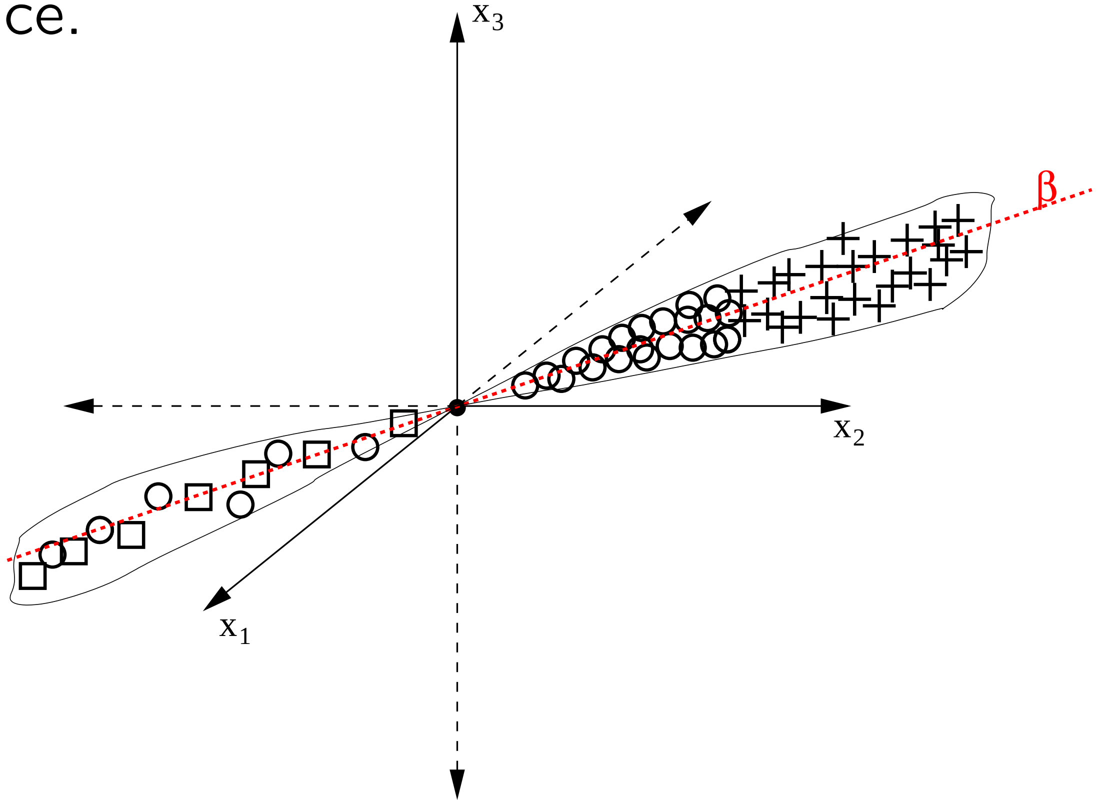
Intelligent Data Analysis

Principal Component Analysis

Peter Tiňo
School of Computer Science
University of Birmingham

Discovering low-dimensional spatial layout in higher dimensional spaces - 1-D/3-D example

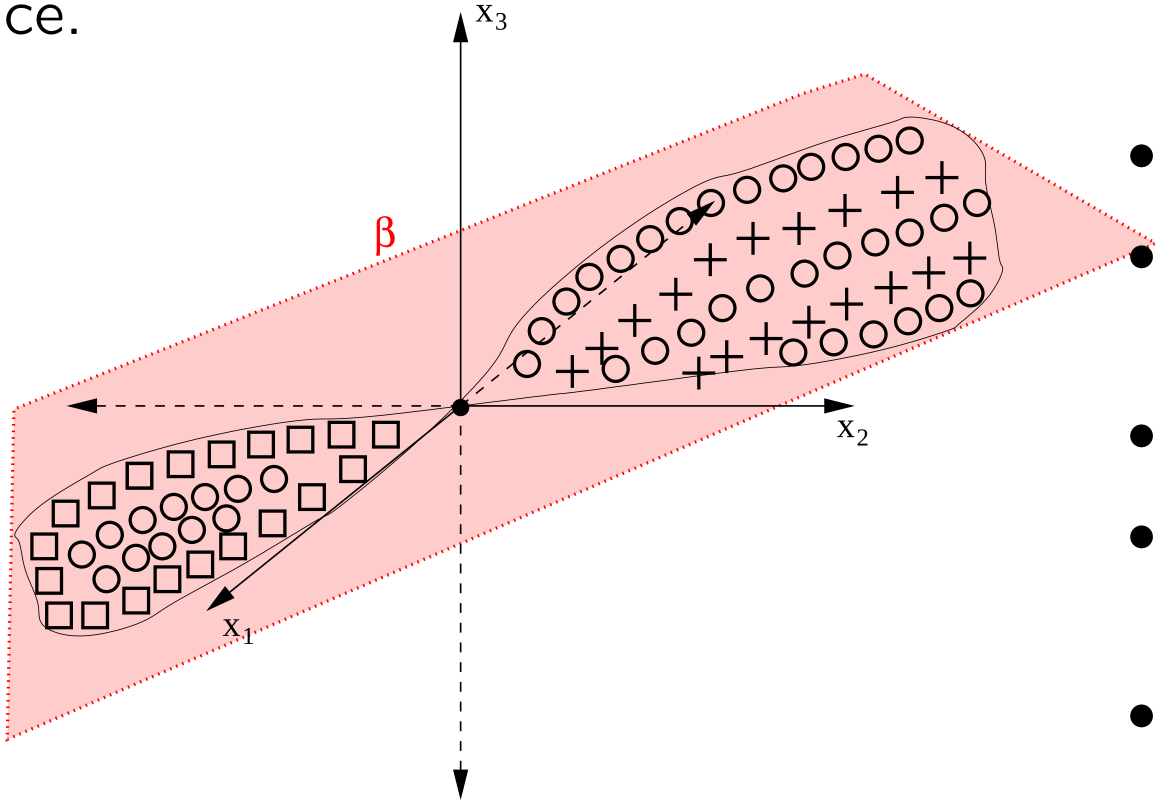
The structure of points $\mathbf{x} = (x_1, x_2, x_3)^T$ in $\mathbb{R}^3$ is inherently 1-dimensional, but the points are (linearly) embedded in a 3-dimensional space.



- 3 types of points $\mathbf{x}$
- What is the best 1-D projection direction?
- Why is β a good choice?
- Try to formalise your intuition ...
- Draw the 1-D projections

Discovering low-dimensional spatial layout in higher dimensional spaces - 2-D/3-D example

The structure of points $\mathbf{x} = (x_1, x_2, x_3)^T$ in $\mathbb{R}^3$ is inherently 2-dimensional, but the points are (linearly) embedded in a 3-dimensional space.



- 3 types of points $\mathbf{x}$
- What is the best 2-D projection direction?
- Why is β a good choice?
- Try to formalise your intuition ...
- Draw the 2-D projections

Random variables (RV)

Consider a random variable X taking on values in $\mathbb{R}$.
$x \in \mathbb{R}$ are realisations of X

N repeated i.i.d. draws from X:
Imagine N independent and identically distributed random variables $X^1, X^2, ..., X^N$. x^i is a realisation of X^i, $i = 1, 2, ..., N$.

Continuous RV:
Realisations are from a continuous subset A of $\mathbb{R}$.
Probability density $p(x)$: $\int_A p(x)\ \mathrm{d}x = 1$.

Discrete RV:
Realisations are from a discrete subset A of $\mathbb{R}$.
Probability distribution $P(x)$: $\sum_{x \in A} P(x) = 1$.

Characterising random variables

Mean of RV X: Center of gravity around which realisations of X happen. First central moment.

$$E[X] = \sum_{x \in A} x \cdot P(X = x) \qquad \text{or} \qquad E[X] = \int_A x \cdot p(x) \; \mathrm{d}x$$

Variance of RV X: (Squared) fluctuations of realisations x around the center of gravity $E[X]$. Second central moment.

$$Var[X] = E[(X - E[X])^2] = \sum_{x \in A} (x - E[X])^2 \cdot P(X = x),$$

or

$$Var[X] = E[(X - E[X])^2] = \int_A (x - E[X])^2 \cdot p(x) \; \mathrm{d}x$$

Estimating central moments of X

N i.i.d. realisations of X:
$x^1, x^2, ..., x^N \in \mathbb{R}$.

$$E[X] \approx \widehat{E[X]} = \frac{1}{N} \sum_{i=1}^{N} x^i$$

$$Var[X] \approx \widehat{Var[X]}_{ML} = \frac{1}{N} \sum_{i=1}^{N} \left(x^i - \widehat{E[X]} \right)^2$$

Unbiased estimation of variance:

$$\frac{1}{N-1} \sum_{i=1}^{N} \left(x^i - \widehat{E[X]} \right)^2$$

Several random variables

Consider 2 RVs X and Y

Still can compute central moments of individual RVs,
i.e. $E[X]$, $E[Y]$ and $Var[X]$, $Var[Y]$.

In addition we can ask whether X and Y are ‘statistically tight together’ in some way

Covariance of RVs X and Y

Co-fluctuations around the means:
Introduce a new random variable $Z = (X - E[X]) \cdot (Y - E[Y])$

$$Cov[X, Y] = E[Z] = E[(X - E[X]) \cdot (Y - E[Y])]$$

Estimating covariance of X and Y

N i.i.d. realisations of (X,Y):
$(x^1,y^1),(x^2,y^2),...,(x^N,y^N) \in \mathbb{R}^2$.

$$Cov[X,Y] \approx \widehat{Cov[X,Y]} = \frac{1}{N}\sum_{i=1}^{N}\left(x^i - \widehat{E[X]}\right)\cdot\left(y^i - \widehat{E[Y]}\right)$$

For centred RVs (means are 0), we have

$$\widehat{Cov[X,Y]} = \frac{1}{N}\sum_{i=1}^{N} x^i y^i$$

Covariance matrix of X, Y

Note that formally

$Var[X] = Cov[X, X]$

and

$\widehat{Var[X]} = \widehat{Cov[X, X]}$.

Covariance matrix summarises the variance/covariance structure in (X, Y):

$$\begin{bmatrix} Var[X] & Cov[X, Y] \\ Cov[Y, X] & Var[Y] \end{bmatrix}$$

Covariance matrix of a vector RV

Consider a vector random variable
$\mathbf{X} = (X_1, X_2, ..., X_n)^T$

Covariance matrix of $\mathbf{X}$ is

$$Cov[\mathbf{X}] = \begin{bmatrix} Var[X_1] & Cov[X_1, X_2] & Cov[X_1, X_3] & ... & Cov[X_1, X_n] \\ Cov[X_2, X_1] & Var[X_2] & Cov[X_2, X_3] & ... & Cov[X_2, X_n] \\ Cov[X_3, X_1] & Cov[X_3, X_2] & Var[X_3] & ... & Cov[X_3, X_n] \\ . & . & . & ... & . \\ . & . & . & ... & . \\ Cov[X_n, X_1] & Cov[X_n, X_2] & Cov[X_n, X_3] & ... & Var[X_n] \end{bmatrix}$$

Note that $Cov[\mathbf{X}]$ is square and symmetric.

Estimating $Cov[\mathbf{X}]$

N i.i.d. realisations of the vector RV $\mathbf{X} = (X_1, X_2, ..., X_n)^T$:
$\mathbf{x}^1 = (x_1^1, x_2^1, ..., x_n^1)^T, \mathbf{x}^2 = (x_1^2, x_2^2, ..., x_n^2)^T, ..., \mathbf{x}^N = (x_1^N, x_2^N, ..., x_n^N)^T$.

Collect the realisations $\mathbf{x}^i$ of $\mathbf{X}$ as columns of the design matrix $\mathcal{X} = [\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N]$.

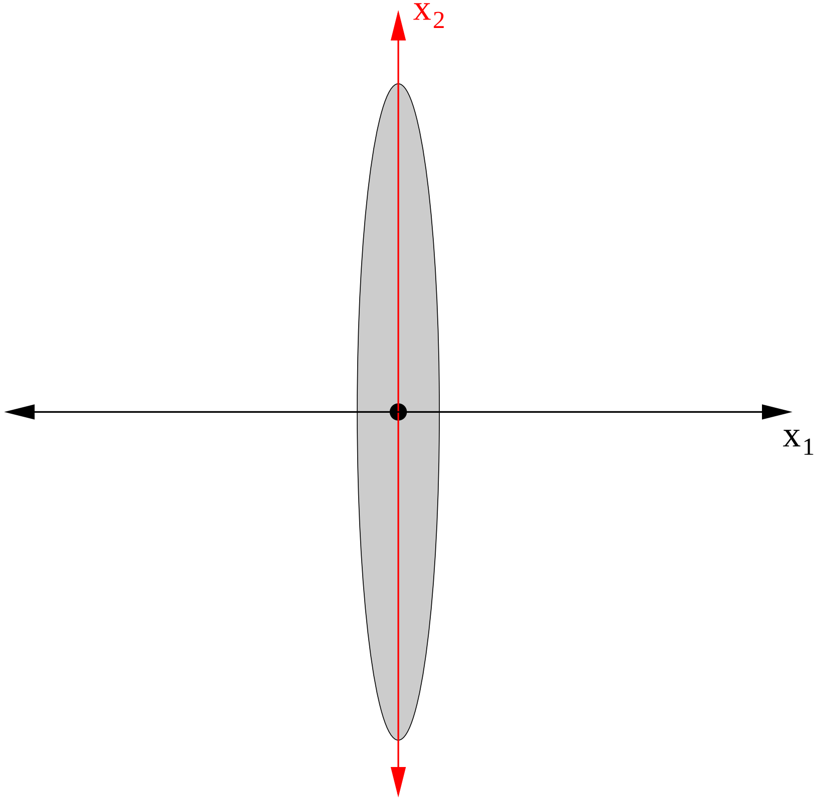
Assume the RV $\mathbf{X}$ is centred ($E[X_i] = 0,\ i = 1, 2, ..., n$).

Then

$$Cov[\mathbf{X}] \approx \widehat{Cov[\mathbf{X}]} = \frac{1}{N} \mathcal{X} \mathcal{X}^T$$

A 2-D example

$$Cov[\mathbf{X}] = \begin{bmatrix} Var[X_1] & Cov[X_1, X_2] \\ Cov[X_2, X_1] & Var[X_2] \end{bmatrix}$$



- $Cov[X_1, X_2] = 0$
- $Var[X_1] << Var[X_2]$
- Model:
  $Var[X_2] = V$,
  $Var[X_1] = \alpha \cdot V$,
  $0 < \alpha << 1$.

2-D example - Continued

Note

$$Cov[\mathbf{X}] \begin{bmatrix} 0 \\ 1 \end{bmatrix} = V \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

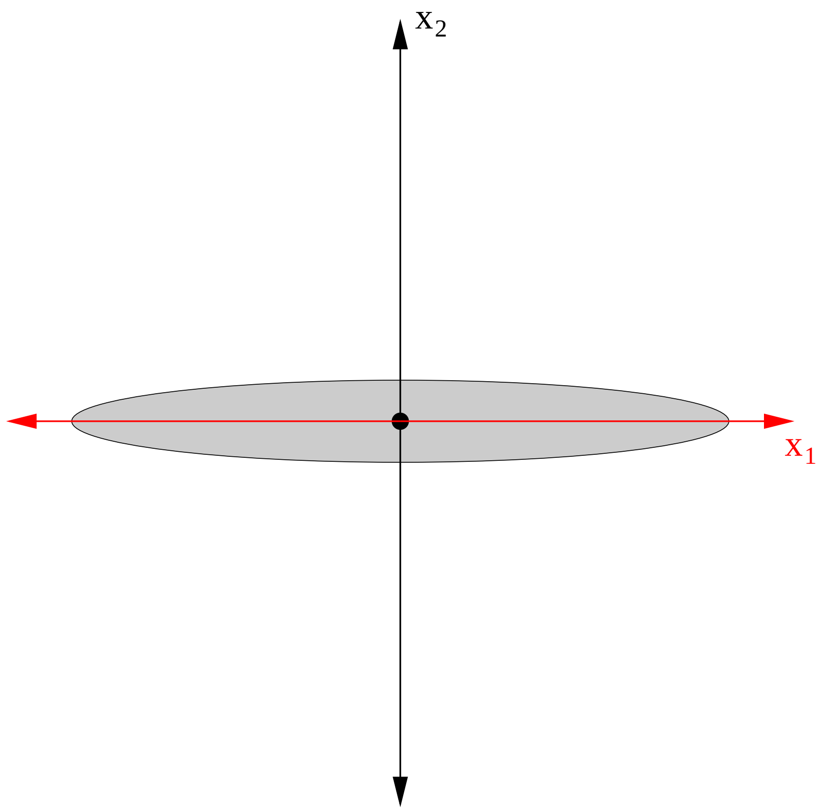
$$Cov[\mathbf{X}] \begin{bmatrix} 1 \\ 0 \end{bmatrix} = (\alpha V) \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Directions of both $(0,1)^T$ and $(1,0)^T$ are preserved by applying $Cov[\mathbf{X}]$ as a linear operator, but since

$$\alpha \cdot V << V,$$

the image of $(1,0)^T$ is much shorter than that of $(0,1)^T$.

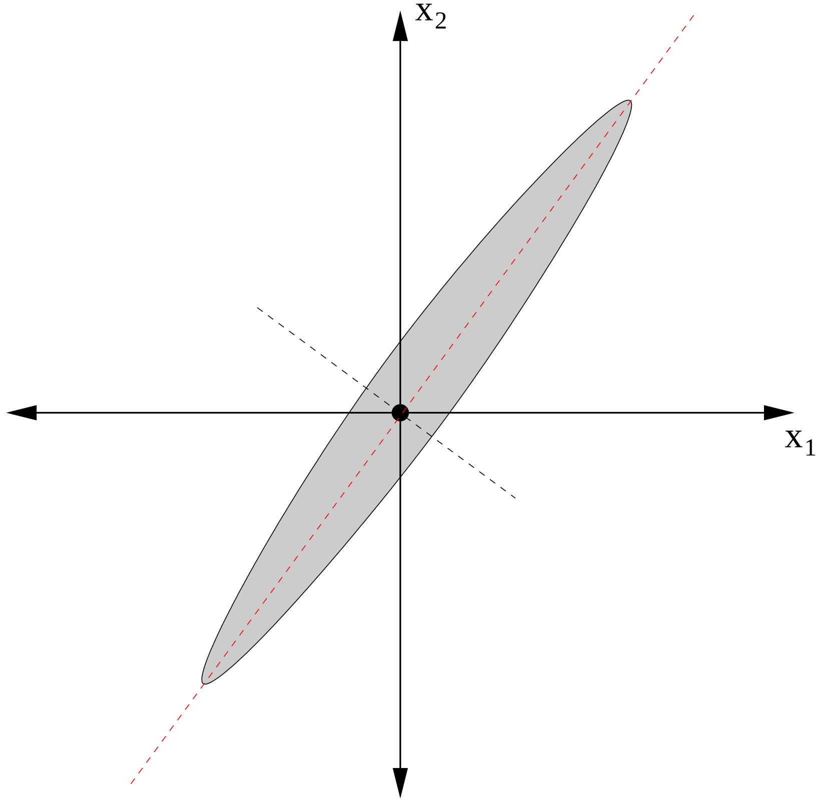
Another 2-D example



- $Cov[X_1, X_2] = 0$
- $Var[X_1] >> Var[X_2]$
- $\alpha >> 1$
- $Cov[\mathbf{X}] \begin{bmatrix} 1 \\ 0 \end{bmatrix} = (\alpha V) \begin{bmatrix} 1 \\ 0 \end{bmatrix}$
- $Cov[\mathbf{X}] \begin{bmatrix} 0 \\ 1 \end{bmatrix} = V \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

This time, the image of $(1, 0)^T$ is much longer than that of $(0, 1)^T$.

Yet another 2-D example



- $Var[X_1] = Var[X_2] = V$
- $Cov[X_1, X_2] = \alpha \cdot V$
- $0 < \alpha < 1$
- $Cov[\mathbf{X}] \begin{bmatrix} 1 \\ 0 \end{bmatrix} = V \begin{bmatrix} 1 \\ \alpha \end{bmatrix}$
- $Cov[\mathbf{X}] \begin{bmatrix} 0 \\ 1 \end{bmatrix} = V \begin{bmatrix} \alpha \\ 1 \end{bmatrix}$

Directions of the coordinate axis are not preserved by the action of $Cov[\mathbf{X}]$.

2-D example - Continued

But

$$Cov[\mathbf{X}] \begin{bmatrix} 1 \\ 1 \end{bmatrix} = (1+\alpha)V \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$Cov[\mathbf{X}] \begin{bmatrix} 1 \\ -1 \end{bmatrix} = (1-\alpha)V \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Note: $(1+\alpha)V > (1-\alpha)V$

Directions invariant to the action of $Cov[\mathbf{X}]$ correspond to the 'principal variance directions'. The corresponding multiplicative constants quantify the extent of variation along the invariant directions.

Eigenvalues and eigenvalues of a symmetric positive definite matrix

Consider an $n \times n$ symmetric positive definite matrix $\mathcal{A}$.

A vector $\mathbf{v} \in \mathbb{R}^n$, such that

$$\mathcal{A}\mathbf{v} = \lambda \mathbf{v}$$

is an eigenvector of $\mathcal{A}$ and the corresponding scalar $\lambda > 0$ is the eigenvalue associated with $\mathbf{v}$.

Eigenvectors (normalized to unit length) – the invariant directions in $\mathbb{R}^n$ when considering the matrix $\mathcal{A}$ as a linear operator on $\mathbb{R}^n$.

Magnitudes of the eigenvalues – quantify the ranges of magnification/contraction along the invariant directions.

PCA for dimensionality reduction

1. Given N centred data points $\mathbf{x}^i = (x_1^i, x_2^i, ..., x_n^i)^T \in \mathbb{R}^n$, $i = 1, 2, ..., N$, construct the design matrix $\mathcal{X} = [\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N]$.

2. Estimate the covariance matrix: $\mathcal{C} = \frac{1}{N}\mathcal{X}\mathcal{X}^T$

3. Compute eigen-decomposition of $\mathcal{C}$.
   All eigenvectors $\mathbf{v}_j$ are normalized to unit length.

4. Select only the eigenvectors $\mathbf{v}_j$, $j = 1, 2, ..., k < n$, with large enough eigenvalues λ_j.

5. Project the data points $\mathbf{x}^i$ to the hyperplane defined by the span of the selected eigenvectors $\mathbf{v}_j$: $\tilde{x}_j^i = \mathbf{v}_j^T \ \mathbf{x}^i$

Amount of variance explained in the projections $\tilde{\mathbf{x}}^i$: $\frac{\sum_{\ell=1}^{k} \lambda_\ell}{\sum_{\ell=1}^{n} \lambda_\ell}$

Data visualization using PCA

Select the 2 eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ with the largest eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n$$

Represent points $\mathbf{x}^i \in \mathbb{R}^n$ by two-dimensional projections $\tilde{\mathbf{x}}^i = (\tilde{x}_1^i, \tilde{x}_2^i)^T$, where

$$\tilde{x}_j^i = \mathbf{v}_j^T \, \mathbf{x}^i, \quad j = 1, 2$$

Plot the projections $\tilde{x}_j^i$ on the computer screen.

You may use other eigenvectors $\mathbf{v}_j$ with large enough eigenvalues λ_j