Principal Component Analysis

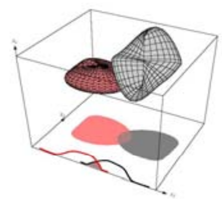
Nirdosh Bhatnagar

Curse of Dimensionality.

- A major problem is *the curse of dimensionality.*
- If the data x lies in high dimensional space, then an enormous amount of data is required to learn distributions or decision rules.
- Example: 50 dimensions. Each dimension has 20 levels. This gives a total of 20^{50} cells. But the no. of data samples will be far less. There will not be enough data samples to learn.

Data Dimensionality

- From a theoretical point of view, increasing the number of features should lead to better performance.
- In practice, the inclusion of more features leads to worse performance (i.e., curse of dimensionality).



- Need exponential number of training examples as dimensionality increases.

Why dimensionality reduction?

- Some features may be irrelevant
- We want to visualize high dimensional data
- “Intrinsic” dimensionality may be smaller than the number of features

Dimension Reduction

- One way to avoid the curse of dimensionality is by projecting the data onto a lower-dimensional space.
- Techniques for dimension reduction:
- Principal Component Analysis (PCA)
- Fisher's Linear Discriminant
- Multi-dimensional Scaling.
- Independent Component Analysis.

Feature reduction algorithms

- Unsupervised
  - Latent Semantic Indexing (LSI): truncated SVD
  - Independent Component Analysis (ICA)
  - Principal Component Analysis (PCA)
  - Canonical Correlation Analysis (CCA)
- Supervised
  - Linear Discriminant Analysis (LDA)
- Semi-supervised
  - Research topic

Application of feature reduction

- Face recognition
- Handwritten digit recognition
- Text mining
- Image retrieval
- Microarray data analysis
- Protein classification

Dimensionality Reduction

- Significant improvements can be achieved by first mapping the data into a *lower-dimensional* space.

$$x = \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_N \end{bmatrix} --> reduce\ dimensionality --> y = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_K \end{bmatrix} \quad (K << N)$$

- Dimensionality can be reduced by:
  - Combining features (linearly or non-linearly)
  - Selecting a subset of features (i.e., feature selection).

- We will focus on feature combinations first.

8

What is Principal Component Analysis?

- Principal component analysis (PCA)
  - Reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables
  - Retains most of the sample's information.
  - Useful for the compression and classification of data.
- By information we mean the variation present in the sample, given by the correlations between the original variables.
  - The new variables, called principal components (PCs), are uncorrelated, and are ordered by the fraction of the total information each retains.

Principal Components Analysis (PCA)

why:

- clarify relationships among variables
- clarify relationships among cases

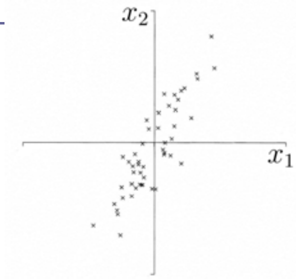
when:

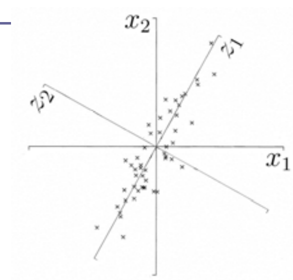
- significant correlations exist among variables

how:

- define new axes (components)
- examine correlation between axes and variables
- find scores of cases on new axes

Geometric picture of principal components (PCs)

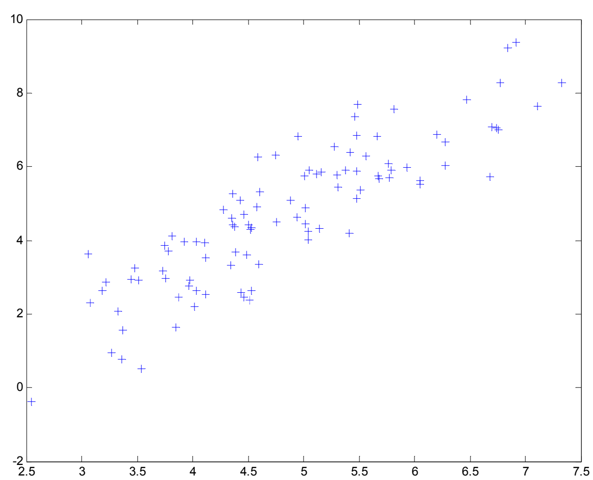




- the 1st PC z_1 is a minimum distance fit to a line in X space
- the 2nd PC z_2 is a minimum distance fit to a line in the plane perpendicular to the 1st PC

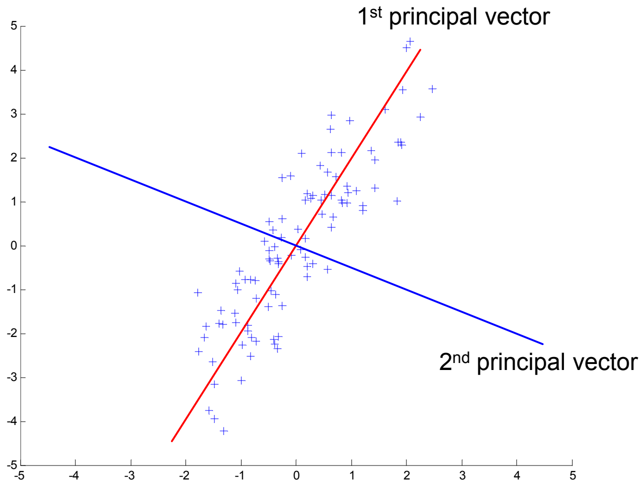
PCs are a series of linear least squares fits to a sample, each orthogonal to all the previous.

2d Data

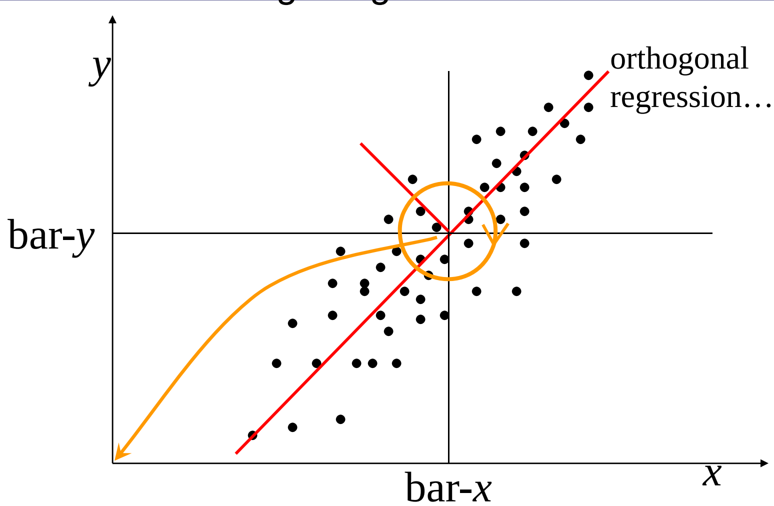


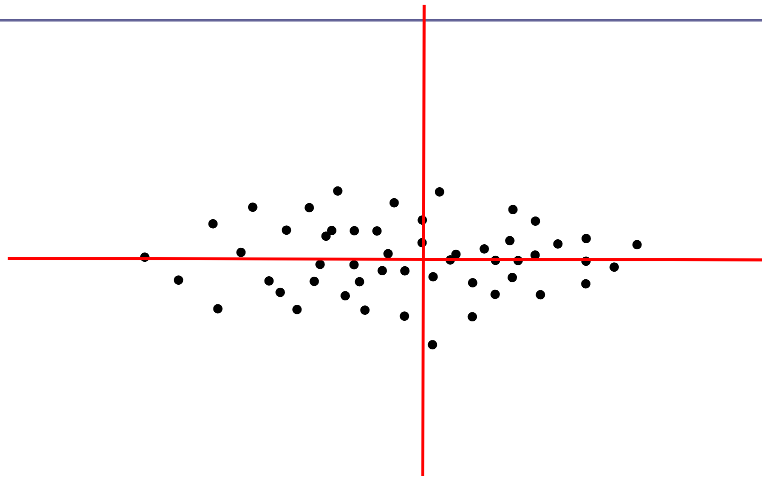
Principal Components

- Gives best axis to project
- Minimum RMS error
- Principal vectors are orthogonal



- imagine a two dimensional scatter of points that show a high degree of correlation …





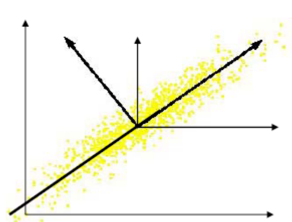
Computing the Components

- Data points are vectors in a multidimensional space
- Projection of vector **x** onto an axis (dimension) **u** is **u.x**
- Direction of greatest variability is that in which the average square of the projection is greatest

– I.e. **u** such that E((**u.x**)2) over all **x** is maximized

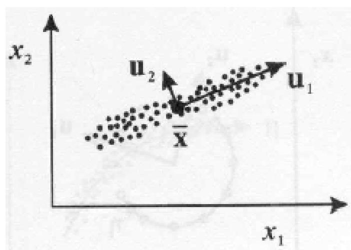
– (we subtract the mean along each dimension, and center the original axis system at the centroid of all data points, for simplicity)

– This direction of **u** is the direction of the first Principal Component



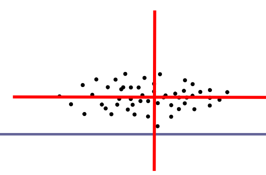
Geometric interpretation

- PCA projects the data along the directions where the data varies the **most**.
- These directions are determined by the eigenvectors of the covariance matrix corresponding to the **largest** eigenvalues.
- The magnitude of the eigenvalues corresponds to the **variance** of the data along the eigenvector directions.



17

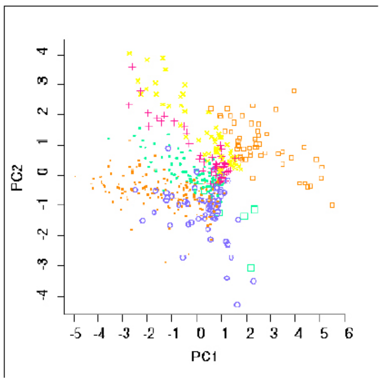
Why bother?



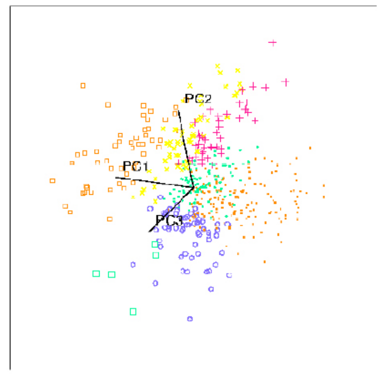
- more “efficient” description
  - 1st var. captures max. variance
  - 2nd var. captures the max. amount of residual variance, at right angles (orthogonal) to the first
- the 1st var. may capture so much of the *information content* in the original data set that we can ignore the remaining axis

1.1 PCA
Example PCA Analysis

477 sporulation genes classified into seven patterns resovled by PCA



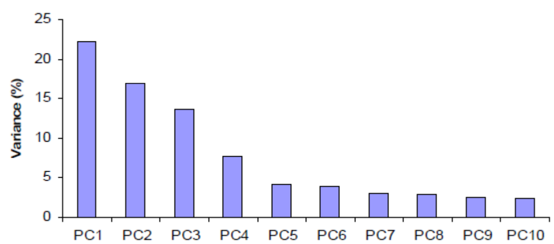
First two PC directions



First three PC directions

Dimensionality Reduction

Can *ignore* the components of lesser significance.

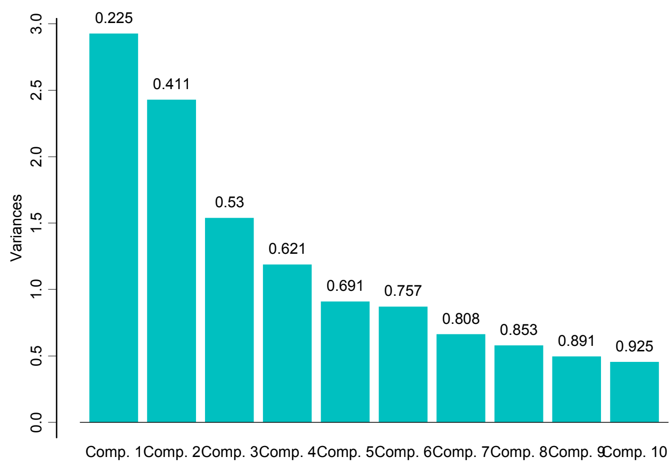


You do lose some information, but if the eigenvalues are small, you don't lose much

– n dimensions in original data
– calculate n eigenvectors and eigenvalues
– choose only the first p eigenvectors, based on their eigenvalues
– final data set has only p dimensions

1.1 PCA

Variance in data explained by the first *n* principle components



PCA Example

- The images of an object under different lighting lie in a low-dimensional space.
- The original images are 256x 256. But the data lies mostly in 3-5 dimensions.
- First we show the PCA for a face under a range of lighting conditions. The PCA components have simple interpretations.
- Then we plot $\frac{\sum_{i=1}^{M} \lambda_i}{\sum_{i=1}^{N} \lambda_i}$ as a function of M for several objects under a range of lighting.

PCA on Faces.

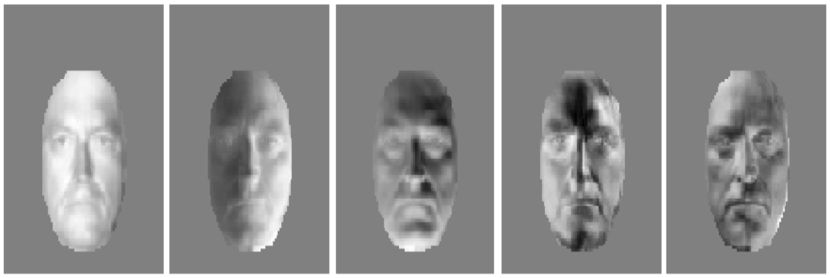


Figure 4: The eigenvectors calculated from the sparse set for the human face. Note that the images were only lit from the right so the eigenvectors are not perfectly symmetric. Observe also that the first three eigenvectors appear to be images of the face illuminated from three orthogonal lighting conditions in agreement with the orthogonal lighting conjecture.

Principal Components Analysis (PCA)

• Principle

– Linear projection method to reduce the number of parameters

– Transfer a set of correlated variables into a new set of uncorrelated variables

– Map the data into a space of lower dimensionality

– Form of unsupervised learning

• Properties

– It can be viewed as a rotation of the existing axes to new positions in the space defined by original variables

– New axes are orthogonal and represent the directions with maximum variability