Projection Operators and Principal Components

Robert M. Haralick

Computer Science, Graduate Center
City University of New York

Outline

Cartesian Products

Definition

The *Cartesian Product* of sets $A_1, \ldots, A_K$ is written as

$$A_1 \times A_2 \times \ldots \times A_K$$

and is defined by

$$A_1 \times A_2 \times \ldots \times A_K = \left\{ \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{pmatrix} \middle| \; x_1 \in A_1, x_2 \in A_2, \ldots, x_K \in A_K \right\}$$

The set A^K is called the *K-fold Cartesian Product* of A.

Euclidean Space

Definition

$\mathbb{R}$ represents the set of all real numbers

Definition

An *N-dimensional Euclidean Space is the set of all N-tuples of real numbers written as* $\mathbb{R}^N$

Definition

The *Dimension* of R^N is N

All the spaces we work with are Euclidean Spaces

Subspace Example

First we have to know what is a space or subspace

A three dimensional Euclidean Space has three kinds of subspaces:

- The zero dimensional point at the origin
- A one dimensional line
  - of infinite extent
  - of arbitrary orientation
  - containing the origin
- A two dimensional plane
  - of infinite extent
  - of arbitrary orientation
  - containing the origin

Scalars and Linear Spaces

Definition

A Scalar is any number from $\mathbb{R}$

Definition

A space $\mathcal{L}$ is called a *Linear Space* if and only if for every scaler α and β

- $x \in \mathcal{L}$ and $y \in \mathcal{L}$ implies that $\alpha x + \beta y \in \mathcal{L}$

Any $x \in \mathcal{L}$ is called a point or vector of $\mathcal{L}$

Subspace

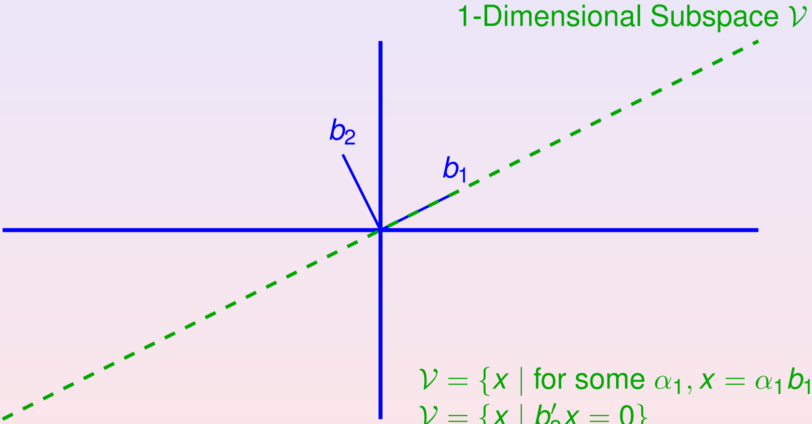
Definition

A subset $\mathcal{V} \subseteq \mathcal{L}$ is called a *Linear subspace* of $\mathcal{L}$ if and only if for every scalars α and β

- $x \in \mathcal{V}$ and $y \in \mathcal{V}$ implies that $\alpha x + \beta y \in \mathcal{V}$

We are only interested in spaces and subspaces that are linear

Representing Subspaces



$$\mathcal{V} = \{x \mid \text{for some } \alpha_1, x = \alpha_1 b_1\}$$
$$\mathcal{V} = \{x \mid b_2' x = 0\}$$

2-Dimensional Space

Linear Combination

Definition

A vector x from a subspace $\mathcal{V}$ is said to be a *linear combination* of vectors $x_1, \ldots, x_K$ if and only if for some scalars $\alpha_1, \ldots, \alpha_K$

$$x = \sum_{n=1}^{N} \alpha_k x_k \tag{1}$$

Linear Independence

Definition

A set of vectors $\{x_1, x_2, \ldots, x_K\}$ from a subspace $\mathcal{V}$ is said to be *Linearly Independent* if and only if for every set of scalars $\{\alpha_1, \ldots, \alpha_K\}$, not all zero,

$$\sum_{k=1}^{K} \alpha_k x_k \neq 0$$

The vectors $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ are linearly independent

Linear Dependence

Definition

A set of vectors $\{x_1, \ldots, x_K\}$ is said to be *Linearly Dependent* if and only if it is not linearly independent.

The vectors $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 \\ 6 \\ 9 \end{pmatrix}$ are linearly dependent

Linear Dependence

Proposition

A set of vectors $\{x_1, x_2, \ldots, x_K\}$ *from a subspace V is Linearly Dependent if for some set of scalars* $\{\alpha_1, \ldots, \alpha_K\}$*, not all zero,*

$$\sum_{k=1}^{K} \alpha_k x_k = 0$$

Span

Definition

The *Span* of a set B,

$$B = \{b_1, \ldots, b_K \mid b_k \in \mathbb{R}^N, k = 1, \ldots, K\}$$

is the set of all linear combinations of vectors from B

- We denote the span of B by Span{B}
- There is no constraint on K relative to N
- Span{B} is a subspace of $\mathbb{R}^N$

Basis

Definition

A set B of vectors

$$B = \{b_1, \ldots, b_K\}$$

is called a *Basis* for the subspace Span{B} if and only if B is a linearly independent set

The vectors $a = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, b = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$ are linearly independent and constitute a basis for the subspace Span$\{a, b\}$

Dimension

Definition

The *Dimension* of a subspace $\mathcal{V}$ is the smallest integer K such that the span of $\{b_1, \ldots, b_K \mid b_k \in \mathcal{V}, k = 1, \ldots, K\}$ satisfies

$$Span\{b_1, ..., b_K\} = \mathcal{V}$$

Proposition

The dimension of a subspace $\mathcal{V}$ is the largest number of vectors from $\mathcal{V}$ such that the vectors are linearly independent.

Inner Product

Definition

The *Inner Product* between two vectors a and b from subspace $\mathcal{V}$ is denoted by $a \cdot b$.

Let $a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_K \end{pmatrix}$ and $b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{pmatrix}$

Then, $a \cdot b = \sum_{k=1}^{K} a_k b_k$
The dot product between vectors a and b can also be written in matrix notation

$$a \cdot b = a' b$$

Orthogonality

Definition

Two vectors $a, b \in \mathcal{V}$ are said to be orthogonal if and only if

$$a \cdot b = 0$$

We express that two vectors a and b are orthogonal by $a \perp b$.

Two spaces $\mathcal{V}$ and $\mathcal{W}$ are said to be orthogonal if and only if for every $v \in \mathcal{V}$ and every $w \in \mathcal{W}$

$$v \cdot w = 0$$

We express that two subspaces $\mathcal{V}$ and $\mathcal{W}$ are orthogonal by $\mathcal{V} \perp \mathcal{W}$

Direct Sum

Definition

Let $\mathcal{V}$ and $\mathcal{W}$ be subspaces of $\mathcal{S}$. The *Direct Sum* of subspaces $\mathcal{V}$ and $\mathcal{W}$, denoted by $\mathcal{V} \oplus \mathcal{W}$, is defined by
$\mathcal{V} \oplus \mathcal{W} = \{x \in S \mid \text{for some } v \in \mathcal{V} \text{ and some } w \in \mathcal{W}, x = v + w\}$

Orthogonal Complement

Definition

Let $\mathcal{V}$ be a subspace of $\mathcal{S}$. The orthogonal complement of $\mathcal{V}$ with respect to $\mathcal{S}$ is denoted by $\mathcal{V}^\perp$ and is defined by

$$\mathcal{V}^\perp = \{w \in \mathcal{S} \mid \text{for every } v \in \mathcal{V},\ w \perp v\}$$

Orthogonal Complement Subspace

Definition

Let $\mathcal{V}$ and $\mathcal{W}$ be two subspaces of $\mathcal{S}$. $\mathcal{W}$ is called the orthogonal complement of $\mathcal{V}$ if and only if $\mathcal{W} = \mathcal{V}^{\perp}$

Proposition

Let $\mathcal{V}$ be a subspace of $\mathcal{S}$. Then

- $\mathcal{V} \perp \mathcal{V}^{\perp}$
- $\mathcal{V} \oplus \mathcal{V}^{\perp} = \mathcal{S}$

Orthogonal Basis

Definition

A basis B for a subspace $\mathcal{V}$ is said to be an orthogonal basis if and only for every $x, y \in B$, $x \neq y$, $x \perp y$

Orthogonal Projection

Definition

Let $\mathcal{V}$ be a subspace of $\mathcal{S}$. Let $x \in \mathcal{S}$ and $x = v + w$ where $v \in \mathcal{V}$ and $w \in \mathcal{V}^{\perp}$. Then v is called the orthogonal projection of x onto $\mathcal{V}$

The orthogonal projection of x is unique.

Proposition

Let $\mathcal{V}$ be a subspace of $\mathcal{S}$. Let $x \in \mathcal{S}$ and $x = v_1 + w_1 = v_2 + w_2$ where $v_1, v_2 \in \mathcal{V}$ and $w_1, w_2 \in \mathcal{V}^{\perp}$ Then $v_1 = v_2$

Proposition

Let $\mathcal{V}$ be a subspace of $\mathcal{S}$ and $x \in \mathcal{S}$. Let $x = v + w$ where $v \in \mathcal{V}$ $w \in \mathcal{V}^{\perp}$. Then $||x||^2 = ||v||^2 + ||w||^2$.

Proof.

Since $x = v + w$,

$$\begin{aligned} ||x||^2 &= ||v + w||^2 \\ &= (v + w)'(v + w) \\ &= v'v + v'w + w'v + w'w \\ &= v'v + w'w \\ &= ||v||^2 + ||w||^2 \end{aligned}$$

$\square$

Projection Operator

Definition

A square matrix P is said to be a projection operator if and only if

$$P^2 = P$$

A square matrix P is said to be an orthogonal projection operator if and only if

$$P^2 = P$$
$$P = P'$$

Projection Operator Examples

Definition

P is called a projection operator if and only if $P^2 = P$

$$\begin{pmatrix} .3 & .7 \\ .3 & .7 \end{pmatrix}\begin{pmatrix} .3 & .7 \\ .3 & .7 \end{pmatrix} = \begin{pmatrix} .3 & .7 \\ .3 & .7 \end{pmatrix}$$

$$\begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} .2 & .4 \\ .4 & .8 \end{pmatrix}\begin{pmatrix} .2 & .4 \\ .4 & .8 \end{pmatrix} = \begin{pmatrix} .2 & .4 \\ .4 & .8 \end{pmatrix}$$

Orthogonal Projection Operator Example

Consider the orthogonal projection operator onto the space spanned by

$$\frac{1}{5}\begin{pmatrix} 3 \\ 4 \end{pmatrix}$$

$$P = \frac{1}{5}\begin{pmatrix} 3 \\ 4 \end{pmatrix}\frac{1}{5}(3\ 4) = \frac{1}{25}\begin{pmatrix} 9 & 12 \\ 12 & 16 \end{pmatrix}$$

$$\frac{1}{25}\begin{pmatrix} 9 & 12 \\ 12 & 16 \end{pmatrix} = \begin{pmatrix} \frac{3}{5} & \frac{-4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \frac{3}{5} & \frac{4}{5} \\ \frac{-4}{5} & \frac{3}{5} \end{pmatrix}$$

Orthogonal Projection Operator

Proposition

Let $\{b_1, b_2, ..., b_K\}$ *be an orthonormal basis for subspace* $\mathcal{V}$*, which is a subspace of an N-dimensional space S. Then the orthogonal projection operator P onto the* K*-dimensional subspace* $\mathcal{V}$ *can be constructed by*

$$P^{N \times N} = B^{N \times K} B'^{\,K \times N}$$

where

$$B = \begin{pmatrix} \vdots & \vdots & & \vdots \\ b_1 & b_2 & \dots & b_K \\ \vdots & \vdots & & \vdots \end{pmatrix}$$

Proposition

Let $b_1, \ldots, b_K$ be an orthonormal basis for the subspace $\mathcal{V}$ Let B be a matrix whose columns are the basis elements. Then, BB' is an orthogonal projection operator onto Col

Proof.

$$(BB')(BB') = B(B'B)B^{=}BB'$$
$$(BB')' = BB'$$

Finally, since any vector x that BB' operates on results in a linear combination of the columns of B, the space that BB' projects to is Col(B) and the columns of B are the orthonormal basis vectors for the subspace $\mathcal{V}$. Hence BB' is the orthogonal projection operator onto the space $\mathcal{V}$. □

Orthogonal Projection Operators are Unique

Proposition

Let Q and P be orthogonal projection operators to the same subspace $\mathcal{V}$. Then $Q = P$

Proof.

Since both P and Q are orthogonal projection operators to the same subspace $\mathcal{V}$, the columns of P and the columns of Q lie in $\mathcal{V}$. Hence $PQ = Q$ and $QP = P$ Since Q is an orthogonal projection operator $Q = Q'$ and $PQ = Q$. Therefore,

$$Q = Q' = (PQ)' = Q'P' = QP = P$$

$\square$

Another Form For Orthogonal Projection Operators

Proposition

Let $b_1, \dots b_K$ be an orthonormal basis for $\mathcal{V}$. Then $\sum_{k=1}^K b_k b_k'$ is an orthogonal projection operator onto the subspace $\mathcal{V}$.

Proof.

$$\sum_{j=1}^K b_j b_j' \sum_{k=1}^K b_k b_k' = \sum_{j=1}^K \sum_{k=1}^K b_j (b_j' b_k) b_k'$$

$$= \sum_{j=1}^K b_k b_k'$$

$$\left(\sum_{k=1}^K b_k b_k' \right)' = \sum_{k=1}^K (b_k b_k')' = \sum_{k=1}^K b_k b_k'$$

It is clear that whenever $\sum_{k=1}^K b_k b_k'$ operates on x, the result is a linear combination of the basis vectors for $\mathcal{V}$. □

Orthogonal Projection Minimizes Error

Theorem

Let $\mathcal{V}$ be a subspace of $\mathcal{S}$. Let $f : \mathcal{S} \to \mathcal{V}$ and $x \in \mathcal{S}$.

$$\min_f (x - f(x))'(x - f(x))$$

is achieved when f is the orthogonal projection operator from $\mathcal{S}$ to $\mathcal{V}$

Proof: Orthogonal Projection Minimizes Error

Proof.

Let $x \in \mathcal{S}$. Then there exists $v \in \mathcal{V}$ and $w \in \mathcal{V}^\perp$ such that $x = v + w$. Consider

$$\begin{aligned}\epsilon^2 &= \|x - f(x)\|^2 \\ &= (x - f(x))'(x - f(x)) \\ &= x'x - (v+w)'f(x) - f(x)'(v+w) + f(x)'f(x)\end{aligned}$$

But $f(x) \in \mathcal{V}$ and $w \in \mathcal{V}^\perp$. Hence $w'f(x) = 0$, therefore

$$\begin{aligned}\epsilon^2 &= x'x - v'f(x) + f(x)'v + f(x)'f(x) \\ &= (v+w)'(v+w) - v'f(x) - f(x)'v + f(x)'f(x) \\ &= v'v + w'w - v'f(x) - f(x)'v + f(x)'f(x) \\ &= (v - f(x))'(v - f(x)) + w'w\end{aligned}$$

ϵ^2 is minimized by making $f(x) = v$, the orthogonal projection of x into $\mathcal{V}$.

Dimensional Reduction by Orthogonal Projection

Corollary

Let $x_1, \ldots, x_K \in \mathcal{S}$. Let $\mathcal{V}$ be a subspace of $\mathcal{S}$. Then

$$\min_{f:\mathcal{S}\to\mathcal{V}} \sum_{k=1}^{K} \|x_k - f(x_k))\|^2$$

is achieved when f is the orthogonal projection operator from $\mathcal{S}$ to $\mathcal{V}$

Proof.

The best f can do for each x_k is for $f(x_k) = v_k$, the orthogonal projection of x_k onto $\mathcal{V}$. Therefore,

$$\min_{f:\mathcal{S}\to\mathcal{V}} \sum_{k=1}^{K} (x_k - f(x_k))'(x_k - f(x_k))$$

is achieved when f is the orthogonal projection operator onto $\mathcal{V}$. □

Trace

Definition

The *Trace* of a $K \times K$ square matrix $A = (a_{ij})$ is defined by

$$Trace(A) = \sum_{k=1}^{K} a_{kk}$$

Proposition

Let $A, B, A_1, \ldots, A_K$ be square $N \times N$ matrices and $\alpha, \beta, \alpha_1, \ldots \alpha_K$ be scalars. Then

- $Trace(AB) = Trace(BA)$
- *Trace is a linear operator*
  - $Trace(\alpha A + \beta B) = \alpha Trace(A) + \beta Trace(B)$
  - $Trace(\sum_{k=1}^{K} \alpha_k A_k) = \sum_{k=1}^{K} \alpha_k Trace(A_k)$

Trace of Orthogonal Projection Operator

Proposition

Let P be an orthogonal projection operator to the M dimensional subspace $\mathcal{V}$. Then $Trace(P) = M$

Proof.

Let $b_1, \ldots, b_M$ be an orthonormal basis for $\mathcal{V}$. Then $P = \sum_{m=1}^{M} b_m b_m'$

$$\begin{aligned} Trace(P) &= Trace(\sum_{m=1}^{M} b_m b_m') \\ &= \sum_{m=1}^{M} Trace(b_m b_m') = \sum_{m=1}^{M} Trace(b_m' b_m) \\ &= \sum_{m=1}^{M} Trace(1) = \sum_{m=1}^{M} 1 = M \end{aligned}$$

$\square$

Kernel and Range

Definition

- The Kernel of a matrix operator A is

$$Kernel(A) = \{x \, | Ax = 0\}$$

- The Range of a matrix operator A is

$$Range(A) = \{y \mid \text{for some } \; x, y = Ax\}$$

Definition

The Column Space of a matrix A is denoted by Col(A) is defined by the Span of its columns.

Proposition

The $Range(A) = Col(A)$

Minkowski Sum

Definition

The Minkowski Sum or simply Sum of two subsets A and B of $\mathcal{S}$ is defined by

$$A \oplus B = \{x \in \mathcal{S} \mid \text{for some } a \in A \text{ and for some } b \in B, x = a + b\}$$

Kernel and Range

Proposition

Let P be a projection operator onto subspace $\mathcal{V}$ of $\mathcal{S}$. Then

$$Range(P) \oplus Ker(P) = \mathcal{S}$$

Proof.

Let $x \in S$. $Px + (I - P)x = Px + x - Px = x$. Certainly $Px \in Range(P)$. Consider $(I - P)x$. $P[(I - P)x] = Px - PPx = Px - Px = 0$ Therefore, by definition of $Kernel(P)$, $(I - P)x \in Kernel(P)$. $\square$

Kernel and Range of Orthogonal Projection Operator

Proposition

Let P be an orthogonal projection operator. Then $Range(P) \perp Kernel(P)$

Proof.

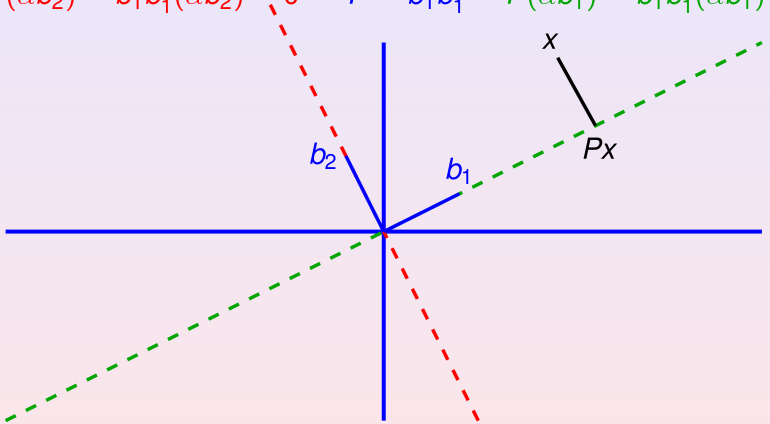
Let $x \in Range(P)$ and $y \in Kernel(P)$. Then for some u, $x = Pu$. Consider $x'y$.

$$x'y = (Pu)'y = u'P'y = u'Py$$

But $y \in Kernel(P)$ so that Py=0. Therefore $x'y = 0$. □

Range And Kernel of Orthogonal Projection Operator

$P(\alpha b_2) = b_1 b_1'(\alpha b_2) = 0$ $\quad P = b_1 b_1'$ $\quad P(\alpha b_1) = b_1 b_1'(\alpha b_1) = \alpha b_1$



$Kernel(P) = \{x \in \mathcal{S} \mid Px = 0\}$

$Range(P) = \{x \in \mathcal{S} \mid \text{for some y}, \; x = P(y)\}$

2-Dimensional Space $\mathcal{S}$

Projection Operator to $\mathcal{V}^\perp$

Proposition

Let P be a projection operator onto the subspace $\mathcal{V}$. (Not necessarily an orthogonal projection operator) Then $I - P$ is the projection operator onto the subspace $\mathcal{V}^\perp$.

Proof.

$$\begin{aligned}(I-P)(I-P) &= I - P - P + P^2 \\ &= I - 2P + P = I - P\end{aligned}$$

$I - P$ is also a projection operator. But what space does it project to?
$\mathcal{V}^\perp = Kernel(P)$. *Let* $x \in \mathcal{V}^\perp$. *Then* $Px = 0$. *Consider*
$(I - P)x = x - Px = x$ $\square$

Orthogonal Projection Operator to $\mathcal{V}^\perp$

Proposition

Let P be the orthogonal projection operator onto the subspace $\mathcal{V}$. Then $I - P$ is the orthogonal projection operator onto the subspace $\mathcal{V}^\perp$.

Proof.

We already know that $(I-P)(I-P) = I - P$. We just have to show that $I - P$ is symmetric and that $I - P$ projects to the $\mathcal{V}^\perp$.

$$(I-P)' = I' - P' = I - P$$

$\mathcal{V}^\perp = Kernel(P)$. *Let* $x \in \mathcal{V}^\perp$. *Then* $Px = 0$. *Consider* $(I-P)x = x - Px = x$ *Every* $x \in \mathcal{V}^\perp$ *gets projected to itself.* □

Covariance Matrix and Expected Value of Squared Length

Definition

The Covariance Matrix of a distribution is defined by

$$\Sigma = E[(x-\mu)(x-\mu)']$$

Proposition

$$Trace\,(\Sigma) = E[||x-\mu||^2]$$

Proof.

$$\begin{aligned}
Trace\,(\Sigma) &= Trace\,(E[(x-\mu)(x-\mu)']) \\
&= E[Trace\,(x-\mu)(x-\mu)')] \\
&= E[Trace\,((x-\mu)'(x-\mu))] \\
&= E[(x-\mu)'(x-\mu)] \\
&= E[||x-\mu||^2]
\end{aligned}$$

Covariance Matrix and Sum of Squared Vector Lengths

Given a sample $x_1, \ldots, x_M$ of N-dimensional vectors, the unbiased estimated of the covariance matrix Σ is given by

$$\Sigma = \frac{1}{M-1} \sum_{m=1}^{M} (x_m - \mu)(x_m - \mu)'$$

where the estimated mean μ is given by

$$\mu = \frac{1}{M} \sum_{m=1}^{M} x_m$$

Then

$$Trace(\Sigma) = \frac{1}{M-1} \sum_{m=1}^{M} ||x_m - \mu||^2$$

Covariance and Sum of Squared Vector Lengths

Proposition

Let Σ be the unbiased estimated covariance matrix. Then

$$Trace(\Sigma) = \frac{1}{M-1}\sum_{m=1}^{M} ||x_m - \mu||^2$$

Proof.

$$Trace(\Sigma) = Trace\left(\frac{1}{M-1}\sum_{m=1}^{M}(x_m-\mu)(x_m-\mu)'\right) = \frac{1}{M-1}Trace\left(\sum_{m=1}^{M}(x_m-\mu)(x_m-\mu)'\right)$$

$$= \frac{1}{M-1}\sum_{m=1}^{M} Trace\left((x_m-\mu)(x_m-\mu)'\right) = \frac{1}{M-1}\sum_{m=1}^{M} Trace\left((x_m-\mu)'(x_m-\mu)\right)$$

$$= \frac{1}{M-1}\sum_{m=1}^{M}(x_m-\mu)'(x_m-\mu) = \frac{1}{M-1}\sum_{m=1}^{M}||x_m-\mu||^2$$

$\square$

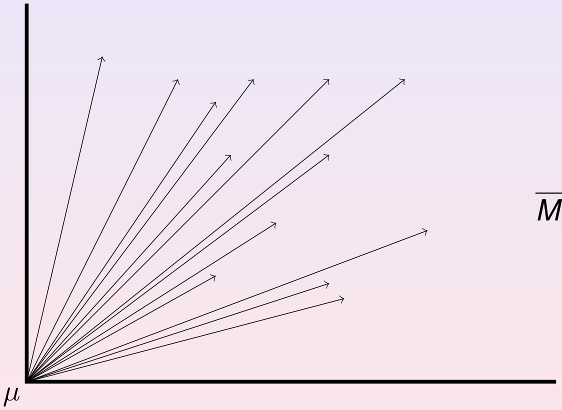
Covariance and Sum of Squared Vector Lengths

The covariance matrix Σ gives information about the spread of the vectors composing it. *Trace*(Σ) is a measure of the total variance. The sum

$$\frac{1}{M-1}\sum_{m=1}^{M}||x_m-\mu||^2$$

is a normalized sum of the squared length of the x_m vectors to the mean μ.

Sum of Squared Vector Lengths



$$\frac{1}{M-1}\sum_{m=1}^{M} ||x_m - \mu||^2$$

Eigen Decomposition of Covariance Matrix

Proposition

Let Σ be the covariance matrix. Let the Eigenvalue Eigenvector decomposition of Σ be $\Sigma = T \Lambda T'$ where $\Lambda = Diagonal(\lambda_1, \ldots, \lambda_N)$ Then,

$$Trace(\Sigma) = \sum_{n=1}^{N} \lambda_n$$

Proof.

$$\begin{aligned} Trace(\Sigma) &= Trace(T \Lambda T') = Trace(\Lambda T T') \\ &= Trace(\Lambda) \\ &= \sum_{n=1}^{N} \lambda_n \end{aligned}$$

Eigen Decomposition of Covariance Matrix

Proposition

Let Σ be the covariance matrix of random vector x. Let the Eigenvalue Eigenvector decomposition of Σ be $\Sigma = T\Lambda T'$ where $\Lambda = Diagonal(\lambda_1, \ldots, \lambda_N)$. Let the n^{th} column of T be t_n. Let $\sigma_n^2 = V[t_n' x]$. Then,

$$\sigma_n^2 = \lambda_n$$

Proof.

$$\begin{aligned}
\sigma_n^2 = V[t_n' x] &= E[(t_n' x - E[t_n' x)^2] \\
&= E[(t_n'(x-\mu))((x-\mu)' t_n)] \\
&= t_n' E[(x-\mu)(x-\mu)'] t_n = t_n' \Sigma t_n \\
&= t_n' T \Lambda T' t_n = (0 \ldots 010 \ldots 0) \Lambda (0 \ldots 010 \ldots 0)' \\
&= \lambda_n
\end{aligned}$$

But if Σ and μ are estimated from the data,

$$Trace(\Sigma) = \frac{1}{M-1}\sum_{n=1}^{N} ||x_m - \mu||^2$$

Therefore, we can conclude that

$$\sum_{n=1}^{N} \sigma_n^2 = \frac{1}{M-1}\sum_{m=1}^{M} ||x_m - \mu||^2$$

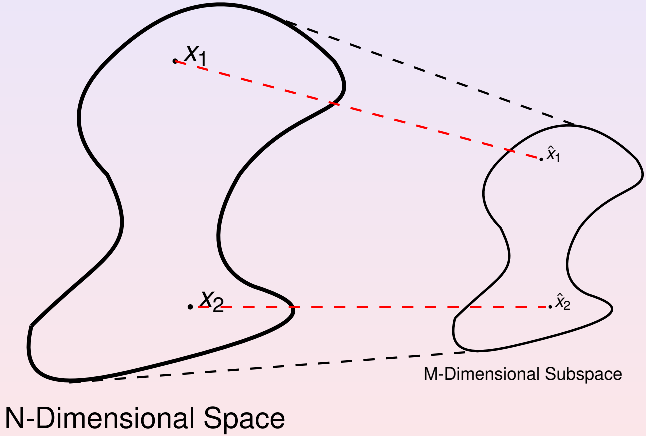
The sum of the Eigenvalues of the Covariance Matrix is the total variance and is equal to the $\frac{1}{M-1}$ of the sum of the squared length of the $(x_m - \mu)$ vectors.

Projecting to Subspaces

Principal Components projects the original data from the larger dimensional space in which it resides to a smaller dimensional space.

- If the decision is to project to a subspace of dimension K, which subspace should be chosen?
- With Principal Components, the K-dimensional space is found that minimizes the sum of the squared distances between the original data vectors and their projection in the K-dimensional subspace.

Space Squeezing: Dimensionality Reduction



Principal Components and Orthogonal Projection Operators

Consider the case for an orthogonal projection operator. It projects a data point or vector to that place in the subspace that is closest to the original point.

Suppose the original data points are N-dimensional. The projection operator projects each point to the closest point to it in the K-dimensional subspace determined by the range of the orthogonal projection operator.

For some subspaces of dimension K the overall distances between the original data points and their projections will be the smallest. This is the one that Principal Components determines.

The Simplest Orthonormal Projection Operators

The simplest othogonal projection operator is a diagonal matrix with some of the entries on the diagonal being 1's and the other entries on the diagonal being 0's.

For example, examine the orthogonal projection operator that projects the data to its first two components, all other components of the projected vector being 0.

$$P = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

Clearly $P = P^2$ and $P = P'$ making it an orthogonal projection operator.

The Orthonormal Matrix

Definition

A square matrix Q is said to be *Orthonormal* if its columns each have norm 1 and each column is orthogonal to every other column.

Proposition

The transpose of an orthonormal matrix is its inverse.

Proof.

Let T be an $N \times N$ orthonormal matrix with columns $t_1, \ldots, t_N$ Then

$$T'T = \begin{pmatrix} \cdots & t_1' & \cdots \\ \cdots & t_2' & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & t_N' & \cdots \end{pmatrix} \begin{pmatrix} \vdots & \vdots & \cdots & \vdots \\ t_1 & t_2 & \cdots & t_N \\ \vdots & \vdots & \cdots & \vdots \end{pmatrix} = I$$

$T'T = I$ so that $T'TT^{-1} = T^{-1}$ Hence,

$T' = T^{-1}$

The General Orthogonal Projection Operator

Proposition

If Q is an orthonormal matrix and P is an orthogonal projection operator, then QPQ' is an orthogonal projection operator.

Proof.

We have to show that QPQ' is idempotent and symmetric. Consider

$$\begin{aligned}(QPQ')(QPQ') &= QP(Q'Q)PQ' \\ &= QPPQ' \\ &= QPQ' \\ (QPQ')' &= QP'Q' \\ &= QPQ'\end{aligned}$$

The General Orthogonal Projection Operator

Proposition

Let P be an orthonormal projection operator. Let Q be an orthonormal matrix. Then QPQ' projects to a subspace of the same dimension as P

Proof.

Since the dimension of the space an orthonormal projection operator projects to is the trace of the operator, we just have to show that $Trace(P) = Trace(QPQ')$

$$\begin{aligned} Trace(QPQ') &= Trace(PQQ') \\ &= Trace(P(QQ')) \\ &= Trace(P) \end{aligned}$$

□

The Orthogonal Projection Operator In Diagonalized Form

Proposition

The form QPQ' can orthogonally project to any given subspace $\mathcal{V}$ with P being a diagonal matrix have ones and zeros on the diagonal.

Proof.

Without loss of generality, we take $P^{N \times N}$ to be a diagonal matrix with the first $M < N$ entries being ones and the remaining diagonal entries zero. The proof is by construction. Let $q_1, \ldots, q_M$ be an orthonormal basis for $\mathcal{V}$. Extend this orthonormal basis to $q_{M+1}, \ldots, q_N$. Define the matrix Q to have columns of $q_1, \ldots, q_N$. Define the orthogonal projection operator P to be a diagonal matrix whose first M diagonal entries are one and all the remaining diagonal entries are zero. $\square$

Proof Continued

Consider QPQ'.

$$
\begin{aligned}
QPQ' &= \begin{pmatrix} \vdots & \vdots & \vdots \\ q_1 & \dots & q_M \\ \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \dots & q_1' & \dots \\ \dots & q_2' & \dots \\ \vdots & \vdots & \vdots \\ \dots & q_M' & \dots \end{pmatrix} \\
&= \sum_{m=1}^{M} q_m q_m'
\end{aligned}
$$

And this is the orthogonal projection operator onto the subspace $\mathcal{V}$

The Principal Component Technique

Let $x_1, \ldots, x_K$ be the observed $N \times 1$ data vectors. First center the data around the mean by subtracting the sample mean vector μ from each of the original data points.

$$\mu = \frac{1}{K} \sum_{k=1}^{K} x_k$$

Define the sample unbiased covariance matrix Σ by

$$\Sigma = \frac{1}{K-1} \sum_{k=1}^{K} (x_k - \mu)(x_k - \mu)'$$

Eigenvector Eigenvalue Decomposition

Σ is an $N \times N$ real symmetric positive semidefinite matrix. Consider the eigenvalue eigenvector decomposition of Σ

$$\Sigma = U \Lambda U'$$

where Λ is a diagonal matrix of eigenvalues and U is an orthonormal matrix. Since Σ is a real symmetric positive semidefinite matrix the eigenvalues are non-negative.

Total Variance

The total variance is given by the trace of Σ. It has the meaning that it is $\frac{1}{K-1}$ times the squared distance between the observed data to the centroid given by the mean.
Note that the trace of Σ is equal to the trace of Λ

$$\begin{aligned} Trace(\Sigma) &= Trace(U\Lambda U') \\ &= Trace(\Lambda U U') \\ &= Trace(\Lambda) \end{aligned}$$

Total Variance

Without loss of generality, we suppose that the diagonal entries are ordered from largest to smallest. Since the eigenvalues are ordered in descending order, the first column of U is that subspace that would have the smallest distance between the observations and the mean vector in the subspace defined by the span of the first eigenvector. Alternatively it is also the subspace whose squared projected lengths is maximal.
The span of the second column of U would be that subspace, orthogonal to the first having the next most smallest squared distance between the observations and the mean vector. And so on.
Because the total variance is fixed, the sum of the squared distances between the data points and their projection are minimized.

Principal Components

Theorem

Let $x_1, \ldots, x_K \in S$ an N-dimensional vector space and Q be an orthogonal projection operator of rank M. Then $\sum_{k=1}^{K} x_k Q x_k$ is maximized when Q projects to the M-Dimensional subspace spanned by the M eigenvectors of $\sum_{k=1}^{K} x_k x_k^{'}$ having largest eigenvalues.

Proof.

Let $\sum_{k=1}^{K} x_k x_k^{'} = TDT^{'}$ and $Q^* = T^{'} QT$. Without loss of generality we assume that the diagonal entries are ordered $d_{ii} \geq d_{jj}$, $i < j$. Then $\max_{Q^*} Trace(Q^* D) = \sum_{m=1}^{M} d_{mm}$, where the maximum is taken over all Q^* satisfying $Q^* = Q^* Q^*$ and $Q^* = Q^{*'}$. Thus, the first M diagonal entries of Q^* are one and the remaining diagonal entries 0. Since $\sum_{i=1}^{N} \sum_{j=1}^{N} q_{ij}^2 = M$, and there are M ones on the diagonal, the remaining elements of Q^* are 0. This implies $Q = TQ^* T^{'}$ is the orthogonal projection operator onto the space spanned by the first M eigenvectors of $\sum_{k=1}^{K} x_k x_k^{'}$ for these are the eigenvectors having largest eigenvalues.

Choice of Best Subspace

In Principal Components the researcher calculates the successive sums of the eigenvalues and compares them to the total sum of the eigenvalues, which is the $Trace(\Lambda)$, and then sets the threshold. Calculate the running sum

$$v_n = \sum_{i=1}^{n} \lambda_n$$

Choice of Best Subspace

Choose the smallest n such that $\frac{v_n}{v_N}$ just exceeds the selected threshold θ. For example the threshold could be set to .85. n is chosen to be the smallest value satisfying

$$\frac{v_n}{V_N} > \theta$$

The subspace projected to is spanned by the first n columns of U. Suppose these n-columns are $u_1, \ldots u_n$ Then the orthogonal projection operator P is defined by

$$P = \sum_{i=1}^{n} u_i u_i'$$

Relative Coordinates

- Suppose every data point is an N-dimensional measurement from space $\mathcal{S}$
- Let $P^{N\times N}$ be a projection operator to M-dimensional subspace $\mathcal{V} \subset \mathcal{S}$
- Suppose $b_1, \ldots b_M$ is any orthonormal basis for $\mathcal{V}$
- The projection operator P is given by

$$P^{N\times N} = \sum_{m=1}^{M} b_m b_m'$$

- $y^{N\times 1} = P^{N\times N} x^{N\times 1}$ is the projection of x into $\mathcal{V}$

Relative Coordinates

- Although y lies in a M-dimensional subspace $\mathcal{V}$, y is an N-dimensional vector
- Since $y \in \mathcal{V}$, we can write $y = \sum_{n=1}^{M} \alpha_m b_m$ since $b_1, \ldots, b_M$ is a basis for $\mathcal{V}$
- The tuple $(\alpha_1, \ldots, \alpha_M)$ is called the relative coordinates of the projection of x
- Let $B^{N \times M}$ be a matrix whose M columns are the basis vectors $b_1, \ldots, b_M$
- The coefficients can be obtained by $(\alpha_1, \ldots, \alpha_M)' = B'x$
- Then the following calculation can produce the orthogonal projection y

$$y^{N \times 1} = B^{N \times M} (\alpha_1, \ldots, \alpha_M)'^{\ M \times 1} = BB'x = Px$$

Principal Components

- Disregarding the class labels, Principle Components selects that K-dimensional subspace having the best fit to the observed measurement vectors
- For each measurement vector, Principal Components computes its relative coordinates in the subspace
- Classification is done using the relative coordinates

Making The Class Assignment

- If the distribution were multivariate normal
- With known mean and known covariance matrix
- d_c^2 has a χ^2 distribution with E_c degrees of freedom
- If the mean is estimated from data with a known covariance matrix d_c^2 has a χ^2 distribution with $E_c - 1$ degrees of freedom
- Assign x to the class c with the smallest squared Mahalanobis distance d_c^2, providing that $d_c^2 < s_{tail,c}$

Problem With Using The Mahalanobis Distance P-value

- It does not include the possibility maximizing economic gain
- Maximizing economic gain is easy with the Discrete Bayes Rule
- The Mahalanobis Distance P-value
  - Produces a real value between 0 and 1
  - The real value has to be converted to an integer to address the class conditional probability table
  - Solution is to quantize the Mahalanobis p-value for each class
- Quantizing
  - Equal Interval Quantizing
  - Equal Probability Quantizing

Equal Interval Quantizing

- Suppose we want K quantizing intervals
- The interval $[0, 1]$ is divided in equal subintervals of size $\frac{1}{K}$
- The quantizing boundaries are $\langle 0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}, 1 \rangle$
- Let p be a p-value
- If $\frac{k}{K} \leq p < \frac{k+1}{K}$ the quantizing index is k
- If $\frac{K-1}{K} \leq p \leq 1$ the quantizing index is K

Equal Probability Quantizing

- Suppose we want K quantizing intervals
- The indexes of the quantizing intervals range are in the set $\{0, 1, \ldots, K-1\}$
- The Training Sequence has Z tuples
- Z is a multiple of K: for some natural integer m, $Z = mK$
- Order the p-values in ascending order $p_{(1)}, p_{(2)}, \ldots p_{(Z)}$
- The quantizing boundaries are $\langle b_0 = 0, b_1, b_2, \ldots, b_{K-1}, b_K = 1 \rangle$
- Where $b_k = p_{(kZ/K)}$, $k \in \{1, \ldots, K-1\}$
- If for some $k \in \{0, \ldots, K-1\}$, $b_k \leq p < b_{k+1}$, the quantizing index is k
- If $p \geq p_{(K-1)}$ the quantizing index is $K-1$

Non-uniform Equal Probability Quantization