Lecture 6: Dimensionality reduction (LDA)

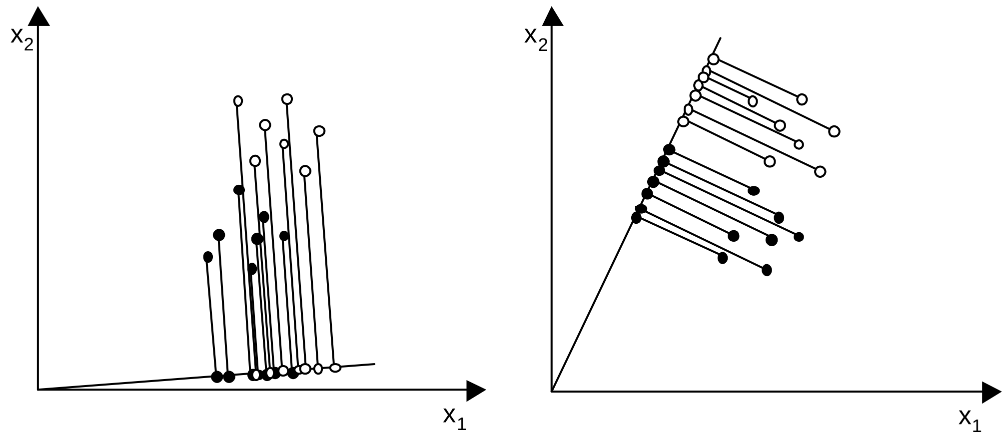
- Linear Discriminant Analysis, two-classes
- Linear Discriminant Analysis, C-classes
- LDA vs. PCA: Coffee discrimination with a gas sensor array
- Limitations of LDA
- Variants of LDA
- Other dimensionality reduction methods

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 values of the *x* and *y* examples are

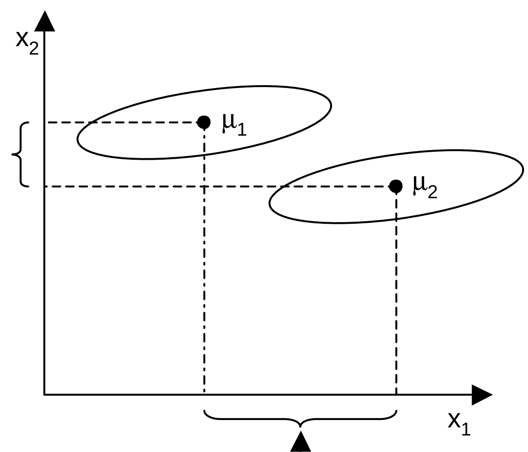
$$\mu_i = \frac{1}{N_i}\sum_{x\in\omega_i} x \quad \text{and} \quad \tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y = \frac{1}{N_i}\sum_{x\in\omega_i} w^T x = w^T \mu_i$$

  - We could 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 projected means is not a very good measure since it does not take into account the standard deviation within the classes

This axis yields better class separability



This axis has a larger distance between means

Linear Discriminant Analysis, two-classes (3)

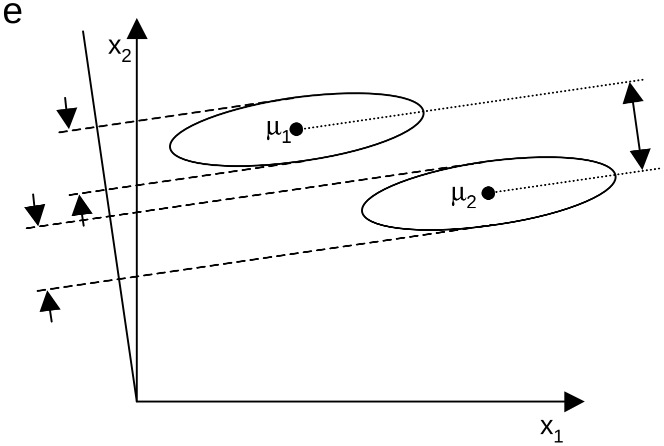
- The solution proposed by Fisher is to maximize a function that represents the difference between the means, normalized by a measure of the within-class scatter
- For each class we define the scatter, an equivalent of the variance, as

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

  - and 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 $w^T x$ 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)

- We need to express J(w) as an explicit function of w in order to find w*
  - For this reason we define the equivalent of the scatter in the projection which, in multivariate feature space become <u>scatter matrices</u>

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

$$S_1 + S_2 = S_W$$

    - The matrix S_W is called the **<u>within-class scatter matrix</u>** and is proportional to the sample covariance matrix
  - The scatter of the projection can be expressed as a function of the scatter matrix in the x feature space

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

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

  - Similarly

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

    - The matrix S_B is called the **<u>between-class scatter</u>** and, since it is the outer product of two vectors, <u>its rank is at most one</u>
- 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

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

  - 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

Linear Discriminant Analysis, C-classes (1)

- **Fisher's LDA generalizes for C-class problems very gracefully**
  - Instead of one discriminant function, we have (C-1) discriminant
  - The projection is from a N-dimensional space onto (C-1) dimensions
- **Derivation**
  - The generalization of the within-class scatter matrix 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 matrix 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 total scatter matrix

  - For the (C-1) class problem we will seek (C-1) projection vectors w_i, which can be arranged by columns into a projection matrix $W=[w_1|w_2|\ldots|w_{C-1}]$ so that

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

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 show that

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

- Recall that we are looking for a projection that, in some sense, maximizes the ratio of between-class to within-class scatter
- Since the projection is not scalar (it has C-1 dimensions), we use the determinant of the scatter matrices into the criterion function, which then becomes

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

  - And we are seeking the projection matrix W* that maximizes this criterion

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

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

NOTES

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

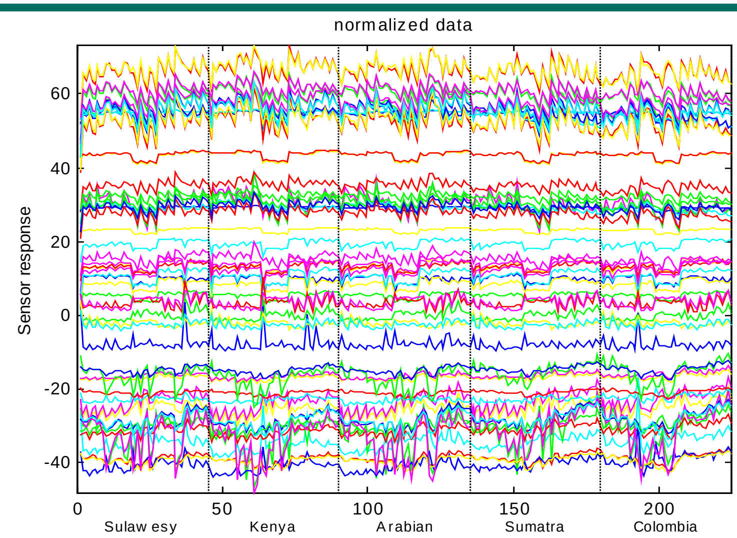
$$\frac{1}{C}\sum_{i=1}^{C} \mu_i = \mu$$

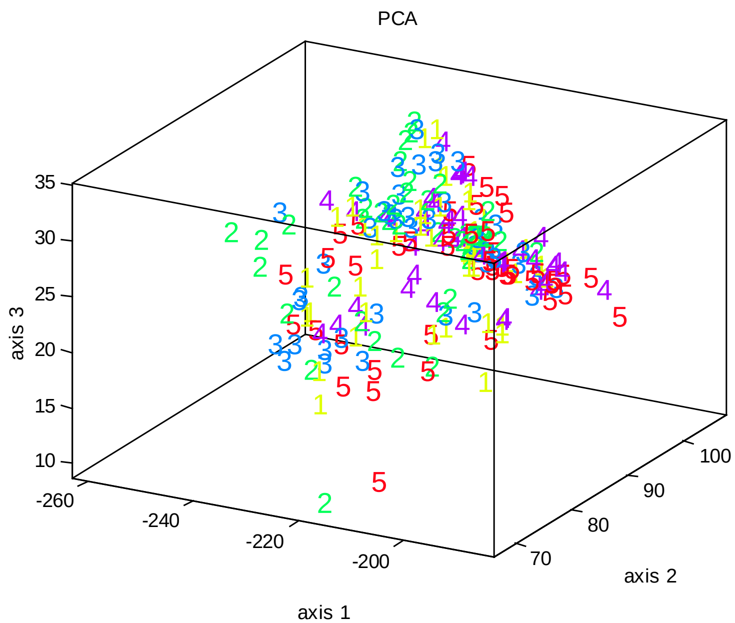
```
rank(A+B)<=rank(A)+rank(B)
rank(AB) <= min{rank(A), rank(B)} for any compatible matrix A,B.
```

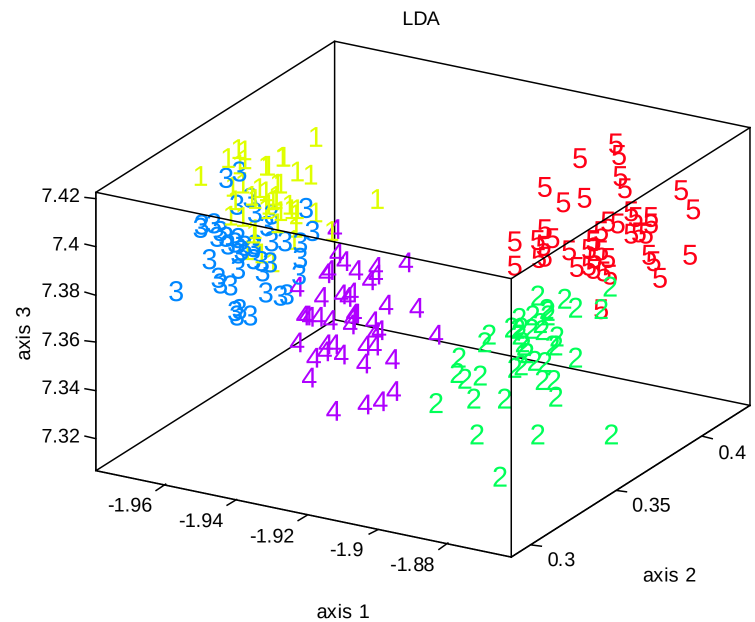
  - <u>Therefore, S_B will be of rank (C-1) or less</u>
  - This means that only (C-1) of the eigenvalues λ_i will be non-zero
- The projections with maximum class separability information are the eigenvectors corresponding to the largest eigenvalues of $S_W^{-1} S_B$
- LDA can be derived as the Maximum 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

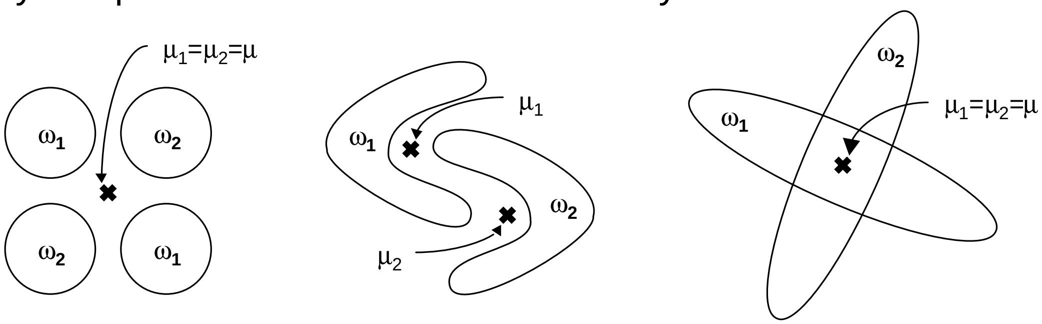




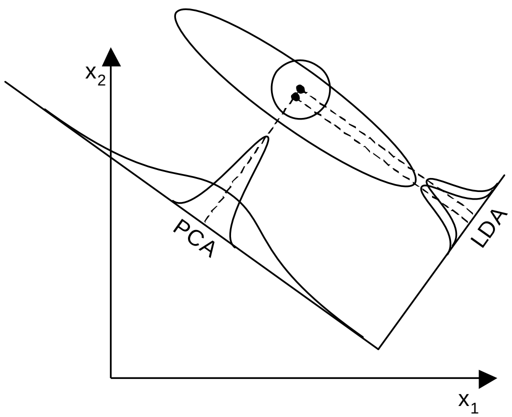


Limitations of LDA

- **LDA produce 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 that 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

- **Exploratory Projection Pursuit (Friedman and Tukey)**
  - EPP seeks a M-dimensional (M=2,3 typically) linear projection of the data that maximizes a measure of "interestingness"
    - Interestingness is based on how much the projected data deviates from normally distributed data in the main body of its distribution
      - 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
    - After an interesting projection has been found, the structure that makes the projection interesting may be removed from the data, and the procedure can be repeated to reveal more of the structure of the dataset
- **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 (MLP and RBF) do 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