Machine learning
Fisher Linear Discriminant (LDA)

Mohammad-Reza A. Dehaqani

dehaqani@ut.ac.ir

Fisher Linear Discriminant

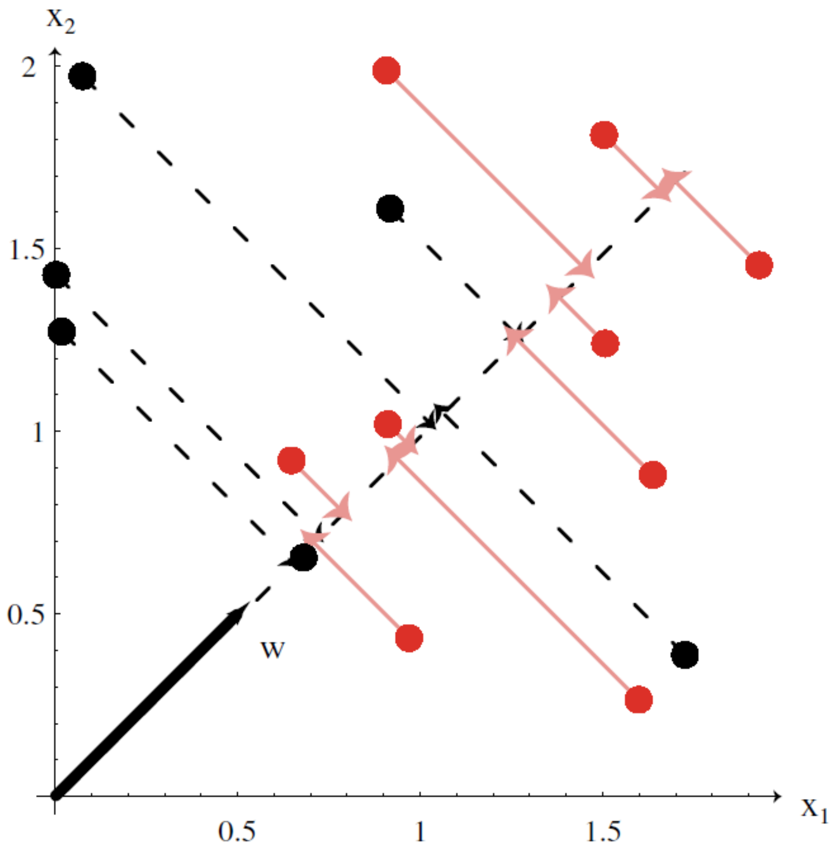
- Whereas PCA seeks **directions** that are **efficient for representation**, **discriminant analysis** seeks directions that are **efficient for discrimination**.
- Given $\mathbf{x_1, . . . , x_n} \in R^d$ divided into two subsets D_1 and D_2 corresponding to the classes ω_1 and ω_2, respectively, the goal is to find a **projection onto a line** defined as

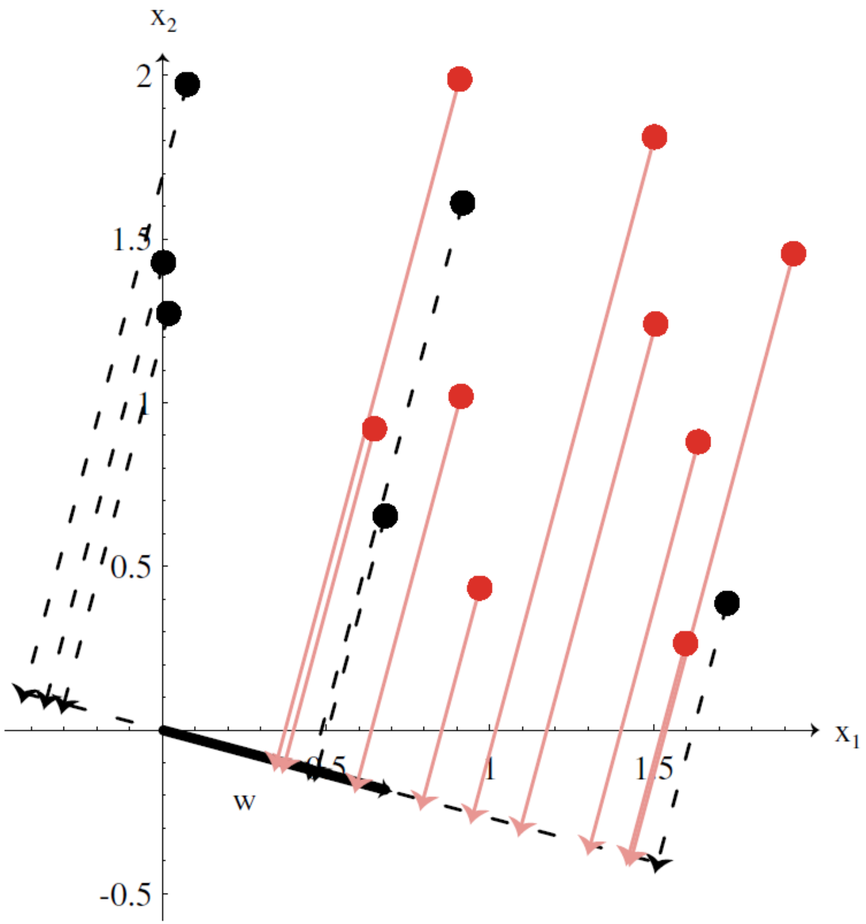
$$y = \mathbf{w}^t\mathbf{x}$$

where the points corresponding to D_1 and D_2 are well separated.

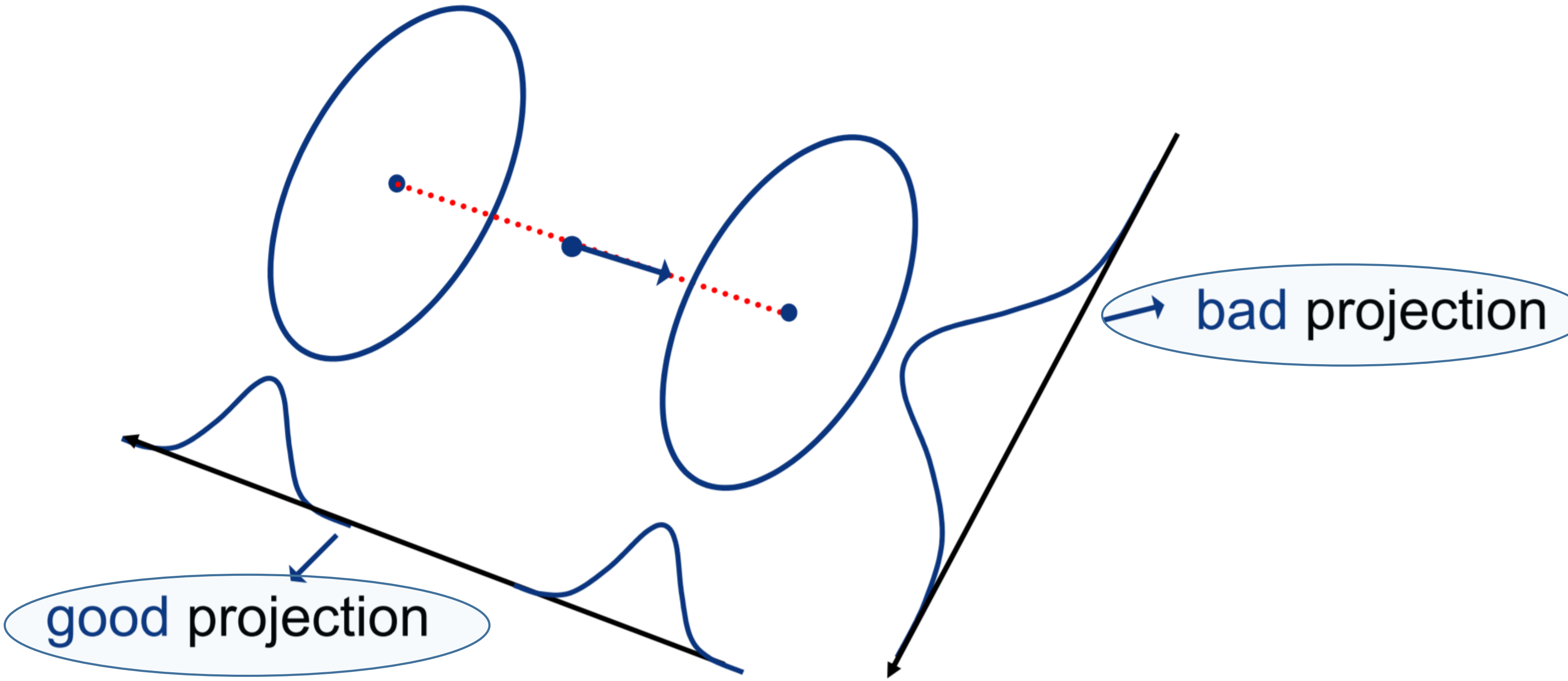
- **Geometrically**, if $||\mathbf{w}|| = 1$, each y_i is the projection of the corresponding $\mathbf{x}_i$ onto a line in the **direction** of $\mathbf{w}$

Linear Discriminant Analysis





Projection of samples onto **two different lines**. The figure on the right shows **greater separation** between the red and black projected points.


bad projection
good projection

Fisher Discriminant Ratio

- The criterion function for the best separation can be defined as

$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

- The sample mean for the projected pointes

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x}, \Rightarrow \tilde{m}_i = \frac{1}{n_i} \sum_{y \in \mathcal{Y}_i} y = \frac{1}{n_i} \sum_{y \in \mathcal{Y}_i} \mathbf{w}^t \mathbf{x} = \mathbf{w}^t \mathbf{m}_i.$$

- **Distance** between the projected means is

$$|\tilde{m}_1 - \tilde{m}_2| = |\mathbf{w}^t(\mathbf{m}_1 - \mathbf{m}_2)|,$$

- We want the difference between the means to be large **relative** to some measure of the **standard deviations**

- Rather than forming **sample variances**, we define the *scatter* for projected **scatter samples**

$$\tilde{s}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{m}_i)^2.$$

$$\boxed{(1/n)(\tilde{s}_1^2 + \tilde{s}_2^2)}$$ estimate of the variance of the pooled data

Fisher's linear discriminant geometric interpretation

- The best projection makes the **difference between the means** as large as possible relative to the **variance**.
- To compute the **optimal w**, we define the **scatter matrices** $\mathbf{S}_i$

$$\mathbf{S}_i = \sum_{\mathbf{x} \in \mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

- The **within-class scatter** matrix $\mathbf{S}_W$ (**symmetric** and **positive semidefinite**, and is usually **nonsingular** if $n > d$)

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2.$$

- And the **between-class scatter** matrix S_B

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t.$$

- (**symmetric** and **positive semidefinite**; since it is the outer product of two vectors, its rank is **at most one**)

Criterion function in terms of $\mathbf{S}_B$ and $\mathbf{S}_W$

- $J(\cdot)$ as an explicit function of **w**,

$$\tilde{s}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{m}_i)^2. \quad \Rightarrow \quad \begin{aligned} \tilde{s}_i^2 &= \sum_{\mathbf{x} \in \mathcal{D}_i} (\mathbf{w}^t\mathbf{x} - \mathbf{w}^t\mathbf{m}_i)^2 \\ &= \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{w}^t(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t\mathbf{w} \\ &= \mathbf{w}^t\mathbf{S}_i\mathbf{w}; \quad \Rightarrow \quad \tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^t\mathbf{S}_W\mathbf{w}. \end{aligned}$$

- Separations of the projected means obeys

$$\begin{aligned} (\tilde{m}_1 - \tilde{m}_2)^2 &= (\mathbf{w}^t\mathbf{m}_1 - \mathbf{w}^t\mathbf{m}_2)^2 \\ &= \mathbf{w}^t\underbrace{(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t}_{\mathbf{S}_B}\mathbf{w} \\ &= \mathbf{w}^t\mathbf{S}_B\mathbf{w}, \end{aligned}$$

$$\Rightarrow \quad \boxed{J(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}}.}$$

Maximization of the Rayleigh quotient

- It appears in many problems in engineering and pattern recognition

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

$\boxed{S_B, S_W, \textit{symmetric positive semidefinite}}$

- This is equivalent to

$$\boxed{\max_w w^T S_B w \quad \text{subject to} \quad w^T S_W w = K}$$

- And can be solved using **Lagrange multipliers**

$$L = w^T S_B w - \lambda\left(w^T S_W w - K\right)$$

The Rayleigh quotient

$$L = w^T S_B w - \lambda\left(w^T S_W w - K\right)$$

- maximize with respect to w

$$\nabla_w L = 2\left(S_B - \lambda S_W\right)w = 0 \quad \Rightarrow \quad S_B w = \lambda S_W w$$

- This is a **generalized** eigenvalue problem that you can solve using any **eigenvalue** routine

$\max\limits_{w} w^T S_B w$ subject to $w^T S_W w = K$ and $S_B w = \lambda S_W w$ hence:

$$\boxed{\left(w^*\right)^T S_B w^* = \lambda \left(w^*\right)^T S_W w^* = \lambda K}$$

which is **maximum** for the **largest** eigenvalue

Two cases

- **Case 1**: S_w invertible (simplifies to a standard eigenvalue problem)

$$S_B w = \lambda S_W w \quad \Rightarrow \quad \boxed{S_W^{-1} S_B w = \lambda w}$$

w* is the **largest eigenvector** of $\boldsymbol{S_w^{-1} S_B}$

- **Case 2**: S_w **not** invertible
    - **Regularize**: $S_w \Rightarrow S_w + \gamma I$

    w* is the eigenvector of largest eigenvalue of $[S_w + \gamma I]^{-1} S_B$

$$S_W = \Phi \Lambda \Phi^T \Rightarrow S_W + \gamma I = \Phi \Lambda \Phi^T + \gamma \Phi I \Phi^T$$
$$= \Phi \left[\Lambda + \gamma I \right] \Phi^T$$

    this makes all **eigenvalues positive** and make S_w **invertible**

- The **max** value is λK, where λ is the largest eigenvalue

LDA in two classes case

- Due to the fact that for any **w**, $\mathbf{S}_B\mathbf{w}$ is always **in the direction** of $\mathbf{m}_1 - \mathbf{m}_2$ ($\mathbf{S}_B$ is quite singular); it is **unnecessary** to solve for the eigenvalues and eigenvectors
- The solution for the **w** that optimizes $J(\cdot)$

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2).$$

- Classification has been **converted** from a *d*-dimensional problem to a hopefully more **manageable** one-dimensional one (many to one reduction)
  - In **theory** can not possibly reduce the minimum achievable error rate if we have a very **large** training set
- **Recall**: when the conditional densities $p(\mathbf{x}|\omega_i)$ are **multivariate normal** with equal covariance matrices $\mathbf{\Sigma}$, we can calculate the **threshold** directly.
  - the optimal decision boundary is $\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0$,

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

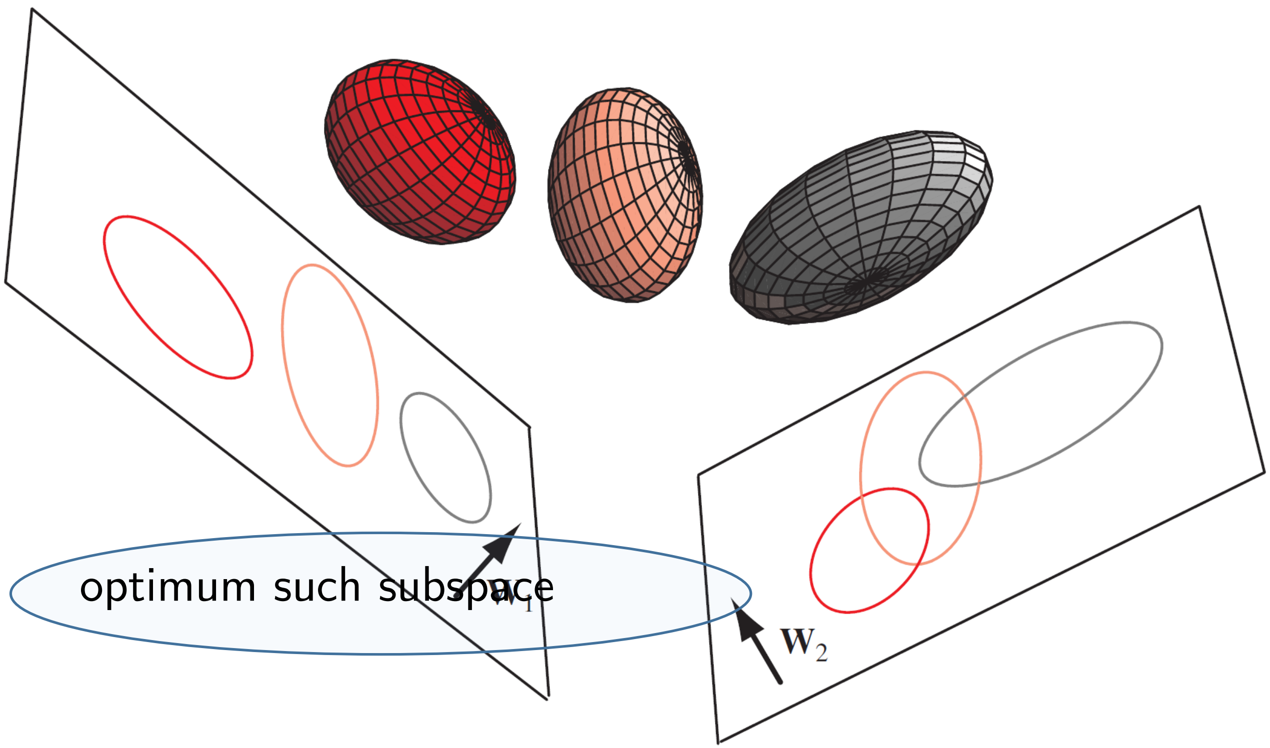
Multiple Discriminant Analysis

- Generalization to **c classes** involves c − 1 discriminant functions where the projection is from a d-dimensional space to a **(c − 1)-dimensional space** (d>c).
- The scatter matrices S_i are computed as

$$\mathbf{S}_W = \sum_{i=1}^{c} \mathbf{S}_i \qquad \mathbf{S}_i = \sum_{\mathbf{x} \in \mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x}.$$

Three-dimensional distributions are projected onto two-dimensional subspaces described by a normal vectors $\mathbf{w}_1$ and $\mathbf{w}_2$



Generalization for $\mathbf{S}_B$

- Generalization for $\mathbf{S}_B$ is not quite so obvious:

  total mean *vector* $$\mathbf{m} = \frac{1}{n}\sum_{\mathbf{x}} \mathbf{x} = \frac{1}{n}\sum_{i=1}^{c} n_i \mathbf{m}_i$$

  total scatter *matrix* $$\mathbf{S}_T = \sum_{\mathbf{x}} (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t.$$

- Then it follows that

$$\begin{aligned}
\mathbf{S}_T &= \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})(\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})^t \\
&= \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t + \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t \\
&= \mathbf{S}_W + \overbrace{\sum_{i=1}^{c} n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t}^{\mathbf{S}_B}.
\end{aligned}$$

$$\boxed{\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B.}$$

Matrix form

- Reduction by $c-1$ **discriminant** functions

$$y_i = \mathbf{w}_i^t \mathbf{x} \qquad i = 1, ..., c-1.$$

- y_i are viewed as components of a vector $\mathbf{y}$ and the weight vectors $\mathbf{w}_i$ are viewed as the columns of a d-by-$(c-1)$ matrix $\mathbf{W}$,

$$\mathbf{y} = \mathbf{W}^t \mathbf{x}.$$

- Mean vectors and scatter matrices

$$\tilde{\mathbf{m}}_i = \frac{1}{n_i} \sum_{\mathbf{y} \in \mathcal{Y}_i} \mathbf{y} \qquad \tilde{\mathbf{S}}_W = \sum_{i=1}^{c} \sum_{\mathbf{y} \in \mathcal{Y}_i} (\mathbf{y} - \tilde{\mathbf{m}}_i)(\mathbf{y} - \tilde{\mathbf{m}}_i)^t$$

$$\tilde{\mathbf{m}} = \frac{1}{n} \sum_{i=1}^{c} n_i \tilde{\mathbf{m}}_i \qquad \tilde{\mathbf{S}}_B = \sum_{i=1}^{c} n_i (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})^t,$$

Criterion function

- it is a straightforward matter to show that

$$\tilde{\mathbf{S}}_W = \mathbf{W}^t \mathbf{S}_W \mathbf{W}$$
$$\tilde{\mathbf{S}}_B = \mathbf{W}^t \mathbf{S}_B \mathbf{W}.$$

- Then, the criterion function becomes

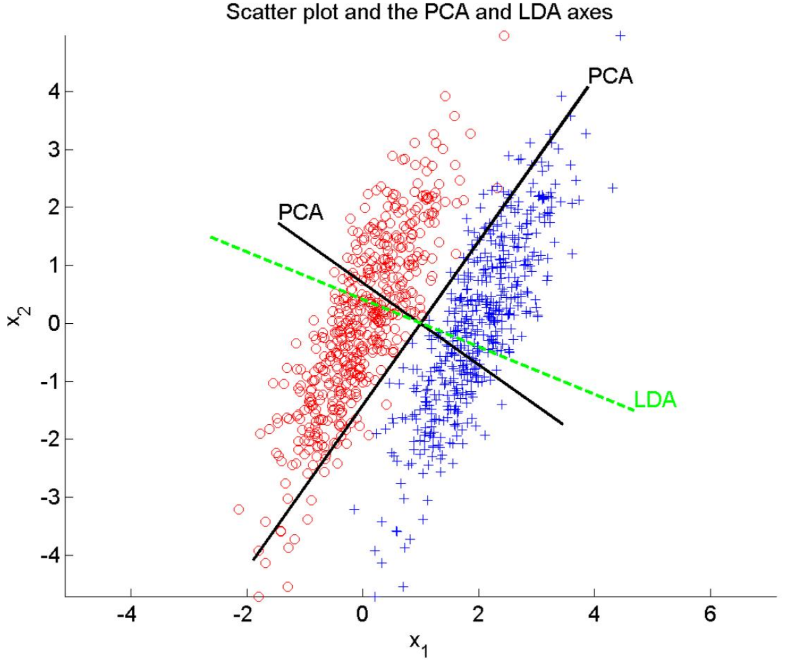
$$J(\mathbf{W}) = \frac{|\tilde{\mathbf{S}}_B|}{|\tilde{\mathbf{S}}_W|} = \frac{|\mathbf{W}^t \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^t \mathbf{S}_W \mathbf{W}|}.$$

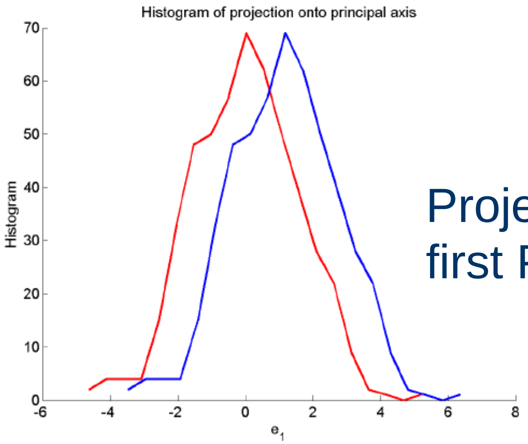
- Where W is the d-by-$(c-1)$ **transformation matrix** and $|\cdot|$ represents the **determinant**.

Transformation matrix

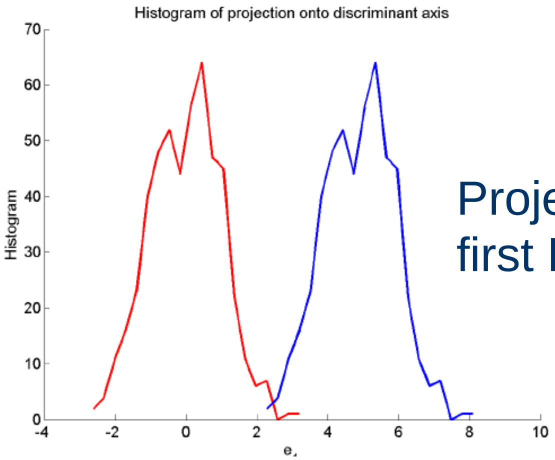
- It can be shown that J(W) is **maximized** when the **columns of W** are the **eigenvectors** of $\boldsymbol{S_w^{-1} S_B}$ having the largest eigenvalues.
- Because S_B is the **sum of c matrices** of rank one or less, and because only c − 1 of these are independent, SB is of rank **c−1 or less**. Thus, no more than c−1 of the **eigenvalues** are **nonzero**.
- Once the transformation from the d-dimensional original feature space to a lower dimensional subspace is done using **PCA or LDA**, parametric or non-parametric methods can be used to train **Bayesian classifiers**.

LDA examples versus PCA





Projection onto the first PCA axis.



Projection onto the first LDA axis

Scatter plot and the PCA and LDA axes for a bivariate sample with two classes. Histogram of the projection onto the first LDA axis shows better separation than the projection onto the first PCA axis