COSC 348: Computing for Bioinformatics

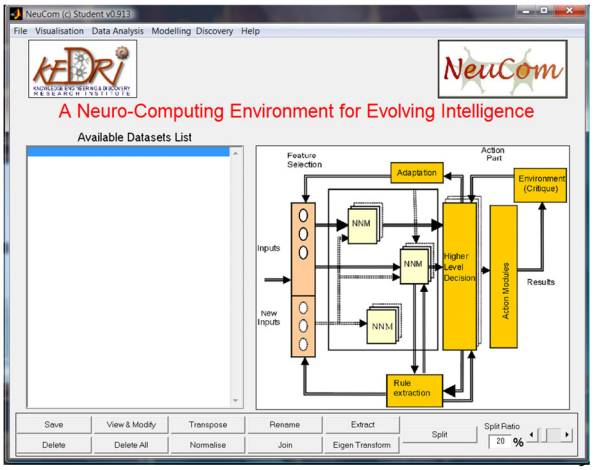
Lecture 19: NeuCom & PCA

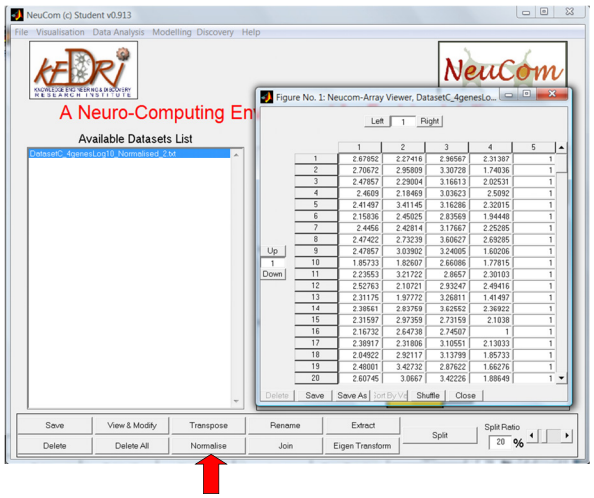
Lubica Benuskova

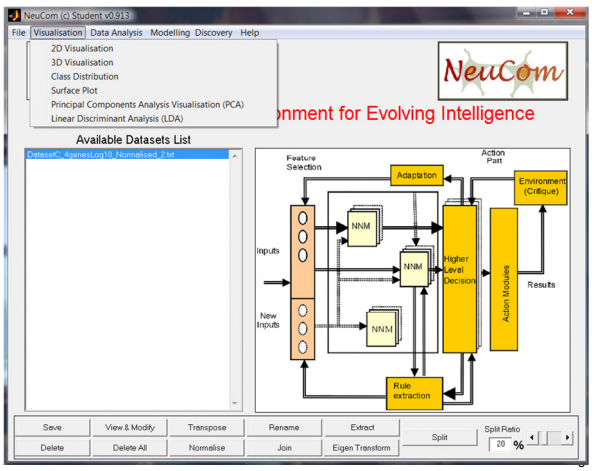
http://www.cs.otago.ac.nz/cosc348/

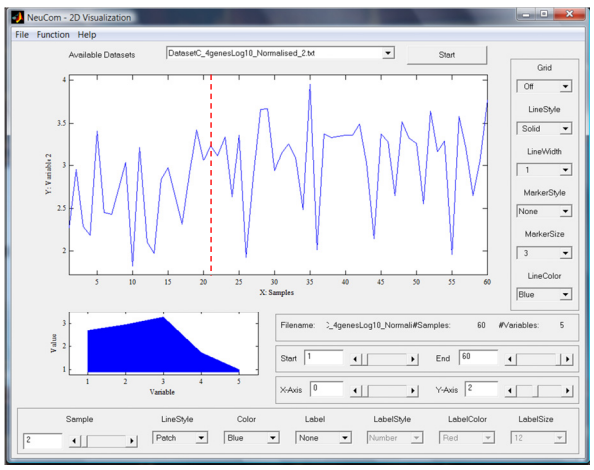
Example: data set C from Pomeroy et al.

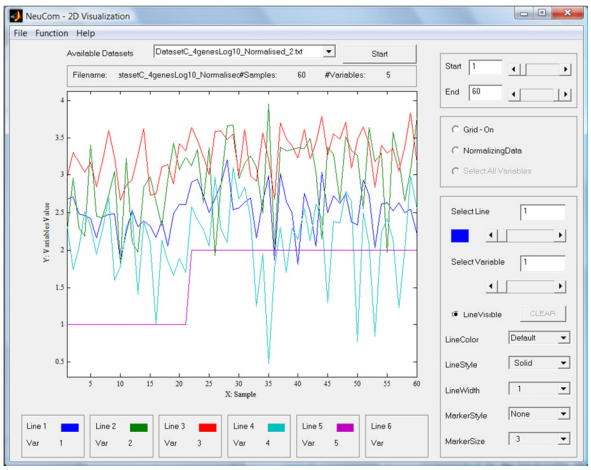
- They wanted to find a genetic profile that distinguishes patients who are alive after treatment ('survivors') compared with those who succumbed to their disease ('failures').
- Minimum follow-up of 24 months for surviving patients, overall median 41.5 months.
- Data set C has 60 samples containing 21 treatment failures and 39 medulloblastoma survivors.
- Let us select the 4 genes with the highest SNR.



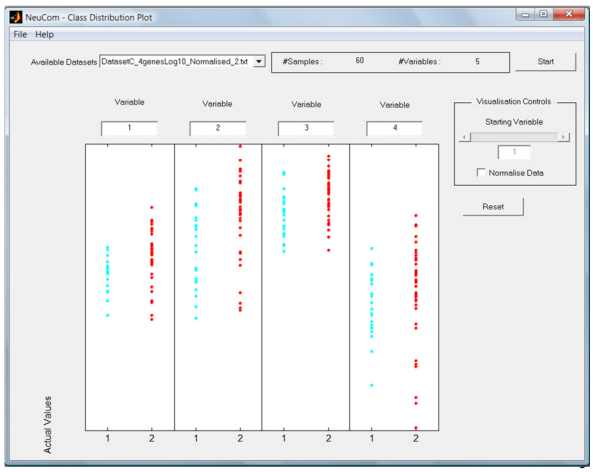




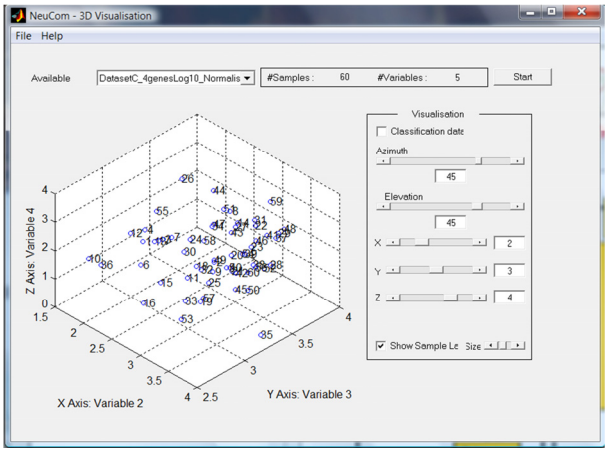




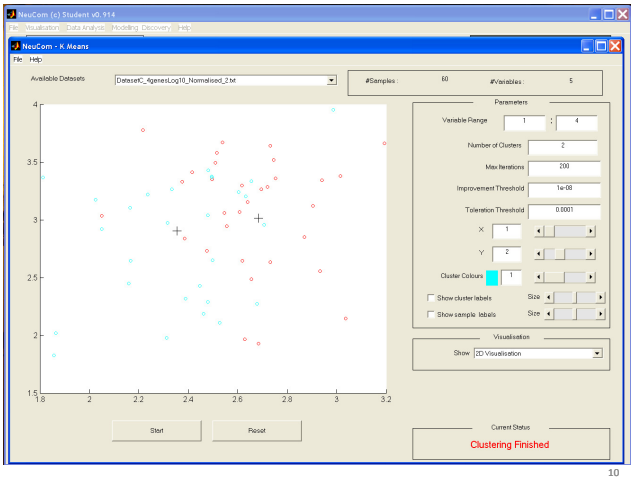
7



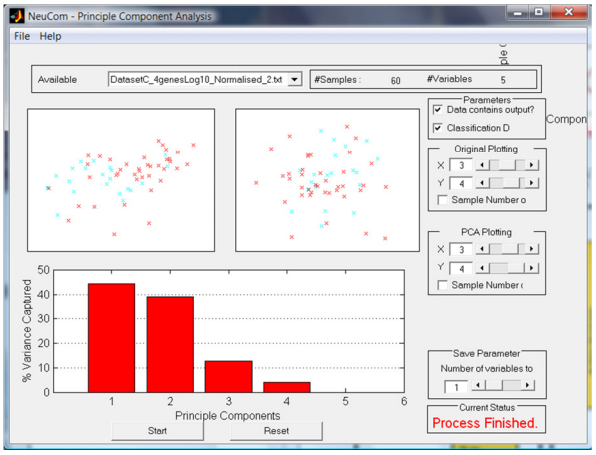
8



9



10



11

Principal component analysis (PCA)

- The goal is to discover a new set of axes against which to represent, describe or evaluate the data
  - For more effective reasoning, insights, or better classification
  - New axis represents a smaller set of factors that are combinations of original gene expressions: hence the dimension reduction
  - Better representation of data without losing much information and reduction of noise
  - Universal data preprocessing method: can build more effective data analyses on the reduced-dimensional space: classification, clustering, pattern recognition.

12

Basic concepts of PCA

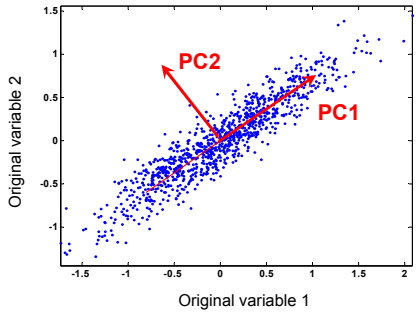
- We want a smaller set of variables that explain most of the variance in the original data in more compact and insightful form.
- If two variables/genes are highly correlated or dependent
  - They are likely to represent highly related phenomena
  - combining them to form a single measure is reasonable
- So we want to combine related variables, and focus on dimensions, along which the observations have high variance
  - Based on an assumption that these directions are interesting, something is happening if the variance is big.

Principal component analysis (PCA)

- Most common form of **factor analysis.**
- Areas of large variance in data are where items can be best discriminated - Areas of greatest "information" in the data
- The new variables/dimensions / components
  - Are linear combinations of the original ones
  - Are uncorrelated with one another and orthogonal in original dimension space
  - Capture as much of the original variance in the data as possible
  - Are called *Principal Components* (PC)

What are these new axes?

- The first principal component PC1 is the direction of greatest variability (i.e. variance) in the data.
- Second PC2 is the next orthogonal direction of greatest variability, and so on … for *N*-dimensional data we can have *N* PCAs



Notions we work with in PCA

- Vectors of (gene expression) values are represented as a cloud of points in a multidimensional space with an axis for each of the *N* variables/genes
- The centroid of the points is defined by the mean of each variable/gene *j* (the sum goes through all *M* samples/subjects):

$$\mu_j = \frac{1}{M}\sum_{i=1}^{M} x_{ji}$$

- The variance of each variable/gene is the average squared deviation of its *M* values around the mean of the *j*th variable/gene:

$$\sigma_j^2 = \frac{1}{M}\sum_{i=1}^{M}\left(x_{ji} - \mu_j\right)^2$$

Covariance matrix

- degree to which the variables, i.e. genes, are linearly correlated is represented by their covariances.

Covariance of variables *k* and *j*

Value of variable *j* in subject *i*

Mean of variable *j*

$$C_{kj} = \frac{1}{M-1}\sum_{i=1}^{M}\left(x_{ki} - \mu_k\right)\left(x_{ji} - \mu_j\right)$$

Sum over all *M* subjects

Value of variable *k* in subject *i*

Mean of variable *k*

Computing the principal components

- We put covariances into **C,** the covariance matrix and then we transform it into a diagonal form:

$$\begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{pmatrix} \Rightarrow \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$

- The diagonal elements of this transformed diagonal matrix are the eigenvalues $\lambda_1,\ldots,\lambda_N$, and we denote the corresponding eigenvectors by $v_1,\ldots, v_N$
  - Eigenvector and eigenvalue is defined by: $\lambda_k v_k = \mathbf{C}\, v_k$
  - Each eigenvalue λ denotes the amount of variability captured along that particular new dimension.

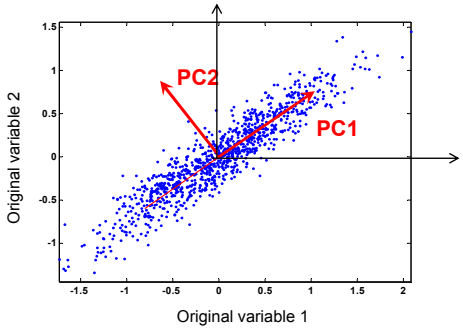
Eigenvectors are principal components

$$Cv_1 = \lambda_1 v_1$$
$$Cv_2 = \lambda_2 v_2$$
$$\vdots$$
$$Cv_k = \lambda_k v_k$$
$$\vdots$$
$$Cv_N = \lambda_N v_N$$

- Eigenvectors for *k* largest eigenvalues are the first *k* principal components.
- The 1st principal component v_1 is the eigenvector for the largest eigenvalue λ_1;
- in the orthogonal space, the eigenvector with second largest eigenvalue is the 2nd PC, etc.

Geometric meaning

- Calculation of PCs geometrically: ***centering followed by rotation*** to align the 1st PC according to the direction of maximal variance.

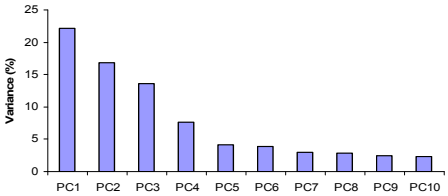


PCA: principle and properties

- 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
- 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 in the data
  - We achieve a better separation of data

Dimensionality reduction

We can *ignore* the components of lesser significance.



We do lose some information, but not much

- *N* dimensions in original data
- calculate *N* eigenvectors and eigenvalues
- choose only the first *k* eigenvectors, based on their eigenvalues
- final data set has only $k < N$ dimensions

Example: data set A from Pomeroy et al.

- Data set A: expression profiles of 42 samples (10 medulloblastomas, 5 CNS AT/RTs, 5 renal and extrarenal rhabdoid tumours, and 8 supratentorial PNETs, as well as 10 non-embryonal brain tumours (malignant glioma) and 4 normal human cerebella).
- SNR was applied to select the differentially expressed genes when compared with normal cerebella.
- They applied PCA to determine whether the different types of tumours could be molecularly distinguished, i.e. whether they are separable.

Example of PCA

- Distribution of data along the first 3 principal component axes. Each new axis is a linear combination of the old axes (i.e. original gene expression values).