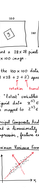
Principal component analysis

Motivation

Very data set has many features, but the geometry that the data points lie on is a manifold of much lower dimensionality than that of the original data space.



Two data velocity vectors are in the dimensional space.

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix}$$

ended = 3D-PCA point using of 1 degree of a 1D-PCA single

In the 1D-PCA data velocity has one $\dot{x}$ (3D $\dot{x}$ $\dot{y}$ $\dot{z}$)

relative translation (latent variables)

The "latent" variables are independent.

Original data $x^{(1)}, x^{(2)}, x^{(3)}$ are correlated

$x^{(1)}$ correlated to $x^{(2)}$ when $\theta = 0$

Principal Components Analysis (PCA)

Used to dimensionality reduction, data compression, feature extraction, data visualization.

Maximum Variance Formulation

$$\begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(N)} \end{pmatrix}$$

Let $D = (x^{(1)}, \dots, x^{(N)})$; where $x^{(i)} \in \mathbb{R}^d$

Project D onto a lower dimensional $M < d$

We are searching the variance of the projected data. We assume M is given a priori.

Consider projecting $x^{(i)} \in \mathbb{R}^d$ onto a one dimensional vector $v_1 \in \mathbb{R}^d$

Without loss of generality, assume $v_1^T v_1 = 1$

Find $x^{(i)}$ in projected data $x_1^{(i)}$ (latent variable)

Variance of projected data

$$v_1 = \frac{1}{N} \sum_{i=1}^{N} \left(v_1^T x^{(i)} - v_1^T \bar{x} \right)^2$$

where $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$

$$v_1 = v_1^T S v_1 \text{ where}$$

$$S = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)} - \bar{x})(x^{(i)} - \bar{x})^T$$

Maximizing $v_1 = v_1^T S v_1$, with respect to the constraint $v_1^T v_1 = 1$

$$L = v_1^T S v_1 + \lambda (1 - v_1^T v_1)$$

$$\frac{\partial L}{\partial v_1} = 2 S v_1 - 2 \lambda v_1 = 0$$

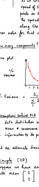
$$\boxed{S v_1 = \lambda v_1} \quad \text{eigenvalue problem}$$

v_1 is an eigenvector of S?

v_1 is an eigenvalue of S?

1. Find S - covariance matrix of transformed D
2. Find λ_1 - the largest eigenvalue of S
3. Find v_1 - the eigenvector corresponding to λ_1 and normalized representation, v_1

$x^{(i)} \in \mathbb{R}^d$ mapped to $\begin{pmatrix} x_1^{(i)} \\ \vdots \\ x_M^{(i)} \end{pmatrix}$



As we write the projection + look at the spread of the projection for that data point on the green line, we can think of it as spread on circumference around the red circle. Arrow where we can take the best optimal radius is by using $\lambda_1 + \lambda_2$.

How many components?

Scree plot

Reconstruction Method PCA

Data distribution is assumed Gaussian. Fully explained by mean & covariance.

Eigenvectors & eigenvalues is the covariance.

Find the eigenvalues from the data. Arrange from largest to smallest.

Example (iris)

Compute an linear transform from a Gaussian distribution with mean $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and covariance $\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$

Subtract the mean.

Now calculate the eigenvalues of the covariance matrix.

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$$

$$|A - \lambda I| = 0$$

$$\begin{vmatrix} 3-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 0$$

$$(3-\lambda)(2-\lambda) - 1 = 0$$

$$(-3\lambda + \lambda^2) - 1 = 0$$

$$\lambda^2 - 3\lambda - 1 = 0$$

$$\lambda = \frac{3 \pm \sqrt{9 - 4 \cdot 1 \cdot -1}}{2} = \begin{cases} 4.1623 \\ 0.3820 \end{cases}$$

Solve for the vectors corresponding to these λs.

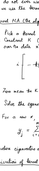
$$\lambda = 4.1623$$

$$(A - \lambda I) v = 0$$

$$\begin{bmatrix} -1.1623 & 1 \\ 1 & -2.1623 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

$$-1.1623 v_1 + 1 v_2 = 0 \qquad v_2 = 0.7236$$

$$1 v_1 - 2.1623 v_2 = 0 \qquad v_1 = 0.1902$$



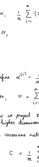
$$\lambda = 0.3820$$

$$(A - \lambda I) v = 0$$

$$\begin{bmatrix} 2.6180 & 1 \\ 1 & 1.6180 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

$$-1.5236 v_1 + 1 v_2 = 0 \qquad v_2 = 1.5236$$

$$1 v_1 + 0.6180 v_2 = 0 \qquad v_1 = -0.6180$$



Variance explained by first component = $\frac{4.1623}{(4.1623 + 0.3820)} = 0.9163$

Variance explained by second component = $\frac{0.3820}{(4.1623 + 0.3820)} = 0.0837$

Example

(1) PCA of Iris data set

(2) Eigenfaces. Given an example of grayscale images of faces (256×256), we can compute faces in the PCA transformed.

(3) Text mining and variance transform of data X

$$\begin{bmatrix} \text{Cat} & \text{Dog} & \text{Bird} & \text{Fish} \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\boxed{S = \frac{X^T X}{(1.1.1.1 - 1)}}$$

Eigenvectors matrix

(4) S - covariance of X

$$S = \frac{1}{n} \sum_{i=1}^{n} x^{(i)} x^{(i)T}$$

Recall PCA

Extract constructed PCA in high dimensional feature space using the kernel trick.

(1) Map each $x^{(i)}$ onto $\phi(x^{(i)})$

where $x^{(i)} \in \mathbb{R}^d$ & $\phi: \mathbb{R}^d \to \mathbb{R}^D$, $D \geq d$

(2) Find principal components in $\mathbb{R}^D$ with the covariance.

(3) The model values projected back to $\mathbb{R}^d$ with the covariance.

We do not want to explicitly compute $\phi(x^{(i)})$

So use the kernel trick.

Recall PCA (the algorithm)

1. Find a kernel
2. Construct K (or the matrix) kernel matrix on the data $x^{(1)}, \dots, x^{(N)}$

$$K = \begin{bmatrix} \phi(x^{(1)}) \\ \vdots \\ \phi(x^{(N)}) \end{bmatrix}$$

3. Form matrix the K matrix to get $\tilde{K}$
4. Solve the eigenvalue problem $\tilde{K} a = \lambda a$, $a \in \mathbb{R}^N$
5. For a new x, we represent (projected) it as

$$y_i(x) = \sum_{n=1}^{N} a_i^{(n)} k(x, x^{(n)}) \quad \text{for } i = 1, \dots, M$$

where eigenvalues are ordered by value.

Derivation of Kernel PCA

We will need to massage the PCA computation in terms of the product.

First, we prove that the eigenvectors lie in the span of the data $\{x^{(1)}, \dots, x^{(N)}\}$

Proof. $S v = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} x^{(i)T} v = \lambda v$

We see the fact that $x^{(i)T} v^{(i)} = (x^{(i)T} v)^{(i)}$ FROM THE IDENTITY

Then, $\frac{1}{N} \sum_{i=1}^{N} x^{(i)} x^{(i)T} v = \lambda v$

$$\sigma_{ij} = \frac{1}{N} \sum_{k=1}^{N} (x_i^{(k)} - \bar{x}_i)(x_j^{(k)} - \bar{x}_j)$$

$\bar{x}_i$ product of i feature & $\bar{x}_j$ product of j feature

Define $x^{(i)T} = \frac{1}{N} (x^{(i)T} v)$

Then, $v = \sum_{i=1}^{N} x^{(i)} a^{(i)}$

We can project $x \in \mathbb{R}^d$ onto $\phi: \mathbb{R}^d \to \mathbb{R}^D$ with a higher dimensional space.

The covariance matrix C of the transformed data is

$$C = \frac{1}{N} \sum_{i=1}^{N} \phi(x^{(i)}) \phi(x^{(i)})^T \quad \text{— (1)}$$

[Assume C is centered]

and we want to solve

$$Cv = \lambda v \qquad v \in \mathbb{R}^D \quad \text{— (2)}$$

for the eigenvectors of C.

We just showed that

$$v = \sum_{i=1}^{N} a^{(i)} \phi(x^{(i)}) \quad \text{— (3)}$$

Substituting (3) in (2) we get

$$Cv = \lambda v$$

$$\left[\frac{1}{N} \sum_{i=1}^{N} \phi(x^{(i)}) \phi(x^{(i)})^T \right] \left[\sum_{j=1}^{N} a^{(j)} \phi(x^{(j)}) \right]$$

$$= \lambda \left[\sum_{j=1}^{N} a^{(j)} \phi(x^{(j)}) \right]$$

Rearranging terms in LHS

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} a^{(j)} \phi(x^{(i)}) \phi(x^{(i)})^T \phi(x^{(j)})$$

dot product

$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} a^{(j)} \phi(x^{(i)}) k(x^{(i)}, x^{(j)})$$

So,

$$\sum_{i=1}^{N} \sum_{j=1}^{N} a^{(j)} \phi(x^{(i)}) k(x^{(i)}, x^{(j)}) = N \lambda \sum_{j=1}^{N} a^{(j)} \phi(x^{(j)})$$

Now we take dot product on both sides with $\phi(x^{(k)})$, getting

$$\sum_{i=1}^{N} \sum_{j=1}^{N} a^{(j)} \phi(x^{(k)}) \phi(x^{(i)}) k(x^{(i)}, x^{(j)}) = N \lambda \sum_{j=1}^{N} a^{(j)} \phi(x^{(k)}) \phi(x^{(j)})$$

dot product

Switching to matrix notation,

$$K^2 a = N \lambda K a$$

We also need to impose the condition that

$$v^T v = 1 \quad \text{which allows us to have}$$

$$a^T K a = 1 \quad \text{— (4)}$$

By multiplying $Ka = \lambda a$ by a we may the normalization condition (4), we get

$$\lambda a^T a = 1$$

For a new point x, the projection with be

$$\phi(x)^T v = \sum_{i=1}^{N} a^{(i)} \phi(x)^T \phi(x^{(i)})$$

dot product

$$= \sum_{i=1}^{N} a^{(i)} k(x, x^{(i)})$$