Data Mining and Machine Learning

Introduction to Data Mining, Vector Data Analysis and Principal Components Analysis (PCA)

Objectives

- To introduce Data Mining
- To outline the techniques that we will study in this part of the course – a Data Mining 'Toolkit'
- To review basic data analysis and to review the notions of mean, variance and covariance
- To explain Principal Components Analysis (PCA)
- To present an example of PCA

What is Data Mining?

- Mining
  - *Digging deep into the earth, to find hidden, valuable materials*
- Data Mining
  - Analysis of large data corpora: biomedical, acoustic, video, text,… to discover structure, patterns and relationships
  - Corpora which are too large for human inspection
  - Patterns and structure may be hidden

Data Mining

- Structure and patterns in large, abstract data sets:
  - Is the data homogeneous or does it consist of several separately identifiable subsets?
  - Are there patterns in the data?
  - If so, do these patterns have an intuitive interpretation?
  - Are there correlations in the data?
  - Is there redundancy in the data?

Data Mining

- In this part of the course we will develop a basic 'data mining toolkit'
  - Subspace projection methods (PCA)
  - Clustering
  - Statistical modelling
  - Sequence analysis
    - Dynamic Programming (DP)

Some example data

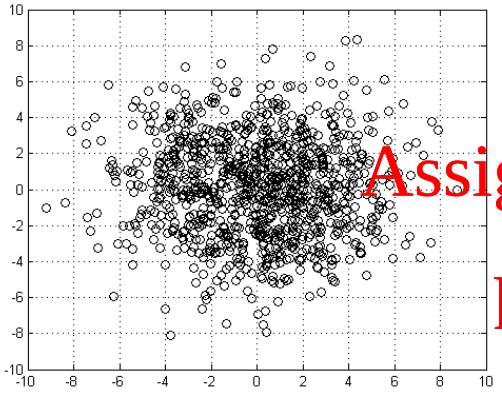


Fig 1: Single, spherical cluster centred at origin

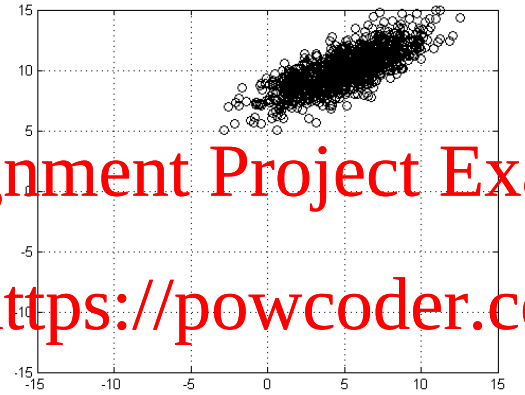


Fig 2: Single, arbitrary elliptical cluster

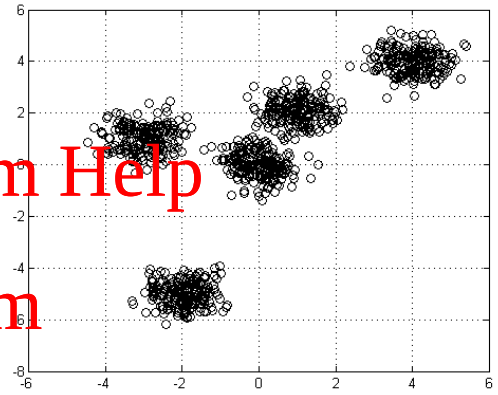
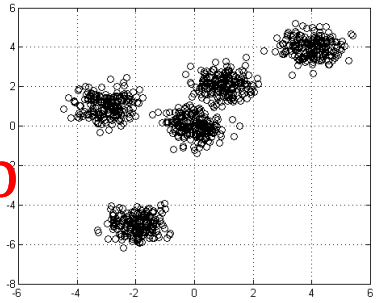


Fig 3: Multiple, arbitrary elliptical clusters

Objectives

- Fig 3 shows "multiple source" data. The data is arranged in a set of "clusters".
- How do we discover the number and locations of the clusters?
- Remember, in real applications there will be many points in a high-dimensional vector space which is difficult to visualise



Objectives

- Fig 1 shows simplest type of data – single source data centred at origin. Equal variance in both dimensions and no covariance.
- Fig 2 is again single source, but the data is correlated and skewed and not centred at the origin.
- How do we convert Fig 2 into Fig 1?
- We will start with this problem
- Solution is a technique called Principal Components Analysis (PCA)

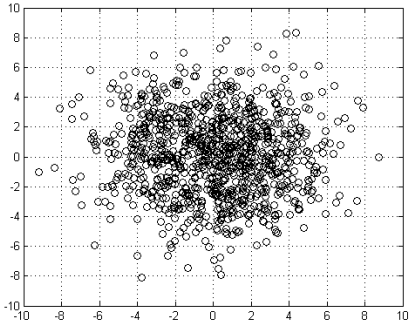


Fig 1

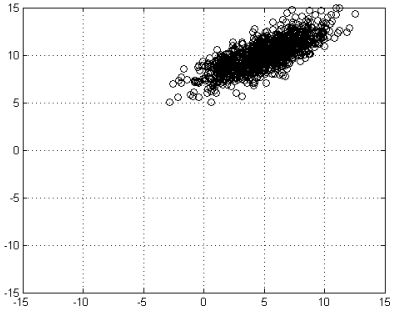
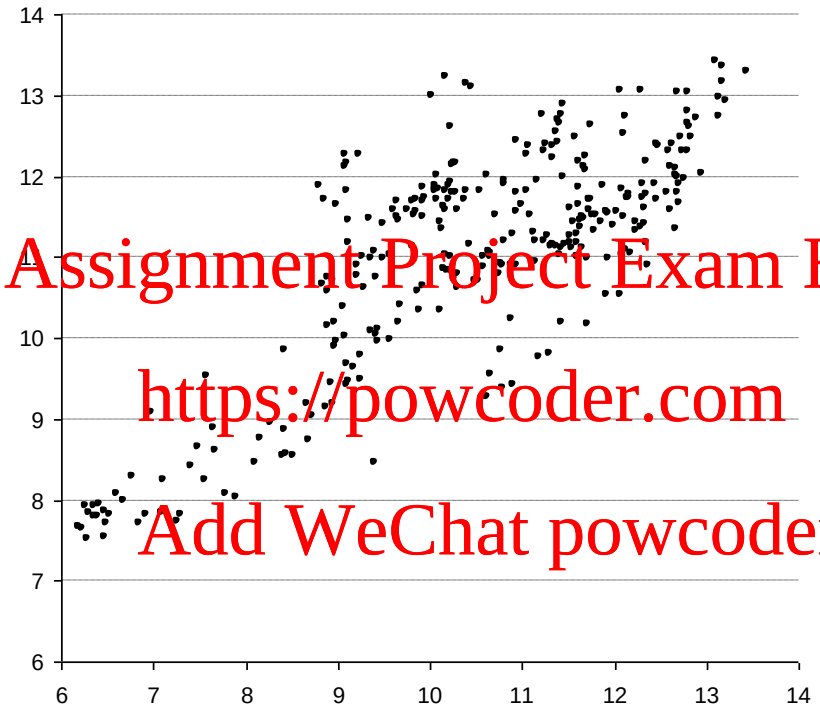


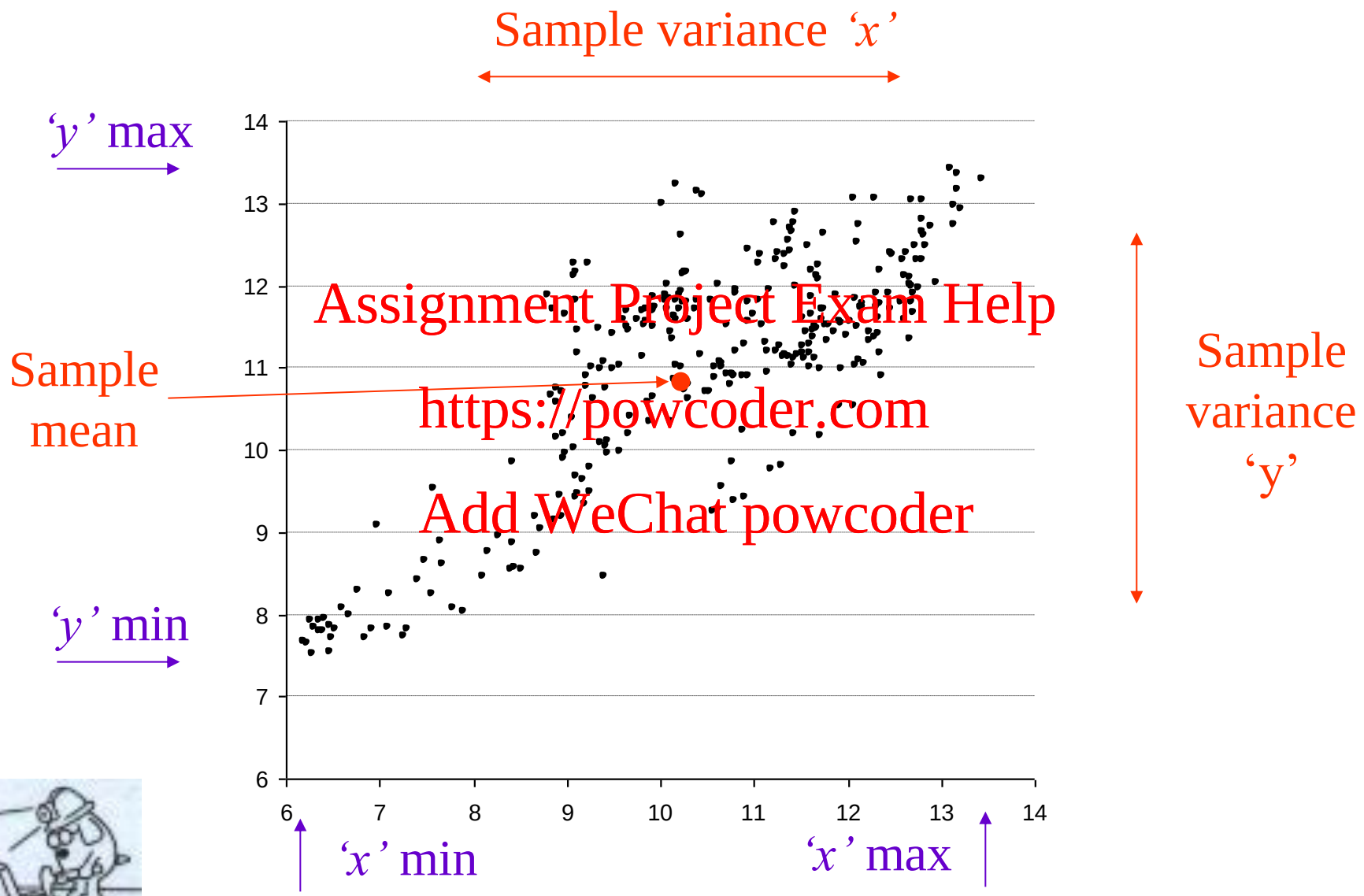
Fig 2

Example from speech processing



Plot of high-frequency energy vs low-frequency energy, for 25 ms speech segments, sampled every 10ms

Basic statistics



Basic statistics

- Denote samples by

$$X = x_1, x_2, \ldots, x_T$$

where $x_t = (x_t^1, x_t^2, \ldots, x_t^N)$

- The sample mean μ (or more correctly $\mu(X)$) vector is given by:

$$\mu^n = \frac{1}{T}\sum_{t=1}^{T} x_t^n$$

$$\mu = \left(\mu^1, \mu^2, \ldots, \mu^n, \ldots, \mu^N\right)$$

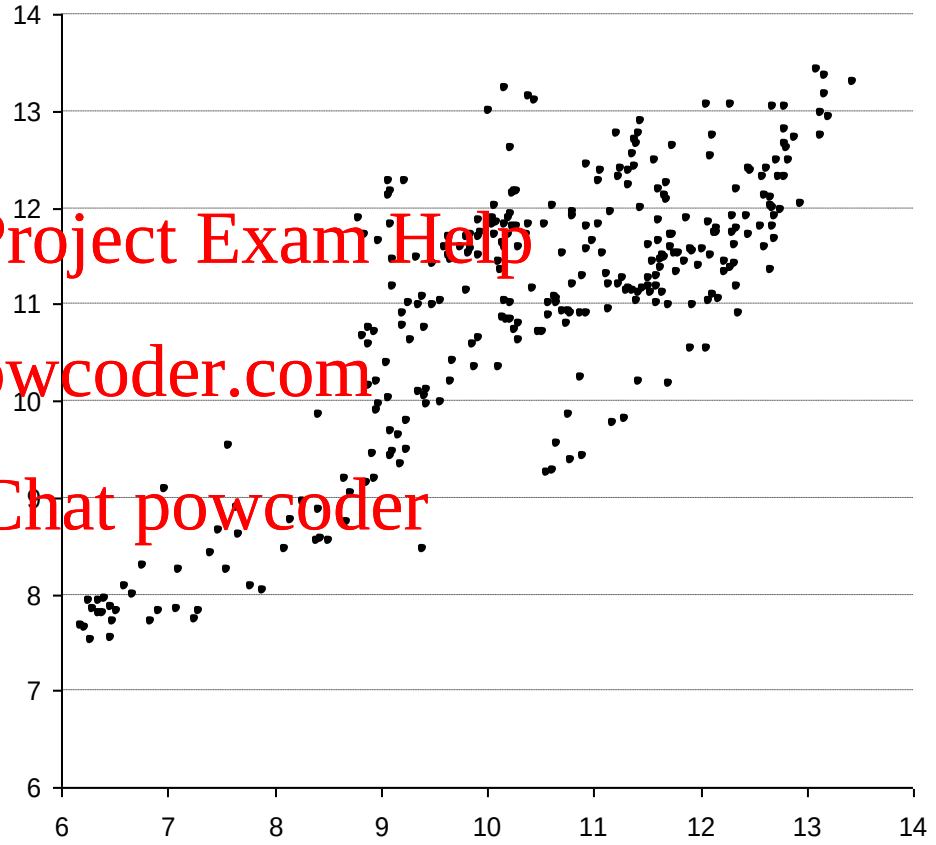
More basic statistics

- The sample variance σ (more correctly $\sigma(X)$) vector is given by:

$$\sigma^n = \frac{1}{T-1}\sum_{t=1}^{T}(x_t^n - \mu^n)^2, \quad \sigma = [\sigma^1, \ldots, \sigma^n]$$

Covariance

- In this data, as the *x* value increases, the *y* value also increases
- This is (positive) co-variance
- If *y* decreases as *x* increases, the result is negative covariance



Definition of covariance

- The covariance between the m^{th} and n^{th} components of the sample data is defined by:

$$\sigma^{m,n} = \frac{1}{T-1}\sum_{t=1}^{T}(x_t^m - \mu^m)(x_t^n - \mu^n),$$

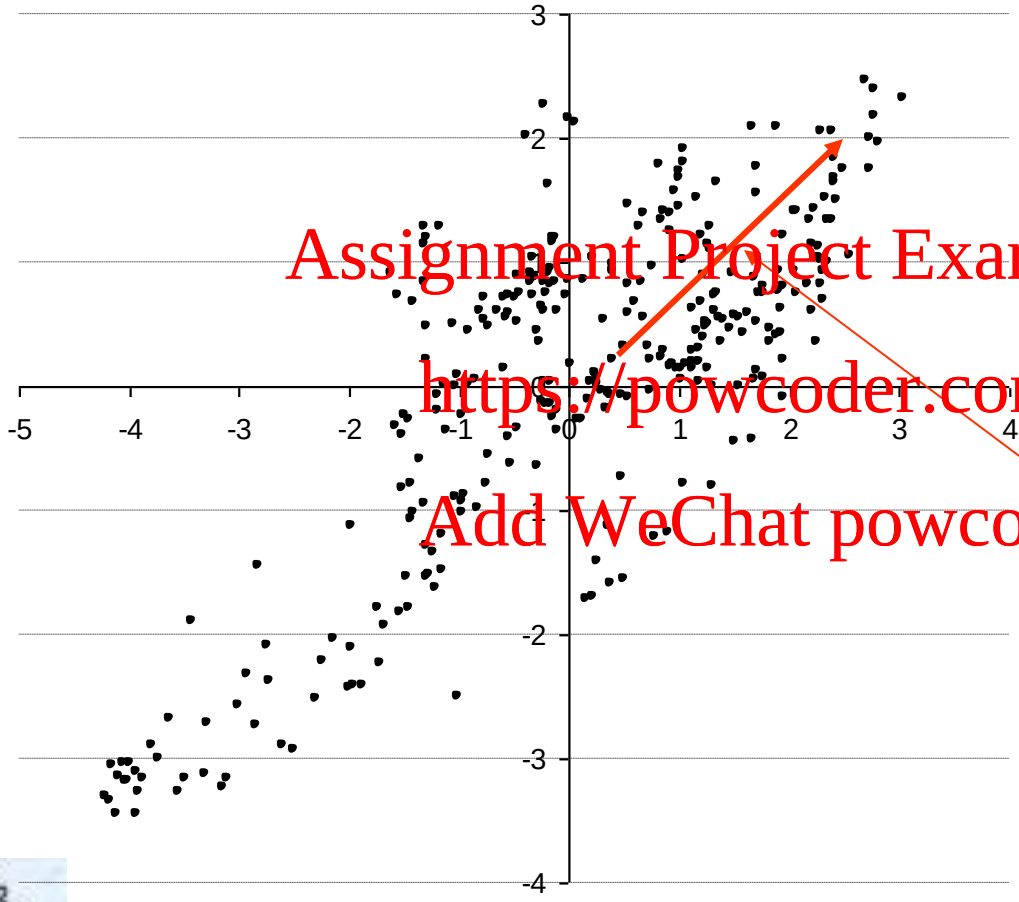
- In practice it is useful to subtract the mean μ from each of the data points x_t. The sample mean is then 0 and

$$\sigma^{m,n} = \frac{1}{T-1}\sum_{t=1}^{T} x_t^m x_t^n,$$

The covariance matrix

$$\sigma = \begin{bmatrix} \sigma^{1,1} & \sigma^{1,2} & \dots & \sigma^{1,n} & \dots & \sigma^{1,N} \\ \sigma^{2,1} & \sigma^{2,2} & \dots & \sigma^{2,n} & \dots & \sigma^{2,N} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma^{m,1} & \dots & \dots & \sigma^{m,n} & \dots & \sigma^{m,N} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma^{N,1} & \dots & \dots & \dots & \dots & \sigma^{N,N} \end{bmatrix}$$

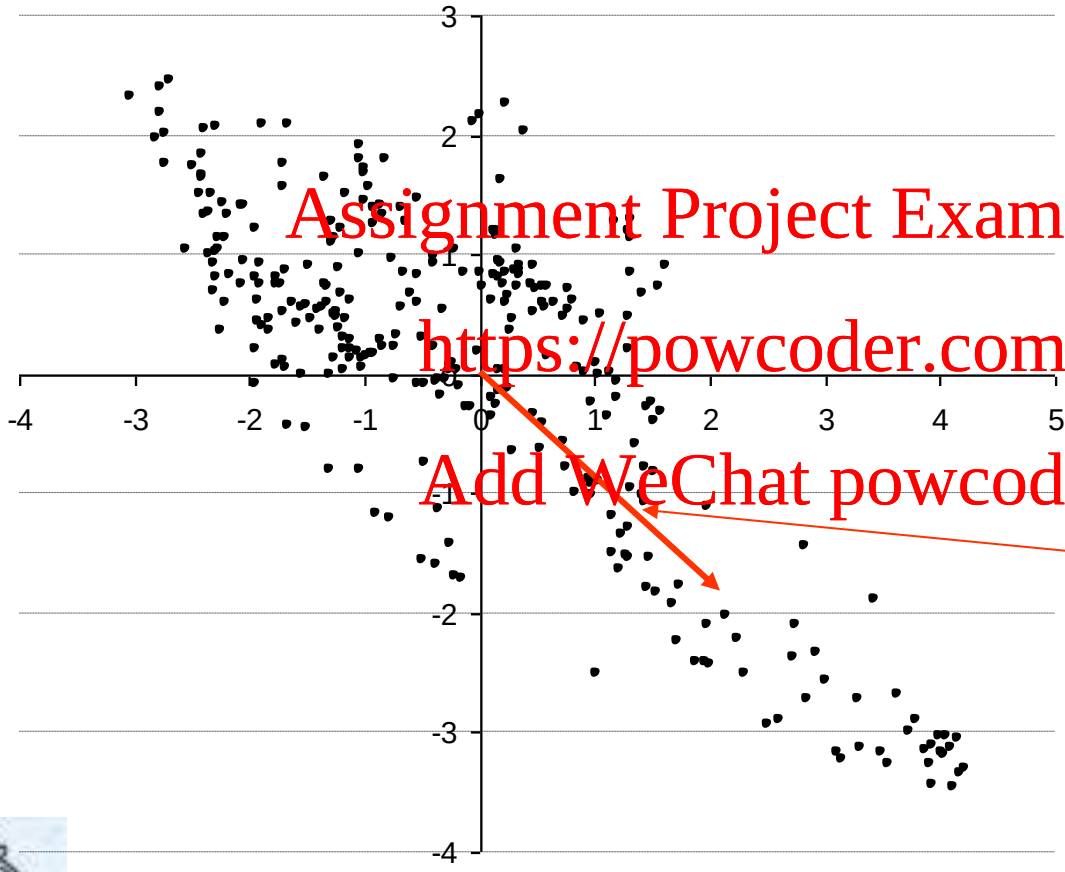
Data with mean subtracted



$$\sigma = \begin{bmatrix} 2.96 & 1.9 \\ 1.9 & 1.97 \end{bmatrix}$$

Implies positive covariance

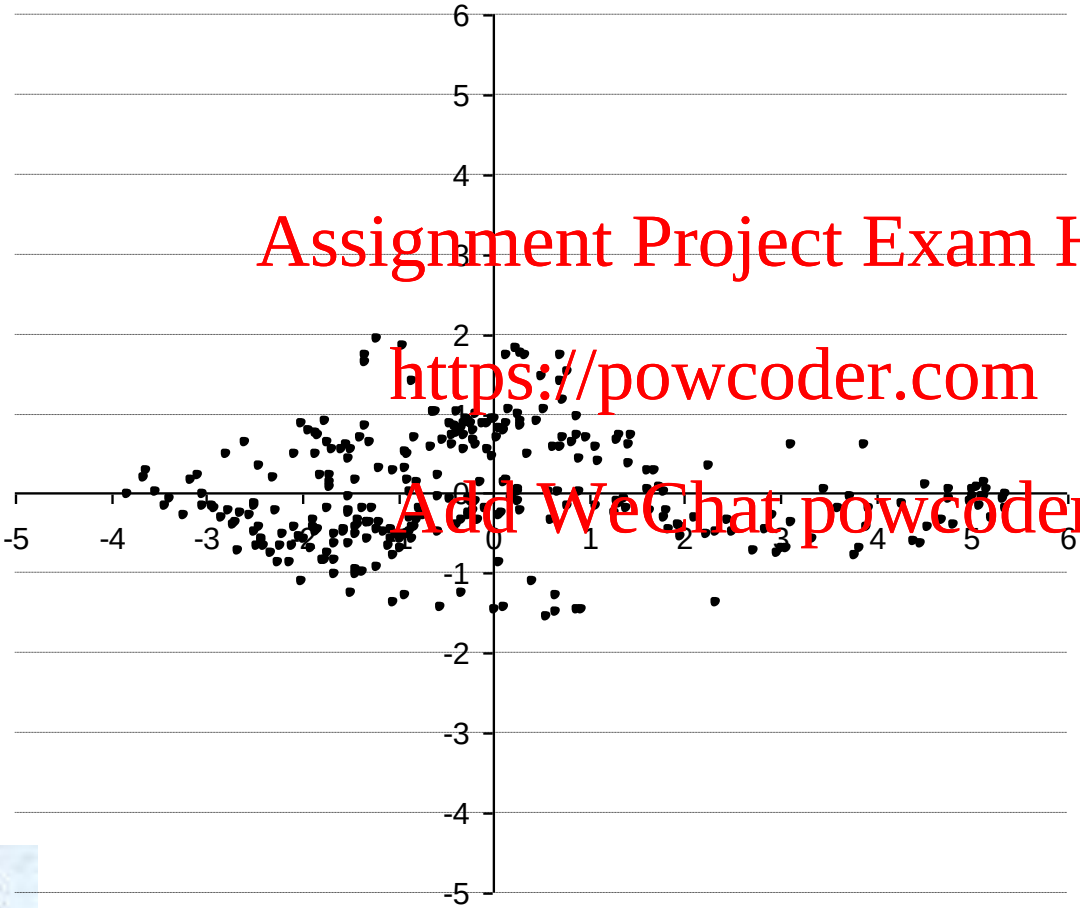
Sample data rotated



$$\sigma = \begin{bmatrix} 2.96 & -1.9 \\ -1.9 & 1.97 \end{bmatrix}$$

Implies negative covariance

Data with covariance removed



$$\sigma = \begin{bmatrix} 4.51 & 0 \\ 0 & 0.48 \end{bmatrix}$$

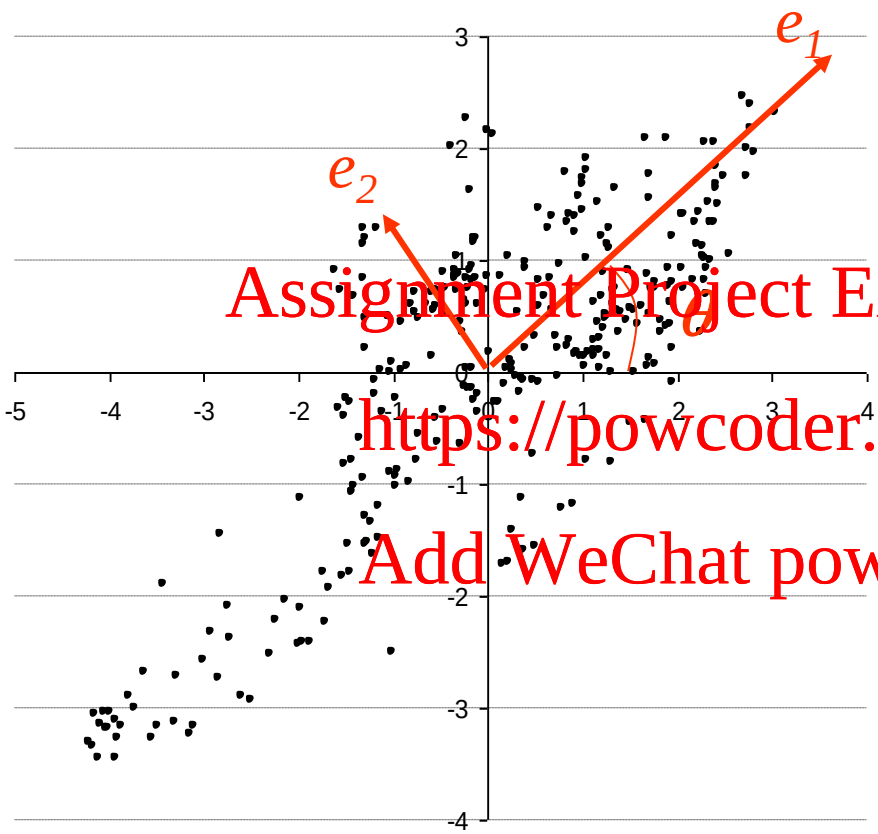
Principal Components Analysis

- PCA is the technique which I used to diagonalise the sample covariance matrix
- The first step is to write the covariance matrix in the form:

$$\Sigma = UDU^T$$

  where D is diagonal and U is a matrix corresponding to a rotation
- You can do this using SVD (see lecture on LSI) or Eigenvalue Decomposition

PCA continued



U implements rotation through angle θ

e_1 is the first column of U

$$e_1 = \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix}$$

d_{11} is the variance in the direction e_1

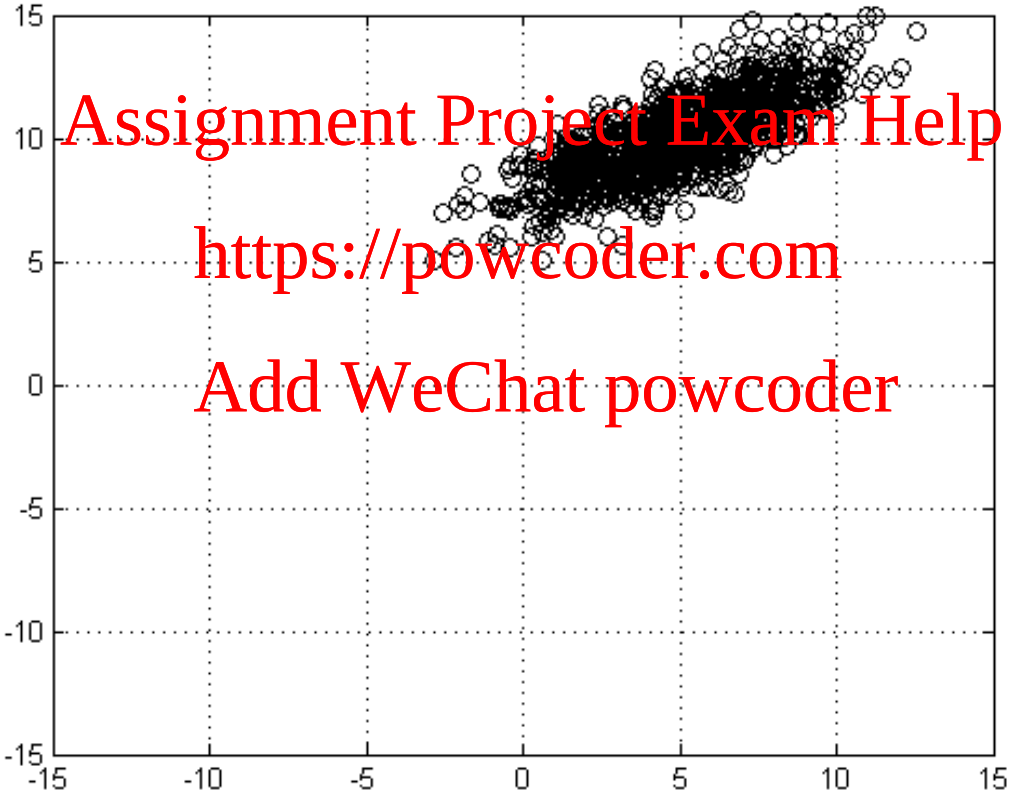
e_2 is the 2nd column of U

d_{22} is the variance in the direction e_2

$$\sigma = UDU^T = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix} \begin{bmatrix} d_{11} & 0 \\ 0 & d_{22} \end{bmatrix} \begin{bmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \end{bmatrix}$$

PCA Example

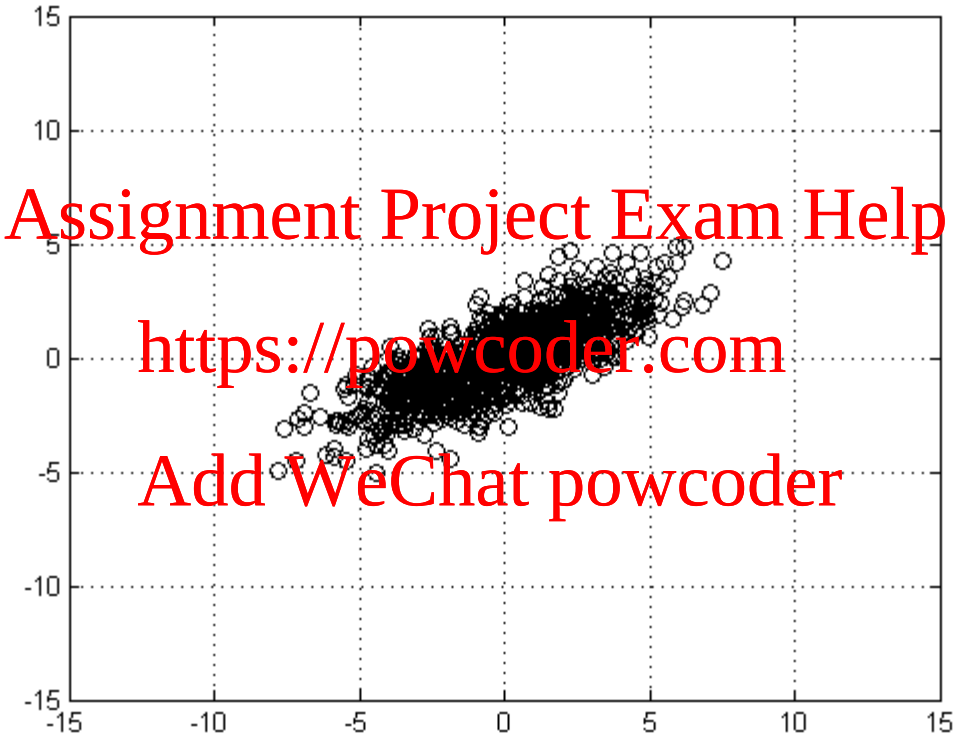
- Abstract data set



PCA Example (continued)

- Step 1: load the data into MATLAB:
  - `A=load('data4');`
- Step 2: Calculate the mean and subtract this from each sample
  - `M=ones(size(A));`
  - `N=mean(A);`
  - `M(:,1)=M(:,1)*N(1);`
  - `M(:,2)=M(:,2)*N(2);`
  - `B=A-M;`
- Plot `B`

PCA Example (continued)



PCA Example (continued)

- Calculate the covariance matrix of B (or A)
  - `S=(B'*B)/size(B,1);`
  - or
  - `S=cov(B);`

$$S = \begin{bmatrix} 6.78 & 3.27 \\ 3.27 & 2.76 \end{bmatrix}$$

- Difficult to deduce much about the data from this covariance matrix

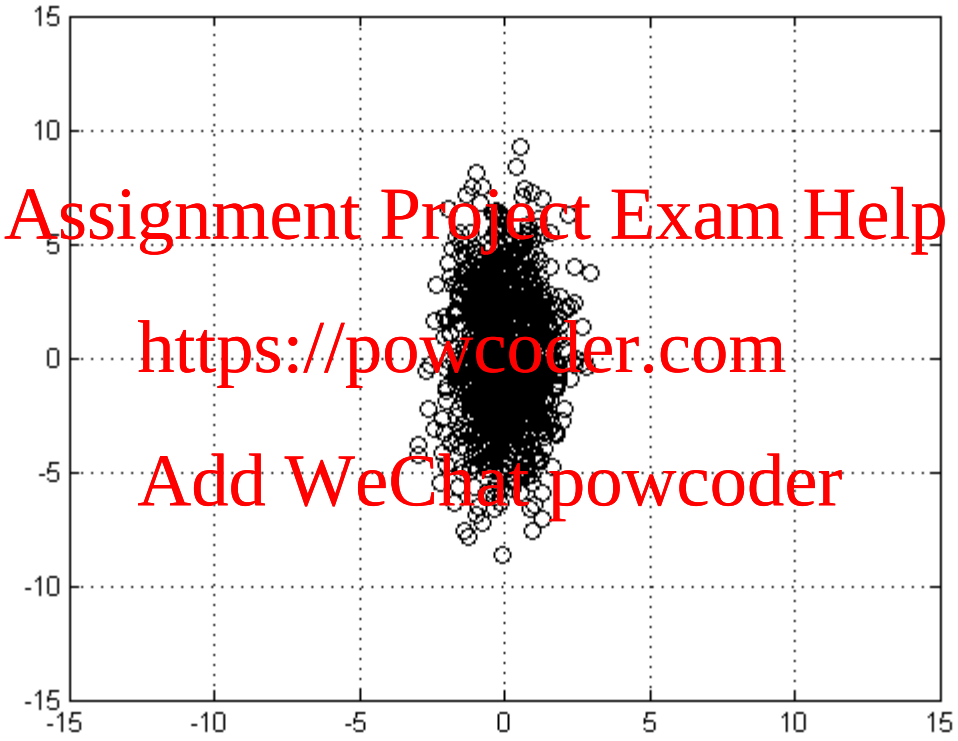
PCA Example (continued)

- Calculate the eigenvalue decomposition of S
  - `[U,E]=eig(S);`

$$U = \begin{bmatrix} 0.4884 & -0.8726 \\ -0.8726 & -0.4884 \end{bmatrix}, E = \begin{bmatrix} 0.2307 & 0 \\ 0 & 8.6079 \end{bmatrix}$$

- After transforming the data using U its covariance matrix becomes E. You can confirm this by plotting the transformed data:

PCA Example (continued)



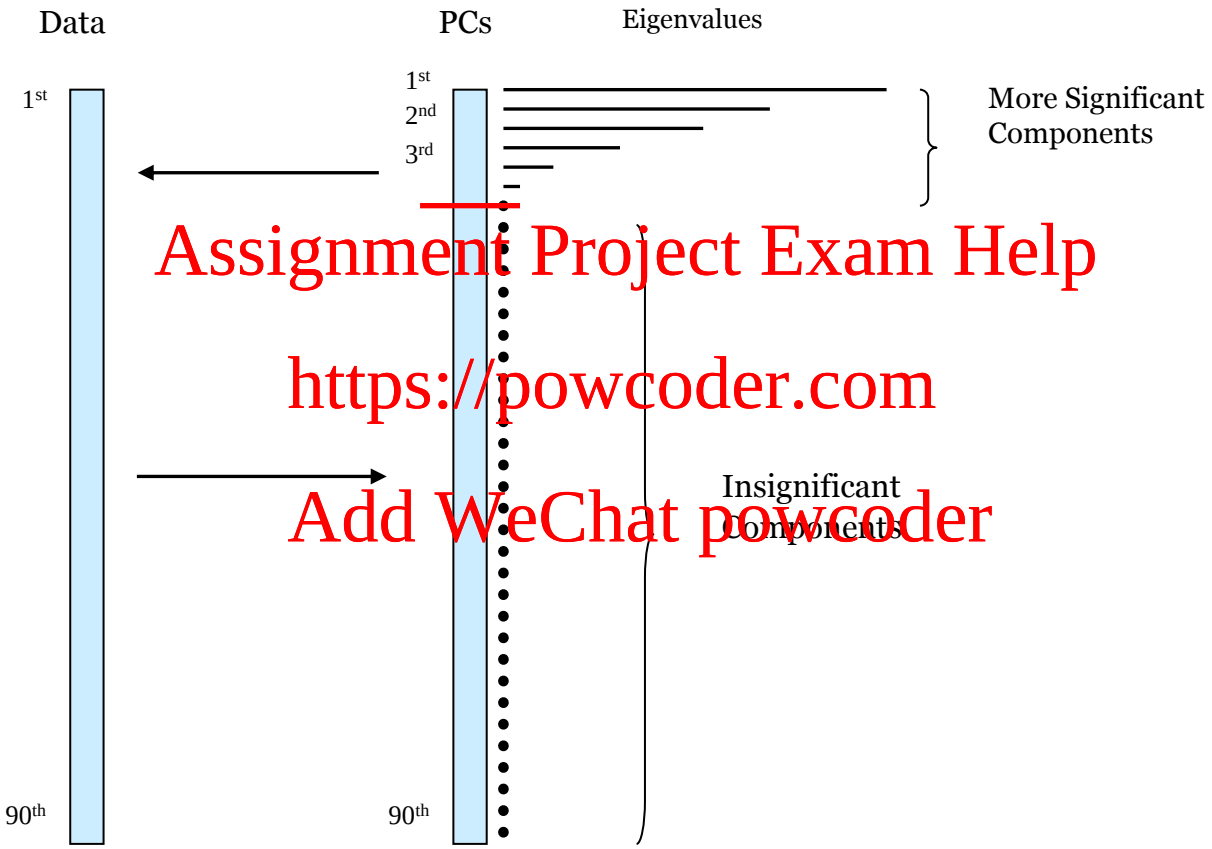
PCA Example (continued)

- After transformation by the matrix U, the covariance matrix has been diagonalized and is now equal to E
  - variance in the x direction is 0.03
  - variance in the y direction is 8.61
- This tells us that most of the variation in the data is contained in the (new) y direction
- There is much less variation in the new x direction, and we could get a 1 dimensional approximation to the data by discarding this dimension
- None of this is obvious from the original covariance matrix

Final notes

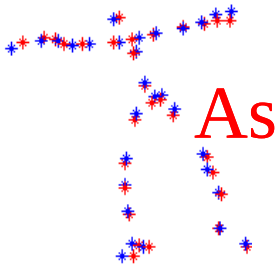
- Each column of U is a principal vector
- The corresponding eigenvalue indicates the variance of the data along that dimension
  - Large eigenvalues indicate significant components of the data
  - Small eigenvalues indicate that the variation along the corresponding eigenvectors may be noise
- It may be advantageous to ignore dimensions which correspond to small eigenvalues and only consider the projection of the data onto the most significant eigenvectors – this way the dimension of the data can be reduced

Eigenvalues



Visualising PCA

Original pattern (blue)



U

Eigenspace

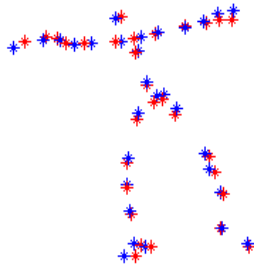
Set coordinates

$n \rightarrow 90$ to zero

Eigenspace

U^{-1}

Reduced pattern (red)



Summary

- Review of basic data analysis (mean, variance and covariance)
- Introduction to Principal Components Analysis (PCA)
- Example of PCA