dependencies among the residuals from a linear regression analysis, so the last eigenvalues will be zero within rounding error.

9. Approximation using principal components

Let us consider approximations of the form $A = [a_1, a_2, \dots, a_n]^\tau$ to the centered data matrix

$$G = [g_1, \dots, g_n]^\tau = [x_1 - \bar{x}, \dots, x_n - \bar{x}]^\tau$$

The error of approximation is the “error sum of squares”:

$$ESS = \sum_{i=1}^n \|g_i - a_i\|^2 = \sum_{i=1}^n (g_i - a_i)^\tau (g_i - a_i) = \|G - A\|_F^2$$

where $\|\dots\|_F$ is the Frobenius norm of a matrix.

Result 11. Among all $n \times p$ matrices A with $\text{rank}(A) \leq m < \min(p, n)$, the ESS is minimized by the choice

$$A = [g_1, \dots, g_n]^\tau \hat{E} \hat{E}^\tau = [\hat{y}_1, \dots, \hat{y}_m] \hat{E}^\tau$$

with $\hat{E} = [\hat{e}_1, \hat{e}_2, \dots, \hat{e}_m]$ where $\hat{e}_1, \dots, \hat{e}_m$ are the first m eigenvectors of S , hence $\text{rank}(\hat{E}) = m$. The i th row of A is

$$a_i^\tau = \hat{y}_{i1} \hat{e}_1^\tau + \hat{y}_{i2} \hat{e}_2^\tau + \dots + \hat{y}_{im} \hat{e}_m^\tau$$

where

$$[\hat{y}_{i1}, \hat{y}_{i2}, \dots, \hat{y}_{im}]^\tau = [g_i^\tau \hat{e}_1, g_i^\tau \hat{e}_2, \dots, g_i^\tau \hat{e}_m]^\tau$$

are the values of the first m sample principal components for the i th observation. Moreover,

$$ESS = (n - 1)(\hat{\lambda}_{m+1} + \dots + \hat{\lambda}_p)$$

where $\hat{\lambda}_{m+1} \geq \dots \geq \hat{\lambda}_p$ are the smallest $p - m$ eigenvalues of S .

This means that the space generated by $\{e_1, \dots, e_m\}$ is the result of a least squares optimization.

Proof.

1. We can write the columns of A^T (denoted as a_1, a_2, \dots, a_n) as linear combinations of a set of m orthonormal vectors l_1, l_2, \dots, l_m in $\mathbb{R}^{p \times 1}$. The resulting $p \times m$ matrix $L = [l_1, l_2, \dots, l_m]$ thus satisfies $L^T L = I_m$. For a fixed L , the closest approximation of $g_i = x_i - \bar{x}$ in the space $\text{vec}\{l_1, \dots, l_m\}$ is its orthogonal projection on that space, given by

$$(g_i^T l_1)l_1 + (g_i^T l_2)l_2 + \dots + (g_i^T l_m)l_m = [l_1, l_2, \dots, l_m] \begin{bmatrix} l_1^T g_i \\ l_2^T g_i \\ \vdots \\ l_m^T g_i \end{bmatrix} = LL^T g_i$$

This is the best approximation because any vector a_i in $\text{vec}\{l_1, \dots, l_m\}$ can be written as $a_i = Lb_i$ for some vector $b_i \in \mathbb{R}^m$, and

$$\begin{aligned} g_i - a_i &= g_i - Lb_i = g_i - LL^T g_i + LL^T g_i - Lb_i \\ &= (I - LL^T)g_i + L(L^T g_i - b_i) \end{aligned}$$

so

$$\begin{aligned} \|g_i - a_i\|^2 &= (g_i - Lb_i)^T (g_i - Lb_i) \\ &= g_i^T (I_p - LL^T)g_i + 0 + (LL^T g_i - Lb_i)^T (LL^T g_i - Lb_i) \end{aligned}$$

where the cross-product is zero because $(I_p - LL^T)L = L - L = 0$. The last term is positive unless b_i is chosen so that $Lb_i = LL^T g_i$ which means that a_i is the projection of g_i .

With the choice $a_i = Lb_i = LL^T g_i$ the ESS becomes

$$\begin{aligned}
\sum_{i=1}^n \|g_i - a_i\|^2 &= \sum_{i=1}^n (g_i - LL^T g_i)^\tau (g_i - LL^T g_i) \\
&= \sum_{i=1}^n g_i^\tau (I_p - LL^T) g_i \\
&= \sum_{i=1}^n g_i^\tau g_i - \sum_{i=1}^n g_i^\tau LL^T g_i
\end{aligned} \tag{9.1}$$

2. We now minimize the ESS over all choices of L , by maximizing the last term in (9.1). By the properties of the trace:

$$\begin{aligned}
\sum_{i=1}^n g_i^\tau LL^T g_i &= \sum_{i=1}^n \text{tr}[g_i^\tau LL^T g_i] \\
&= \sum_{i=1}^n \text{tr}[LL^T g_i^\tau g_i] \\
&= (n-1) \text{tr}[LL^T S] \\
&= (n-1) \text{tr}[L^T S L]
\end{aligned}$$

The best choice for L thus maximizes the sum of the diagonal elements of the $m \times m$ matrix $L^T S L$. We know already that $l_1 = \hat{e}_1$ maximizes $l_1^\tau S l_1$, and for l_2 perpendicular to \hat{e}_1 , $l_2^\tau S l_2$ is maximized by \hat{e}_2 .

We find $\hat{L} = [\hat{e}_1, \hat{e}_2, \dots, \hat{e}_m] = \hat{E}$ and $A^\tau = E E^\tau [g_1, g_2, \dots, g_n]$, so

$$A = [g_1, g_2, \dots, g_n]^\tau E E^\tau.$$

With this choice, the h th diagonal element of $\hat{L}^\tau S \hat{L}$ is

$$\hat{e}_h^\tau S \hat{e}_h = \hat{e}_h^\tau (\hat{\lambda}_h \hat{e}_h) = \hat{\lambda}_h$$

so

$$\text{tr}[\hat{L}^\tau S \hat{L}] = \hat{\lambda}_1 + \hat{\lambda}_2 + \dots + \hat{\lambda}_m.$$

Also,

$$\sum_{i=1}^n g_i^T g_i = \text{tr}\left[\sum_{i=1}^n g_i g_i^T\right] = (n-1)\text{tr}(S) = (n-1)(\hat{\lambda}_1 + \hat{\lambda}_2 + \dots + \hat{\lambda}_p)$$

With $L = \hat{L}$ in (9.1) the ESS follows.

□

10. The connection between PCA and orthogonal regression

Let us consider a set of m orthonormal column vectors l_1, \dots, l_m in $\mathbb{R}^{p \times 1}$ (as in the proof of the previous result). The m -dimensional subspace through the origin determined by l_1, l_2, \dots, l_m is then

$$\text{vec}\{l_1, \dots, l_m\} = \{Lb; b \in \mathbb{R}^{m \times 1}\}$$

Translating this subspace to pass through some point $c \in \mathbb{R}^{p \times 1}$ yields the affine subspace

$$H = c + \text{vec}\{l_1, \dots, l_m\} = \{c + Lb; b \in \mathbb{R}^{m \times 1}\}$$

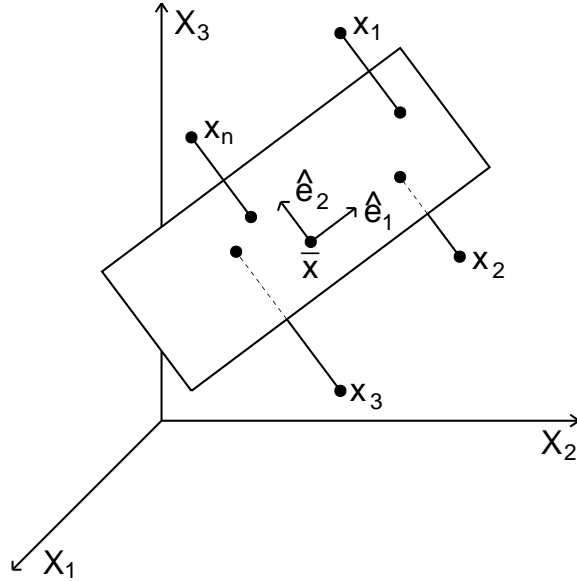
We are going to select the m -dimensional affine subspace H that minimizes the sum of squared distances between the observations x_i and H . This is **orthogonal regression**. The word 'orthogonal' comes from the fact that the distance $d(x_i, H)$ is measured in the direction orthogonal to H (whereas in classical regression we use the vertical distance, i.e. the absolute residual $|r_i|$).

Suppose that we approximate x_i by $c + Lb_i$ with $\sum_{i=1}^n b_i = 0$. (If $\sum_{i=1}^n b_i = n\bar{b} \neq 0$, use $c + Lb_i = (c + L\bar{b}) + L(b_i - \bar{b}) = c^* + Lb_i^*$ instead.) Then

$$\begin{aligned} & \sum_{i=1}^n (x_i - c - Lb_i)^\tau (x_i - c - Lb_i) \\ &= \sum_{i=1}^n (x_i - \bar{x} - Lb_i + \bar{x} - c)^\tau (x_i - \bar{x} - Lb_i + \bar{x} - c) \\ &= \sum_{i=1}^n (g_i - Lb_i)^\tau (g_i - Lb_i) + n(\bar{x} - c)^\tau (\bar{x} - c) \\ &\stackrel{(*)}{\geq} \sum_{i=1}^n (g_i - \hat{E}\hat{E}^\tau g_i)^\tau (g_i - \hat{E}\hat{E}^\tau g_i) + n(\bar{x} - c)^\tau (\bar{x} - c) \\ &\geq \sum_{i=1}^n (g_i - \hat{E}\hat{E}^\tau g_i)^\tau (g_i - \hat{E}\hat{E}^\tau g_i) \end{aligned}$$

where (*) holds by result 11, since $\text{rank}[Lb_1, \dots, Lb_n] \leq m$. If we take $c = \bar{x}$ the lower bound is reached, so the best affine subspace passes through the

sample mean. The subspace is thus determined by the first m eigenvectors of S , namely $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_m$. Moreover, the coefficient of \hat{e}_k is $\hat{e}_k^T(x_i - \bar{x}) = \hat{y}_{ik}$, which is the k th sample principal component of the i th observation x_i .



The name 'orthogonal regression' is often used when fitting a hyperplane, i.e. $m = p - 1$. In this case we can write the approximating hyperplane H as

$$H = \bar{x} + e_p^\perp \tag{10.1}$$

where e_p^\perp is the orthogonal complement of the last eigenvector e_p . In this case, the sum of squared distances to H becomes

$$\sum_{i=1}^n d^2(x_i, H) = ESS = (n - 1)\hat{\lambda}_p = (n - 1) \text{var}_{i=1}^n(y_{ip}).$$

Making use of (10.1) we could also determine the principal components by repeatedly applying orthogonal regression: the first step yields e_p , after which we project the data on the $(p - 1)$ dimensional space e_p^\perp . In this space we apply orthogonal regression again, yielding e_{p-1} , and so on.

Remark. In two dimensions ($p = 2$), the orthogonal regression line therefore corresponds to the first principal component.