Computational functional genomics

(Spring 2005: Lecture 8)

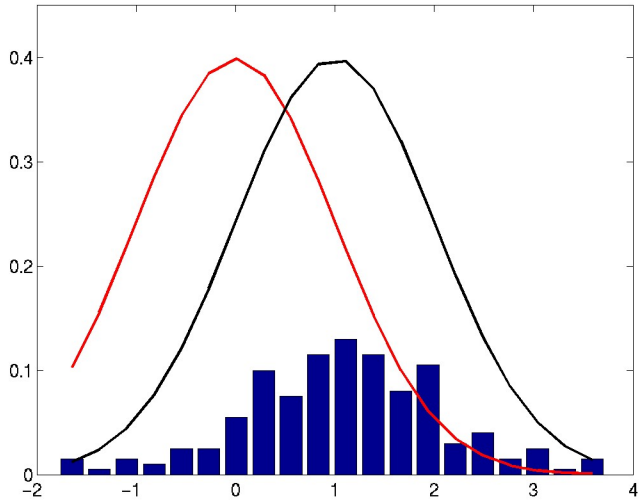
David K. Gifford
(Adapted from a lecture by Tommi S. Jaakkola)
MIT CSAIL

Topics

- Basic clustering methods
  - hierarchical
  - k-means
  - mixture models
- Multi-variate gaussians
- Principle Component Analysis

Simple mixture models

- Instead of representing clusters only in terms of their centroids, we can assume that each cluster corresponds to some distribution of examples such as Gaussian
- Two clusters, two Gaussian models $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$



- The partial assignment of examples to clusters should be based on the probabilities that the models assign to the examples

Simple mixture model clustering

(for cluster models with fixed covariance)

- The procedure:
  1. Pick k arbitrary centroids
  2. Assign examples to clusters based on the relative likelihoods that the cluster models assign to the examples
  3. Adjust the centroids to the weighted means of the examples
  4. Goto step 2 (until little change)

Simple mixture models

- We can also adjust the covariance matrices in the Gaussian cluster models
- Ideas how?
- In this case the clusters can become more elongated

Simple mixture models

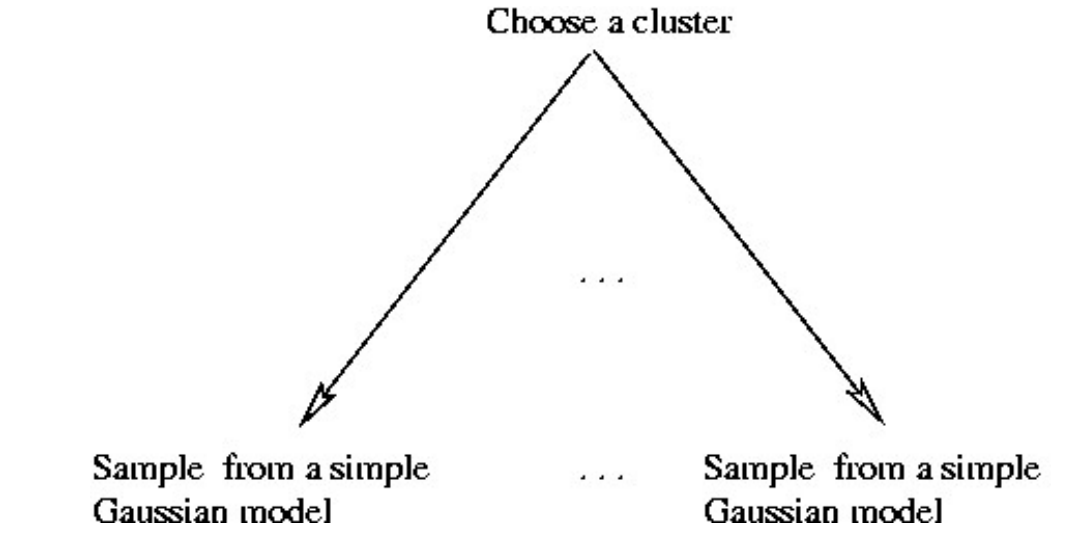
- A generative model perspective:

  We are fitting a generative model to the observed data via the maximum likelihood principle

Simple mixture models

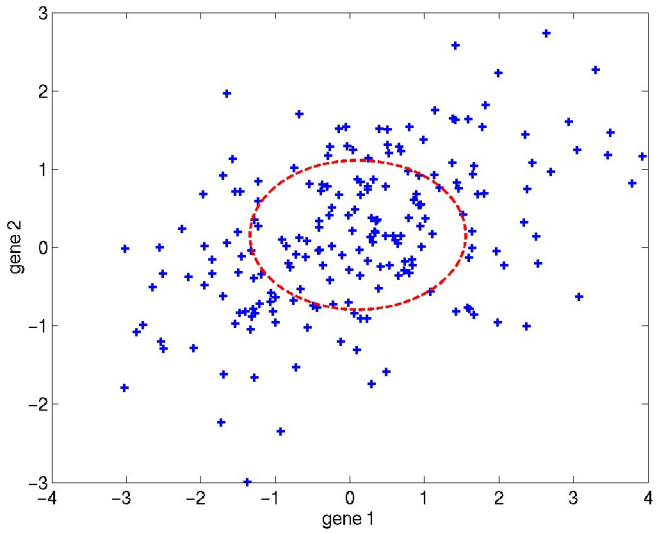
- A generative model perspective:

  We are fitting a generative model to the observed data via the maximum likelihood principle

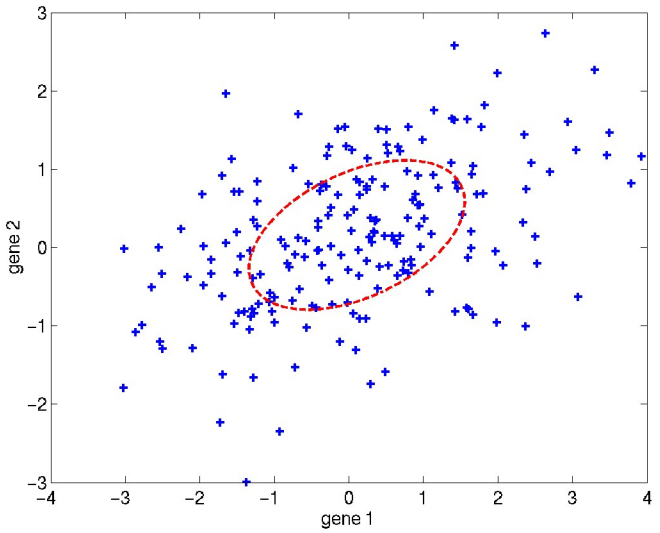


Statistical tests: example

- The alternative hypothesis H_1 is more expressive in terms of explaining the observed data



null hypothesis



alternative hypothesis

- We need to find a way of testing whether this difference is significant

Test statistic

- Likelihood ratio statistic

$$T(X^{(1)}, \ldots, X^{(n)}) = 2 \log \frac{P(X^{(1)}, \ldots, X^{(n)} | \hat{H}_1)}{P(X^{(1)}, \ldots, X^{(n)} | \hat{H}_0)} \tag{1}$$

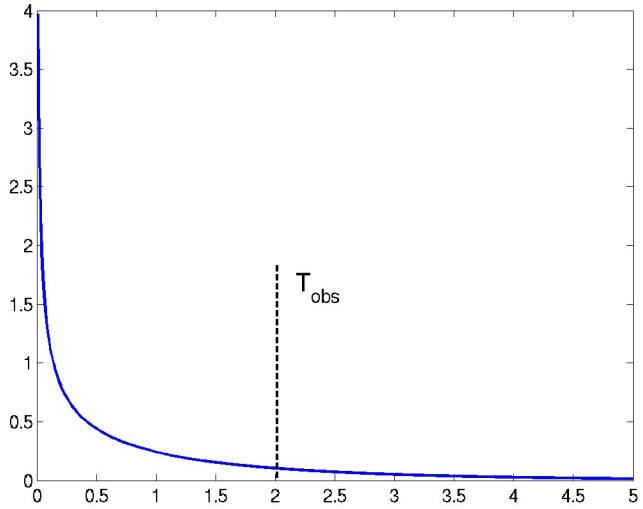
  Larger values of T imply that the model corresponding to the null hypothesis H_0 is much less able to account for the observed data

- To evaluate the P-value, we also need to know the sampling distribution for the test statistic

  In other words, we need to know how the test statistic $T(X^{(1)}, \ldots, X^{(n)})$ varies if the null hypothesis H_0 is correct

Test statistic cont'd

- For the likelihood ratio statistic, the sampling distribution is χ^2 with degrees of freedom equal to the difference in the number of free parameters in the two hypotheses



- Once we know the sampling distribution, we can compute the P-value

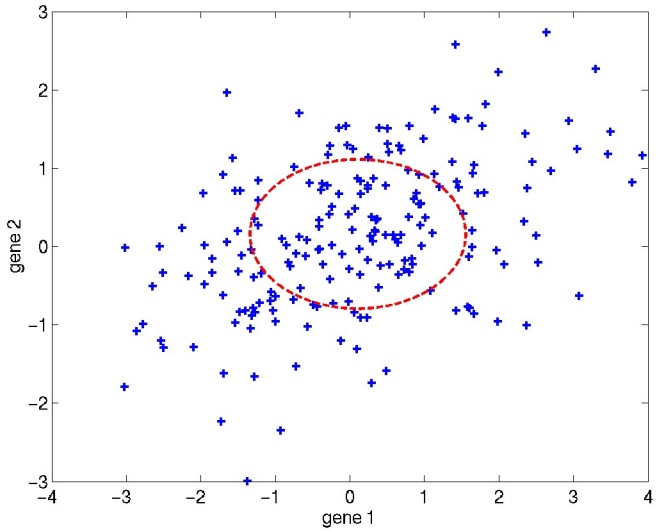
$$p = Prob(\, T(X^{(1)}, \ldots, X^{(n)}) \geq T_{obs} \,|\, H_0 \,) \tag{2}$$

Degrees of freedom

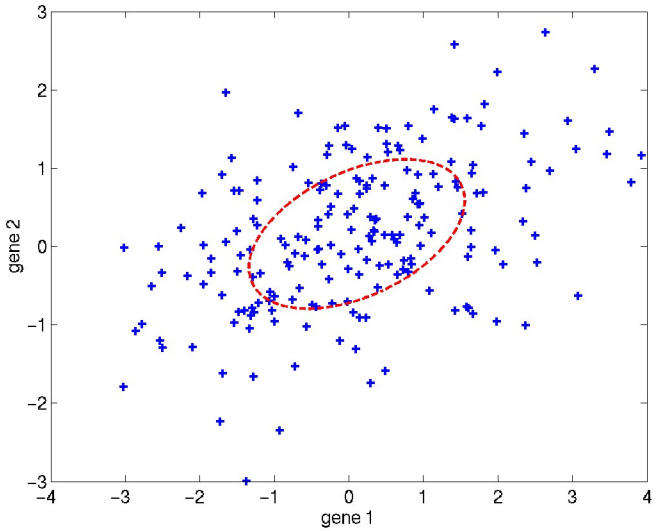
- How many degrees of freedom do we have in the two models?

$$H_0: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right)$$

$$H_1: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$



H_0



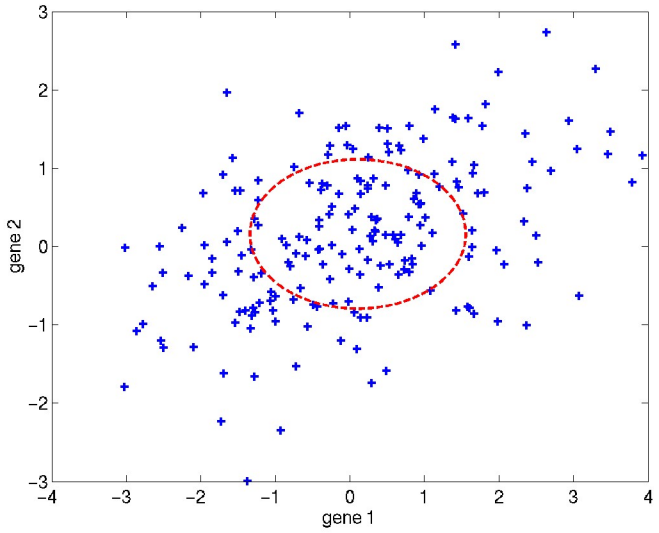
H_1

Degrees of freedom

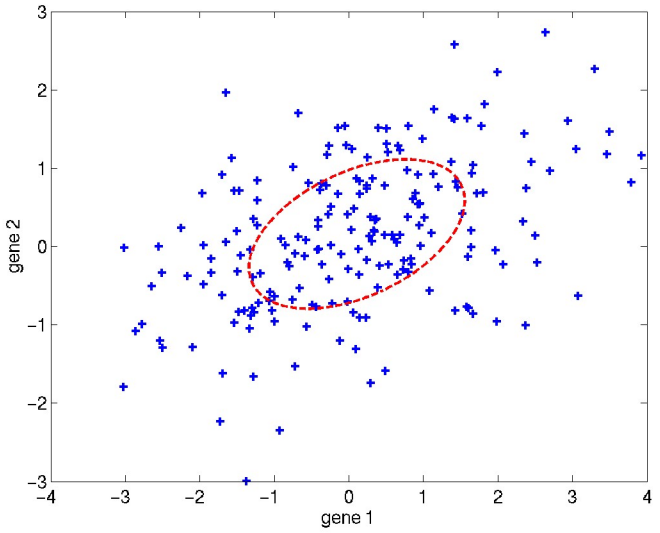
- How many degrees of freedom do we have in the two models?

$$H_0: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right)$$

$$H_1: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$



H_0



H_1

- The observed data overwhelmingly supports H_1

Significance of clusters

- For k-means each value of k implies a a different number of degrees of freedom
  - A gene cluster has a centroid
  - A centroid contains n values, where n is the number of experiments.
  - Thus in this case we have $k \times n$ degrees of freedom
- Random vector X_i models the expression of gene i over n experiments. $\mu_{Clust(i)}$ is the centroid of the cluster of gene i

$$H_0 : X_i \sim N(\mu_{Clust(i)}, \Sigma) \quad \|Range(Clust)\| = (j-1) \tag{3}$$

$$H_1 : X_i \sim N(\mu_{Clust(i)}, \Sigma) \quad \|Range(Clust)\| = j \tag{4}$$

Principle Component Analysis (PCA)

- How can we discover vector components that describe our data?
  1. To discover hidden factors that explain the data
  2. Similar to cluster centroids
  3. To reduce the dimensionality of our data

Multi-Variate Gaussian Review

- Recall multi-variate Gaussians:

$$
\begin{aligned}
Z_i &\sim N(0,1) && (5)\\
X &= AZ + \mu && (6)\\
\Sigma &= E[(X-\mu)(X-\mu)^T] && (7)\\
&= E[(AZ)(AZ)^T] && (8)\\
&= E[AZZ^TA^T] && (9)\\
&= AE[ZZ^T]A^T && (10)\\
&= AA^T && (11)
\end{aligned}
$$

- A multivariate Gaussian model

$$
\begin{aligned}
p(x|\theta) &= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{ -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) \} && (12)\\
X &\sim N(\mu,\Sigma) && (13)
\end{aligned}
$$

  where μ is the mean vector and Σ is the covariance matrix

Principle Component Analysis (PCA)

- Consider the variance of X projected onto vector v

$$\begin{aligned} Var(v^T X) &= E[(v^T X)^2] - E[v^T X]^2 && (14)\\ &= v^T E[XX^T]v - v^T E[X]E[X^T]v && (15)\\ &= v^T (E[XX^T] - E[X]E[X^T])v && (16)\\ &= v^T \Sigma v && (17) \end{aligned}$$

- We would like to pick v_i to maximize the variance with the constraint $v_i^T v_i = 1$. Each v_i will be orthogonal to all of the other v_i
- The v_i are called the eigenvectors of Σ and λ_i^2 are the eigenvalues:

$$\begin{aligned} \Sigma v_i &= \lambda_i^2 v_i && (18)\\ v_i^T \Sigma v_i &= v_i^T \lambda_i^2 v_i && (19)\\ v_i^T \Sigma v_i &= \lambda_i^2 v_i^T v_i && (20)\\ v_i^T \Sigma v_i &= \lambda_i^2 && (21) \end{aligned}$$

Principle Component Analysis (PCA)

- How do we find the eigenvectors v_i?
- We use singular value decomposition to decompose Σ into an orthogonal rotation matrix U and a diagonal scaling matrix S:

$$\begin{aligned} \Sigma &= USU^T && (22) \\ \Sigma U &= (USU^T)U && (23) \\ &= US && (24) \end{aligned}$$

- The columns of U are the v_i, and S is the diagonal matrix of eigenvalues λ_i^2

Principle Component Analysis (PCA)

- How do we interpret eigenvectors and eigenvalues with respect to our orginal transform A?

$$X = AZ + \mu \tag{25}$$

- A is:

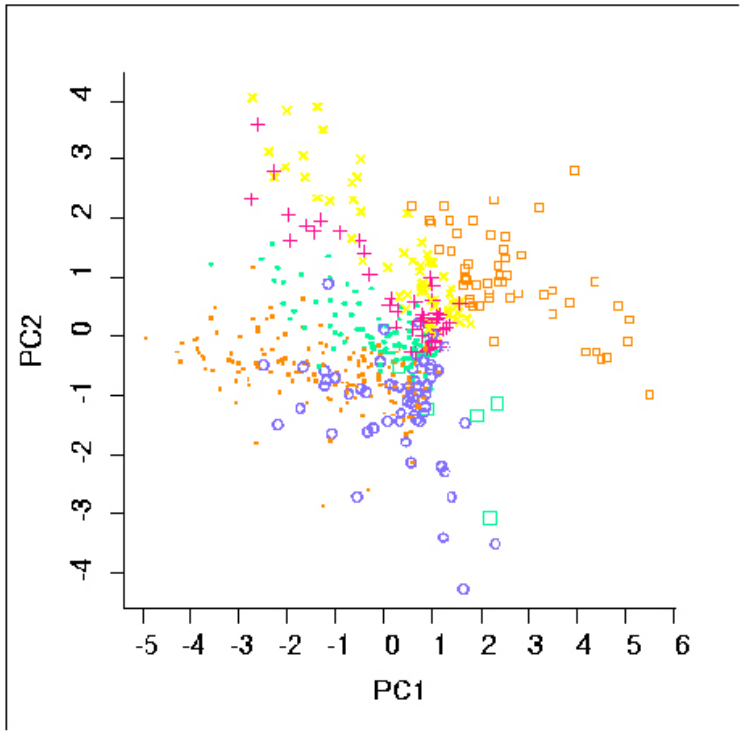
$$A = US^{1/2} \tag{26}$$
$$\Sigma = AA^T \tag{27}$$
$$\Sigma = USU^T \tag{28}$$

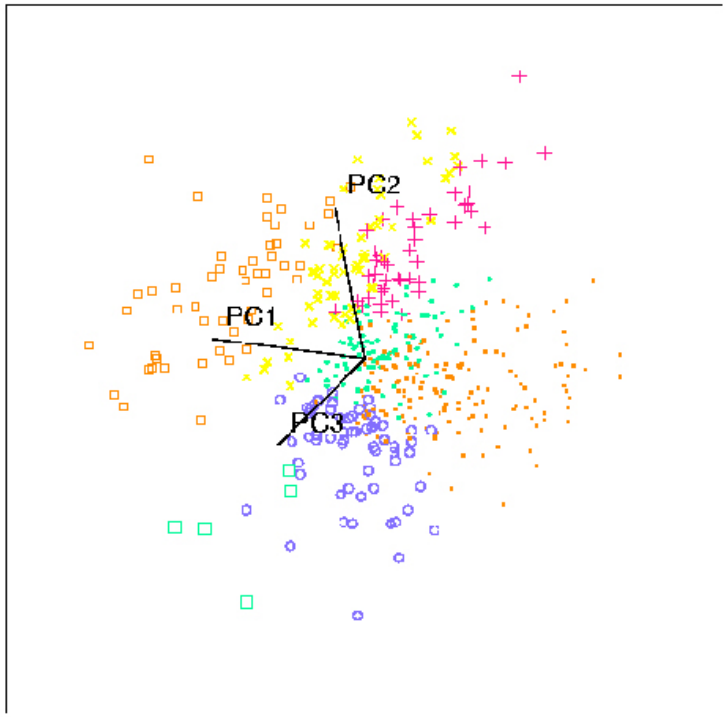
- Thus, the transformation A scales by $S^{1/2}$ and rotates by U independent Gaussians to make X

$$Z_i \sim N(0,1) \tag{29}$$
$$X = US^{1/2}Z + \mu \tag{30}$$

Example PCA Analysis

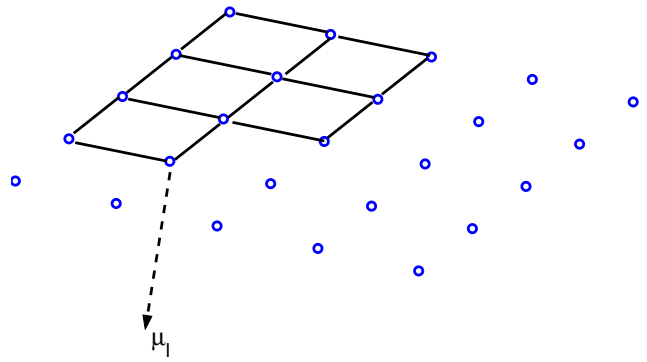
477 sporulation genes classified into seven patterns resovled by PCA





Self-organizing maps

- We want to cluster the data while preserving certain predefined topographic (neighborhood) relations among the clusters
- First we have to specify the desired cluster topology or grid



- Each grid point l has a cluster centroid μ_l associated with it (initially chosen at random)

  We have to update the cluster centroids somehow while preserving the topographic organization of the data

Self-organizing maps cont'd

- For each training example $\mathbf{x}_i$

  1. find the cluster centroid μ_l closest to $\mathbf{x}_i$

$$\mu_{l^*} = \arg\min_{l \in grid} d(\mathbf{x}_i, \mu_l) \tag{31}$$

  2. move the centroid *as well as nearby centroids in the grid* towards the training point

$$\mu_l \leftarrow \mu_l + \epsilon \, \Lambda(l^*, l) \, (\mathbf{x}_i - \mu_l) \tag{32}$$

  where $\Lambda(l^*, l)$ is a neighborhood function (decreases with increasing distance in the original grid), e.g.,

$$\Lambda(l^*, l) = \exp(-\|r_{l^*} - r_l\|^2/2) \tag{33}$$

  where r_{l^*} and r_l are the corresponding grid points.