Lecture 15:

Linear Discriminant Analysis

Many thanks to Carlos Thomaz who authored the original version of these slides

The story so far:

Let us start with a data set which we can write as a matrix:

$$D = \begin{pmatrix} x_{1,1} & x_{1,2} & & x_{1,n} \\ x_{2,1} & x_{2,2} & & x_{2,n} \\ x_{3,1} & & & \\ & & & \\ x_{N,1} & & & x_{N,n} \end{pmatrix}$$

Each row is one data point, each column is a variable, but take care sometimes the transpose is used

The mean adjusted data matrix

We form the mean adjusted data matrix by subtracting the mean of each variable

$$U = \begin{pmatrix} x_{1,1} - \mu_1 & x_{1,2} - \mu_2 & & x_{1,n} - \mu_n \\ x_{2,1} - \mu_1 & x_{2,2} - \mu_2 & & x_{2,n} - \mu_n \\ x_{3,1} - \mu_1 & & & \\ & & & \\ x_{N,1} - \mu_1 & & & x_{N,n} - \mu_n \end{pmatrix}$$

μ_i is the mean of the data items in column i of D

Covariance Matrix

The covariance matrix can be formed from the product:

$$\Sigma = (1/(N-1))\, U^T U$$

Here is one of the worlds great mysteries:

why is it 1/(N-1) not 1/N?

Alternative notation for Covariance

Covariance is also expressed as an outer product:

$$\Sigma = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \bar{x})(x_j - \bar{x})^T$$

Sum over the data points

nx1 column vector

1xn row vector

nxn matrix

Scatter Matrix

A scatter matrix is un-normalised, using the data matrix formulation:

$$S = U^T U$$

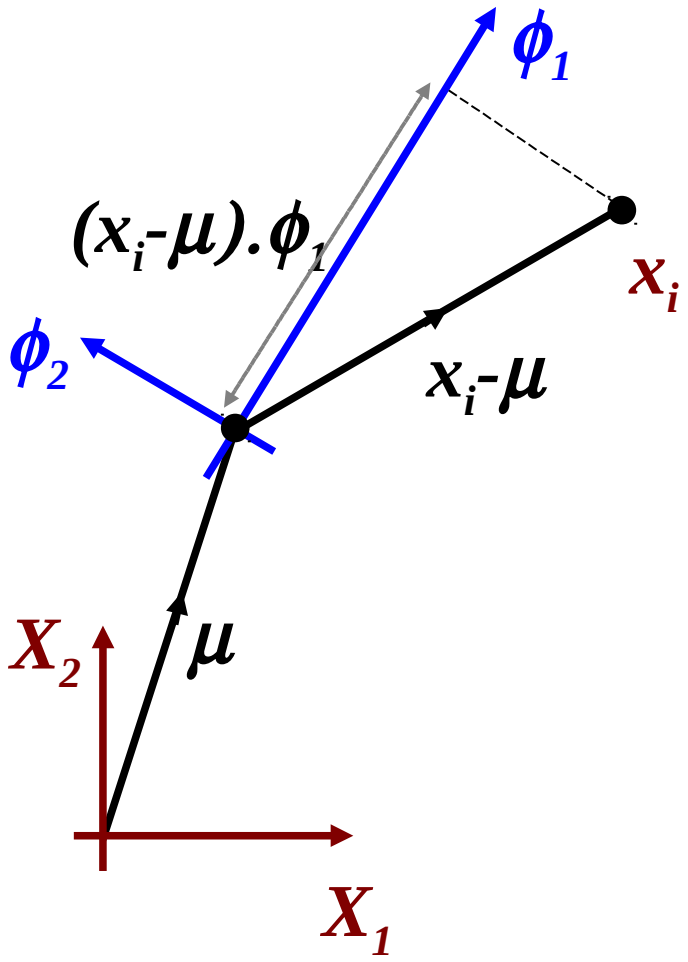
is a scatter matrix but

$$\Sigma = S_{cov} = U^T U/(N-1)$$

is the corresponding covariance matrix

All the projection basis can be calculated on either co-variance or scatter matrices

Projection



A projection is a transformation of data points from one axis system to another.

Finding the mean adjusted data matrix is equivalent to moving the origin to the centre of the data. Projection is then carried out by a dot product

Projection Matrix

A full projection is defined by a matrix in which each column is a vector defining the direction of one of the new axes.

$$\Phi = [\ \phi_1, \phi_2, \phi_3, \ \ldots \phi_m\ ]$$

Each basis vector has the dimension of the original data space. The projection of data point x_i is:

$$y_i = (x_i - \mu)\,\Phi$$

Projecting every point

The projection of the data in mean adjusted form can be written:

$$U\,\Phi = (\Phi^T U^T)^T$$

Projection of the covariance matrix is

$$\begin{aligned}\Phi^T \Sigma\, \Phi &= \Phi^T (1/(N-1))\, U^T\, U\Phi \\ &= (1/(N-1))\, \Phi^T U^T U\Phi \\ &= (1/(N-1))\, (U\Phi)^T U\Phi\end{aligned}$$

Projecting every point

The projection of the data in mean adjusted form can be written:

$$U\,\Phi = (\Phi^T U^T)^T$$

Projection of the covariance matrix is

$$\Phi^T \Sigma \Phi = \Phi^T (1/(N-1))\, U^T U \Phi$$

$$= (1/(N-1))\, \Phi^T U^T U \Phi$$

$$= (1/(N-1))\, (U\Phi)^T (U\Phi)$$

which is the covariance matrix of the projected points

Orthogonal and Orthonormal

For most practical cases we expect the projection to be orthogonal

(all the new axes are at right angles to each other)

and orthonormal

(all the basis vectors defining the axes are unit length) thus:

$$\Phi^T \Phi = I$$

PCA

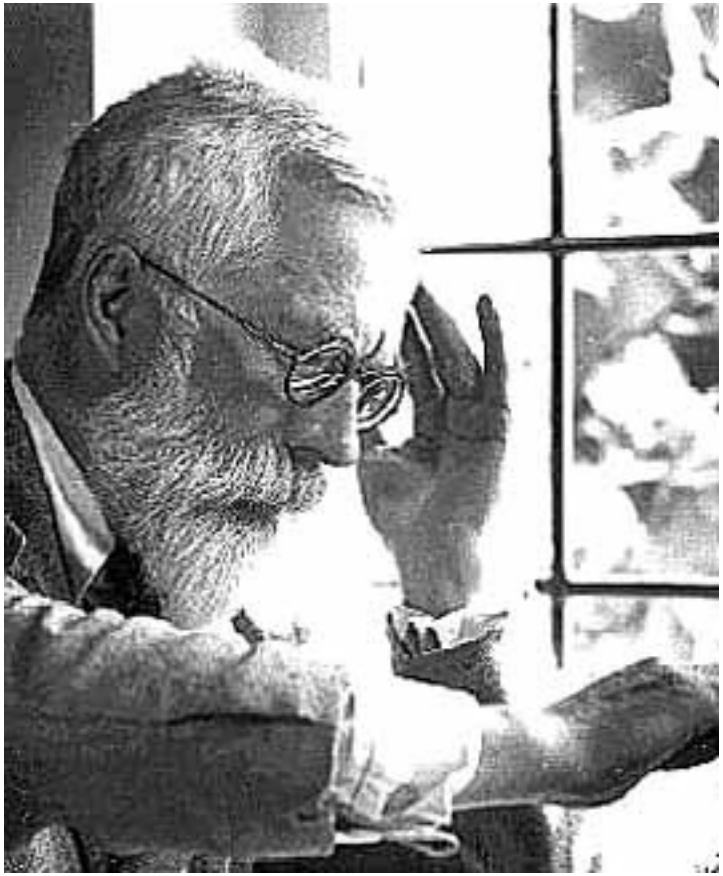
The PCA projection is the one that diagonalises the covariance matrix. That is it transforms the points such that they are independent of each other.

$$\Phi^T \Sigma \Phi = \Lambda$$

We will look at another projection today called the LDA.

Introduction to LDA

Ronald A. Fisher, 1890-1962



"The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data."

1936

Introduction to LDA

What is LDA?

Linear Discriminant Analysis, or simply LDA, is a well-known feature extraction technique that has been used successfully in many statistical pattern recognition problems.

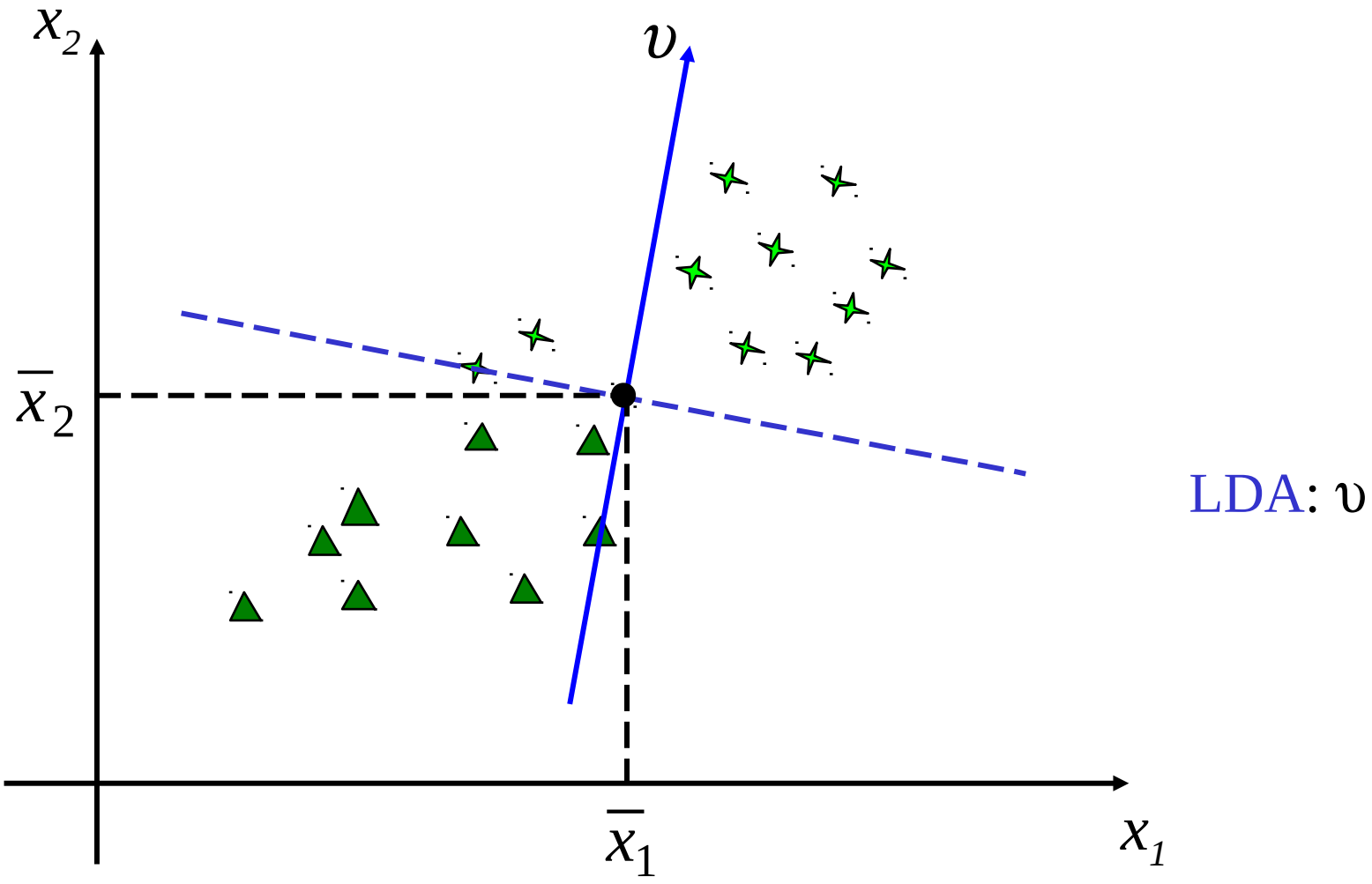
LDA is sometimes called Fisher Discriminant Analysis (FDA).

Motivation

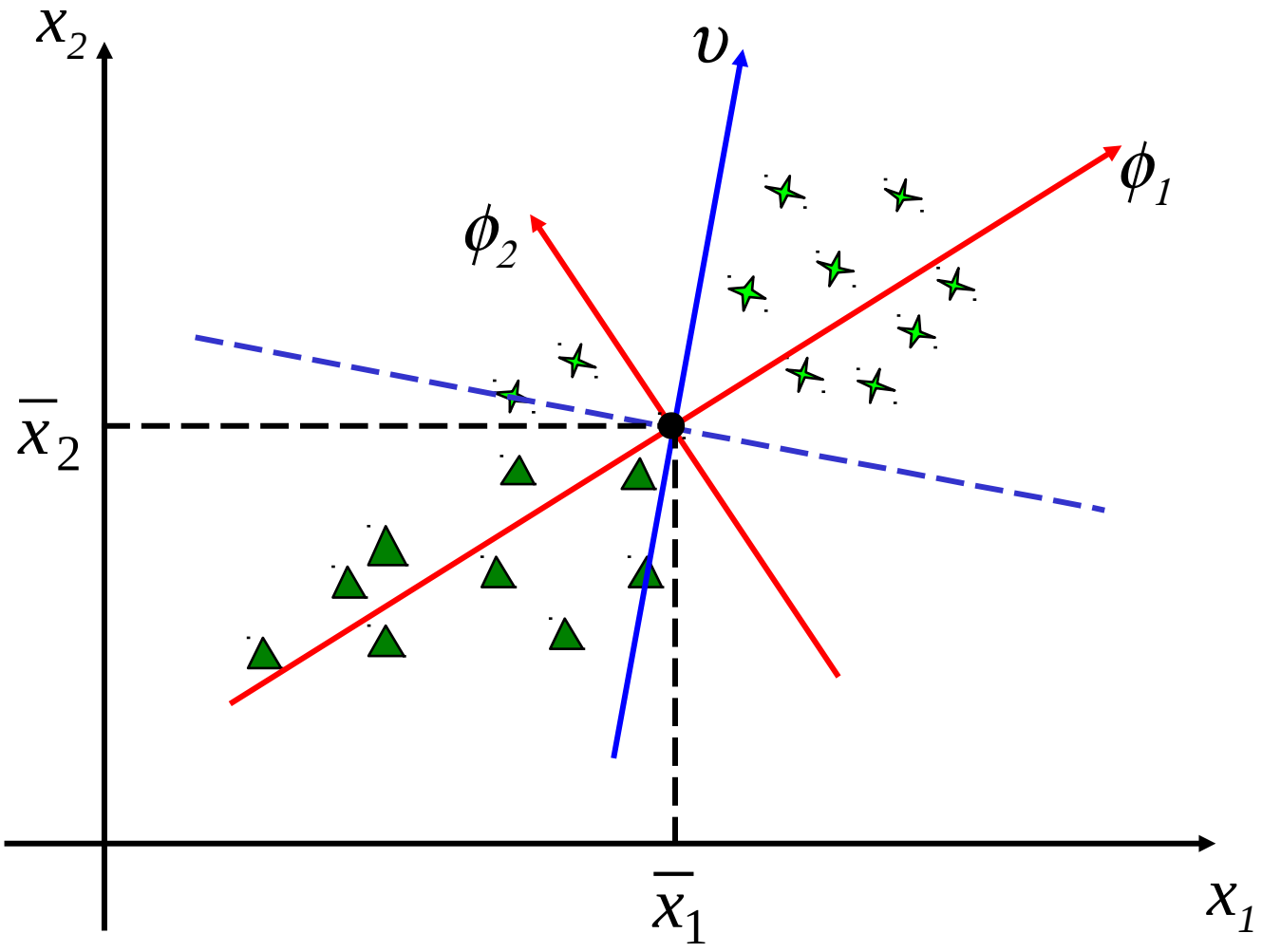
The primary purpose of LDA is to separate samples of distinct groups by transforming then to a space which maximises their between-class separability while minimising their within-class variability.

It assumes implicitly that the true covariance matrices of each class are equal because the same within-class scatter matrix is used for all the classes considered.

Geometric Idea



Geometric Idea



PCA: (ϕ_1, ϕ_2)

LDA: υ

Method

Each class mean and class covariance, and the grand mean vector are given respectively by:

$$\bar{x}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{i,j}$$

$$S_i = \frac{1}{(N_i - 1)} \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T$$

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{g} N_i \bar{x}_i = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{N_i} x_{i,j} \qquad N = N_1 + N_2 + \ldots + N_g$$

Method

Let the between-class scatter matrix S_b be defined as

$$S_b = \sum_{i=1}^{g} N_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$$

and the within-class scatter matrix S_w be defined as

$$S_w = \sum_{i=1}^{g} (N_i - 1) S_i = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T$$

where $x_{i,j}$ is an n-dimensional data point j from class π_i, N_i is the number of training examples from class π_i, and g is the total number of classes or groups.

Method (cont.)

The main objective of LDA is to find a projection matrix Φ_{lda} that maximises the ratio of the determinant of S_b to the determinant of S_w (Fisher's criterion), that is

$$\Phi_{lda} = \arg\max_{\Phi} \frac{\left|\Phi^T S_b \Phi\right|}{\left|\Phi^T S_w \Phi\right|}$$

Intuition

The determinant of the co-variance matrix tells us how much variance a class has.

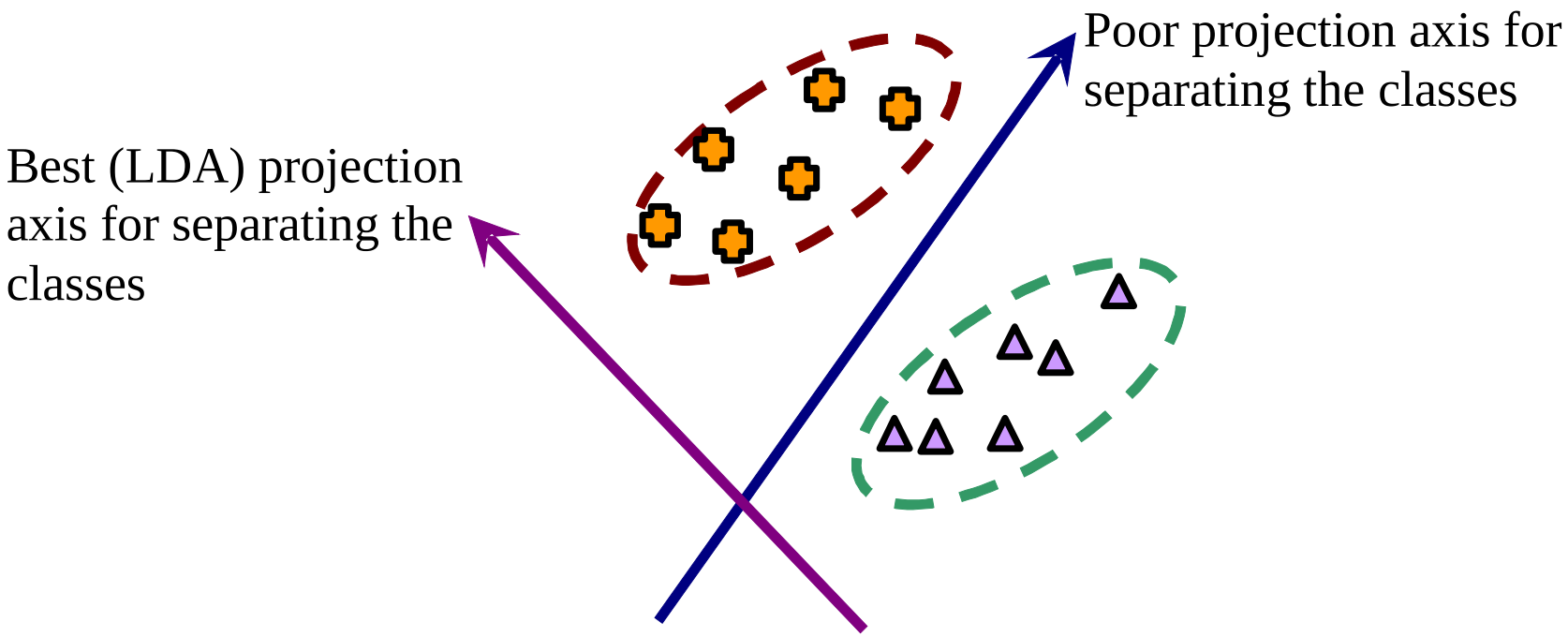
Consider the co-variance matrix in the PCA (diagonal) projection - the determinant is just the product of the diagonal elements which are the individual variable variances.

The determinant has the same value under any ortho-normal projection.

Intuition (cont)

So Fisher's criterion tries to find the projection that:

- Maximises the variance of the class means
- Minimises the variance of the individual classes



Method (cont.)

It has been shown that P_{lda} is in fact the solution of the following eigensystem problem:

$$S_b \Phi - S_w \Phi \Lambda = 0$$

Multiplying by the inverse of S_w

$$S_w^{-1} S_b \Phi - S_w^{-1} S_w \Phi \Lambda = 0$$

$$S_w^{-1} S_b \Phi - \Phi \Lambda = 0$$

$$S_w^{-1} S_b \Phi = \Phi \Lambda$$

Standard LDA

If S_w is a non-singular matrix then the Fisher's criterion is maximised when the projection matrix Φ_{lda} is composed of the eigenvectors of

$$S_w^{-1} S_b$$

with at most (g-1) nonzero corresponding eigenvalues.

(since there are only g points to estimate S_b)

Classification Using LDA

The LDA is an axis projection.

Once the projection is found all the data points can be transformed to the new axis system along with the class means and covariances.

Allocation of a new point to a class can be done using a distance measure such as the Mahalanobis distance.

LDA versus PCA

LDA seeks directions that are efficient for *discriminating* data whereas PCA seeks directions that are efficient for *representing* data.

The directions that are discarded by PCA might be exactly the directions that are necessary for distinguishing between groups.

Limited Sample Size Problem

The performance of the standard LDA can be seriously degraded if there are only a limited number of total training observations N compared to the dimension of the feature space n.

Since S_w is a function of $(N - g)$ or fewer linearly independent vectors, its rank is $(N - g)$ or less. Therefore, S_w is a singular matrix if N is less than $(n+g)$, or might be unstable unless $N >> n$.

Two-stage feature extraction technique

First the *n*-dimensional training samples from the original vector space are projected to a lower dimensional space using PCA

Then LDA is applied next to find the best linear discriminant features on that PCA subspace. This is often called the Most Discriminant Features (MDF) method.

$$\Phi_{lda} = \arg\max_{\Phi} \frac{\left|\Phi^T \Phi_{pca}^T S_b \Phi_{pca} \Phi\right|}{\left|\Phi^T \Phi_{pca}^T S_w \Phi_{pca} \Phi\right|}$$

Two-stage feature extraction technique (cont.)

Thus, the Fisher's criterion is maximised when the projection matrix Φ_{lda} is composed of the eigenvectors of

$$(\Phi^T_{pca}\, S_w\, \Phi_{pca})^{-1}(\Phi^T_{pca}\, S_b\, \Phi_{pca})$$

with at most $(g - 1)$ nonzero corresponding eigenvalues. Therefore the singularity of S_w is overcome if the number of principal components (p)

$$g \leq p \leq (N - g)$$