Microarray analysis 3

CS426, week 12
Golan Yona

How to detect patterns in expression arrays

Clustering

Dimensionality reduction

Statistical and graphical models (e.g. Bayesian networks)

Dimensionality reduction

Microarray data is high dimensional

- Not all features are important, yet they can affect the analysis
- Number of examples needed grows exponentially with the number of features ("curse of dimensionality").
- Clustering algorithms do not perform well on high dimensional data (problems of convergence and validity)

Dimensionality reduction algorithms (embedding algorithms) can help to detect meaningful clusters by reducing "effectively" the dimension of the data set.

Can also help to visualize data.

Gain insight into the structure of data.

Two main types..

Embeddings by linear projections

- PCA (Principal Component Analysis)
- ICA (Independent Component Analysis)

Embeddings by non-linear projections

- MDS (Multi-Dimensional Scaling)

Principal Component Analysis

PCA seeks a linear projection that best represents the data in a *least-square error* sense.

Reduces the dimension by projecting the vectors on the "interesting subspace using a linear combination of features

Main advantages:

- **Simple to compute**
- **Analytically tractable**
- **Reversible**

How does it work?

We are given a sample set $D=\{v_1,v_2,..,v_n\}$ **where each** v_i **is a vector of dimension** *d*

(e.g. a set of expression profiles)

- Transform the data such that the mean is the zero vector
- Compute the $d \times d$ covariance matrix of the data
- Compute the eigenvectors of the covariance matrix
- Order the eigenvectors based on the eigenvalues (largest first)
- To embed in dimension *d'* pick the top *d'* eigenvectors and project the data on these vectors

Eigengenes and eigenarrays

Given a $n \times m$ matrix of n genes and m experiments

Eigengenes - the eigenvectors of the $m \times m$ covariance matrix

Eigenarrays – the eigenvectors of the $n \times n$ *covariance matrix*

The meaning: