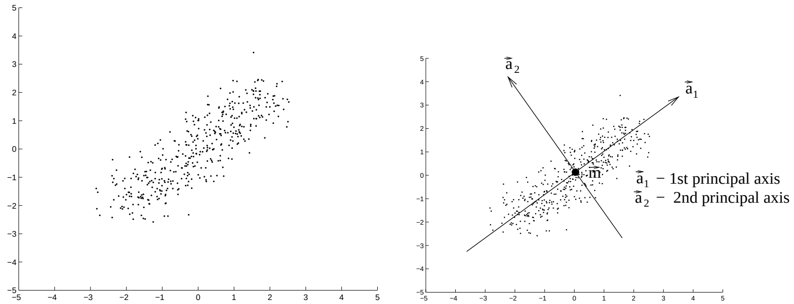
Principal Component Analysis

Robert B. Fisher
School of Informatics
University of Edinburgh

? Component Analysis

Given a set of D dimension points $\{\vec{x}_i\}$ with mean $\vec{m}$



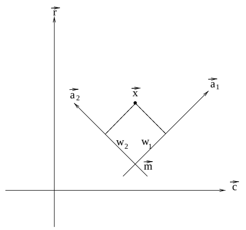
Find a new set of D perpendicular coordinate axes $\{\vec{a}_j\}$ such that

$$\vec{x}_i = \vec{m} + \sum_j w_{ij}\vec{a}_j$$

Ie. point $\vec{x}_i$ represented as a mean plus weighted sum of axis directions

Transforming points to the new representation

Transforming ? is easy as $\vec{a}_k \cdot \vec{a}_j = 0$ and $\vec{a}_k \cdot \vec{a}_k = 1$ for $k \neq j$

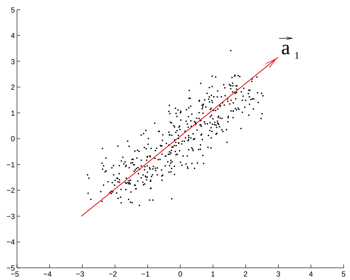


Computing w_{ik} is easy:

$$\vec{a}_k \cdot (\vec{x}_i - \vec{m}) = \vec{a}_k \cdot \sum_j w_{ij}\vec{a}_j = \sum_j w_{ij}\vec{a}_k \cdot \vec{a}_j = w_{ik}$$

How to do PCA I

1. Choose axis $\vec{a}_1$ as the direction of the most ? in the dataset:



2. Project each $\vec{x}_i$ onto a $D-1$ dimensional subspace perpendicular to $\vec{a}_1$ (ie removing the component of variation in direction $\vec{a}_1$) to give $\vec{x}_i'$
3. Calculate the axis $\vec{a}_2$ as the direction of the most remaining variation in $\{\vec{x}_i'\}$

4. Project each $\vec{x}_i'$ onto a $D-2$ dimension $\boxed{?}$
5. Continue like this until all D new axes $\vec{a}_i$ are found.

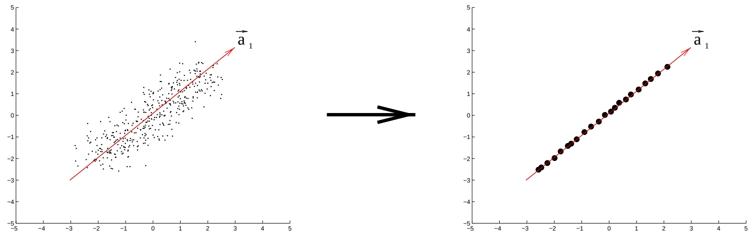
Why PCA

Many possible axis sets $\{\vec{a}_i\}$

$\boxed{?}$ chooses axis directions $\vec{a}_i$ in order of largest remaining variation

Gives an ordering on dimensions from most to least significant

Allows us to omit low significance axes. Eg, projecting $\vec{a}_2$ gives:



How to Do PCA II

Via Eigenanalysis
Given N D-dimensional points $\{\vec{x}_i\}$

1. Mean $\vec{m} = \frac{1}{N}\sum_i \vec{x}_i$
2. Compute $\boxed{?}$ matrix
   $\mathrm{S} = \sum_i (\vec{x}_i - \vec{m})(\vec{x}_i - \vec{m})'$
3. Compute Singular Value Decomposition (SVD): S = U D V', where D is a diagonal matrix and U' U = V' V = I
4. PCA: i^{th} column of V is axis $\vec{a}_i$ (i^{th} eigenvector of S)
   d_{ii} of D is a measure of significance (i^{th} eigenvalue)

What We Have Learned

1. Using $\boxed{?}$ to find the 'natural' axes of a dataset
2. Algorithm for computing PCA