PRINCIPAL COMPONENT ANALYSIS

Prof. Dan A. Simovici

UMB

What is PCA?

Principal component analysis (PCA) is a dimensionality reduction technique that aims to create a few new, uncorrelated linear combinations of the variables of an experiments that “explain” the major parts of the data variability.

Definition

Let

$$X = \begin{pmatrix} \mathbf{u}_1' \\ \vdots \\ \mathbf{u}_m' \end{pmatrix} = (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n) \in \mathbb{R}^{m \times n}$$

be a data matrix. The principal components of X are the eigenvectors of the covariance matrix $cov(X) = \frac{1}{m-1}\hat{X}'\hat{X}$, or equivalently, of the scatter matrix $S(X) = \hat{X}'\hat{X}$.

The covariance matrix $cov(X)$ is a scalar multiple of the Gram matrix of the columns $\hat{v}_1, \ldots, \hat{v}_n$ of the centered data matrix $\hat{X}$.

Let $W = (\mathbf{w}_1 \ \mathbf{w}_2 \ \cdots \ \mathbf{w}_k) \in \mathbb{R}^{n \times k}$ be a matrix with $n \leqslant k$ and $rank(W) = k$. Note that

- the columns $\mathbf{w}_1, \ldots, \mathbf{w}_k$ are linearly independent and constitute a basis of the subspace $\text{Ran}(A) = \{\mathbf{t} \in \mathbb{R}^n \mid \mathbf{t} = W\mathbf{u}, \mathbf{u} \in \mathbb{R}^k\}$;
- the Gram matrix of W is the matrix $G_W = (g_{ij}) \in \mathbb{R}^{k \times k}$ defined by $g_{ij} = \mathbf{w}_i'\mathbf{w}_j$ for $1 \leqslant i, j \leqslant k$;
- $G_W = W'W$ and G_W is a symmetric matrix.

Theorem

Let $W = (\mathbf{w}_1\ \mathbf{w}_2\ \cdots\ \mathbf{w}_k) \in \mathbb{R}^{n\times k}$ *be a matrix, where* $k \leqslant n$*.*
If $\mathit{rank}(W) = k$*, then the Gram matrix* G_W *is positive definite.*

Proof

Suppose that $rank(W) = k$, that is, its set of columns is linearly independent. Let $\mathbf{x} \in \mathbb{R}^k$. We have $\mathbf{x}' G_W \mathbf{x} = \mathbf{x}' W' W \mathbf{x} = (W\mathbf{x})' W \mathbf{x} =\parallel W\mathbf{x} \parallel_2^2$. Therefore, if $\mathbf{x}' G_W \mathbf{x} = 0$, we have $W\mathbf{x} = \mathbf{0}$, which is equivalent to $x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n = 0$. Since $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is linearly independent it follows that $x_1 = \cdots = x_m = 0$, so $\mathbf{x} = \mathbf{0}$. Thus, G_W is indeed, positive definite.

Note: The Gram matrix of an arbitrary sequence of vectors is positive semidefinite.

Definition

Let $W = (\mathbf{w}_1\ \mathbf{w}_2\ \cdots\ \mathbf{w}_k) \in \mathbb{R}^{n \times k}$ be a matrix, where $k \leqslant n$.
The Gramian of W is the number $\det(G_W)$.

Theorem

Let $W = (\mathbf{w}_1\ \mathbf{w}_2\ \cdots\ \mathbf{w}_k) \in \mathbb{R}^{n \times k}$ *be a matrix, where* $k \leqslant n$*. The set of columns of* W *linearly independent if and only if* $\det(G_W) \neq 0$*.*

Proof

Suppose that $\det(G_W) \neq 0$ and that W is not linearly independent. In other words, the numbers $a_1, \ldots, a_k$ exists such that at least one of them is not 0 and $a_1\mathbf{w}_1 + \cdots + a_k\mathbf{w}_k = \mathbf{0}$. This implies the equalities

$$a_1(\mathbf{w}_1, \mathbf{w}_j) + \cdots + a_k(\mathbf{w}_k, \mathbf{w}_j) = \mathbf{0},$$

for $1 \leqslant j \leqslant k$, so the system $G_W\mathbf{a} = \mathbf{0}$ has a non-trivial solution in $a_1, \ldots, a_k$. This implies $\det(G_W) = 0$, which contradicts the initial assumption.

Conversely, suppose that the set of columns of W is linearly independent and $\det(G_W) = 0$. Then, the linear system

$$a_1(\mathbf{w}_1, \mathbf{w}_j) + \cdots + a_k(\mathbf{w}_k, \mathbf{w}_j) = \mathbf{0},$$

for $1 \leqslant j \leqslant k$, has a non-trivial solution in $a_1, \ldots, a_k$. If $\mathbf{w} = a_1\mathbf{w}_1 + \cdots a_k\mathbf{w}_k$, this amounts to $(\mathbf{w}, \mathbf{w}_i) = 0$ for $1 \leqslant i \leqslant k$. This, in turn, implies $(\mathbf{w}, \mathbf{w}) =\parallel \mathbf{w} \parallel_2^2 = 0$, so $\mathbf{w} = 0$, which contradicts the linear independence of W.

Orthogonal Projection

Definition

Let T be an m-dimensional subspace of $\mathbb{R}^n$ and let $\{\mathbf{t}_1, \ldots, \mathbf{t}_m\}$ be an orthonormal basis of this subspace.
The orthogonal projection of the vector $\mathbf{x} \in \mathbb{R}^n$ on the subspace T is the vector

$$\mathbf{t} = (\mathbf{x}, \mathbf{t}_1)\mathbf{t}_1 + \cdots + (\mathbf{x}, \mathbf{t}_m)\mathbf{t}_m.$$

Theorem

Let T be an m-dimensional subspace of $\mathbb{R}^n$ and let $\mathbf{x} \in \mathbb{R}^n$. The vector $\mathbf{y} = \mathbf{x} - proj_T(\mathbf{x})$ belongs to the subspace $T^\perp$ that is perpendicular on T.

Proof: Let $B_T = \{\mathbf{t}_1, \ldots, \mathbf{t}_m\}$ be an orthonormal basis of T. Note that

$$\begin{aligned} (\mathbf{y}, \mathbf{t}_j) &= (\mathbf{x}, \mathbf{t}_j) - \left(\sum_{i=1}^{m} (\mathbf{x}, \mathbf{t}_i)\mathbf{t}_i, \mathbf{t}_j \right) \\ &= (\mathbf{x}, \mathbf{t}_j) - \sum_{i=1}^{m} (\mathbf{x}, \mathbf{t}_i)(\mathbf{t}_i, \mathbf{t}_j) = 0, \end{aligned}$$

due to the orthogonality of the basis B_T. Therefore, $\mathbf{y}$ is orthogonal on every linear combination of B_T, that is on the subspace T.

Theorem

Let T be an m-dimensional subspace of $\mathbb{R}^n$ having the orthonormal basis $\{\mathbf{t}_1, \ldots, \mathbf{t}_m\}$. The orthogonal projection $proj_T$ is given by $proj_T(\mathbf{x}) = B_T B_T' \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n$, where $B_T \in \mathbb{R}^{n \times m}$ is the matrix $B_T = (\mathbf{t}_1 \ \cdots \ \mathbf{t}_m) \in \mathbb{R}^{n \times m}$.

Proof: We can write

$$\text{proj}_T(\mathbf{x}) = \sum_{i=1}^{m} \mathbf{t}_i(\mathbf{t}_i' \mathbf{x}) = (\mathbf{t}_1 \ \cdots \ \mathbf{t}_m) \begin{pmatrix} \mathbf{t}_1' \\ \vdots \\ \mathbf{t}_m' \end{pmatrix} \mathbf{x} = B_T B_T' \mathbf{x}.$$

Projection Matrix

Since the basis $\{\mathbf{t}_1, \ldots, \mathbf{t}_m\}$ is orthonormal, we have $B'_T B_T = I_m$. Observe that the matrix $B_T B'_T \in \mathbb{R}^{n \times n}$ is symmetric and idempotent because

$$(B_T B'_T)(B_T B'_T) = B_T(B'_T B_T)B'_T = B_T B'_T.$$

For an m-dimensional subspace T of $\mathbb{R}^n$ we denote by $P_T = B_T B'_T \in \mathbb{R}^{n \times n}$, where B_T is a matrix of an orthonormal basis of T as defined before. P_T is the projection matrix of the subspace T.
Note that every non-zero subspace T, the matrix P_T is a symmetric matrix.

Let T be an m-dimensional subspace of $\mathbb{R}^n$ having the orthonormal basis $\{\mathbf{t}_1, \ldots, \mathbf{t}_m\}$.
If $B_T = (\mathbf{t}_1 \ \cdots \ \mathbf{t}_m) \in \mathbb{R}^{n \times m}$, then for every $\mathbf{x} \in \mathbb{R}^n$ we have the decomposition $\mathbf{x} = P_T\mathbf{x} + Q_T\mathbf{x}$, where $P_T = B_T B_T'$ and $Q_T = I_n - P_T$, $P_T\mathbf{x} \in T$ and $Q_T\mathbf{x} \in T^{\perp}$.
Observe that

$$\begin{aligned} Q_T^2 &= (I_n - P_T P_T')(I_n - P_T P_T') \\ &= I_n - P_T P_T' - P_T P_T' + P_T P_T' P_T P_T' = Q_T, \end{aligned}$$

so Q_T is an idempotent matrix. The matrix Q_T is the projection matrix on the subspace $T^{\perp}$.

The Purpose of the Algorithm

- the Gram-Schmidt algorithm constructs an orthonormal basis for a subspace U of $\mathbb{R}^n$, starting from an arbitrary basis of $\{\mathbf{t}_1, \ldots, \mathbf{t}_m\}$ of U;
- the orthonormal basis is constructed sequentially such that the subspace generated by the first k vectors, $\langle \mathbf{w}_1, \ldots, \mathbf{w}_k \rangle$ equals $\langle \mathbf{t}_1, \ldots, \mathbf{t}_k \rangle$ for $1 \leqslant k \leqslant m$.

Gram-Schmidt Orthogonalization Algorithm

Algorithm 3.1: Gram-Schmidt Algorithm

Data: A basis $\{\mathbf{t}_1, \ldots, \mathbf{t}_m\}$ for a subspace U of $\mathbb{R}^n$

Result: An orthonormal basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ for U

```
$W = O_{n,m}$;
$W(:,1) = W(:,1) + \frac{1}{\|U(:,1)\|_2} U(:,1)$;
for $k = 2$ to $m$ do
    $P = I_n - W(:,1:(k-1))W(:,1:(k-1))'$;
    $W(:,k) = W(:,k) + \frac{1}{\|PU(:,k)\|_2} PU(:,k)$;
end
return $W = (\mathbf{w}_1 \ \cdots \ \mathbf{w}_m)$;
```

Correctness of Gram-Schmidt Algorithm

Theorem

Let $(\mathbf{w}_1, \ldots, \mathbf{w}_m)$ *be the sequence of vectors constructed by the Gram-Schmidt algorithm starting from the basis* $\{\mathbf{t}_1, \ldots, \mathbf{t}_m\}$ *of an m-dimensional subspace* U *of* $\mathbb{R}^n$.

The set $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ *is an orthogonal basis of* U *and* $\langle \mathbf{w}_1, \ldots, \mathbf{w}_k \rangle = \langle \mathbf{t}_1, \ldots \mathbf{t}_k \rangle$ *for* $1 \leqslant k \leqslant m$.

Proof

W is initialized as $O_{n,m}$. Its columns will contain eventually the vectors of the orthonormal basis $\mathbf{w}_1, \ldots, \mathbf{w}_m$. The argument is by induction on $k \geqslant 1$. The base case, $k = 1$, is immediate. Suppose that the statement of the theorem holds for k, that is, the set $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ is an orthonormal basis for $U_k = \langle \mathbf{t}_1, \ldots, \mathbf{t}_k \rangle$ and constitutes the set of the initial k columns of the matrix W, that is, $W_k = W(:, 1 : k)$. Then, $P_k = I_n - W_k W_k'$ is the projection matrix on the subspace $U_k^{\perp}$, so $P_k \mathbf{t}_k$ is orthogonal on every $\mathbf{w}_i$, where $1 \leqslant i \leqslant k$. Therefore, $\mathbf{w}_{k+1} = W(:, (k+1))$ is a unit vector orthogonal on all its predecessors $\mathbf{w}_1, \ldots, \mathbf{w}_k$, so $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ is an orthonormal set.

Proof (cont'd)

The equality $\langle \mathbf{t}_1, \ldots, \mathbf{t}_k \rangle = \langle \mathbf{w}_1, \ldots, \mathbf{w}_k \rangle$ clearly holds for $k = 1$. Suppose that it holds for k. Then, we have

$$\begin{aligned} \mathbf{w}_{k+1} &= \frac{1}{\parallel P_k \mathbf{t}_{k+1} \parallel_2} \left(\mathbf{t}_{k+1} - W_k W_k' \mathbf{t}_{k+1} \right) \\ &= \frac{1}{\parallel P_k \mathbf{t}_{k+1} \parallel_2} \left(\mathbf{t}_{k+1} - (\mathbf{w}_1 \ \cdots \ \mathbf{w}_k) W_k' \mathbf{t}_{k+1} \right). \end{aligned}$$

Since $\mathbf{w}_1 \ldots, \mathbf{w}_k$ belong to the subspace $\langle \mathbf{t}_1, \ldots, \mathbf{t}_k \rangle$ (by inductive hypothesis) it follows that $\mathbf{w}_{k+1} \in \langle \mathbf{t}_1, \ldots, \mathbf{t}_k, \mathbf{t}_{k+1} \rangle$, so $\langle \mathbf{w}_1, \ldots, \mathbf{w}_{k+1} \rangle \subseteq \langle \mathbf{t}_1, \ldots, \mathbf{t}_k \rangle$. For the converse inclusion, since

$$\mathbf{t}_{k+1} = \parallel P_k \mathbf{t}_{k+1} \parallel_2 \mathbf{w}_{k+1} + (\mathbf{w}_1 \ \cdots \ \mathbf{w}_k) W_k' \mathbf{t}_{k+1},$$

it follows that $\mathbf{t}_{k+1} \in \langle \mathbf{w}_1, \ldots, \mathbf{w}_k, \mathbf{w}_{k+1} \rangle$. Thus, $\langle \mathbf{t}_1, \ldots, \mathbf{t}_k, \mathbf{t}_{k+1} \rangle \subseteq \langle \mathbf{w}_1, \ldots, \mathbf{w}_k, \mathbf{w}_{k+1} \rangle$.

Parseval's Equality

Let W be the subspace of $\mathbb{R}^n$ generated by the orthonormal set $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$.

- if $\mathbf{x} = \sum_{i=1}^{k} x_i \mathbf{w}_i$, then $\mathbf{x}'\mathbf{w}_i = x_i$ for $1 \leqslant i \leqslant k$; the equality

$$\mathbf{x} = \sum_{i=1}^{k} (\mathbf{x}'\mathbf{w}_i)\mathbf{w}_i$$

  is the Fourier expansion of $\mathbf{x}$ with respect to the orthonormal set W.
- We have Parseval's Equality

$$\| \mathbf{x} \|^2 = (\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{k} (\mathbf{x}, \mathbf{w}_i)^2.$$

Ky Fan's Theorem

Theorem

Let $A \in \mathbb{R}^{n\times n}$ *be a symmetric matrix such that spec*$(A) = \{\lambda_1, \ldots, \lambda_n\}$*, where* $\lambda_1 \geqslant \cdots \geqslant \lambda_n$ *and let* $V = (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n)$ *whose columns consists of the corresponding unit eigenvectors of* A*. For any positive integer* $q \leqslant n$*, the sums*

$$\sum_{i=1}^{q} \lambda_i \text{ and } \sum_{i=1}^{q} \lambda_{n+1-i}$$

are, respectively, the maximum and minimum of $\sum_{j=1}^{q} \mathbf{x}_j' A \mathbf{x}_j$*, where* $\{\mathbf{x}_1, \ldots, \mathbf{x}_q\}$ *is an orthonormal set of vectors in* $\mathbb{R}^n$*.*
The maximum is obtained by choosing the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_q$ *as the first* q *columns of* V*; the minimum is obtained by assigning to* $\mathbf{x}_1, \ldots, \mathbf{x}_q$ *the last* q *columns of* V*.*

Proof

Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be an orthonormal set of eigenvectors of A and let $\mathbf{x}_i = \sum_{p=1}^{n} b_{pi}\mathbf{v}_p$ the expression of $\mathbf{x}_i$ in the basis of orthogonal eigenvectors, where $b_{pi} = \mathbf{v}'_p\mathbf{x}_i = \mathbf{x}'_i\mathbf{v}_p$. In matrix form these equalities can be written as $X = (\mathbf{v}_1 \ \ldots \ \mathbf{v}_n)B$, where B is the orthonormal matrix

$$B = \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ldots & \vdots \\ b_{n1} & \cdots & b_{nn} \end{pmatrix}.$$

Note that $b_{pi} = \mathbf{x}_i\mathbf{v}'_p = \mathbf{x}'_i\mathbf{v}_p$. By Parseval's equality

$$\sum_{p=1}^{n} b_{pj}^2 = 1$$

because $\| \mathbf{x}_i \| = 1$.

Proof (cont'd)

We have

$$\begin{aligned} \mathbf{x}_j' A \mathbf{x}_j &= \mathbf{x}_j' A \sum_{p=1}^{n} b_{pj} \mathbf{v}_p = \sum_{p=1}^{n} b_{pj} \mathbf{x}_j' A \mathbf{v}_p \\ &= \sum_{p=1}^{n} b_{pj} \mathbf{x}_j' \lambda_p \mathbf{v}_p = \sum_{p=1}^{n} b_{pj}^2 \lambda_p \\ &= \lambda_q \sum_{p=1}^{n} b_{pj}^2 + \sum_{p=1}^{q} (\lambda_p - \lambda_q) b_{pj}^2 + \sum_{j=q+1}^{n} (\lambda_j - \lambda_q) b_{pj}^2. \end{aligned}$$

Since $\parallel \mathbf{x}_j \parallel = 1$ this implies

$$\mathbf{x}_j' A \mathbf{x}_j \leqslant \lambda_q + \sum_{p=1}^{q} (\lambda_p - \lambda_q) b_{pj}^2.$$

Therefore,

$$\sum_{i=1}^{q} \lambda_i - \sum_{j=1}^{q} \mathbf{x}_j' A \mathbf{x}_j \geqslant \sum_{i=1}^{q} (\lambda_i - \lambda_q) \left(1 - \sum_{j=1}^{q} b_{ij}^2\right).$$

Again, by Parseval's equality, we have $\sum_{j=1}^{q} b_{ij}^2 \leqslant \| \mathbf{x}_i \|^2 = 1$, so $\sum_{i=1}^{q} (\lambda_i - \lambda_q) \left(1 - \sum_{j=1}^{q} b_{ij}^2\right) \geqslant 0$. The left member of previous inequality becomes 0, when $\mathbf{x}_i = \mathbf{v}_i$, so $\sum_{j=1}^{q} \mathbf{x}_j' A \mathbf{x}_j \leqslant \sum_{i=1}^{q} \lambda_i$.
The maximum of $\sum_{j=1}^{q} \mathbf{x}_j' A \mathbf{x}_j$ is obtained when $\mathbf{x}_j = \mathbf{v}_j$ for $1 \leqslant j \leqslant q$, that is, when X consists of the first q columns of V that correspond to eigenvectors of the top k largest eigenvalues.
The argument for the minimum is similar.

Projections on a unidimensional space

Let $\mathbf{t} \in \mathbb{R}^n$ be a unit vector. The projection of a vector $\mathbf{w} \in \mathbb{R}^n$ on the subspace $\langle \mathbf{t} \rangle$ generated by $\mathbf{t}$ is given by

$$\mathsf{proj}_{\langle \mathbf{t} \rangle}(\mathbf{w}) = \mathbf{t}\mathbf{t}'\mathbf{w}.$$

To simplify the notation we shall write $\mathsf{proj}_{\mathbf{t}}$ instead of $\mathsf{proj}_{\langle \mathbf{t} \rangle}$.

Inertia of a Sequence of Vectors

Definition

Let $U = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$ be a sequence of vectors in $\mathbb{R}^n$. The inertia of this sequence relative to a vector $\mathbf{z} \in \mathbb{R}^n$ is the number

$$I_{\mathbf{z}}(U) = \sum_{j=1}^{m} \parallel \mathbf{u}_j - \mathbf{z} \parallel_2^2 .$$

We seek to evaluate the inertia $I_{\mathbf{0}}(\mathrm{proj}_{\mathbf{t}}(\hat{X}'))$ of the projections on the subspace generated by the unit vector $\mathbf{t} \in \mathbb{R}^n$.
Since $\hat{X}' = (\mathbf{u}_1 \ \cdots \ \mathbf{u}_m)$, by the definition of the inertia, we have:

$$\begin{aligned}
I_{\mathbf{0}}(\mathrm{proj}_{\mathbf{t}}(\hat{X}')) &= \sum_{j=1}^{m} \parallel \mathbf{t}\mathbf{t}'\mathbf{u}_j \parallel_2^2 = \sum_{j=1}^{m} \mathbf{u}_j'\mathbf{t}\mathbf{t}'\mathbf{t}\mathbf{t}'\mathbf{u}_j \\
&= \sum_{j=1}^{m} \mathbf{u}_j'\mathbf{t}\mathbf{t}'\mathbf{u}_j \\
&\quad \text{(because } \mathbf{t}'\mathbf{t} = 1\text{)} \\
&= \sum_{j=1}^{m} \mathbf{t}'\mathbf{u}_j\mathbf{u}_j'\mathbf{t} \\
&\quad \text{(because both } \mathbf{u}_j'\mathbf{t} \text{ and } \mathbf{t}'\mathbf{u}_j \text{ are scalars)} \\
&= \mathbf{t}'X'X\mathbf{t}.
\end{aligned}$$

The necessary condition for the existence of extreme values of this inertia as a function of **t** is

$$\begin{aligned} \nabla\left(I_{\mathbf{0}}(\mathrm{proj}_{\mathbf{t}}(\hat{X}')) + \lambda(1-\mathbf{t}'\mathbf{t})\right) &= \nabla\left(\mathbf{t}'\hat{X}'\hat{X}\mathbf{t} + \lambda(1-\mathbf{t}'\mathbf{t})\right) \\ &= 2\hat{X}'\hat{X}\mathbf{u} - 2\lambda\mathbf{t} = \mathbf{0}, \end{aligned}$$

where λ is a Lagrange multiplier.

This implies $\hat{X}'\hat{X}\mathbf{t} = \lambda\mathbf{t}$. In other words, to achieve extreme values of the inertia $I_{\mathbf{0}}(\text{proj}_{\mathbf{t}}(\hat{X}'))$, $\mathbf{t}$ must be chosen as a eigenvector of the covariance matrix of $\hat{X}$, that is, as a principal direction of $\hat{X}$.

Courant-Fisher Theorem

Theorem

Let $A \in \mathbb{R}^{n\times n}$ *be a symmetric matrix having the eigenvalues* $\lambda_1 \geqslant \cdots \geqslant \lambda_n$*. We have*

$$\lambda_k = \min_{dim(S)=n-k+1} \max_{\mathbf{x}} \{\mathbf{x}' A \mathbf{x} \mid \mathbf{x} \in S \text{ and } \parallel \mathbf{x} \parallel_2 = 1\},$$

and

$$\lambda_k = \max_{dim(S)=k} \min_{\mathbf{x}} \{\mathbf{x}' A \mathbf{x} \mid \mathbf{x} \in S \text{ and } \parallel \mathbf{x} \parallel_2 = 1\},$$

where S *ranges over the subspaces of* $\mathbb{R}^n$*.*

Since A is a symmetric matrix there exists an orthogonal matrix U and a diagonal matrix D such that $A = U'DU$ and the diagonal elements of D are the eigenvalues of A, that is, $D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$.
We prove initially that

$$\lambda_k = \min_{dim(S)=n-k+1} \max_{\mathbf{x}} \{\mathbf{x}'D\mathbf{x} \mid \mathbf{x} \in S \text{ and } \parallel \mathbf{x} \parallel_2 = 1\}.$$

For $\dim(S) = n - k + 1$ define $\tilde{S} = \{\mathbf{y} \in \mathbb{R}^n \mid y \in S \text{ and } \parallel \mathbf{y} \parallel = 1\}$ and $\hat{S} = \{y \in S \subseteq \langle \mathbf{e}_1, \ldots, \mathbf{e}_k \rangle\}$.

We have $S \cap \langle \mathbf{e}_1, \ldots, \mathbf{e}_k \rangle \neq 0$ because otherwise, we would have $\dim(S \cap \langle \mathbf{e}_1, \ldots, \mathbf{e}_k \rangle) = n + 1$. Therefore, $\hat{S}$ consists of vectors of $\tilde{S}$ having the form

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

such that $\sum_{i=1}^{k} y_i^2 = 1$. So, of $\dim(S) = n - k + 1$ we have

$$\mathbf{y}' D \mathbf{y} = \sum_{i=1}^{k} \lambda_i |y_i|^2 \geqslant \lambda_k \sum_{i=1}^{k} |y_i|^2 = \lambda_k$$

for all $\mathbf{y} \in \hat{S}$.

Since $\hat{S} \subseteq \tilde{S}$ it follows that $\max_{\mathbf{y}\in\tilde{S}} \mathbf{y}'D\mathbf{y} \geqslant \max_{\mathbf{y}\in\hat{S}} \mathbf{y}'D\mathbf{y} \geqslant \lambda_k$, so

$$\min_{dim(S)=n-k+1} \max_{\mathbf{x}} \{\mathbf{x}'D\mathbf{x} \mid \mathbf{x} \in S \text{ and } \parallel \mathbf{x} \parallel_2 = 1\} \geqslant \lambda_k.$$

Let now S be $S = \langle \mathbf{e}_1, \ldots, \mathbf{e}_{k-1} \rangle^{\perp}$. Clearly, $\dim(S) = n - k + 1$. A vector $\mathbf{y} \in S$ has the form

$$\mathbf{y} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ y_k \\ \vdots \\ y_n \end{pmatrix}$$

Therefore,

$$\mathbf{y}' D \mathbf{y} = \sum_{i=k}^{n} \lambda_i |y_i|^2 \leq \lambda_i \sum_{i=k}^{n} |y_i|^2 = \lambda_i$$

for all $\mathbf{y} \in \{\mathbf{y} \in S \mid \parallel \mathbf{y} \parallel_2 = 1\}$. This implies

$$\min_{dim(S)=n-k+1} \max_{\mathbf{x}} \{\mathbf{x}' D \mathbf{x} \mid \mathbf{x} \in S \text{ and } \parallel \mathbf{x} \parallel_2 = 1\} \leqslant \lambda_k,$$

which yields the desired equality.

The matrices A and D have the same eigenvalues. Also $\mathbf{x}'A\mathbf{x} = \mathbf{x}'A\mathbf{x} = \mathbf{x}'U'DU\mathbf{x} = (U\mathbf{x})'D(U\mathbf{x})$ and $\| U\mathbf{x} \|_2 = \| \mathbf{x} \|_2$, because U is a unitary matrix. This yields the first equality of the theorem.

Corollary

Let $A \in \mathbb{R}^{n \times n}$ *be a symmetric matrix having the eigenvalues* $\lambda_1 \geqslant \cdots \geqslant \lambda_n$*. If* $\mathbf{u}_1, \ldots, \mathbf{u}_k$ *are eigenvectors that correspond to* $\lambda_1, \ldots, \lambda_k$*, respectively, then a unit vector* $\mathbf{x}$ *that maximizes* $\mathbf{x}' A \mathbf{x}$ *and belongs to the subspace orthogonal to the subspace generated by the first* k *eigenvectors of* A *is an eigenvector that corresponds to* λ_{k+1}*.*

Proof

Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be the eigenvectors of A and let $\mathbf{x} \in \langle \mathbf{u}_1, \ldots, \mathbf{u}_k \rangle^{\perp}$ be a unit vector. We have $\mathbf{x} = \sum_{j=k+1}^{n} a_j \mathbf{u}_j$, and $\sum_{j=k+1}^{n} a_j^2 = 1$ which implies

$$\mathbf{x}' A \mathbf{x} = \sum_{j=k+1}^{n} \lambda_j a_j^2 = \lambda_{k+1}.$$

This, in turn, implies $a_{k+1} = 1$ and $a_{k+2} = \cdots = a_n = 0$, so $\mathbf{x} = \mathbf{u}_{k+1}$.

The principal directions of a data sample matrix $\hat{X} \in \mathbb{R}^{m \times n}$ can be obtained directly from the data sample matrix X by applying the previous Corollary.

Suppose that the eigenvalues of $\hat{X}'\hat{X}$ are

$$\lambda_1 \geqslant \cdots \geqslant \lambda_n.$$

The first principal direction $\mathbf{t}_1$ of X which corresponds to the largest eigenvalue of $\hat{X}'\hat{X}$ is

$$\begin{aligned} \mathbf{t}_1 &= \arg\max_{\mathbf{t}} \left\{ \mathbf{t}'\hat{X}\hat{X}'\mathbf{t} \mid \mathbf{t} \in \mathbb{R}^n, \| \mathbf{t} \|_2 = 1 \right\} \\ &= \arg\max_{\mathbf{t}} \left\{ \| \hat{X}'\mathbf{t} \|_2^2 \mid \| \mathbf{t} \|_2 = 1 \right\}. \end{aligned}$$

Suppose that we computed the principal directions $\mathbf{t}_1, \ldots, \mathbf{t}_k$ of $\hat{X}$. Then, $\mathbf{t}_{k+1} \in \mathbb{R}^n$ is a unit vector $\mathbf{t}$ that maximizes $\mathbf{t}'\hat{X}\hat{X}'\mathbf{t} = \parallel \hat{X}'\mathbf{t} \parallel_2^2$ and belongs to the subspace orthogonal to the subspace generated by the first k principal directions of $\hat{X}$, that is,

$$\mathbf{t}_{k+1} = \arg\max_{\mathbf{t}} \left\{ \parallel \hat{X}'\mathbf{t} \parallel_2^2 \mid \mathbf{t} \in \mathbb{R}^n, \parallel \mathbf{t} \parallel_2 = 1, \mathbf{t} \in \langle \mathbf{t}_1, \ldots, \mathbf{t}_k \rangle^{\perp} \right\}.$$

For $\mathbf{z} \in \mathbb{R}^n$ we have

$$(I - \sum_{j=1}^{k} \mathbf{t}_j \mathbf{t}_j')\mathbf{z} = \mathbf{z} - \text{proj}_{\langle \mathbf{t}_1, \ldots, \mathbf{t}_k \rangle} \mathbf{z} \in \langle \mathbf{t}_1, \ldots, \mathbf{t}_k \rangle^{\perp}.$$

Therefore, $\mathbf{x} \in \langle \mathbf{t}_1, \ldots, \mathbf{t}_k \rangle^{\perp}$, is equivalent to $\mathbf{x} = (I - \sum_{j=1}^{k} \mathbf{t}_j \mathbf{t}_j')\mathbf{x}$.

Thus,

$$\mathbf{t}_{k+1} = \arg\max_{\mathbf{t}} \left\{ \left\| \hat{X}(I - \sum_{j=1}^{k} \mathbf{t}_j \mathbf{t}_j')\mathbf{t} \right\|_2 \mid \| \mathbf{t} \|_2 = 1 \right\},$$

for $0 \leqslant k \leqslant n-1$. This technique allows finding the principal directions of $\hat{X}$ by solving a sequence of optimization problems involving the matrix $\hat{X}$.

A vector $\mathbf{x} \in V$ can be represented in the k-dimensional subspace W by a vector $\mathbf{w} = a_1\mathbf{w}_1 + \cdots + a_k\mathbf{w}_k$. The error of this representation is

$$\epsilon = \left\| \mathbf{x} - \sum_{i=1}^{k} a_i \mathbf{w}_i \right\|.$$

For a set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ the total error is

$$J(\mathbf{w}_1, \ldots, \mathbf{w}_k, a_{11}, \ldots, a_{mk}) = \sum_{j=1}^{m} \left\| \mathbf{x}_j - \sum_{i=1}^{k} a_{ji} \mathbf{w}_i \right\|^2.$$

A Simpler form for J

$$\begin{aligned} & J(\mathbf{w}_1, \ldots, \mathbf{w}_k, a_{11}, \ldots, a_{mk}) \\ = & \sum_{j=1}^{m} \parallel \mathbf{x}_j \parallel^2 - 2\sum_{j=1}^{m}\sum_{i=1}^{k} a_{ji}\mathbf{x}_j'\mathbf{w}_i + \sum_{j=1}^{m}\sum_{i=1}^{k} a_{ji}^2. \end{aligned}$$

The partial derivative with respect to a_{pq} are

$$\frac{\partial J}{\partial a_{pq}} = -2\mathbf{x}_p'\mathbf{w}_q + 2a_{pq} = 0,$$

so $a_{pq} = \mathbf{x}_p'\mathbf{w}_q$.

For the optimal values of a_{pq} the function J becomes

$$\begin{aligned} & J(\mathbf{w}_1, \ldots, \mathbf{w}_k) \\ = \; & \sum_{j=1}^{m} \parallel \mathbf{x}_j \parallel^2 - 2 \sum_{j=1}^{m} \sum_{i=1}^{k} (\mathbf{x}_j' \mathbf{w}_i)(\mathbf{x}_j' \mathbf{w}_i) + \sum_{j=1}^{m} \sum_{i=1}^{k} \sum_{j=1}^{m} \parallel \mathbf{x}_j \parallel^2 \\ = \; & \sum_{j=1}^{m} \parallel \mathbf{x}_j \parallel^2 - \sum_{j=1}^{m} \sum_{i=1}^{k} (\mathbf{x}_j' \mathbf{w}_i)^2. \end{aligned}$$

The last expression of J can be simplified since

$$(\mathbf{x}_j'\mathbf{w}_i)^2 = (\mathbf{x}_j'\mathbf{w}_i)(\mathbf{x}_j'\mathbf{w}_i) = (\mathbf{w}_i'\mathbf{x}_j)(\mathbf{x}_j'\mathbf{w}_i) = \mathbf{w}_i'(\mathbf{x}_j\mathbf{x}_j')\mathbf{w}_i.$$

Thus, we have

$$\begin{aligned} & J(\mathbf{w}_1, \ldots, \mathbf{w}_k) \\ = & \sum_{j=1}^{m} \parallel \hat{\mathbf{u}}_j \parallel^2 - \sum_{j=1}^{m}\sum_{i=1}^{k} (\mathbf{x}_j'\mathbf{w}_i)^2 \\ = & \sum_{j=1}^{m} \parallel \mathbf{x}_j \parallel^2 - \sum_{i=1}^{k} \mathbf{w}_i' \left(\sum_{j=1}^{m} \mathbf{x}_j\mathbf{x}_j' \right) \mathbf{w}_i. \end{aligned}$$

Let $\hat{X}$ be a centered data matrix

$$\hat{X} = \begin{pmatrix} \mathbf{u}_1' - \tilde{U}' \\ \vdots \\ \mathbf{u}_m' - \tilde{U}' \end{pmatrix} = \begin{pmatrix} \hat{u}_1' \\ \vdots \\ \hat{u_m}' \end{pmatrix}.$$

By applying this evaluation of J to the data set $\{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_m\}$ we have:

$$\begin{aligned} & J(\mathbf{w}_1, \ldots, \mathbf{w}_k) \\ = & \sum_{j=1}^{m} \parallel \mathbf{x}_j \parallel^2 - \sum_{i=1}^{k} \mathbf{w}_i' \left(\sum_{j=1}^{m} \hat{\mathbf{u}}_j \hat{\mathbf{u}}_j' \right) \mathbf{w}_i \\ = & \sum_{j=1}^{m} \parallel \hat{\mathbf{u}}_j \parallel^2 - \sum_{i=1}^{k} \mathbf{w}_i' S \mathbf{w}_i, \end{aligned}$$

where $S = \sum_{j=1}^{m} \hat{\mathbf{u}}_j \hat{\mathbf{u}}_j'$ is the scatter matrix of the data set $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$.

Optimizing J is equivalent to optimizing $\sum_{i=1}^{k} \mathbf{w}_i' S \mathbf{w}_i$ with the equality constraints $\mathbf{w}_i \mathbf{w}_i = 1$ for $1 \leqslant i \leqslant k$.
We use the Lagrangean

$$L(\mathbf{w}_1, \ldots, \mathbf{w}_k) = \sum_{i=1}^{k} \mathbf{w}_i' S \mathbf{w}_i - \sum_{j=1}^{k} \lambda_j (\mathbf{w}_j' \mathbf{w}_j - 1).$$

The portion of ∇L that corresponds to the components of $\mathbf{w}_p$ is denoted by $\frac{\partial L}{\partial \mathbf{w}_p}$.
We have the optimality conditions

$$\frac{\partial L(\mathbf{w}_1, \ldots, \mathbf{w}_k)}{\partial \mathbf{w}_p} = 2S\mathbf{w}_p - 2\lambda_p \mathbf{w}_p = \mathbf{0}.$$

Thus, $(\lambda_p, \mathbf{w}_p)$ must be an eigenpair of S and

$$J(\mathbf{w}_1, \ldots, \mathbf{w}_k) = \sum_{j=1}^{m} \parallel \hat{\mathbf{u}}_j \parallel^2 - \sum_{i=1}^{k} \lambda_i \parallel \mathbf{w}_i \parallel^2 .$$

To minimize J we take the basis the k eigenvectors of S corresponding to the k largest eigenvalues of S. This is tantamount to projecting the data set on the subspace that has the largest variance.

Example

The FAO data set shows the protein and fat consumption for 37 European countries in grams per person per day.
The sample matrix $X \in \mathbb{R}^{37 \times 2}$ is obtained from the second and third columns of this table that correspond to the variables *prot* and *fat*.

Example

code	prot	fat
AL	97	87
AT	107	155
BY	88	97
BE	97	164
BA	86	67
BG	79	101
HR	74	97
CY	99	133
CZ	95	121
DK	108	135
EE	88	96
FI	105	127
FR	117	164
GE	77	58
DE	99	142
GR	117	146
HU	90	145
IS	128	143
IE	115	135
IT	113	158
LV	87	116
LT	112	105
LU	124	164
MK	72	102
MT	116	110
MD	73	59
NL	103	135
NO	104	144
PL	100	113
PT	114	137
RO	110	107
RU	92	87
YU	75	116
SK	72	108
SI	102	131
ES	109	152
CH	91	152

The centered matrix $\hat{X} = H_{37}X$

Rows 1 to 20:

-1.2432	-34.8649
8.7568	33.1351
-10.2432	-24.8649
-1.2432	42.1351
-12.2432	-54.8649
-19.2432	-20.8649
-24.2432	-24.8649
0.7568	11.1351
-3.2432	-0.8649
9.7568	13.1351
-10.2432	-25.8649
6.7568	5.1351
18.7568	42.1351
-21.2432	-63.8649
0.7568	20.1351
18.7568	24.1351
-8.2432	23.1351
29.7568	21.1351
16.7568	13.1351
14.7568	36.1351

Rows 21 to 37

-11.2432	-5.8649
13.7568	-16.8649
25.7568	42.1351
-26.2432	-19.8649
17.7568	-11.8649
-25.2432	-62.8649
4.7568	13.1351
5.7568	22.1351
1.7568	-8.8649
15.7568	15.1351
11.7568	-14.8649
-6.2432	-34.8649
-23.2432	-5.8649
-26.2432	-13.8649
3.7568	9.1351
10.7568	30.1351
-7.2432	30.1351

The standard deviations of the two columns are 15.52 and 28.95. Since the magnitudes of the sample variances are substantial and quite distinct we normalize the data by dividing the columns of X be their respective sample variances.

The Adjusted Matrix $\hat{\hat{X}}$

Rows 1 to 20:

-0.0801	-1.2041
0.5642	1.1444
-0.6599	-0.8588
-0.0801	1.4552
-0.7888	-1.8949
-1.2398	-0.7206
-1.5619	-0.8588
0.0488	0.3846
-0.2090	-0.0299
0.6286	0.4537
-0.6599	-0.8933
0.4353	0.1774
1.2085	1.4552
-1.3687	-2.2057
0.0488	0.6954
1.2085	0.8336
-0.5311	0.7990
1.9172	0.7300
1.0796	0.4537
0.9507	1.2480

Rows 20 to 37:

-0.7244	-0.2026
0.8863	-0.5825
1.6594	1.4552
-1.6908	-0.6861
1.1440	-0.4098
-1.6264	-2.1712
0.3065	0.4537
0.3709	0.7645
0.1132	-0.3062
1.0152	0.5227
0.7575	-0.5134
-0.4022	-1.2041
-1.4975	-0.2026
-1.6908	-0.4789
0.2420	0.3155
0.6930	1.0408
-0.4667	1.0408

The scatter matrix that corresponds to this matrix is

$$S = \hat{\hat{X}}'\hat{\hat{X}} = \begin{pmatrix} 36.0001 & 23.0754 \\ 23.0754 & 36.0001 \end{pmatrix}$$

Its eigenvectors are

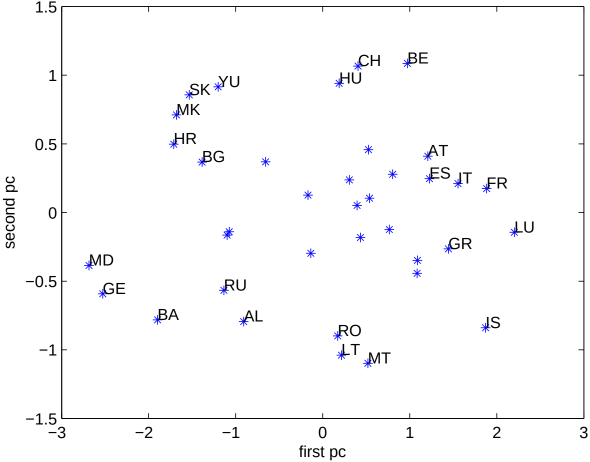
$$\begin{pmatrix} 0.7071 \\ 0.7071 \end{pmatrix} \text{ and } \begin{pmatrix} -0.7071 \\ 0.7071 \end{pmatrix},$$

corresponding to the eigenvalues 59.0755 and 12.9247, respectively.

Both coefficients in the first column (which represents the first principal component) are equal and positive, which means that the first principal component is a weighted average of the two variables.
The second principal component corresponds to a weighted difference of the original variables. The coordinates of the data in the new coordinate system is defined by the matrix scores.

The first two principal components of the FAO dataset

These scores have been ploted below.



- PCA is intended for accurate data representations;
- PCA identifies the directions of maximum and minimum variance.