LECTURE 10: Linear Discriminant Analysis

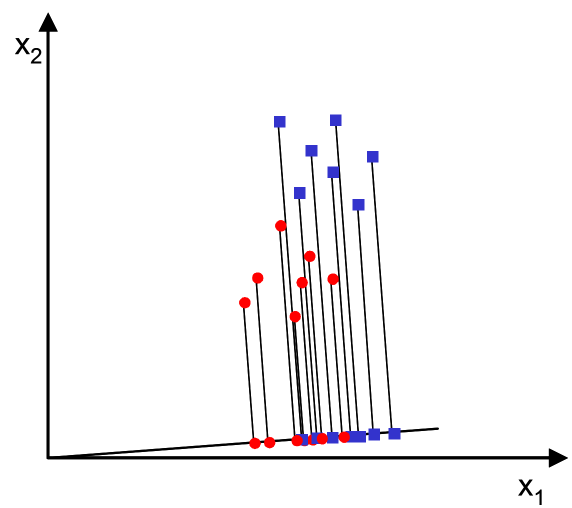
- Linear Discriminant Analysis, two classes
- Linear Discriminant Analysis, C classes
- LDA vs. PCA example
- Limitations of LDA
- Variants of LDA
- Other dimensionality reduction methods

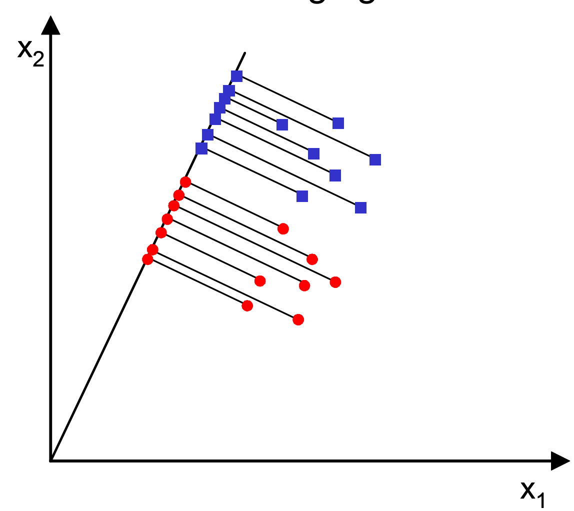
Linear Discriminant Analysis, two-classes (1)

- **The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible**
  - Assume we have a set of D-dimensional samples $\{x^{(1}, x^{(2}, \ldots, x^{(N}\}$, N_1 of which belong to class ω_1, and N_2 to class ω_2. We seek to obtain a scalar ***y*** by projecting the samples ***x*** onto a line

$$y = w^T x$$

  - Of all the possible lines we would like to select the one that maximizes the separability of the scalars
    - This is illustrated for the two-dimensional case in the following figures





Linear Discriminant Analysis, two-classes (2)

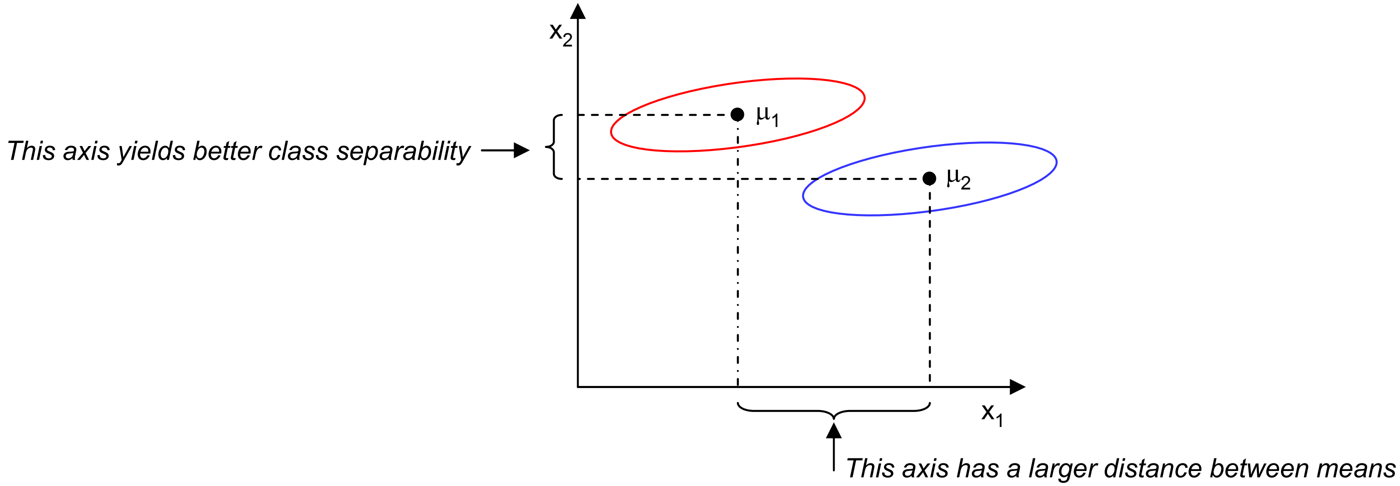
- **In order to find a good projection vector, we need to define a measure of separation between the projections**
  - The mean vector of each class in ***x*** and ***y*** feature space is

$$\mu_i = \frac{1}{N_i}\sum_{x\in\omega_i} x \quad \text{and} \quad \tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y = \frac{1}{N_i}\sum_{x\in\omega_i} w^T x = w^T \mu_i$$

  - We could then choose the distance between the projected means as our objective function

$$J(w) = \left|\tilde{\mu}_1 - \tilde{\mu}_2\right| = \left|w^T(\mu_1 - \mu_2)\right|$$

  - However, the distance between the projected means is not a very good measure since it does not take into account the standard deviation within the classes



Linear Discriminant Analysis, two-classes (3)

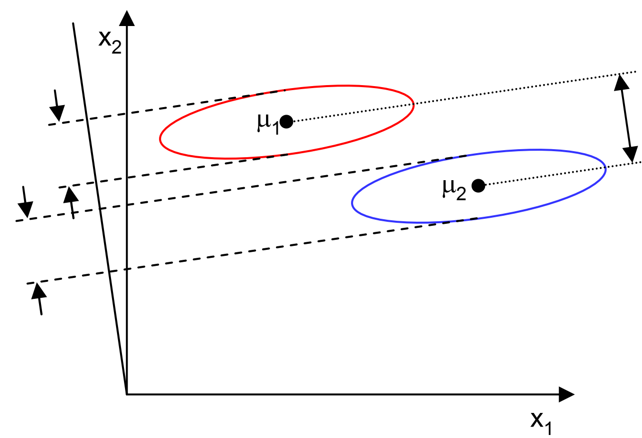
- **The solution proposed by Fisher is to maximize a function that represents the difference between the means, normalized by a measure of the within-class scatter**
  - For each class we define the scatter, an equivalent of the variance, as

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} \left(y - \tilde{\mu}_i\right)^2$$

    - where the quantity $\left(\tilde{s}_1^2 + \tilde{s}_2^2\right)$ is called the within-class scatter of the projected examples
  - The Fisher linear discriminant is defined as the linear function $\mathbf{w}^T\mathbf{x}$ that maximizes the criterion function

$$J(w) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

  - Therefore, we will be looking for a projection where examples from the same class are projected very close to each other and, at the same time, the projected means are as farther apart as possible



Linear Discriminant Analysis, two-classes (4)

- In order to find the optimum projection w*, we need to express J(w) as an explicit function of w
- We define a measure of the scatter in multivariate feature space **x**, which are scatter matrices

$$S_i = \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$$

$$S_1 + S_2 = S_W$$

  - where S_W is called the **within-class scatter matrix**

- The scatter of the projection **y** can then be expressed as a function of the scatter matrix in feature space **x**

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2 = \sum_{x \in \omega_i} (w^T x - w^T \mu_i)^2 = \sum_{x \in \omega_i} w^T (x - \mu_i)(x - \mu_i)^T w = w^T S_i w$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = w^T S_W w$$

- Similarly, the difference between the projected means can be expressed in terms of the means in the original feature space

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T \mu_1 - w^T \mu_2)^2 = w^T \underbrace{(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T}_{S_B} w = w^T S_B w$$

  - The matrix S_B is called the **between-class scatter**. Note that, since S_B is the outer product of two vectors, its rank is at most one

- We can finally express the Fisher criterion in terms of S_W and S_B as

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Linear Discriminant Analysis, two-classes (5)

- To find the maximum of J(w) we derive and equate to zero

$$\frac{d}{dw}\left[J(w)\right]=\frac{d}{dw}\left[\frac{w^T S_B w}{w^T S_W w}\right]=0 \Rightarrow$$

$$\Rightarrow \left[w^T S_W w\right]\frac{d\left[w^T S_B w\right]}{dw}-\left[w^T S_B w\right]\frac{d\left[w^T S_W w\right]}{dw}=0 \Rightarrow$$

$$\Rightarrow \left[w^T S_W w\right]2S_B w-\left[w^T S_B w\right]2S_W w=0$$

- Dividing by $w^T S_W w$

$$\frac{\left[w^T S_W w\right]}{\left[w^T S_W w\right]}S_B w-\frac{\left[w^T S_B w\right]}{\left[w^T S_W w\right]}S_W w=0 \Rightarrow$$

$$\Rightarrow S_B w - J S_W w = 0 \Rightarrow$$

$$\Rightarrow S_W^{-1} S_B w - J w = 0$$

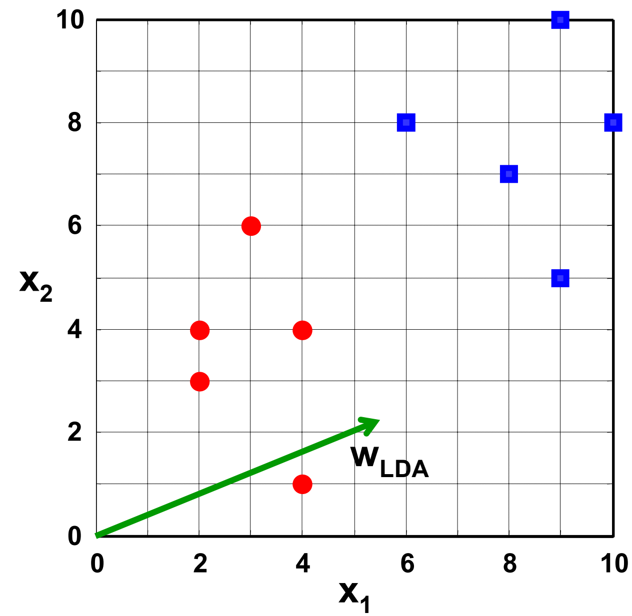
- Solving the generalized eigenvalue problem ($S_W^{-1}S_B w=Jw$) yields

$$\boxed{w^* = \underset{w}{\operatorname{argmax}}\left\{\frac{w^T S_B w}{w^T S_W w}\right\}=S_W^{-1}\left(\mu_1-\mu_2\right)}$$

  - This is know as **Fisher's Linear Discriminant** (1936), although it is not a discriminant but rather a specific choice of direction for the projection of the data down to one dimension

LDA example

- **Compute the Linear Discriminant projection for the following two-dimensional dataset**
  - $X1=(x_1,x_2)=\{(4,1),(2,4),(2,3),(3,6),(4,4)\}$
  - $X2=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$



- **SOLUTION (by hand)**
  - The class statistics are:

$$S_1=\begin{bmatrix}0.80 & -0.40\\ -0.40 & 2.60\end{bmatrix};\ S_2=\begin{bmatrix}1.84 & -0.04\\ -0.04 & 2.64\end{bmatrix}$$

$$\mu_1=\begin{bmatrix}3.00 & 3.60\end{bmatrix};\qquad \mu_2=\begin{bmatrix}8.40 & 7.60\end{bmatrix}$$

  - The within- and between-class scatter are

$$S_B=\begin{bmatrix}29.16 & 21.60\\ 21.60 & 16.00\end{bmatrix};\ S_W=\begin{bmatrix}2.64 & -0.44\\ -0.44 & 5.28\end{bmatrix}$$

  - The LDA projection is then obtained as the solution of the generalized eigenvalue problem

$$S_W^{-1}S_B v=\lambda v\Rightarrow\left|S_W^{-1}S_B-\lambda I\right|=0\Rightarrow\begin{vmatrix}11.89-\lambda & 8.81\\ 5.08 & 3.76-\lambda\end{vmatrix}=0\Rightarrow\lambda=15.65$$

$$\begin{bmatrix}11.89 & 8.81\\ 5.08 & 3.76\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix}=15.65\begin{bmatrix}v_1\\ v_2\end{bmatrix}\Rightarrow\begin{bmatrix}v_1\\ v_2\end{bmatrix}=\begin{bmatrix}0.91\\ 0.39\end{bmatrix}$$

  - Or directly by

$$w^*=S_W^{-1}(\mu_1-\mu_2)=\begin{bmatrix}-0.91 & -0.39\end{bmatrix}^T$$

Linear Discriminant Analysis, C-classes (1)

- **Fisher's LDA generalizes very gracefully for C-class problems**
  - Instead of one projection **y**, we will now seek (C-1) projections $[y_1,y_2,\ldots,y_{C-1}]$ by means of (C-1) projection vectors w_i, which can be arranged by columns into a projection matrix $W=[w_1|w_2|\ldots|w_{C-1}]$:

$$y_i = w_i^T x \Rightarrow y = W^T x$$

- **Derivation**
  - The generalization of the within-class scatter is

$$S_W = \sum_{i=1}^{C} S_i$$

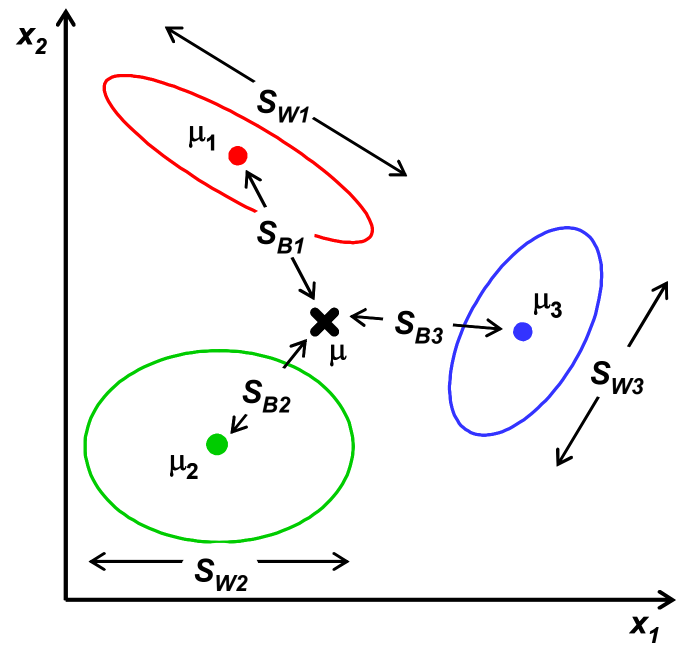
$$\text{where } S_i = \sum_{x\in\omega_i}(x-\mu_i)(x-\mu_i)^T \text{ and } \mu_i = \frac{1}{N_i}\sum_{x\in\omega_i} x$$

  - The generalization for the between-class scatter is

$$S_B = \sum_{i=1}^{C} N_i(\mu_i-\mu)(\mu_i-\mu)^T$$

$$\text{where } \mu = \frac{1}{N}\sum_{\forall x} x = \frac{1}{N}\sum_{x\in\omega_i} N_i\mu_i$$

    - where $S_T=S_B+S_W$ is called the <u>total scatter matrix</u>



Linear Discriminant Analysis, C-classes (2)

- Similarly, we define the mean vector and scatter matrices for the projected samples as

$$\tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y \qquad \tilde{S}_W = \sum_{i=1}^{C}\sum_{y\in\omega_i}(y-\tilde{\mu}_i)(y-\tilde{\mu}_i)^T$$

$$\tilde{\mu} = \frac{1}{N}\sum_{\forall y} y \qquad \tilde{S}_B = \sum_{i=1}^{C} N_i(\tilde{\mu}_i-\tilde{\mu})(\tilde{\mu}_i-\tilde{\mu})^T$$

- From our derivation for the two-class problem, we can write

$$\tilde{S}_W = W^T S_W W$$
$$\tilde{S}_B = W^T S_B W$$

- Recall that we are looking for a projection that maximizes the ratio of between-class to within-class scatter. Since the projection is no longer a scalar (it has C-1 dimensions), we then use the determinant of the scatter matrices to obtain a scalar objective function:

$$J(W) = \frac{|\tilde{S}_B|}{|\tilde{S}_W|} = \frac{|W^T S_B W|}{|W^T S_W W|}$$

- And we will seek the projection matrix W* that maximizes this ratio

Linear Discriminant Analysis, C-classes (3)

- It can be shown that the optimal projection matrix W* is the one whose columns are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigenvalue problem

$$W^* = \left[w_1^* \mid w_2^* \mid \cdots \mid w_{C-1}^*\right] = \operatorname{argmax}\left\{\frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|}\right\} \Rightarrow \left(S_B - \lambda_i S_W\right) w_i^* = 0$$

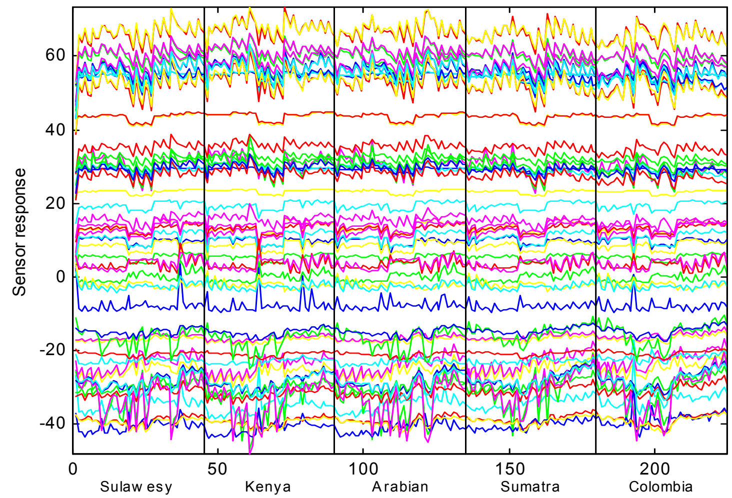
- **NOTES**
  - S_B is the sum of C matrices of rank one or less and the mean vectors are constrained by

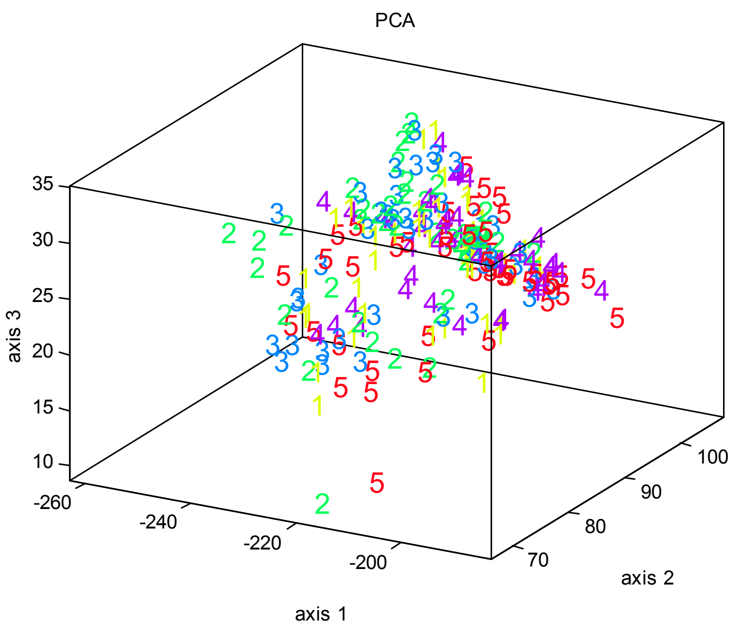
$$\frac{1}{C}\sum_{i=1}^{C}\mu_i = \mu$$

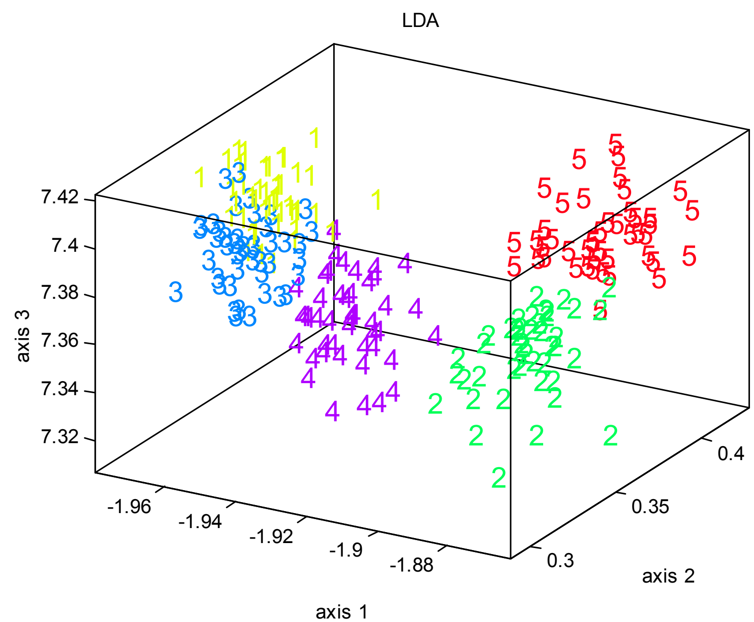
    - Therefore, S_B will be of rank (C-1) or less
    - This means that only (C-1) of the eigenvalues λ_i will be non-zero
  - The projections with maximum class separability information are the eigenvectors corresponding to the largest eigenvalues of $S_W^{-1}S_B$
  - LDA can be derived as the Maximum Likelihood method for the case of normal class-conditional densities with equal covariance matrices

LDA Vs. PCA: Coffee discrimination with a gas sensor array

- **These figures show the performance of PCA and LDA on an odor recognition problem**
  - Five types of coffee beans were presented to an array of chemical gas sensors
  - For each coffee type, 45 “sniffs” were performed and the response of the gas sensor array was processed in order to obtain a 60-dimensional feature vector
- **Results**
  - From the 3D scatter plots it is clear that LDA outperforms PCA in terms of class discrimination
  - This is one example where the discriminatory information is not aligned with the direction of maximum variance

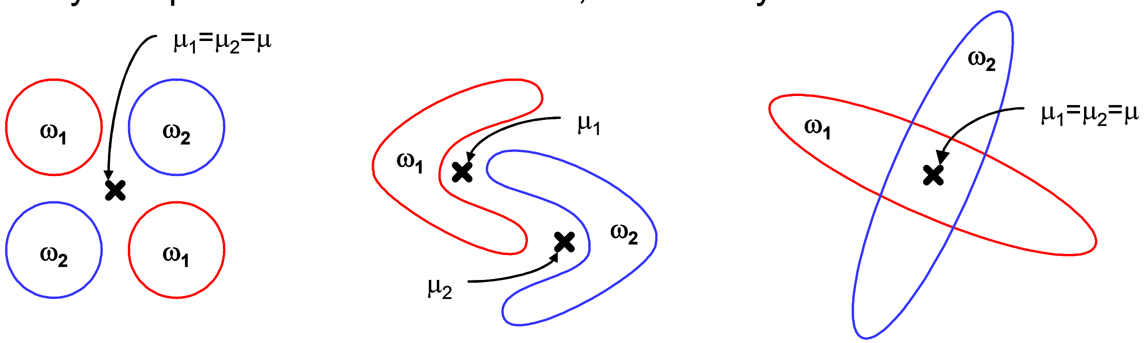




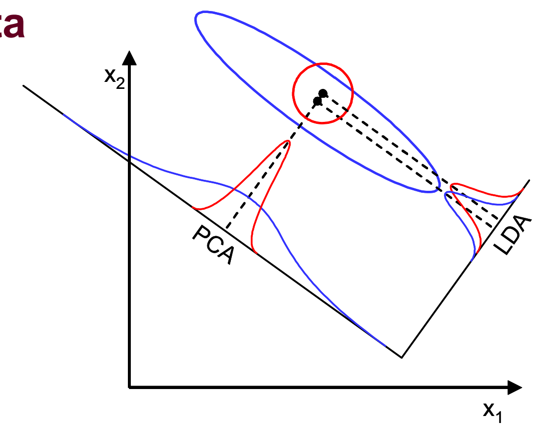


Limitations of LDA

- **LDA produces at most C-1 feature projections**
  - If the classification error estimates establish that more features are needed, some other method must be employed to provide those additional features
- **LDA is a parametric method since it assumes unimodal Gaussian likelihoods**
  - If the distributions are significantly non-Gaussian, the LDA projections will not be able to preserve any complex structure of the data, which may be needed for classification



- **LDA will fail when the discriminatory information is not in the mean but rather in the variance of the data**



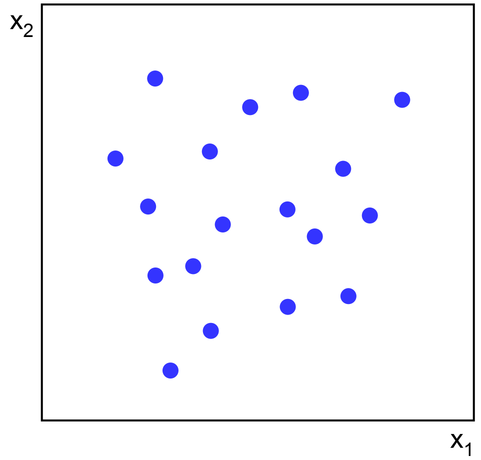
Variants of LDA

- **Non-parametric LDA (Fukunaga)**
  - NPLDA removes the unimodal Gaussian assumption by computing the between-class scatter matrix S_B using local information and the K Nearest Neighbors rule. As a result of this
    - The matrix S_B is full-rank, allowing us to extract more than (C-1) features
    - The projections are able to preserve the structure of the data more closely
- **Orthonormal LDA (Okada and Tomita)**
  - OLDA computes projections that maximize the Fisher criterion and, at the same time, are pair-wise orthonormal
    - The method used in OLDA combines the eigenvalue solution of $S_W^{-1}S_B$ and the Gram-Schmidt orthonormalization procedure
    - OLDA sequentially finds axes that maximize the Fisher criterion in the subspace orthogonal to all features already extracted
    - OLDA is also capable of finding more than (C-1) features
- **Generalized LDA (Lowe)**
  - GLDA generalizes the Fisher criterion by incorporating a cost function similar to the one we used to compute the Bayes Risk
    - The effect of this generalized criterion is an LDA projection with a structure that is biased by the cost function
    - Classes with a higher cost C_{ij} will be placed further apart in the low-dimensional projection
- **Multilayer Perceptrons (Webb and Lowe)**
  - It has been shown that the hidden layers of multi-layer perceptrons (MLP) perform <u>non-linear discriminant analysis</u> by maximizing $Tr[S_B S_T^{\dagger}]$, where the scatter matrices are measured at the output of the last hidden layer

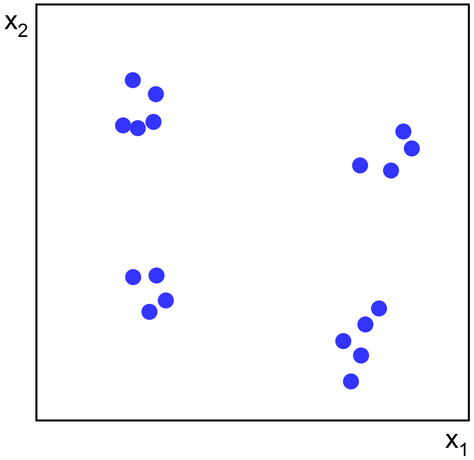
Other dimensionality reduction methods (1)

- **Exploratory Projection Pursuit (Friedman and Tukey)**
  - EPP seeks an M-dimensional (M=2,3 typically) linear projection of the data that maximizes a measure of "interestingness"
  - Interestingness is measured as <u>departure from multivariate normality</u>
    - This measure is not the variance and is commonly scale-free. In most proposals it is also affine invariant, so it does not depend on correlations between features . [Ripley, 1996]
  - In other words, EPP seeks projections that separate clusters as much as possible and keeps these clusters compact, a similar criterion as Fisher's, but EPP does NOT use class labels
  - Once an interesting projection is found, it is important to remove the structure it reveals to allow other interesting views to be found more easily

UNINTERESTING



INTERESTING



Other dimensionality reduction methods (2)

- **Sammon's Non-linear Mapping (Sammon)**
  - This method seeks a mapping onto an M-dimensional space that preserves the inter-point distances of the original N-dimensional space
    - This is accomplished by minimizing the following objective function

$$E(d,d') = \sum_{i \neq j} \frac{\left[d(P_i,P_j) - d(P_i',P_j')\right]^2}{d(P_i,P_j)}$$

- The original method did not obtain an explicit mapping but only a lookup table for the elements in the training set
- Recent implementations using artificial neural networks (MLPs and RBFs) <u>do</u> provide an explicit mapping for test data and also consider cost functions (Neuroscale)
- Sammon's mapping is closely related to Multi-Dimensional Scaling (MDS), a family of multivariate statistical methods commonly used in the social sciences