Discriminant Analysis

Fisher Linear Discriminant
Multiple Discriminant Analysis
Fisherfaces

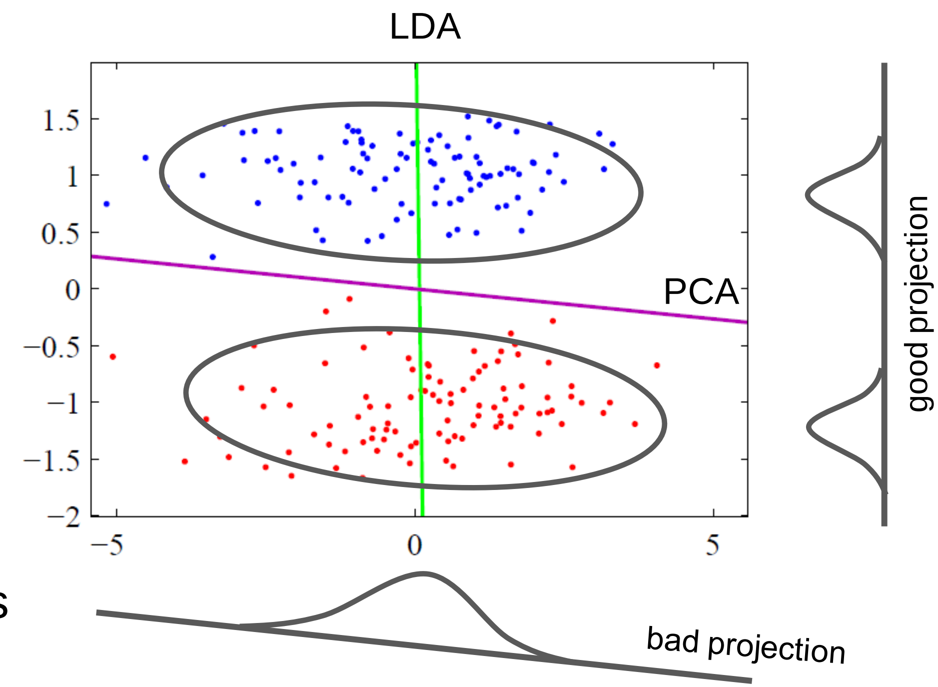
Tae-Kyun Kim
Senior Lecturer
http://www.iis.ee.ic.ac.uk/ComputerVision/

Further reading:
P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, TPAMI, 1997.

Motivation

Projection that best separates the data in a least-squares sense

- PCA finds components that are useful for representing data
- However no reason to assume that components are useful for discriminating between data in different classes
- Pooling data may discard essential directions
- PCA finds the direction for maximum data variance (unsupervised), while LDA (Linear Discriminant Analysis) or MDA finds the direction that optimally separates data of different classes (discriminative or supervised)



Fisher Linear Discriminant

- 2-class problem
- Projecting data from D dimensions onto a line

Difficult to generalize into more classes.

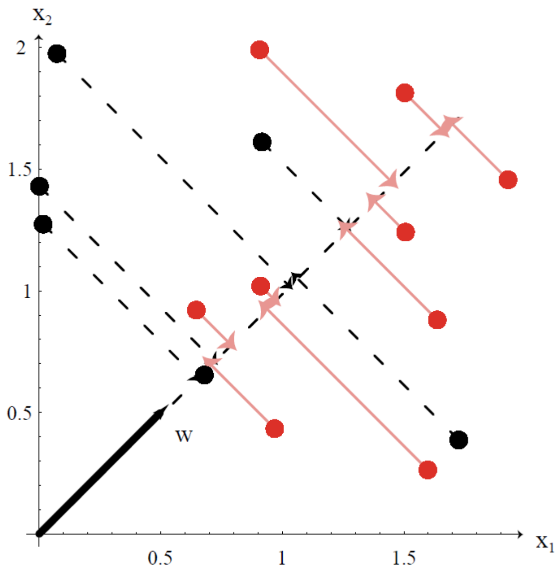
- Set of N D-dimensional samples $\mathbf{x}_1, \dots, \mathbf{x}_N$
  N_1 in the subset labelled c_1
  N_2 in the subset labelled c_2

- We wish to form a linear combination of the components of $\mathbf{x}$ as
  $y = \mathbf{w}^T\mathbf{x}$
  and a corresponding set N of samples $y_1, \dots, y_N$

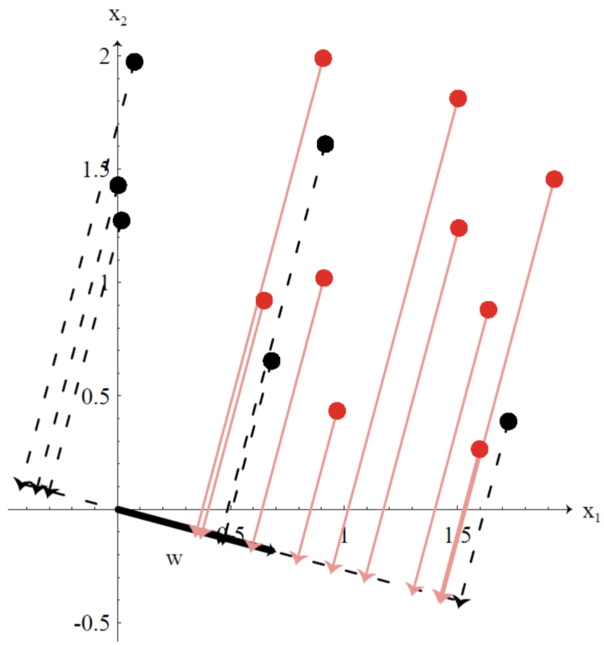
Fisher Liner Discriminant: two-dimensional example

– Projection of same set of two-class samples onto two different lines in the direction marked **w**.

Classes mixed



Better separation



Finding best direction **w**

– Class mean in D-dimensional space:

$$\mathbf{m}_i = \frac{1}{N_i}\sum_{\mathbf{x}\in c_i}\mathbf{x}$$

– Class mean of projected points:

$$\widetilde{\mathbf{m}}_i = \frac{1}{N_i}\sum_{y\in c_i} y = \frac{1}{N_i}\sum_{\mathbf{x}\in c_i}\mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{m}_i$$

– Distance between projected means is

$$|\widetilde{\mathbf{m}}_1 - \widetilde{\mathbf{m}}_2| = |\mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)|$$

Criterion for Fisher Linear Discriminant

– Rather than forming sample variances, define scatter for the projected samples

$$\boxed{\tilde{s}_i^2 = \sum_{y \in c_i} (\mathbf{y} - \tilde{\mathbf{m}}_i)^2}$$

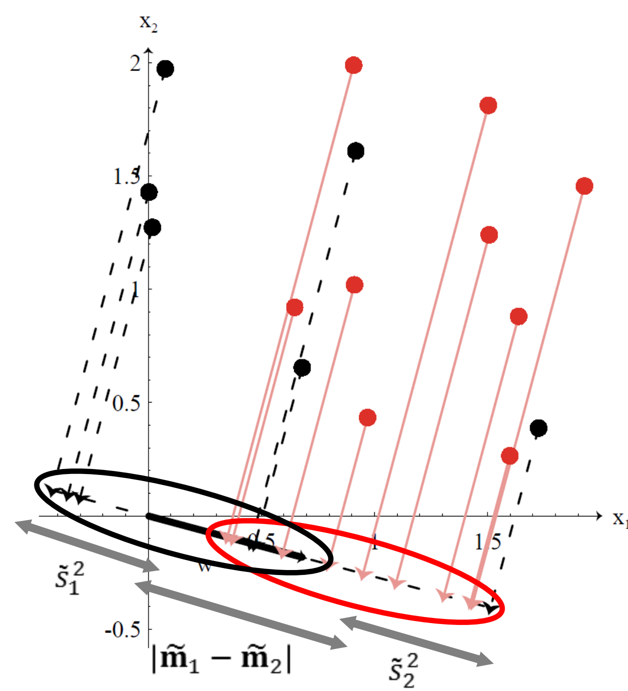
– Thus $(1/N)(\tilde{s}_1^2 + \tilde{s}_2^2)$ is an estimate of the variance of the pooled (or projected) data.

– Total within-class scatter is $\tilde{s}_1^2 + \tilde{s}_2^2$.

– Find that linear function $\mathbf{w}^T\mathbf{x}$ for which

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

is maximum and independent of ||**w**||.



Scatter Matrices

- To obtain $J(\cdot)$ as an explicit function of $\mathbf{w}$, we define scatter matrices $\mathbf{S}_i$ and $\mathbf{S}_W$

$$\mathbf{S}_i = \sum_{\mathbf{x}\in c_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$$

And Within-Class Scatter Matrix $\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$.

- We can then write

$$\tilde{s}_i^2 = \sum_{\mathbf{x}\in c_i} \left(\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{m}_i\right)^2$$

$$= \sum_{\mathbf{x}\in c_i} \mathbf{w}^T(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T\mathbf{w}$$

$$= \mathbf{w}^T\mathbf{S}_i\mathbf{w}$$

Therefore, $\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^T(\mathbf{S}_1 + \mathbf{S}_2)\mathbf{w} = \mathbf{w}^T\mathbf{S}_W\mathbf{w}$ ≥ 0

- The within-class scatter matrix $\mathbf{S}_W$ is proportional to the sample covariance matrix for the D-dimensional data.
- It is symmetric and positive semidefinite, is usually nonsingular if N>D.

Scatter Matrices

– Similarly, the separation of the projected means is

$$|\tilde{\mathbf{m}}_1 - \tilde{\mathbf{m}}_2|^2 = \left(\mathbf{w}^T\mathbf{m_1} - \mathbf{w}^T\mathbf{m_2}\right)^2$$
$$= \mathbf{w}^T(\mathbf{m_1} - \mathbf{m_2})(\mathbf{m_1} - \mathbf{m_2})^T\mathbf{w}$$
$$= \mathbf{w}^T\mathbf{S}_B\mathbf{w}$$

Where Between-Class Scatter Matrix $\boldsymbol{S}_B = (\mathbf{m_1} - \mathbf{m_2})(\mathbf{m_1} - \mathbf{m_2})^T$.

– The between-class scatter matrix $\boldsymbol{S}_B$ is also symmetric and positive semidefinite.

– Its rank is at most one, since it is the outer product of two vectors.

Criterion Function in terms of Scatter Matrices & Final form of Fisher Discriminant

– The criterion function is written as

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

– This is well known the generalised Rayleigh quotient.

– Maximizing the ratio is equivalent to maximizing the numerator while keeping the denominator constant, i.e.

$$\max_{\mathbf{w}} \mathbf{w}^T \mathbf{S}_B \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T \mathbf{S}_W \mathbf{w} = \mathrm{K}$$

– This can be accomplished using Lagrange multipliers as

$$L = \mathbf{w}^T \mathbf{S}_B \mathbf{w} + \lambda(\mathrm{K} - \mathbf{w}^T \mathbf{S}_W \mathbf{w})$$

maximize *L* with respect to both **w** and λ.

Optimisation for Fisher Discriminant

– Setting the gradient of

$$L = \mathbf{w}^T(\mathbf{S}_B - \lambda\mathbf{S}_W)\mathbf{w} + \lambda \mathrm{K}$$

with respect to **w** to zero, we get

$$2(\mathbf{S}_B - \lambda\mathbf{S}_W)\mathbf{w} = 0$$

then

$$\boxed{\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}}$$

– This is a generalized eigenvalue problem.

– The solution is easy, when $\mathbf{S}_W$ is nonsingular:

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w}$$

where $\mathbf{w}$ and λ are the eigenvector and eigenvalue of $\mathbf{S}_W^{-1}\mathbf{S}_B$.

Optimisation for Fisher Discriminant

- In our particular case, using the definition of $\mathbf{S}_B$

$$\mathbf{S}_W^{-1}(\mathbf{m_1}-\mathbf{m}_2)(\mathbf{m_1}-\mathbf{m}_2)^T\mathbf{w}=\lambda\mathbf{w}$$

- Noting that $(\mathbf{m_1}-\mathbf{m_2})^T\mathbf{w}=\alpha$ is a scalar. This can be written as

$$\mathbf{S}_W^{-1}(\mathbf{m_1}-\mathbf{m}_2)=\frac{\lambda}{\alpha}\mathbf{w}$$

- Since we don't care about the magnitude of $\mathbf{w}$

$$\boxed{\mathbf{w}=\mathbf{S}_W^{-1}(\mathbf{m_1}-\mathbf{m}_2)}$$

Multiple Discriminant Analysis

- c-class problem
- Generalization of Fisher's Linear Discriminant function involves M discriminant functions $\mathbf{w}_\mathrm{i}$, $\mathrm{i} = 1, \dots, \mathrm{M}$.
- Projection is from a D-dimensional space to a M-dimensional space.

- The Within-Class and Between-Class Scatter Matrices are defined as

$$\mathbf{S}_W = \sum_{i=1}^{c} \mathbf{S}_i$$

where $\mathbf{S}_i = \sum_{\mathbf{x} \in c_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$,

$$\mathbf{S}_B = \sum_{i=1}^{c} (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T .$$

- The desired projections are found as generalised eigenvectors:

$$\mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{S}_W \mathbf{w}_i, \quad i = 1, \dots, \mathrm{M}$$

for eigenvalues λ_i.

- If $\mathbf{S}_W$ has full rank, the solutions are generalized eigenvectors of $\mathbf{S}_W^{-1} \mathbf{S}_B$ with largest M eigenvalues.

Fisherfaces: Recognition Using Class Specific Linear Projection

- Let us consider N sample images $\{\mathbf{x}_n\}$, n = 1,...,N and $\mathbf{x}_n \in R^D$ in an D-dimensional image space, and assume that each image belongs to one of c classes $\{c_i\}$, i = 1,...,c.
- We consider a linear transformation mapping the D-dimensional image space into an M-dimensional feature space, where M < D.
- The feature vectors $\mathbf{y}_n \in R^M$ are defined by the following linear transformation:

$$\mathbf{y}_n = \mathbf{W}^T \mathbf{x}_n$$

where $\mathbf{W} \in R^{DxM}$ is a matrix with orthonormal columns.

Eigenfaces

- The total scatter matrix $\mathbf{S}_T$ (or the covariance matrix) is defined as

$$\mathbf{S}_T = \sum_n (\mathbf{x}_n - \mathbf{m})(\mathbf{x}_n - \mathbf{m})^T$$

where $\mathbf{m} \in R^D$ is the mean image of all samples.

Fisherfaces: Recognition Using Class Specific Linear Projection

– After applying the linear transformation $\mathbf{W}^T$, the scatter matrix of the feature vectors $\mathbf{y}_n \in R^M$, n = 1,…,N, is $\mathbf{W}^T\mathbf{S}_T\mathbf{W}$

– In PCA, the projection $\mathbf{W}_{opt}$ is chosen to maximize the determinant of the total scatter matrix of the projected samples, i.e.,

$$\mathbf{W}_{\mathrm{opt}} = \arg\max_{\mathbf{W}} \left|\mathbf{W}^T\mathbf{S}_T\mathbf{W}\right| = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{\mathrm{M}}]$$

where $\mathbf{w}_{\mathrm{i}}$ is the set of D-dimensional eigenvectors of $\mathbf{S}_T$ corresponding to the M largest eigenvalues.

– A drawback of this approach is that the scatter being maximized is due not only to the between-class scatter that is useful for classification, but also to the within-class scatter that, for classification purposes, is to be minimized.

Fisherfaces: Recognition Using Class Specific Linear Projection

Fisherfaces

- Since the learning set is labelled, we use this information to build a more reliable method for reducing the feature space dimensionality.
- Using class specific linear methods for dimensionality reduction and NN classifiers in the reduced feature space, one may get better recognition rates than with the Eigenface method.
- Fisher's Linear Discriminant (FLD) is a class specific method that selects **W** in such a way that the ratio of the between-class scatter and the withinclass scatter is maximized.
- Let the between-class scatter matrix be defined as

$$\mathbf{S}_B = \sum_{i=1}^{c} (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T ,$$

the within-class scatter matrix be defined as

$$\mathbf{S}_W = \sum_{i=1}^{c} \sum_{\mathbf{x} \in c_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$$

where $\mathbf{m}_i$ is the mean image of class c_i , and N_i is the number of samples in class c_i.

Fisherfaces: Recognition Using Class Specific Linear Projection

- If $\boldsymbol{S}_W$ is nonsingular, the optimal projection $\mathbf{W}_{\text{opt}}$ is chosen as the matrix with orthonormal columns which maximizes the ratio of the determinant of the between-class scatter matrix to the determinant of the within-class scatter matrix of the projected samples, i.e.,

$$\mathbf{W}_{\text{opt}} = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^T \boldsymbol{S}_B \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_W \mathbf{W}|} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_\mathrm{M}]$$

where $\mathbf{w}_\mathrm{i}$ is the set of generalized eigenvectors of $\boldsymbol{S}_B$ and $\mathbf{S}_W$ corresponding to the M largest eigenvalues.

$$\mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{S}_W \mathbf{w}_i, \quad \mathrm{i} = 1, \dots, \mathrm{M}$$

- Note that there are at most c - 1 nonzero generalized eigenvalues, and so an upper bound on M is c - 1.
- In the face recognition problem, the within-class scatter matrix $\mathbf{S}_W \in \mathrm{R}^{DxD}$ is often singular.
- This stems from the fact that the rank of $\mathbf{S}_W$ is at most N - c, and, in general, N is smaller than D.

Fisherfaces: Recognition Using Class Specific Linear Projection

- In order to overcome the singular $\boldsymbol{S}_W$, we propose an alternative to the criterion.
- This method, which we call Fisherfaces, avoids the problem by projecting the image set to a lower dimensional space.
- We use PCA to reduce the dimension of the feature space M_{pca} (<=N–c), and then apply the standard FLD to reduce the dimension to M_{lda} (<=c–1).
- More formally, $\mathbf{W}_{opt}$ is given by

$$\mathbf{W}_{\text{opt}}^T = \mathbf{W}_{\text{lda}}^T \mathbf{W}_{\text{pca}}^T$$

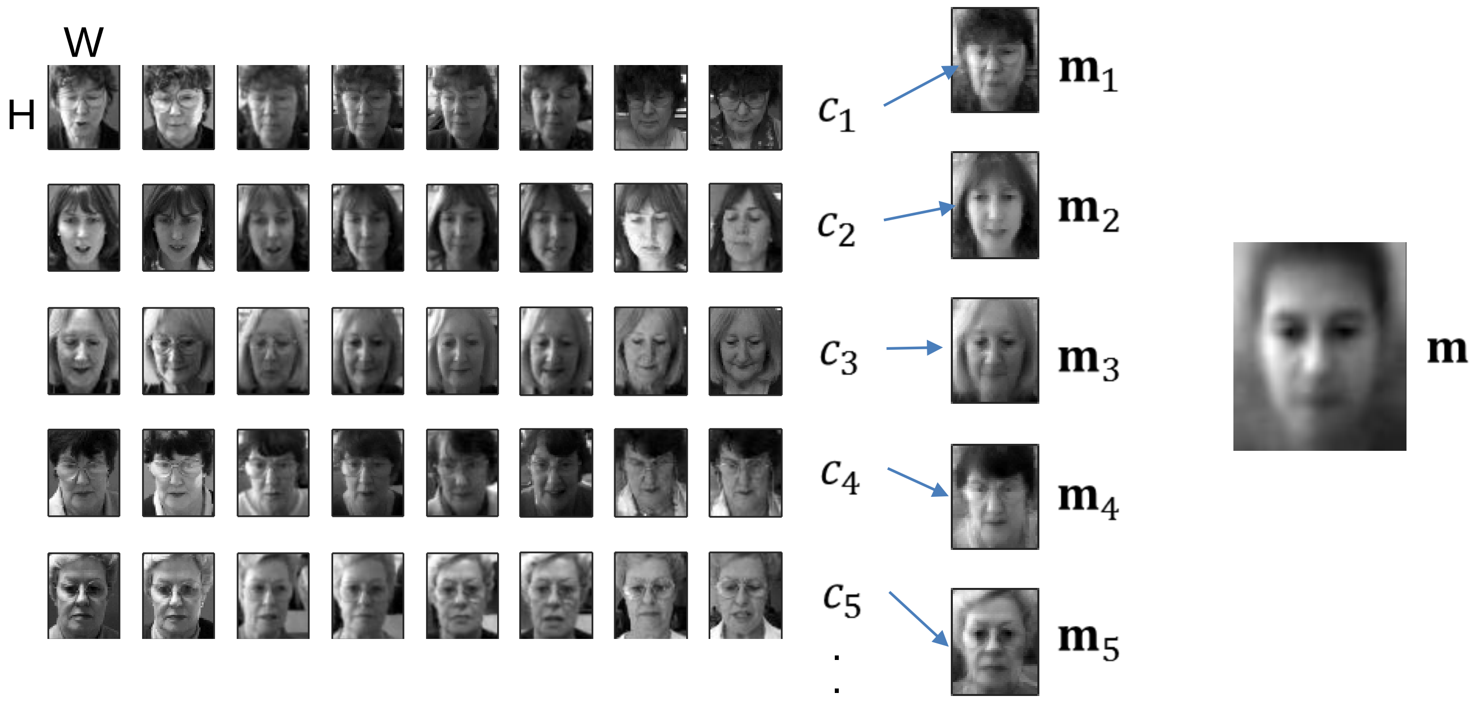
$$\mathbf{W}_{\text{pca}} = \arg\max_{\mathbf{W}} \left| \mathbf{W}^T \boldsymbol{S}_T \mathbf{W} \right|$$

$$\mathbf{W}_{\text{lda}} = \arg\max_{\mathbf{W}} \frac{\left| \mathbf{W}^T \mathbf{W}_{\text{pca}}^T \boldsymbol{S}_B \mathbf{W}_{\text{pca}} \mathbf{W} \right|}{\left| \mathbf{W}^T \mathbf{W}_{\text{pca}}^T \boldsymbol{S}_W \mathbf{W}_{\text{pca}} \mathbf{W} \right|}$$

- There are other ways of reducing the withinclass scatter while preserving between-class scatter e.g. Direct LDA, Null LDA, etc.

Procedures: Fisherfaces

- Collect training images $\mathbf{x}_n$ of c classes (c=26, N=208,. D=2576)
- Compute the class means $\mathbf{m}_i$, i=1,…,c and the global mean $\mathbf{m}$

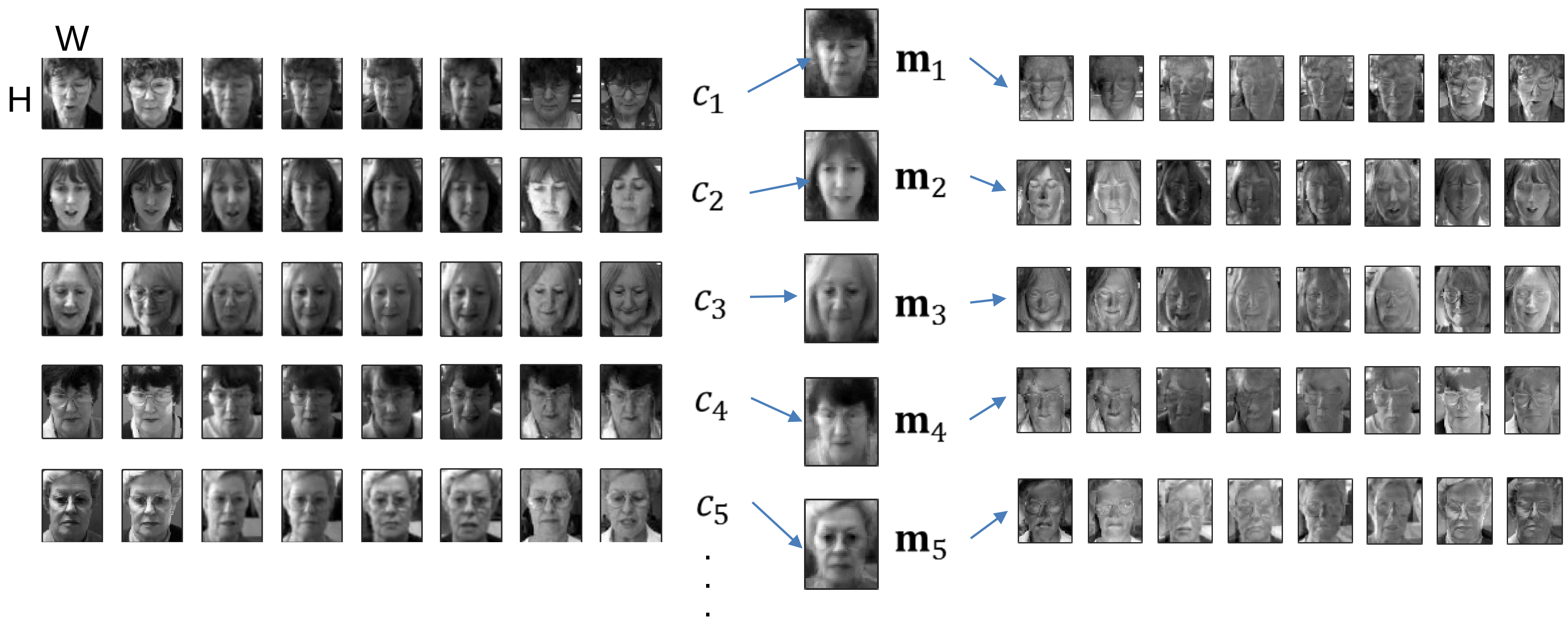


Procedures: Fisherfaces

- Compute $\mathbf{m}_i - \mathbf{m}$, and $\mathbf{S}_B$

Procedures: Fisherfaces

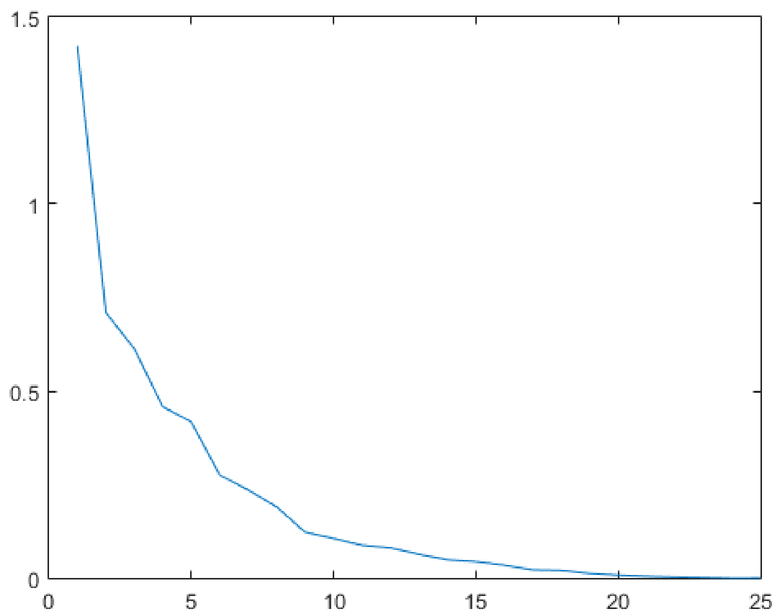
– Compute $\mathbf{x} - \mathbf{m}_i$, and $\mathbf{S}_W$



Procedures: Fisherfaces

- rank(Sw) = 182 (=N – c), rank(Sb) = 25 (=c – 1)
- Perform PCA to get $\mathbf{W}_{\text{pca}}$ (Mpca=25), and compute $\mathbf{W}_{\text{pca}}^T \mathbf{S}_B \mathbf{W}_{\text{pca}}$ and $\mathbf{W}_{\text{pca}}^T \mathbf{S}_W \mathbf{W}_{\text{pca}}$.
- Get the generalized eigenvectors of $(\mathbf{W}_{\text{pca}}^T \mathbf{S}_W \mathbf{W}_{\text{pca}})^{-1}(\mathbf{W}_{\text{pca}}^T \mathbf{S}_B \mathbf{W}_{\text{pca}})$ with largest Mlda eigenvalues.

more discriminative

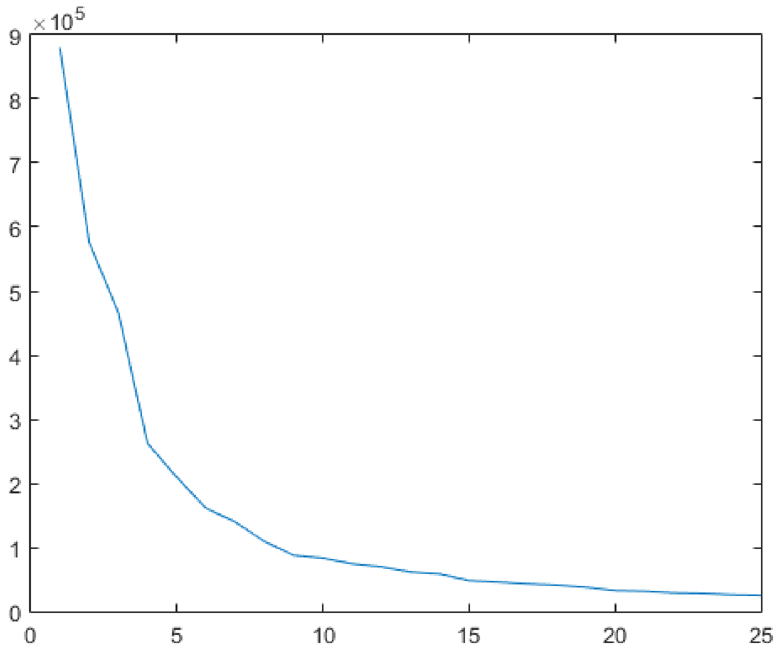


Generalized eigenvalues

Top 25 generalized eigenvectors

Comparison to Eigenfaces

more generative.

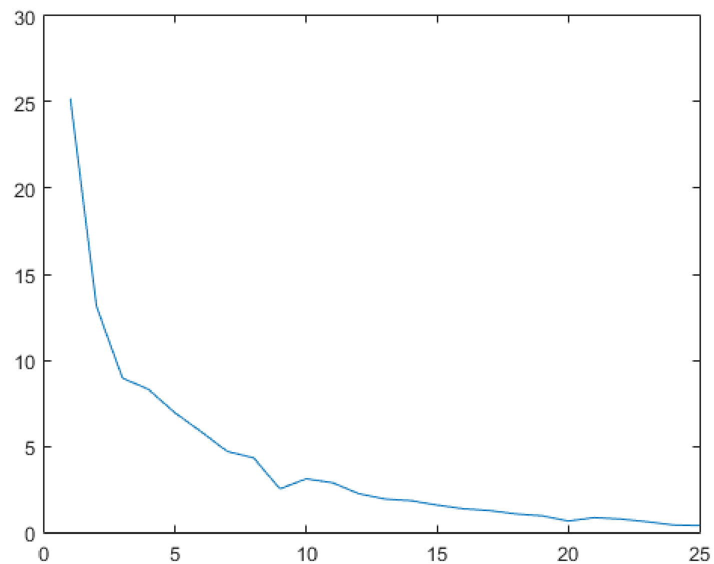


Eigenvalues

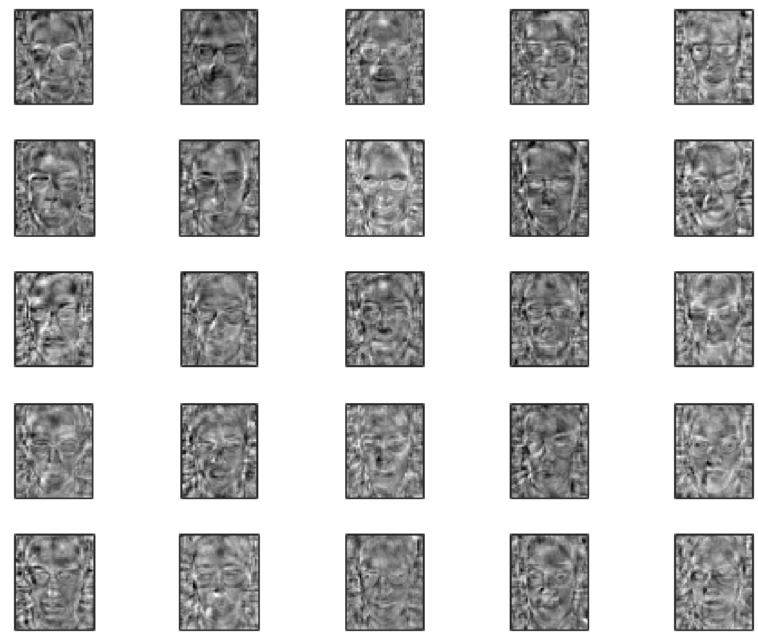
Top 25 eigenvectors

Procedures: Fisherfaces

- rank(Sw) = 182 (=N – c), rank(Sb) = 25 (=c – 1)
- Perform PCA to get $\mathbf{W}_{\text{pca}}$ (Mpca=150), and compute $\mathbf{W}_{\text{pca}}^T \mathbf{S}_B \mathbf{W}_{\text{pca}}$ and $\mathbf{W}_{\text{pca}}^T \mathbf{S}_W \mathbf{W}_{\text{pca}}$.
- Get the generalized eigenvectors of $(\mathbf{W}_{\text{pca}}^T \mathbf{S}_W \mathbf{W}_{\text{pca}})^{-1}(\mathbf{W}_{\text{pca}}^T \mathbf{S}_B \mathbf{W}_{\text{pca}})$ with largest Mlda eigenvalues.

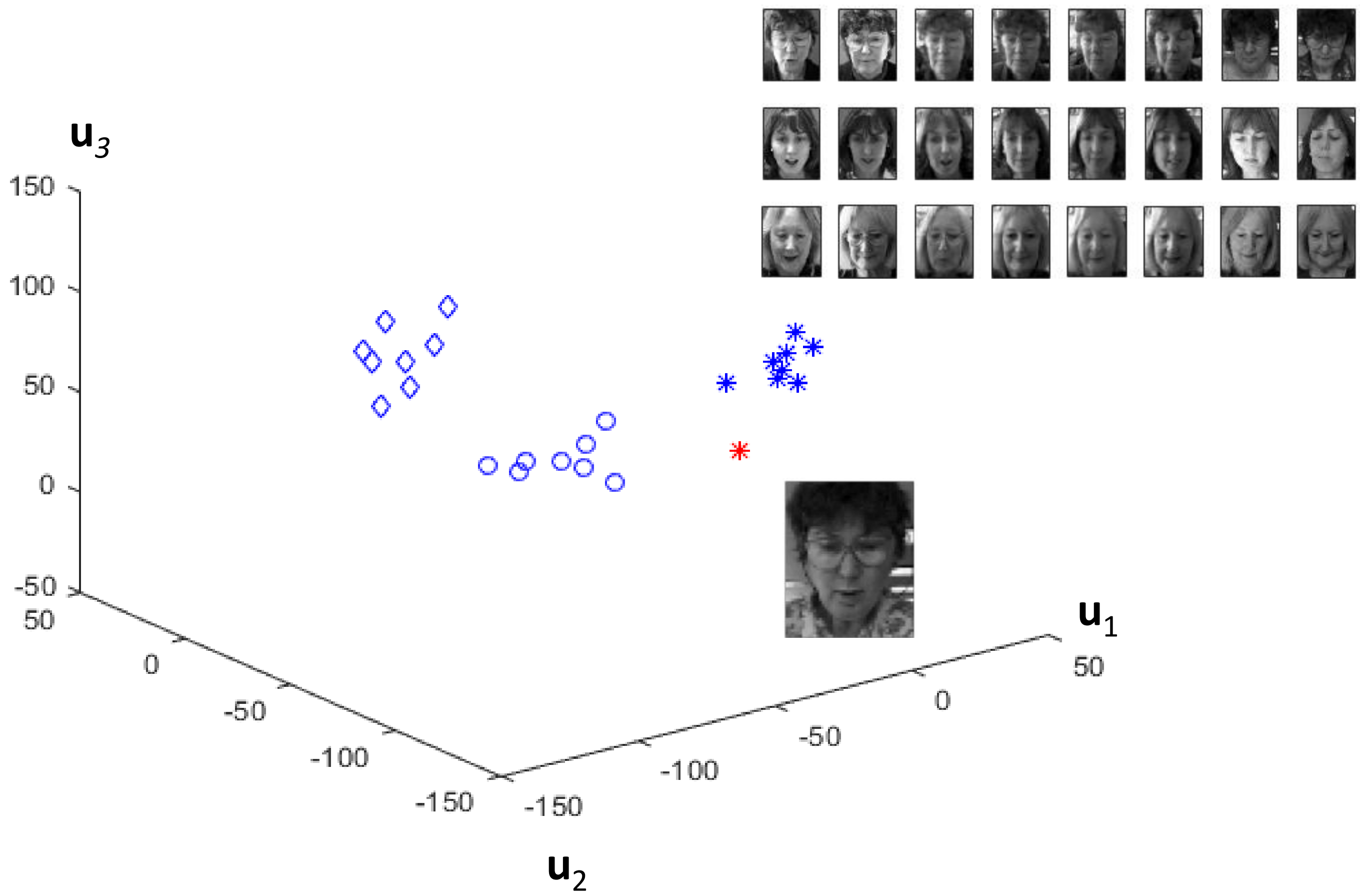


Generalized eigenvalues



Top 25 generalized eigenvectors

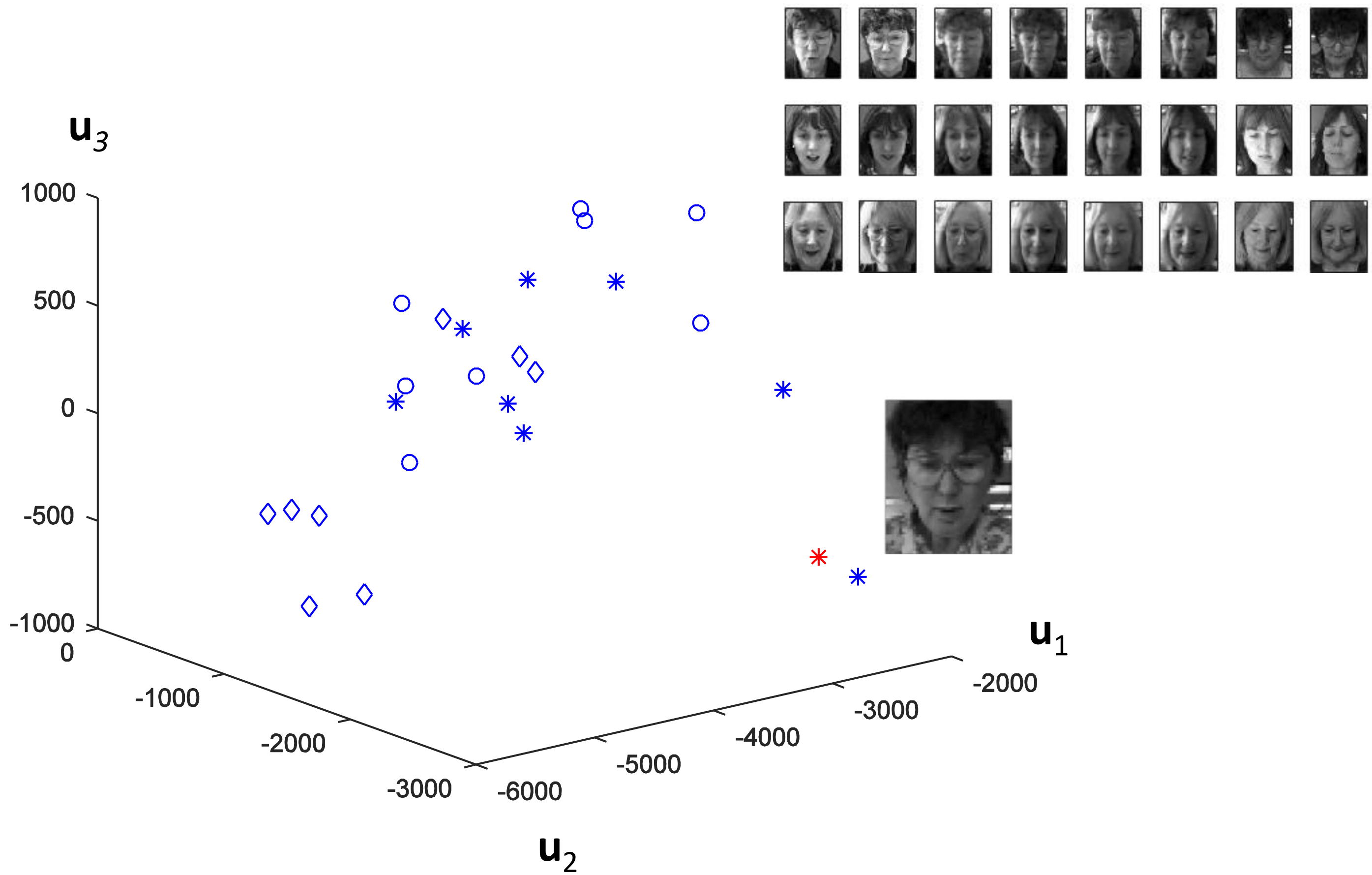
Procedures: Fisherfaces



Face images in 3-dimensional fisher-subspace

24 training images of 3 different face classes (star, diamond, circle, “in blue”) are projected. A query image projection is “in red”.

Comparison to Eigenfaces



Face images in 3-dimensional eigen-subspace

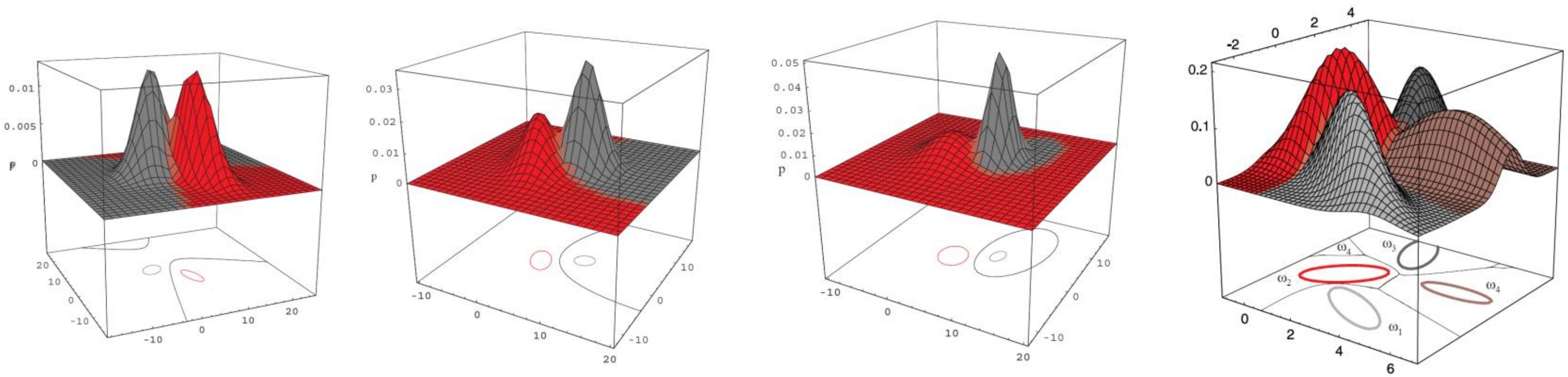
24 training images of 3 different face classes (star, diamond, circle, “in blue”) are projected. A query image projection is “in red”.

Relation to Optimal Bayesian Decision Theory

Bayes Decision Theory

- Fundamental statistical approach to statistical pattern classification
- Quantifies trade-offs between classification using probabilities and costs of decisions
- Assumes all relevant probabilities are known

- (Σ_i (data covariance matrix of class i)= arbitrary) Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics

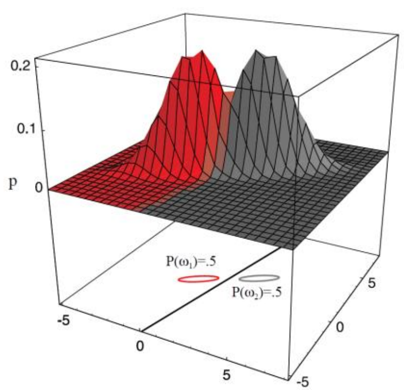


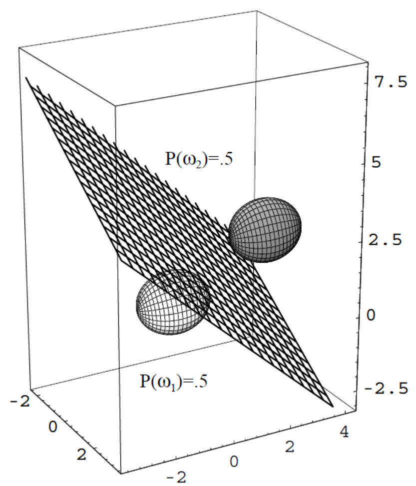
Relation to Optimal Bayesian Decision Theory

– ($\Sigma_i = \Sigma$) For a classification problem with Gaussian classes of equal covariance $\Sigma_i = \Sigma$, the Bayes decision boundaries (or the discriminant function) is the plane of normal

$$\mathbf{w} = \Sigma^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

– The hyperplane is generally not orthogonal to the line between the means.





Relation to Optimal Bayesian Decision Theory

- If $\Sigma_1 = \Sigma_2$, this is also the FLD solution.
- In FLD, $\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$, $\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$

- This gives some interpretations of FLD/LDA
  - It is optimal if and only if the classes are Gaussian and have equal covariance.
  - Better than PCA, but not necessarily good enough.
  - A classifier on the LDA feature, is equivalent to the BDR after the approximation of the data by two Gaussians with equal covariance.
  - The extension from two-classes to multiple classes in LDA is ad-hoc. done for a particular purpose only.

Tae-Kyun Kim, PhD dissertation (Discriminant Analysis of Patterns in Images, Image Ensembles, and Videos), Univ. of Cambridge, 2007.