Principal Component Analysis
+
Multidimensional Scaling

From Kernels slides …

Since $Ax-\lambda x=0$ for an eigenvector x and matrix A, then to find the eigenvalues of A, we need to find solutions to it's **characteristic equation**, which is

$$\det(A-\lambda I)=0$$

where det is the **determinant** of the matrix A.

Once we find the **eigenvalues**, we can find the **eigenvectors**.

…….

Those eigenvectors with the highest eigenvalues capture the most variance in the data.

PCA focuses precisely on these highly representative eigenvectors.

PCA - Example

- Dataset: Thousands of genes probed in 5 conditions (time points relative to treatment)
- The expression profile of each gene is presented by the vector of its expression levels: $X = (X_1, X_2, X_3, X_4, X_5)$
- Imagine each gene X as a point in a 5-dimentional space.
- Each direction/axis corresponds to a specific condition
- Genes with similar profiles are close to each other in this space
- PCA- Project this dataset to 2 dimensions, preserving as much information as possible

The covariance matrix

PCA finds the principal axis to represent the data by finding the eigenvectors and the eigenvalues to the covariance matrix C.

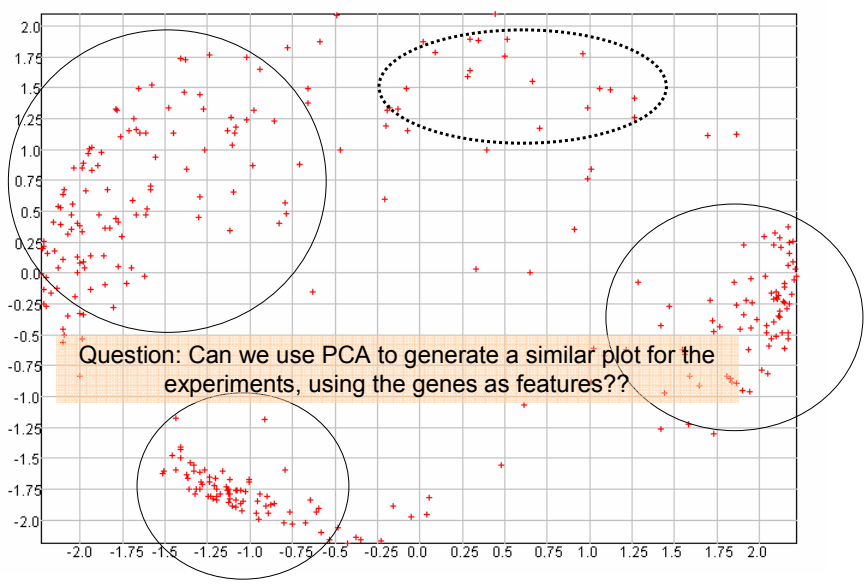
If your input matrix is X with m features and n data points, and your data is centered such that the features for every datapoint sum up to zero, then the covariance matrix is defined to be:

$$C_{i,j} = \frac{1}{m}\langle x_i, x_j \rangle$$

where xi and xj are the feature vectors in the data.
(In the gene example we had before, we had x1, x2, x3, x4, x5 and so the covariance matrix would be 5x5)

PCA – Example

Visual estimation of the number of clusters in the **genes**



Question: Can we use PCA to generate a similar plot for the experiments, using the genes as features??

PCA – van Erk

http://courses.cs.tamu.edu/rgutier/cpsc689_f07/l9.pdf

PCA - ricardo

http://www.bigcat.nl/public/presentations/pca_basics.pdf

MDS - Multidimensional Scaling

- Multi-Dimensional Scaling (MDS) is a general technique for displaying n-dimensional data in 2D.
- It preserves the notion of “nearness”, and therefore clusters of items in n-dimensions still look like clusters on a 2-dim plot.

Multidimensional Scaling

given: - a set of n objects
- the dissimilarities δ_{ij} between them

find: points on the plane whose distances d_{ij} are as close as possible to the δ_{ij}

minimize: $$\text{STRESS} = \left[\frac{\sum_{i,j} \left(d_{ij} - \delta_{ij} \right)^2}{\sum_{i,j} \delta_{ij}^{\,2}} \right]^{1/2}$$ [Kruskal 1964]

MDS applications

Nuclear Physics: Pattern Recognition in Nuclear Gamma-ray Spectra John A. Cameron 163

Sociometry: Deriving Sociograms via Asymmetric Multidimensional Scaling Linda M. Collins 179

Market Research: Consumer Preference and Perception Donna L. Hoffman & William D. Perreault 199

Political Science: Using the Generalized Euclidian Model to Study Ideological Shifts in the U.S. Senate Douglas V. Easterling 221

Psychology: An Application of Principal Directions Scaling to Auditory Pattern Perception Mari Riess Jones & Robert MacCallum 259

Psychology: The Subjective Attributes of Natural Categories-- An Application of a Constrained Generalized Euclidean Model Barbara H. Forsyth