

L15: Microarray analysis (Classification)

Silly Quiz

- Social networking site:
- How can you find people with interests similar to yours?

Gene Expression Data

- Gene Expression data:
 - Each row corresponds to a gene
 - Each column corresponds to an expression value
- Can we separate the experiments into two or more classes?
- Given a training set of two classes, can we build a classifier that places a new experiment in one of the two classes.

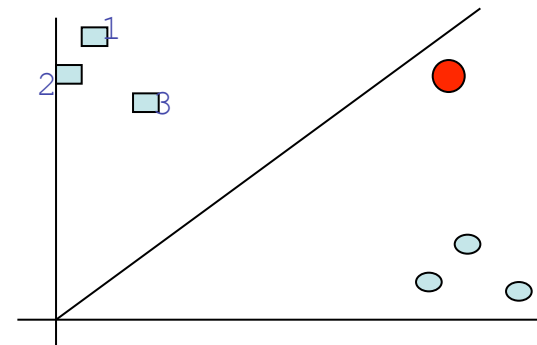
A grid representing gene expression data. The columns are labeled S_1 , S_2 , and S . The row is labeled g . The grid is divided into three vertical sections: the first two columns (S_1 and S_2) are light blue, the next three columns are green, and the final column (S) is white.

	S_1	S_2				S
g						

Formalizing Classification

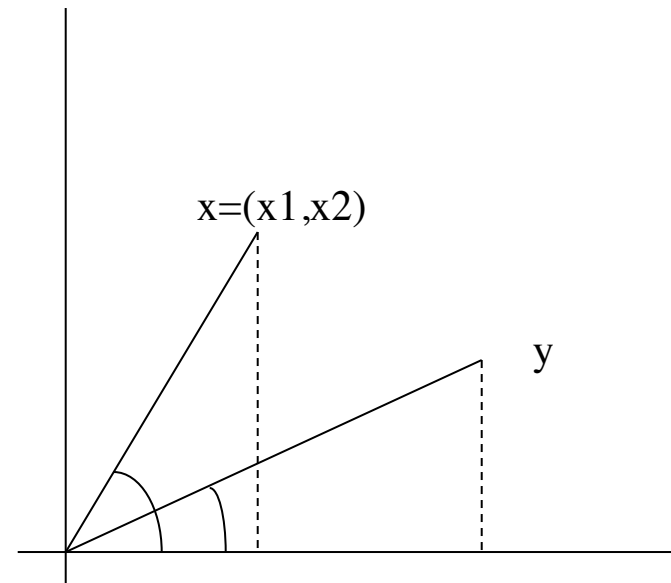
- Classification problem: Find a surface (hyperplane) that will separate the classes
- Given a new sample point, its class is then determined by which side of the surface it lies on.
- How do we find the hyperplane? How do we find the side that a point lies on?

	1	2	3	4	5	6	
g1	1	.9	.8	.1	.2	.1	
g2	.1	0	.2	.8	.7	.9	



Basic geometry

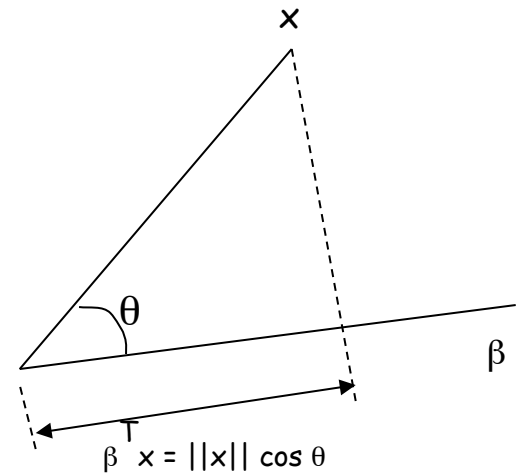
- What is $\|x\|_2$?
- What is $x/\|x\|$
- Dot product?



$$\begin{aligned}x^T y &= x_1 y_1 + x_2 y_2 \\ &= \|x\| \cdot \|y\| \cos \theta_x \cos \theta_y + \|x\| \cdot \|y\| \sin(\theta_x) \sin(\theta_y) \\ &= \|x\| \cdot \|y\| \cos(\theta_x - \theta_y)\end{aligned}$$

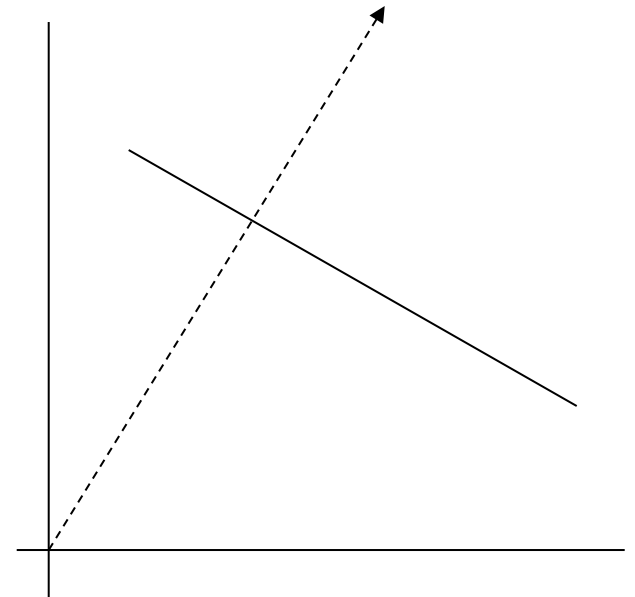
Dot Product

- Let β be a unit vector.
 - $\|\beta\| = 1$
- Recall that
 - $\beta^T \mathbf{x} = \|\mathbf{x}\| \cos \theta$
- What is $\beta^T \mathbf{x}$ if \mathbf{x} is orthogonal (perpendicular) to β ?



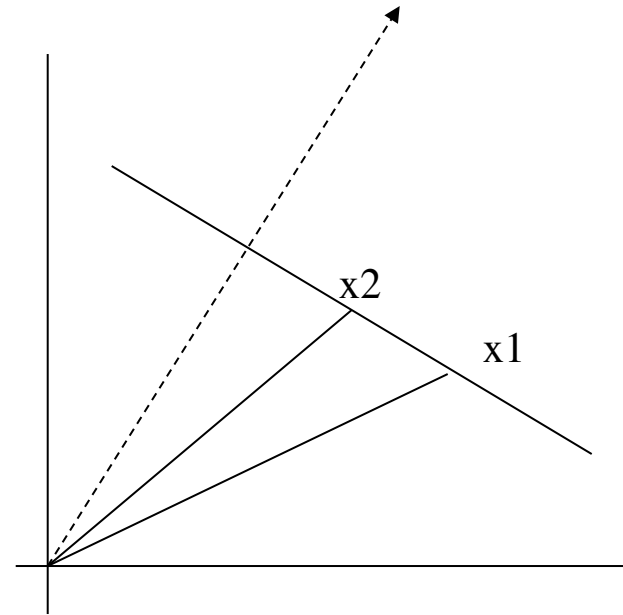
Hyperplane

- How can we define a hyperplane L ?
- Find the unit vector that is perpendicular (normal to the hyperplane)



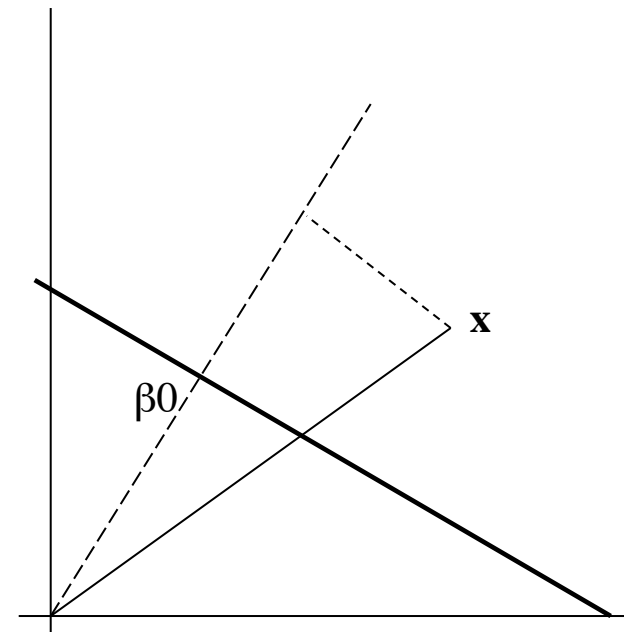
Points on the hyperplane

- Consider a hyperplane L defined by unit vector β , and distance β_0
- Notes;
 - For all $x \in L$, $x^T \beta$ must be the same, $x^T \beta = \beta_0$
 - For any two points x_1, x_2 ,
 - $(x_1 - x_2)^T \beta = 0$



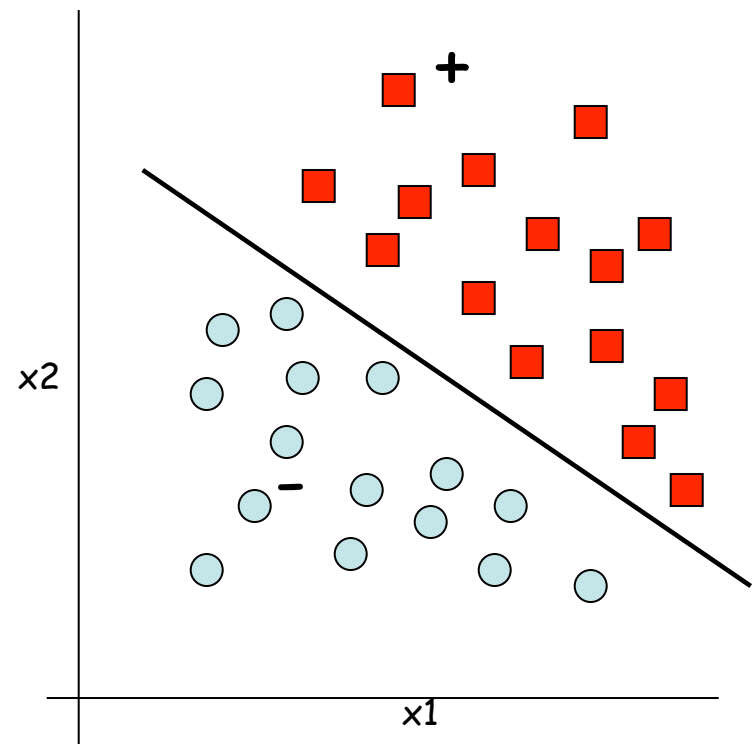
Hyperplane properties

- Given an arbitrary point x , what is the distance from x to the plane L ?
 - $D(x,L) = (\beta^T x - \beta_0)$
- When are points x_1 and x_2 on different sides of the hyperplane?



Separating by a hyperplane

- Input: A training set of +ve & -ve examples
- Goal: Find a hyperplane that separates the two classes.
- Classification: A new point x is +ve if it lies on the +ve side of the hyperplane, -ve otherwise.
- The hyperplane is represented by the line
- $\{x: -\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0\}$

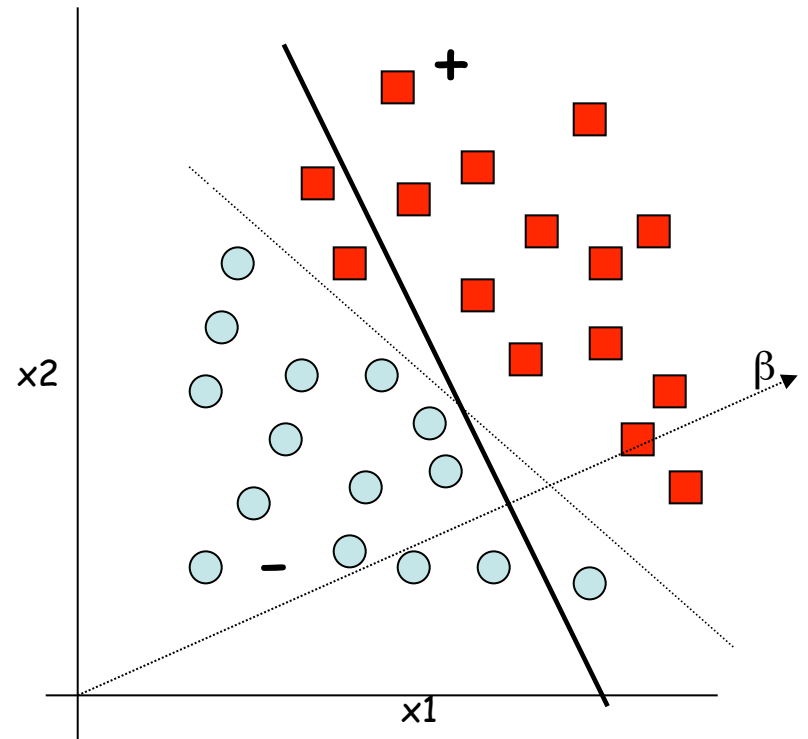


Error in classification

- An arbitrarily chosen hyperplane might not separate the test. We need to minimize a mis-classification error
- Error: sum of distances of the misclassified points.
- Let $y_i = -1$ for +ve example i ,
 - $y_i = 1$ otherwise.

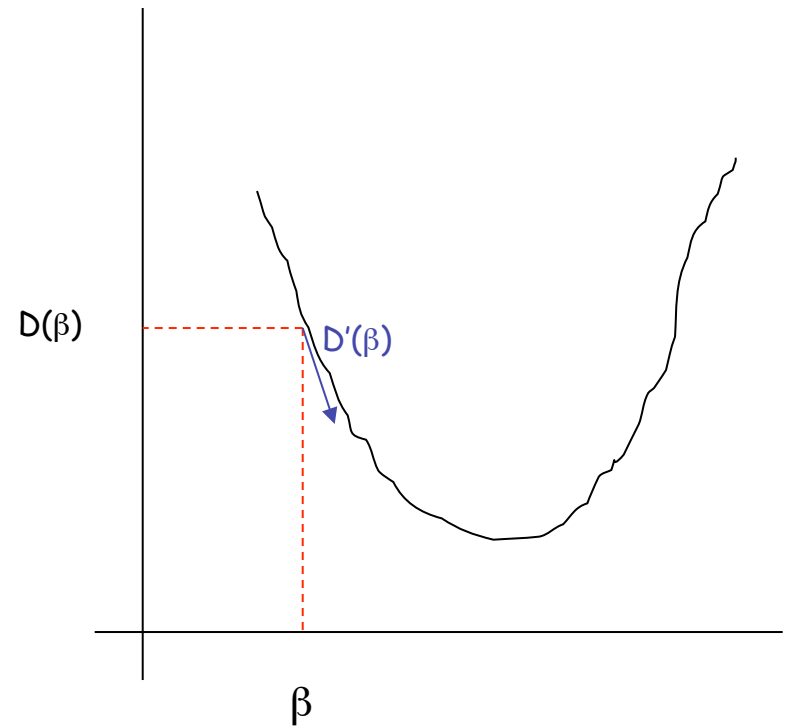
$$D(\beta, \beta_0) = \sum_{i \in M} y_i (x_i^T \beta + \beta_0)$$

- Other definitions are also possible.



Gradient Descent

- The function $D(\beta)$ defines the error.
- We follow an iterative refinement. In each step, refine β so the error is reduced.
- Gradient descent is an approach to such iterative refinement.



November 09 $\beta \leftarrow \beta - \rho \cdot D'(\beta)$

Rosenblatt's perceptron learning algorithm

$$D(\beta, \beta_0) = \sum_{i \in M} y_i (x_i^T \beta + \beta_0)$$

$$\frac{\partial D(\beta, \beta_0)}{\partial \beta} = \sum_{i \in M} y_i x_i$$

$$\frac{\partial D(\beta, \beta_0)}{\partial \beta_0} = \sum_{i \in M} y_i$$

$$\Rightarrow \text{Update rule : } \begin{pmatrix} \beta \\ \beta_0 \end{pmatrix} = \begin{pmatrix} \beta \\ \beta_0 \end{pmatrix} - \rho \begin{pmatrix} \sum_{i \in M} y_i x_i \\ \sum_{i \in M} y_i \end{pmatrix}$$

Classification based on perceptron learning

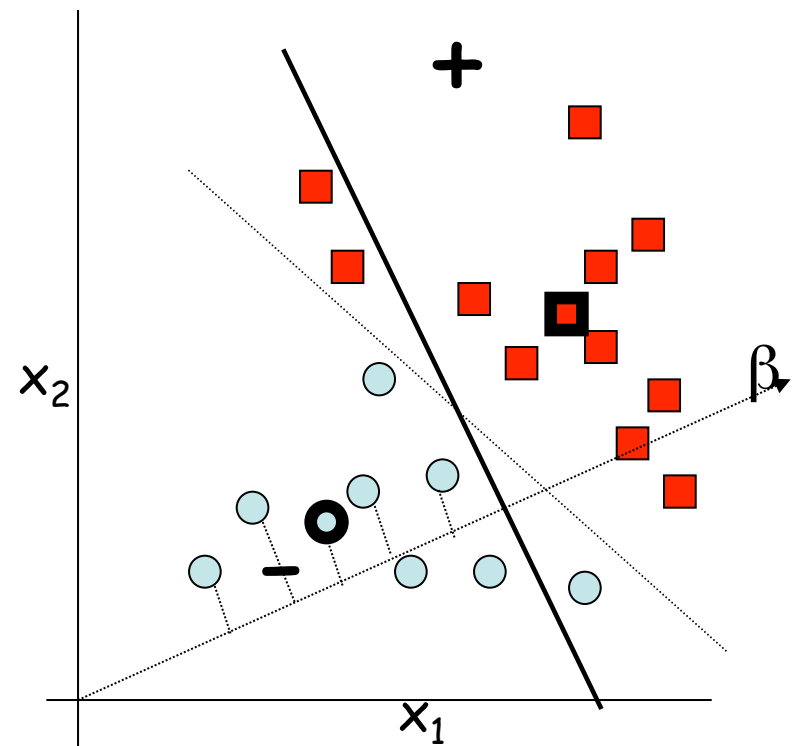
- Use Rosenblatt's algorithm to compute the hyperplane $L=(\beta,\beta_0)$.
- Assign x to class 1 if $f(x) \geq 0$, and to class 2 otherwise.

Perceptron learning

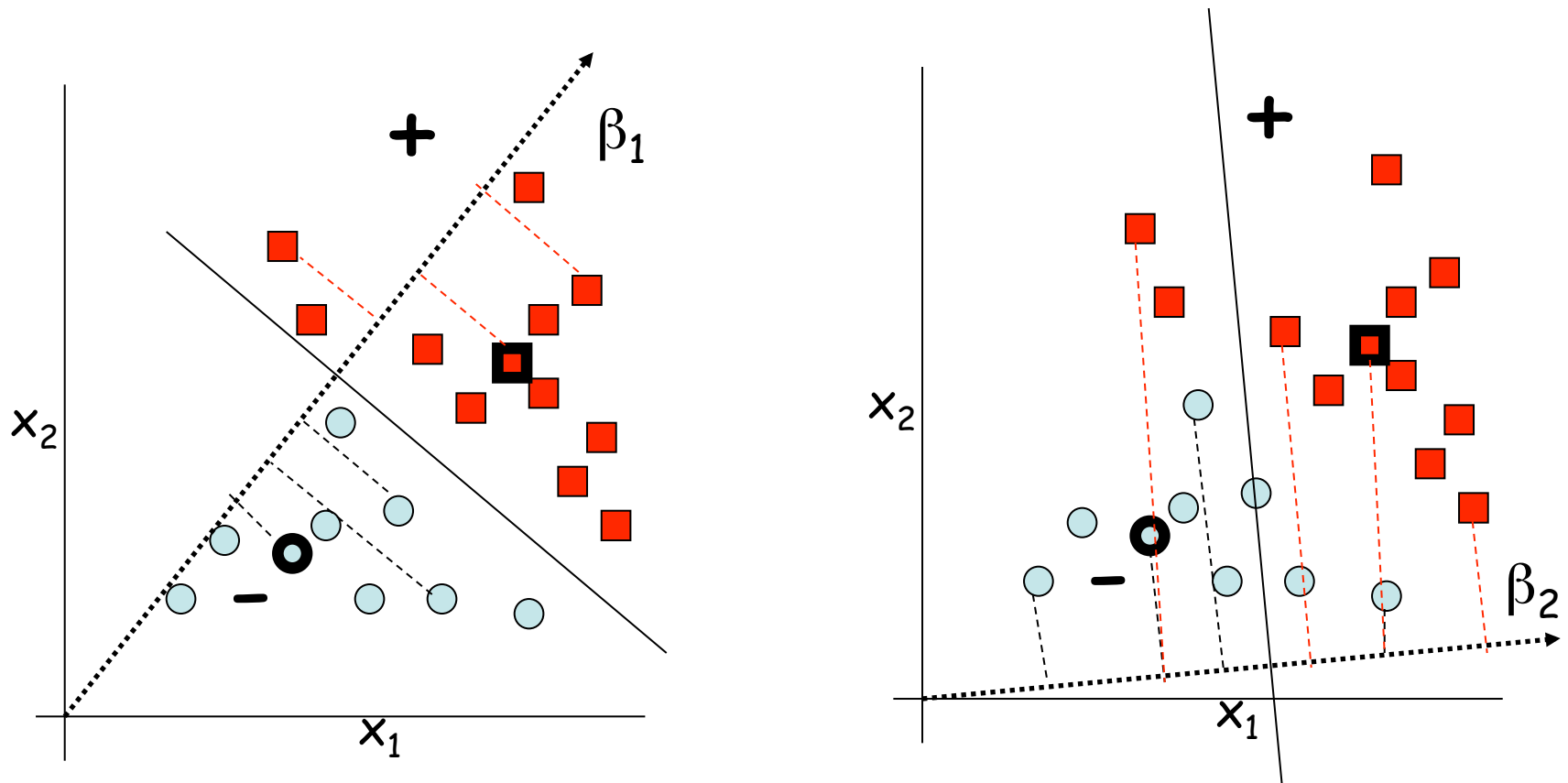
- If many solutions are possible, it does not choose between solutions
- If data is not linearly separable, it does not terminate, and it is hard to detect.
- Time of convergence is not well understood

Linear Discriminant analysis

- Provides an alternative approach to classification with a linear function.
- Project all points, including the means, onto vector β .
- We want to choose β such that
 - Difference of projected means is large.
 - Variance within group is small



Choosing the right β



- β_1 is a better choice than β_2 as the variance within a group is small, and difference of means is large.
- How do we compute the best β ?

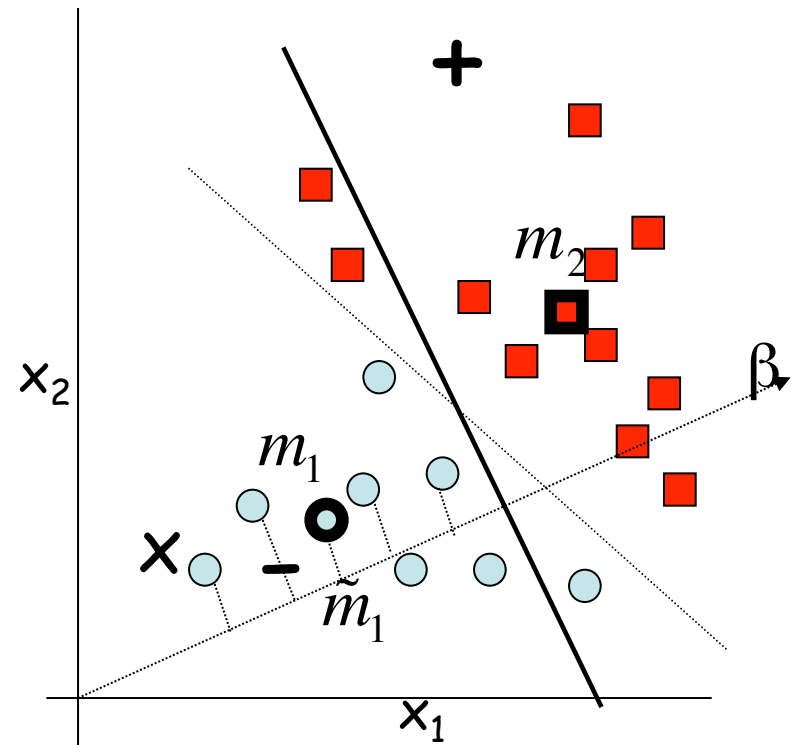
Linear Discriminant analysis

- Fisher Criterion

$$\text{Max}_{\beta} \frac{(\text{difference of projected means})}{(\text{sum of projected variance})}$$

LDA cont'd

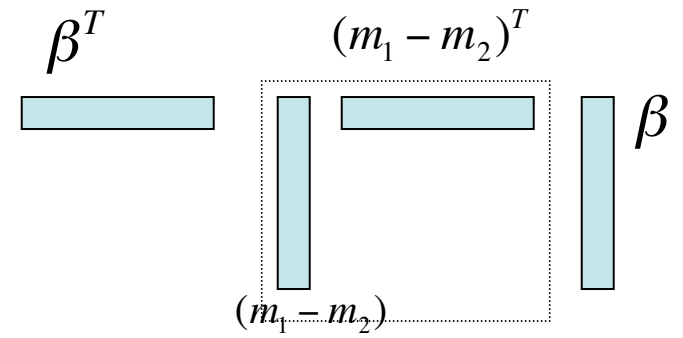
- What is the projection of a point x onto β ?
 - Ans: $\beta^T x$
- What is the distance between projected means?



$$|\tilde{m}_1 - \tilde{m}_2|^2 = |\beta^T (m_1 - m_2)|^2$$

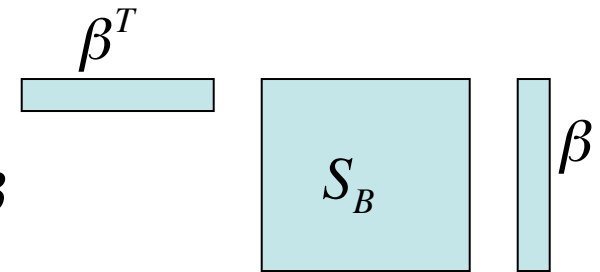
LDA Cont'd

$$\begin{aligned}
 |\tilde{m}_1 - \tilde{m}_2|^2 &= |\beta^T (m_1 - m_2)|^2 \\
 &= \beta^T (m_1 - m_2)(m_1 - m_2)^T \beta \\
 &= \beta^T S_B \beta
 \end{aligned}$$



scatter within sample: $\tilde{s}_1^2 + \tilde{s}_2^2$

$$\text{where, } \tilde{s}_1^2 = \sum_y (\tilde{x} - \tilde{m}_1)^2 = \sum_{x \in D_1} (\beta^T (x - m_1))^2 = \beta^T S_1 \beta$$



$$\tilde{s}_1^2 + \tilde{s}_2^2 = \beta^T (S_1 + S_2) \beta = \beta^T S_w \beta$$

Fisher Criterion

November 09

$$\max_{\beta} \frac{\beta^T S_B \beta}{\beta^T S_w \beta}$$

LDA

$$\text{Let } \max_{\beta} \frac{\beta^T S_B \beta}{\beta^T S_w \beta} = \lambda$$

$$\text{Then, } \beta^T (S_B \beta - \lambda S_w \beta) = 0$$

$$\Rightarrow \lambda S_w \beta = S_B \beta$$

$$\Rightarrow \lambda \beta = S_w^{-1} S_B \beta$$

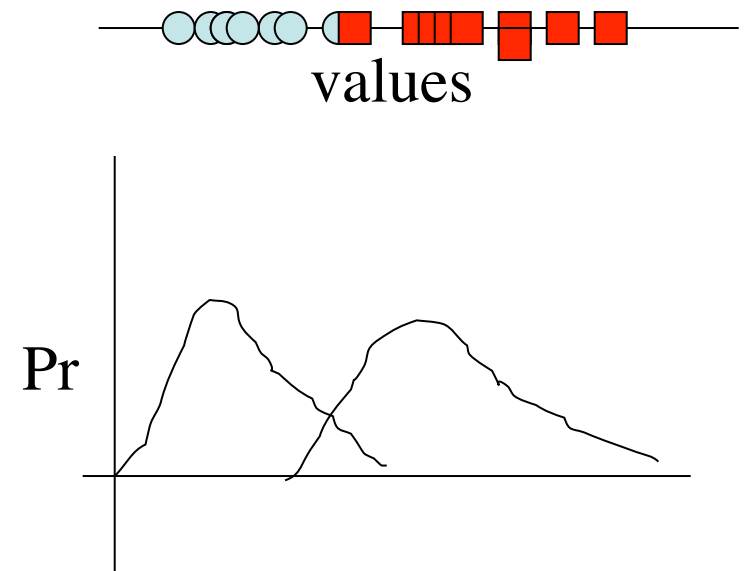
$$\Rightarrow \beta = S_w^{-1} (m_1 - m_2)$$

Therefore, a simple computation (Matrix inverse) is sufficient to compute the 'best' separating hyperplane

End of Lecture 15

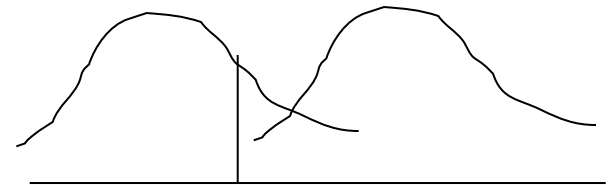
Maximum Likelihood discrimination

- Consider the simple case of single dimensional data.
- Compute a distribution of the values in each class.



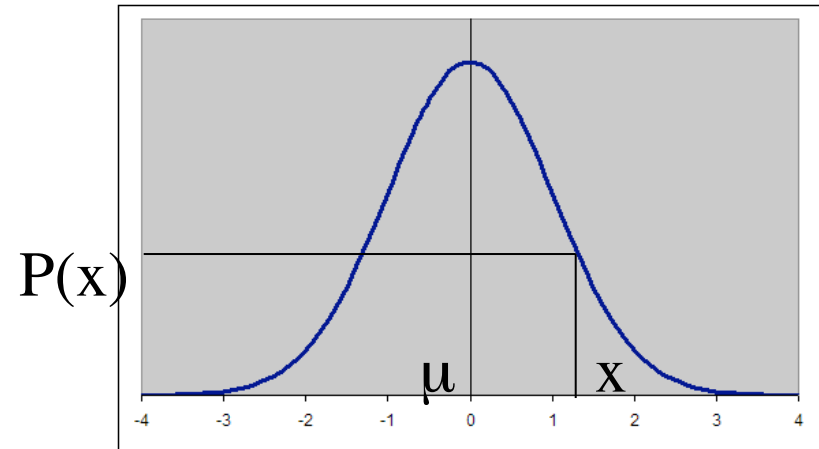
Maximum Likelihood discrimination

- Suppose we knew the distribution of points in each class ω_i .
 - We can compute $\Pr(x|\omega_i)$ for all classes i , and take the maximum
- The true distribution is not known, so usually, we assume that it is Gaussian



ML discrimination

- Use a Bayesian approach to identify the class for each sample



$$\Pr(\omega_i | x) = \frac{\Pr(x | \omega_i) \Pr(\omega_i)}{\sum_j \Pr(x | \omega_j) \Pr(\omega_j)}$$

$$\begin{aligned} g_i(x) &= \ln(\Pr(x | \omega_i)) + \ln(\Pr(\omega_i)) \\ &\cong \frac{-(x - \mu_i)^2}{2\sigma_i^2} + \ln(\Pr(\omega_i)) \end{aligned}$$

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ML discrimination recipe (1 dimensional case)

- We know the distribution for each class, but not the parameters
- Estimate the mean and variance for each class.
- For a new point x , compute the discrimination function $g_i(x)$ for each class i .
- Choose $\operatorname{argmax}_i g_i(x)$ as the class for x

ML discrimination

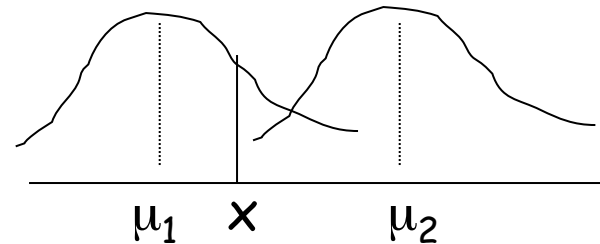
- Suppose all the points were in 1 dimension, and all classes were normally distributed.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

$$\Pr(\omega_i | x) = \frac{\Pr(x | \omega_i) \Pr(\omega_i)}{\sum_j \Pr(x | \omega_j) \Pr(\omega_j)}$$

$$\begin{aligned} g_i(x) &= \ln(\Pr(x | \omega_i)) + \ln(\Pr(\omega_i)) \\ &\cong \frac{-(x - \mu_i)^2}{2\sigma_i^2} - \ln(\sigma_i) + \ln(\Pr(\omega_i)) \end{aligned}$$

$$\text{Choose } \operatorname{argmin}_i \left(\frac{(x - \mu_i)^2}{2\sigma_i^2} + \ln(\sigma_i) - \ln(\Pr(\omega_i)) \right)$$



ML discrimination (multi-dimensional case)

Sample mean, $\hat{\mu} = \frac{1}{n} \sum_i \vec{x}_i$

Covariance matrix $= \hat{\Sigma} = \frac{1}{n-1} \sum_k \left(\vec{x}_k - \hat{\mu} \right) \left(\vec{x}_k - \hat{\mu} \right)^T$

ML discrimination (multi-dimensional case)

$$p(\mathbf{x} | \omega_i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m})^T \Sigma^{-1}(\mathbf{x} - \mathbf{m})\right)$$

$$g_i(\mathbf{x}) = \ln(p(\mathbf{x} | \omega_i)) + \ln P(\omega_i)$$

Compute $\arg \max_i g_i(\mathbf{x})$

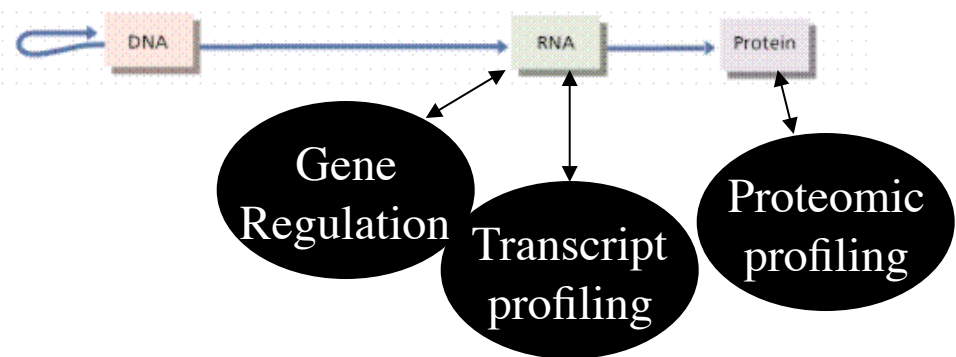
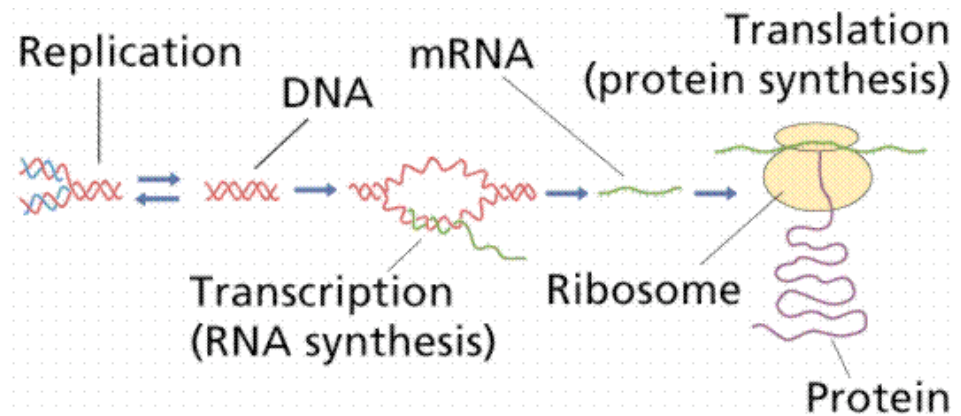
Supervised classification summary

- Most techniques for supervised classification are based on the notion of a separating hyperplane.
- The 'optimal' separation can be computed using various combinatorial (perceptron), algebraic (LDA), or statistical (ML) analyses.

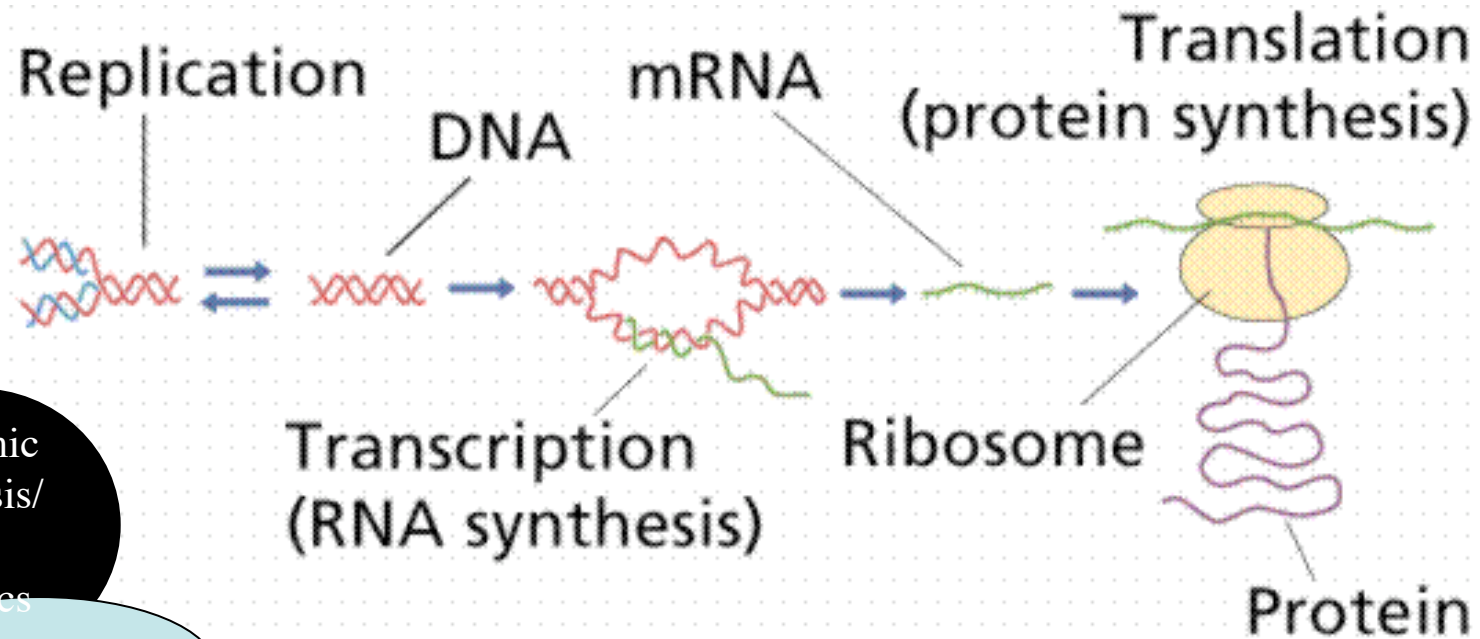
Review of micro-array analysis

The dynamic picture of the cellular activity

- Each Cell is continuously active,
 - Genes are being transcribed into RNA
 - RNA is translated into proteins
 - Proteins are PT modified and transported
 - Proteins perform various cellular functions
- Can we probe the Cell dynamically?
 - Which transcripts are active?
 - Which proteins are active?
 - Which proteins interact?



Other static analysis is possible



Genomic Analysis/
Pop. Genetics

Assembly

Sequence Analysis

November 09



Gene Finding

Bafna

Protein Sequence Analysis

ncRNA

Silly Quiz

- Who are these people, and what is the occasion?



November 09

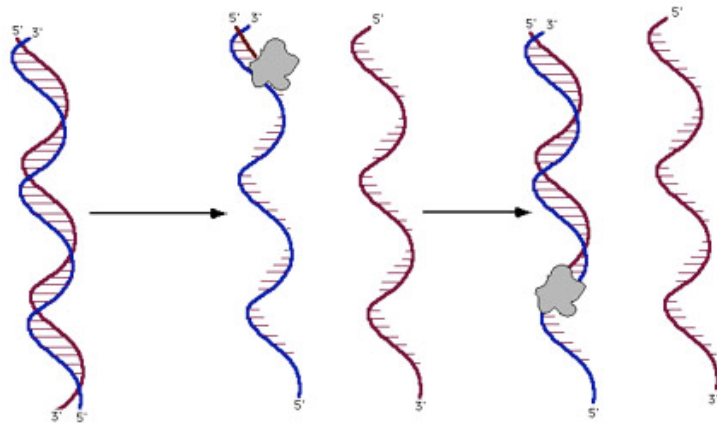
Bafn

Genome Sequencing and Assembly

November 09

Bafna

DNA Sequencing



- DNA is double-stranded
- The strands are separated, and a polymerase is used to copy the second strand.
- Special bases terminate this process early.

Sequencing

- A break at T is shown here.
- Measuring the lengths using electrophoresis allows us to get the position of each T
- The same can be done with every nucleotide. Fluorescent labeling can help separate different nucleotides

DNA Polymerase reads the template strand and synthesizes a new second strand to match:

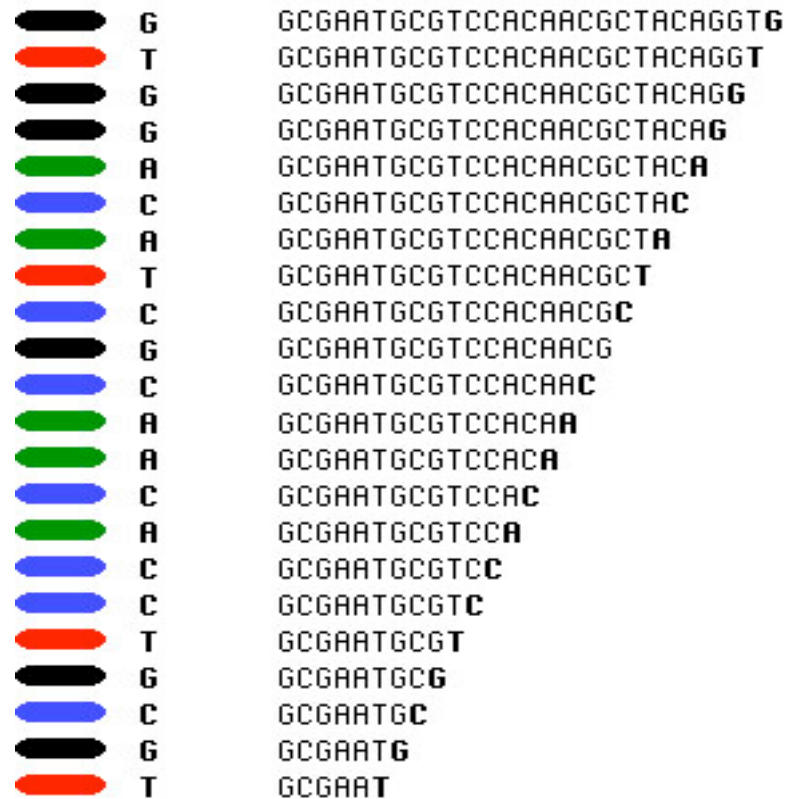


IF 5% of the T nucleotides are actually dideoxy T, then each strand will terminate when it gets a ddT on its growing end:



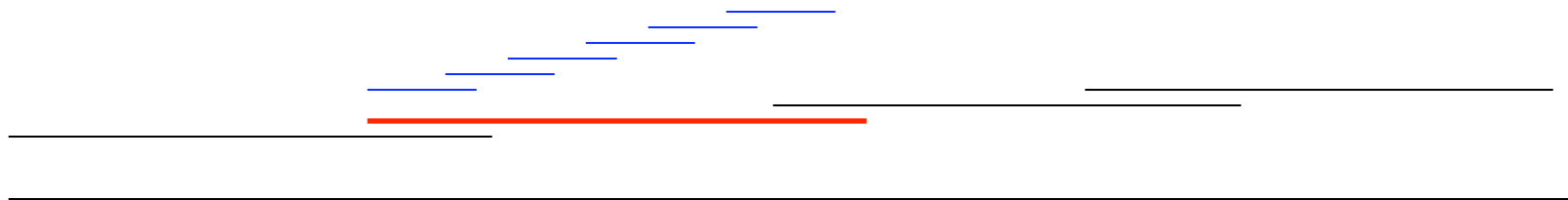
- Automated detectors 'read' the terminating bases.
- The signal decays after 1000 bases.

Gel:

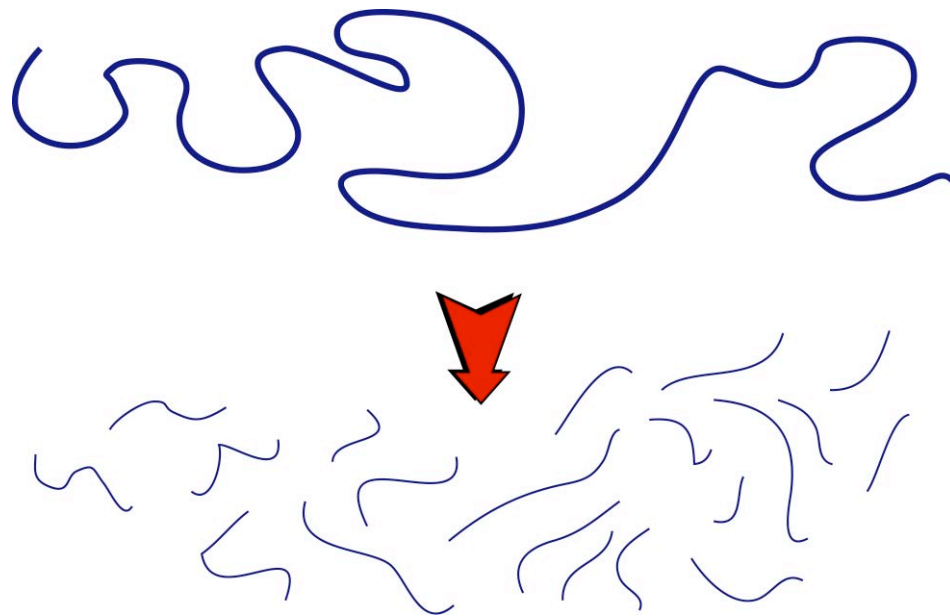


Sequencing Genomes: Clone by Clone

- Clones are constructed to span the entire length of the genome.
- These clones are ordered and oriented correctly (Mapping)
- Each clone is sequenced individually

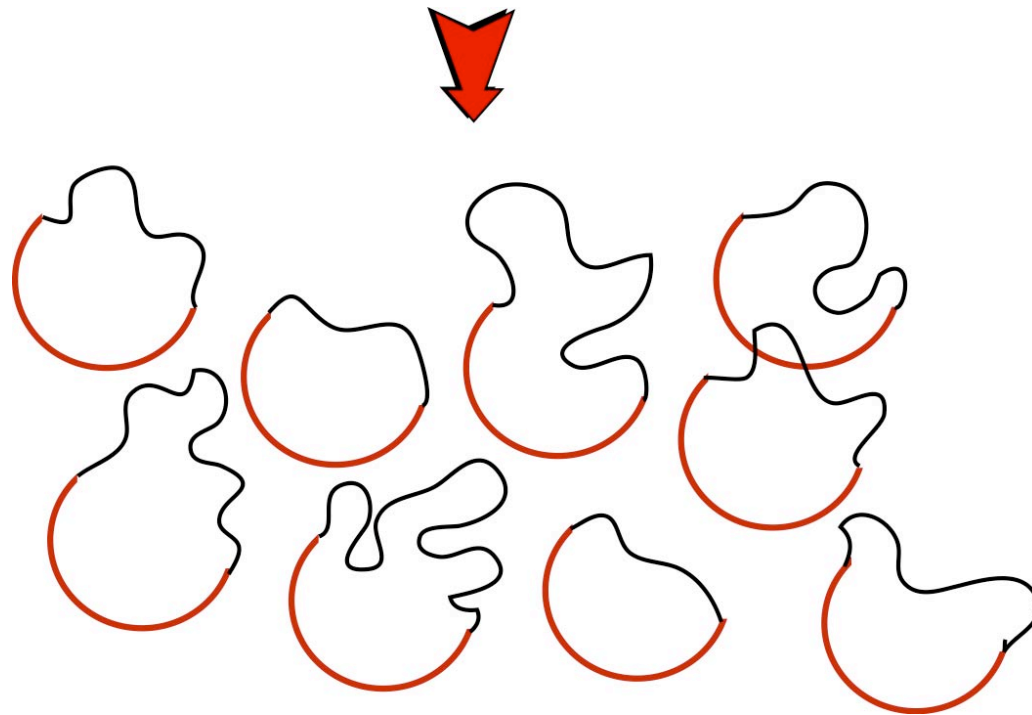


Shotgun Sequencing



- Shotgun sequencing of clones was considered viable
- However, researchers in 1999 proposed shotgunning the entire genome.

Library



- Create vectors of the sequence and introduce them into bacteria. As bacteria multiply you will have many copies of the same clone.

Whole Genome Shotgun

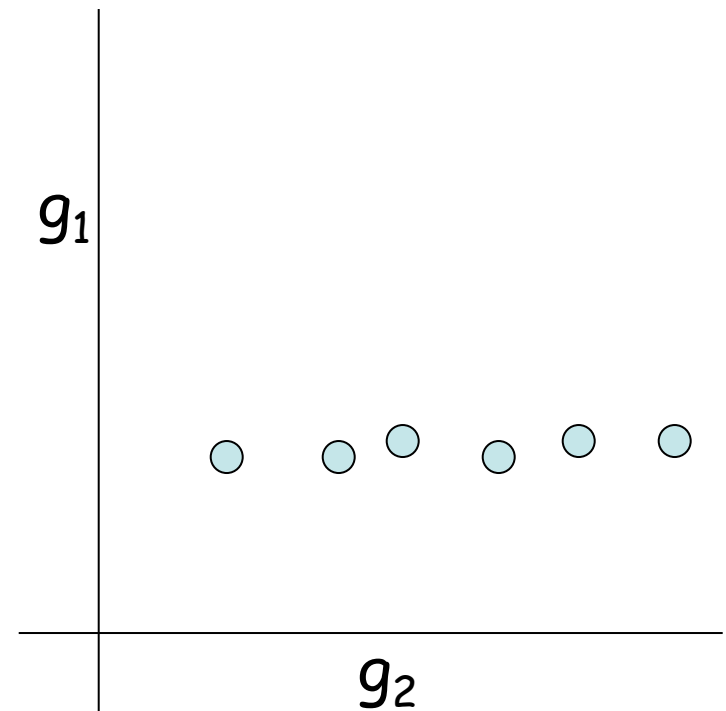
- Break up the entire genome into pieces
- Sequence ends, and assemble using a computer
- LW statistics & Repeats argue against the success of such an approach



Alternative: build a roadmap of the genome, with physical clones mapped for each region. Sequence each of the clones, and put them together

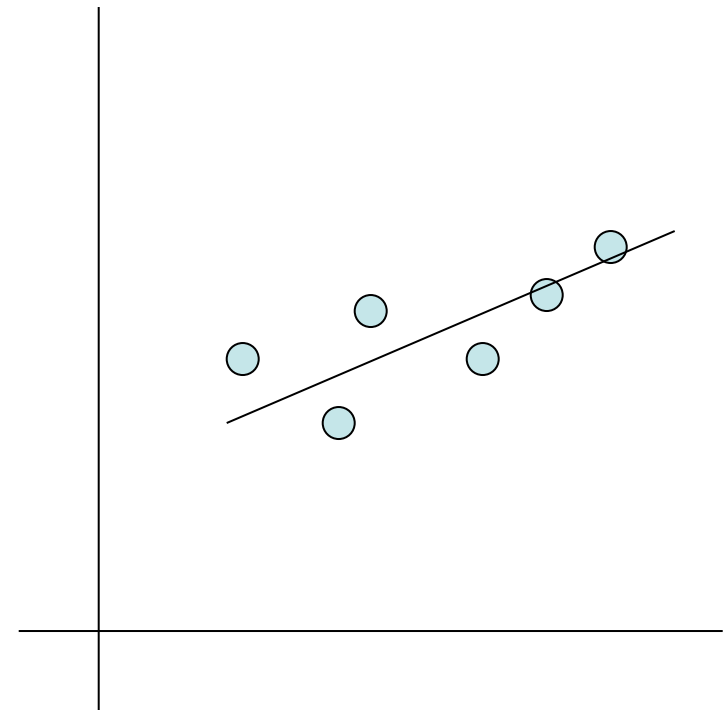
PCA: motivating example

- Consider the expression values of 2 genes over 6 samples.
- Clearly, the expression of g_1 is not informative, and it suffices to look at g_2 values.
- Dimensionality can be reduced by discarding the gene g_1



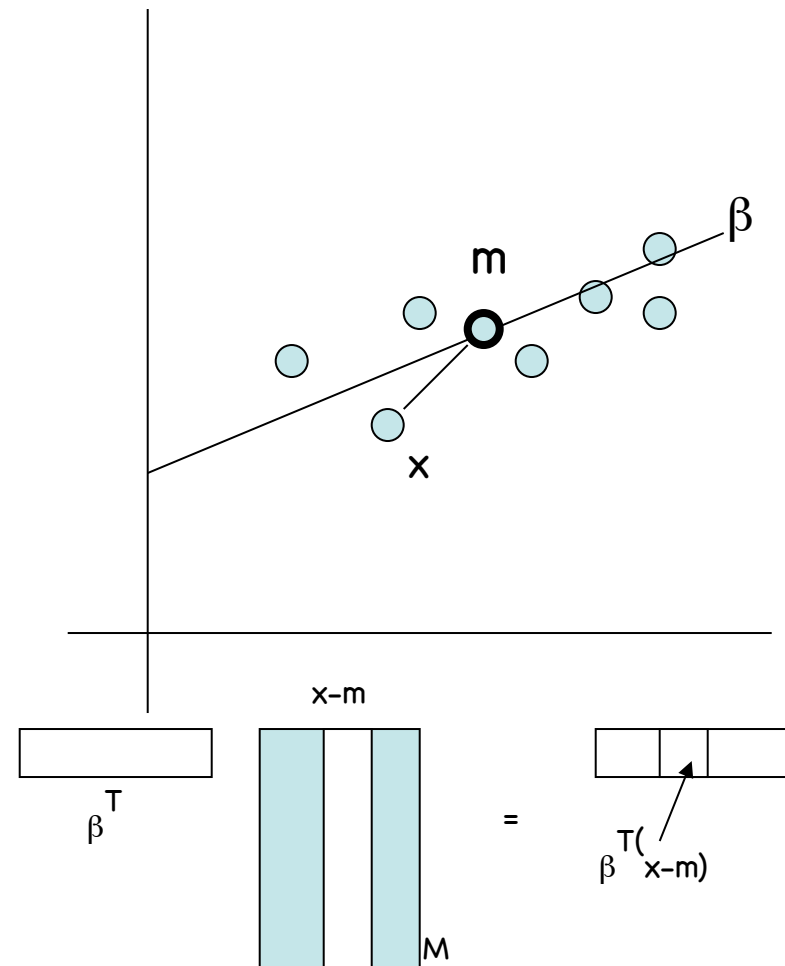
Principle Components Analysis

- Consider the expression values of 2 genes over 6 samples.
- Clearly, the expression of the two genes is highly correlated.
- Projecting all the genes on a single line could explain most of the data.
- This is a generalization of "discarding the gene".



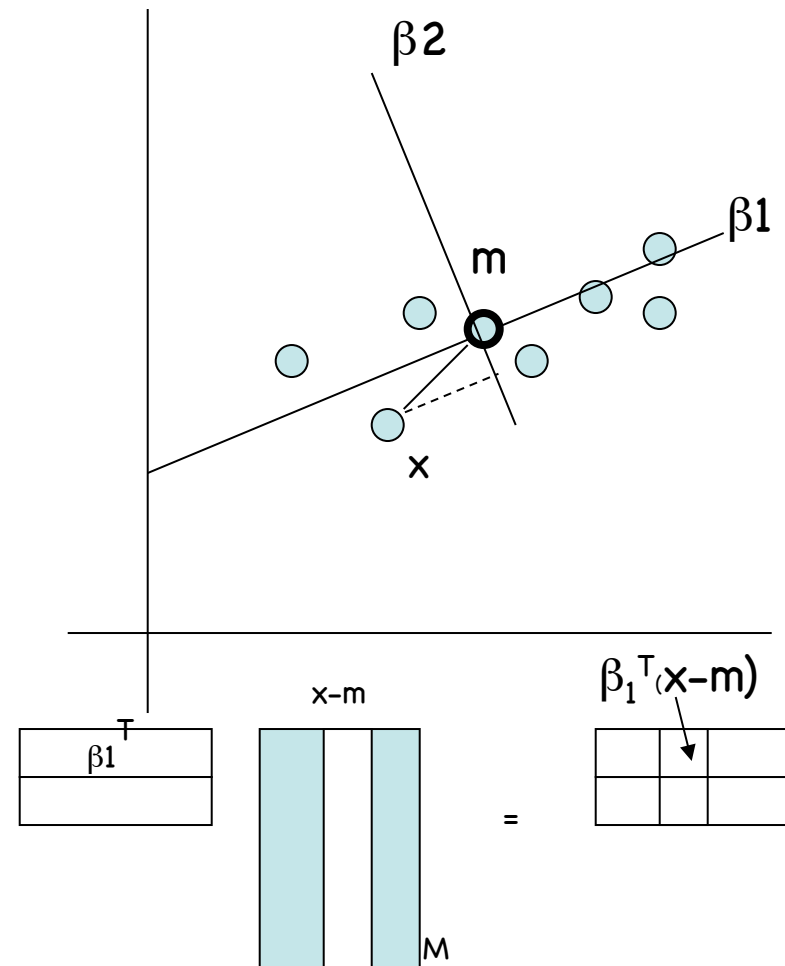
Projecting

- Consider the mean of all points m , and a vector emanating from the mean
- Algebraically, this projection on β means that all samples x can be represented by a single value $\beta^T(x-m)$



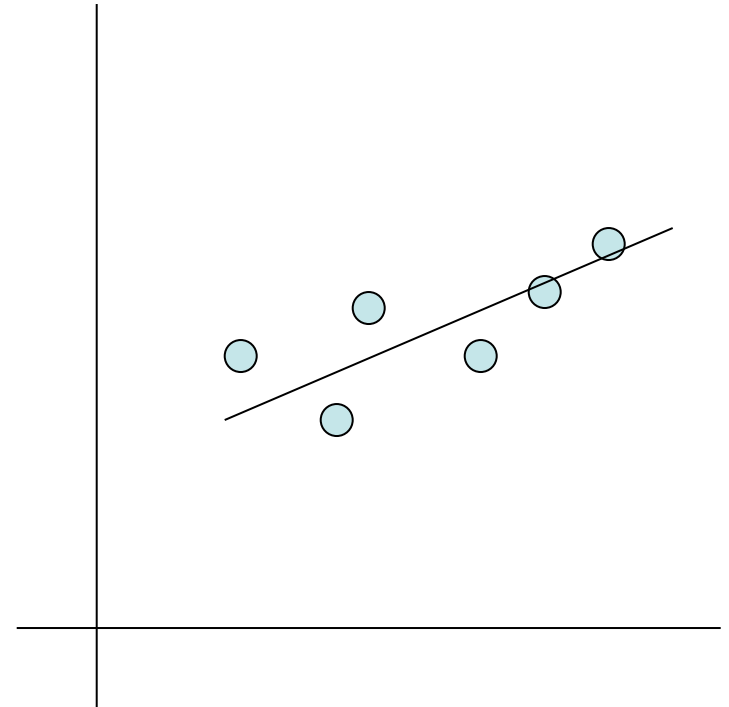
Higher dimensions

- Consider a set of 2 (k) orthonormal vectors $\beta_1, \beta_2 \dots$
- Once projected, each sample means that all samples x can be represented by 2 (k) dimensional vector
 - $\beta_1^T(x-m), \beta_2^T(x-m)$



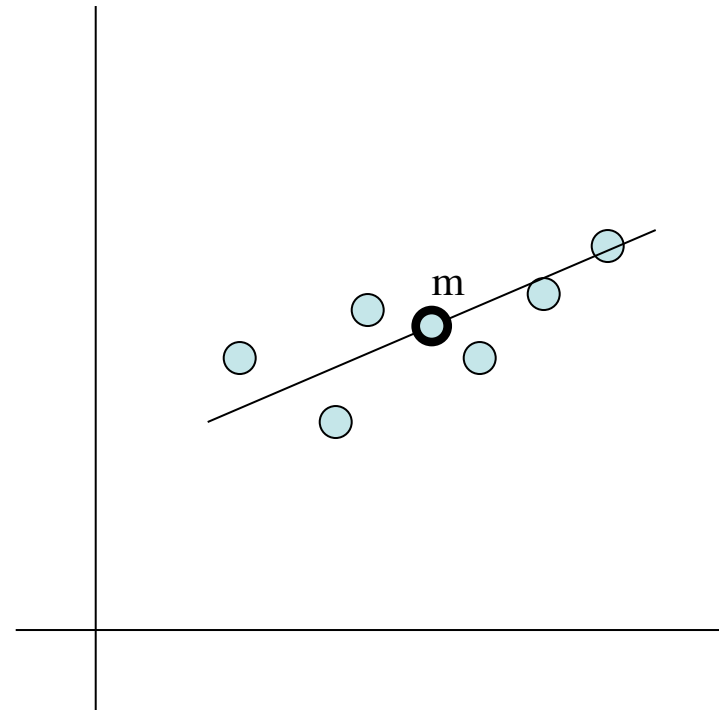
How to project

- The generic scheme allows us to project an m dimensional surface into a k dimensional one.
- How do we select the k 'best' dimensions?
- The strategy used by PCA is one that maximizes the variance of the projected points around the mean



PCA

- Suppose all of the data were to be reduced by projecting to a single line β from the mean.
- How do we select the line β ?

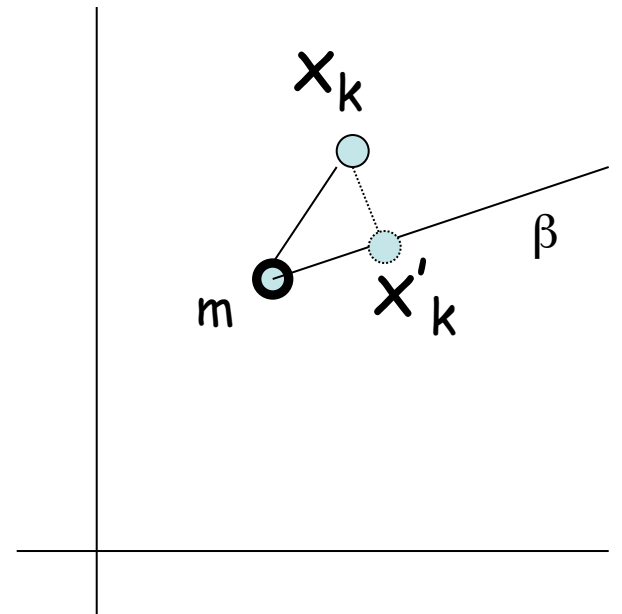


PCA cont'd

- Let each point x_k map to $x'_k = m + a_k \beta$. We want to minimize the error

$$\sum_k \|x_k - x'_k\|^2$$

- Observation 1: Each point x_k maps to $x'_k = m + \beta^T(x_k - m)\beta$
- ($a_k = \beta^T(x_k - m)$)



Proof of Observation 1

$$\begin{aligned} & \min_{a_k} \|x_k - x'_k\|^2 \\ &= \min_{a_k} \|x_k - m + m - x'_k\|^2 \\ &= \min_{a_k} \|x_k - m\|^2 + \|m - x'_k\|^2 - 2(x'_k - m)^T (x_k - m) \\ &= \min_{a_k} \|x_k - m\|^2 + a_k^2 \beta^T \beta - 2a_k \beta^T (x_k - m) \\ &= \min_{a_k} \|x_k - m\|^2 + a_k^2 - 2a_k \beta^T (x_k - m) \end{aligned}$$

Differentiating w.r.t a_k

$$2a_k - 2\beta^T (x_k - m) = 0$$

$$a_k = \beta^T (x_k - m)$$

$$\Rightarrow a_k^2 = a_k \beta^T (x_k - m)$$

$$\Rightarrow \|x_k - x'_k\|^2 = \|x_k - m\|^2 - \beta^T (x_k - m)(x_k - m)^T \beta$$

Minimizing PCA Error

$$\begin{aligned} & \sum_k \|x_k - x'_k\|^2 \\ &= C - \sum_k \beta^T (x_k - m)(x_k - m)^T \beta = C - \beta^T S \beta \end{aligned}$$

- To minimize error, we must maximize $\beta^T S \beta$
- By definition, $\lambda = \beta^T S \beta$ implies that λ is an eigenvalue, and β the corresponding eigenvector.
- Therefore, we must choose the eigenvector corresponding to the largest eigenvalue.

PCA steps

$$1. m = \frac{1}{n} \sum_{j=1}^n x_j$$

$$2. h^T = [1 \ 1 \ \dots \ 1]$$

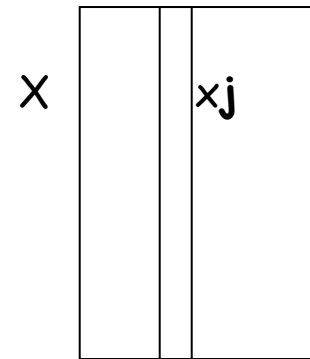
$$3. M = X - mh^T$$

$$4. S = MM^T = \sum_{j=1}^n (x_j - m)(x_j - m)^T$$

$$5. B^T SB = \Lambda$$

6. Return $B^T M$

- X = starting matrix with n columns, m rows



November 09

Bafna