Learning from Data: Dimensionality Reduction

Amos Storkey, School of Informatics
University of Edinburgh

Semester 1, 2004

Dimensionality Reduction

- **Goal**: to construct new representations of the data that capture its underlying structure

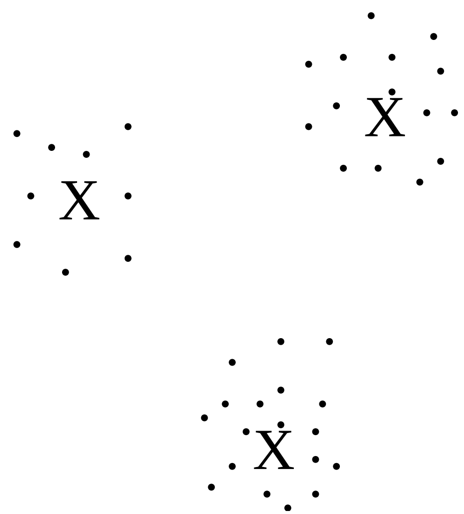
- Presumed that the the inherent (useful) structure of the data does not fill the whole of the space.

- Don't forget the size of these spaces. 4000 data points. 12 attributes. Many quadrants of the space must have 0 data points in them (2^{12} quadrants in all).

- Often choose attributes with some conceptual overlap.

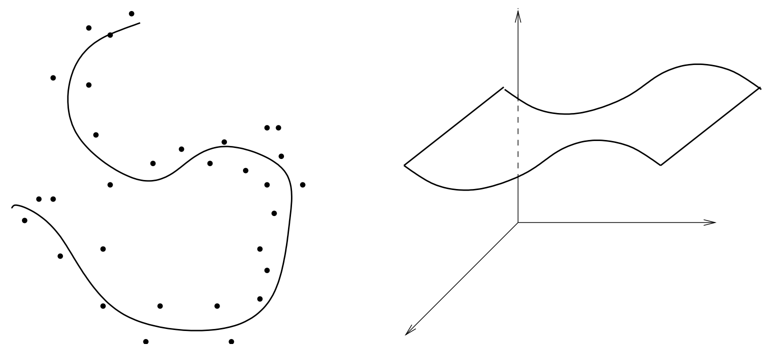
Lower Dimensional Structures

- Some lower dimensional structures in a higher-dimensional space e.g.
- Cluster centres (points in 0-d)



Lower Dimensional Structures

- Some lower dimensional structures in a higher-dimensional space e.g.
- Lower-dimensional manifolds, e.g. lines, sheets (1-d, 2-d)



Linear dimensionality reduction

- If lines or surfaces are linear manifolds.
- Straight lines, Flat sheets.
- Want to find the positions of those flat sheets
- This is linear dimensionality reduction.

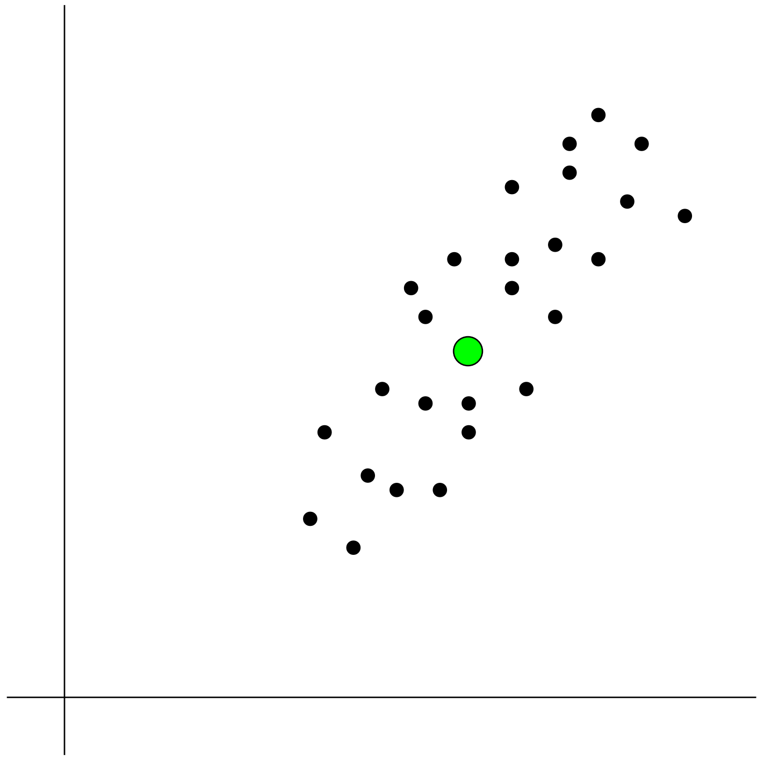
Exploratory data analysis

- Related idea, understand structure in data.
- See what you get if you reduce dimensionality to visualisable levels.

Covariance Matrix: Variance

- Let $\langle \; \rangle$ denote an average
- Suppose we have a random vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$
- $\langle \mathbf{x} \rangle$ denotes the mean of $\mathbf{x}$, $(\mu_1, \mu_2, \ldots \mu_d)^T$
- $\sigma_{ii} = \langle (x_i - \mu_i)^2 \rangle$ is the variance of component i (gives a measure of the "spread" of component i)

Covariance Matrix: Illustration



Covariance Matrix: Calculation

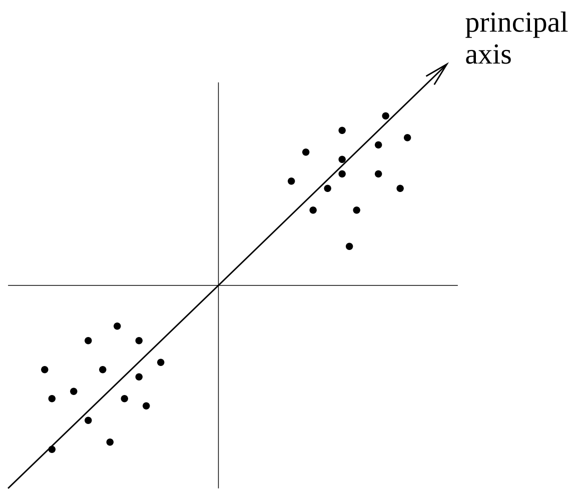
- $\sigma_{ij} = \langle (x_i - \mu_i)(x_j - \mu_j) \rangle$ is the covariance between components i and j
- In d-dimensions there are d variances and $d(d-1)/2$ covariances which can be arranged into a *covariance matrix* C

$$C = \langle (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \rangle$$

- Covariance matrix is symmetric
- E.g. Weight and Height
- Highly correlated variables say the same thing, there is redundancy to be removed

Principal Components Analysis

- A linear dimensionality reduction technique



One view of PCA

- If you want to use a single number to describe a whole vector drawn from a known distribution, pick the projection of the vector onto the direction of maximum variation (variance)
- Assume $\langle \mathbf{x} \rangle = \mathbf{0}$
- $y = \mathbf{w}.\mathbf{x}$
- Choose $\mathbf{w}$ to maximise $\langle y^2 \rangle$, subject to $\mathbf{w}.\mathbf{w} = 1$
- Solution: $\mathbf{w}$ is the eigenvector corresponding to the largest eigenvalue of $C = \langle \mathbf{x}\mathbf{x}^T \rangle$

More Generally

- Want to write

$$\mathbf{x}_i = c + \sum_{k=1}^{M} w_i^k \mathbf{b}^k + \boldsymbol{\epsilon}_i$$

- The vectors $\{\mathbf{b}^k, k = 1, \ldots, M\}$ are orthonormal. That is

$$(\mathbf{b}^i)^T \mathbf{b}^j = \delta^{ij}$$

- Want to choose the set $\{\mathbf{b}^k, k = 1, \ldots, M\}$ to minimise the size of the error terms $\boldsymbol{\epsilon}_i$.
- I.e. Min $\sum_i \boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i$.

Solution

- Solution is to choose $\mathbf{b}$ to be given by:
  - Calculating the *sample* mean and covariance of the data:

$$m = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k, \quad \text{and} \quad S = \frac{1}{N-1}\sum_{k=1}^{N}(\mathbf{x}_k - m)(\mathbf{x}_k - m)^T$$

  - Calculating the eigenvalues λ_i of the sample covariance matrix (use `eig` in Matlab).
  - Ordering λ_i in descending order, and finding the M largest eigenvalues
  - Setting $\mathbf{b}^k$ to be the eigenvector corresponding to the kth largest eigenvalue.

Solution

- Then the span of the vectors $\mathbf{b}_i$ are the *principal subspace*
- Set $\mathbf{c} = \mathbf{m}$
- $w_i^k = (\mathbf{b}^k)^T(\mathbf{x}_i - \mathbf{m})$ is the lower dimensional representation of data point $\mathbf{x}_i$. This is the projection to the principal linear manifold.
- For details of the derivation see the handout.
- Fraction of total variation explained by using M principal components is

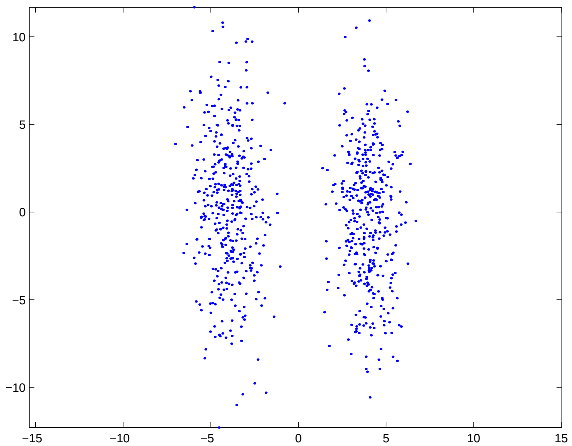
$$\frac{\sum_{i=1}^{M} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \leq 1$$

Example

- Handwritten Characters
- See handout.
- Can summarise much of data using principal components.
- Captures the essence of the character.

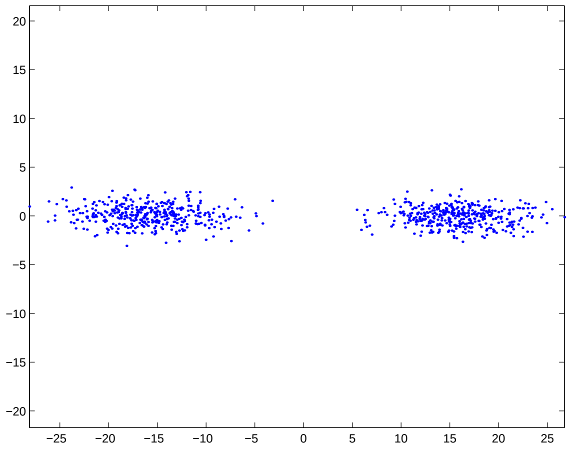
Issues

- Inherent dimensionality?
- Usefulness.
- Scaling dependent.



Issues

- Inherent dimensionality?
- Usefulness.
- Scaling dependent.



Summary

- Dimensionality reduction
- Linear manifolds
- Covariance matrix
- PCA as finding largest eigenvalues