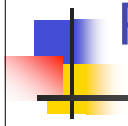


Principal Component Analysis

Rezarta Islamaj Dogan

Resources

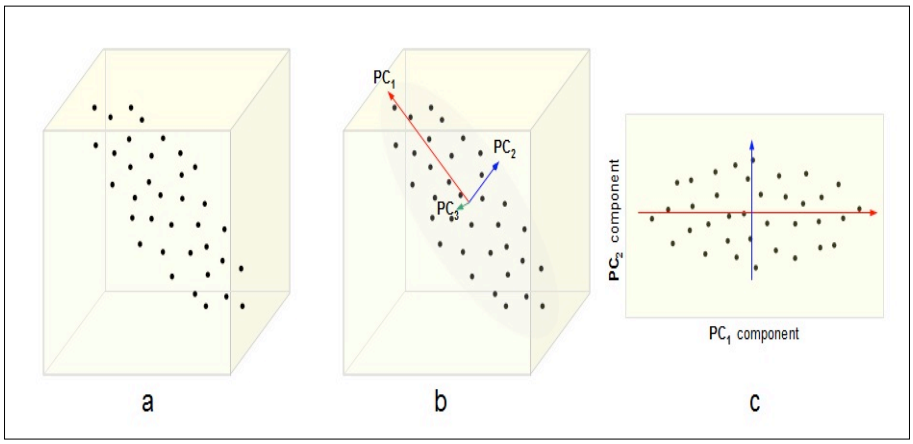
- A tutorial on principal component analysis, derivation, discussion and singular value decomposition,

  Jon Shlens
  - www.dgp.toronto.edu/~aranjan/tuts/pca.pdf
- A tutorial on principal components analysis, Lindsay I Smith
  - www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Goal of PCA

- Given a dataset D, described by n variables, describe this dataset with a smaller set of new variables.
- The new set of variables are linear combinations of the originals
- They are called principal components.

Geometric interpretation



Applications of PCA

- Exploratory data analysis
  - PCA is used for making 2,3-dimensional plots of the data for visual examination and interpretation:

Detection of outliers
Identification of clusters

Applications of PCA

- Exploratory data analysis
- Data preprocessing, dimensionality reduction
  - Data is often described by more variables then necessary for building the best model. Specific techniques exist for selecting a “good” subset of variables. PCA is one of them.

Applications of PCA

- Exploratory data analysis
- Data preprocessing, dimensionality reduction

Reducing nr of variables generally leads to loss of information
PCA makes this loss minimal

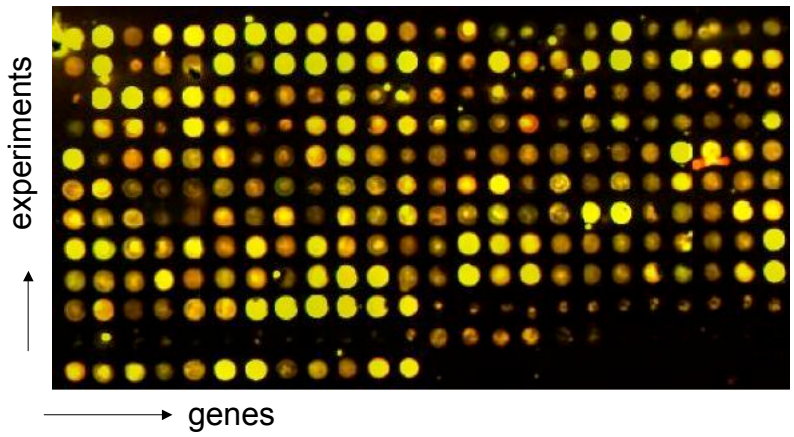
Applications of PCA

- Exploratory data analysis
- Data preprocessing, dimensionality reduction
- Data compression, data reconstruction
  - (lossy) data compression technique
  - The table describing the data with first k-principal components is smaller than original data table

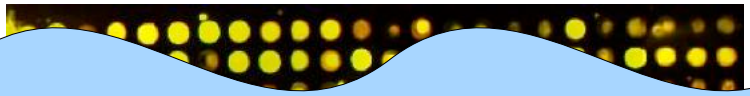
Examples

- neuroscience
- computer graphics
- meteorology
- oceanography
- gene expression
- etc

Microarray example



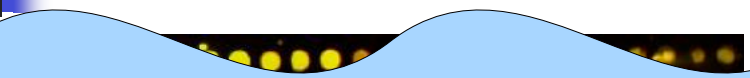
Microarray example



Reduce dimensionality

- In gene expression experiments, thousands of variables
- It is useful to collapse the genes into a smaller set of principal components.
- This makes plots easier to interpret, which can help to identify structure in the data.

Microarray example



Principal Components

- New variables,
- Linear combinations of the original gene data variables

- Looking at which genes or gene families have a large contribution to a principal component can be an indication of shared function of behavior, similar to the inferences that can be made using clustering.

Principal Components

- Can be used to determine how many real dimensions there are in the data.
  - If a small number of components accounts for most of the variation in the data, then the other components can be thought of as noise variables.

Properties of principal components

- Number
  - Same as original data variables
- Orthogonality
- Uncorrelatedness
  - Pairwise orthogonal and uncorrelated
- Ordering
  - Can be ordered by decreasing order of importance
  - The first k-principal components span the best k-dimensional subspace

Properties of principal components

- Number
  - Same as original data variables
- Orthogonality
- Uncorrelatedness
  - Pairwise orthogonal and uncorrelated
- Ordering

PCA assumes the directions with the largest variances are the most “important” or in other words, most principal

nce
k-

Principal components

- Change of basis
  - Is there another basis, which is a linear combination of the original basis, that best re-expresses our data set?
- Mean and variance
- Large variances are important

How PCA works

Data D, number of variables n

1. Select a normalized direction in the n-dimensional space so that the variance in D is maximized, save it as p_1
2. For i=2..n

   Find another direction perpendicular to all (p_1..p_{i-1}) directions, along which variance is maximized, save as p_i
3. The ordered set of p's are the principal components