Discriminant Analysis
Fisherfaces

Tae-Kyun (T-K) Kim
KAIST, Imperial College London
https://sites.google.com/view/tkkim/

Some references

ICML 2018

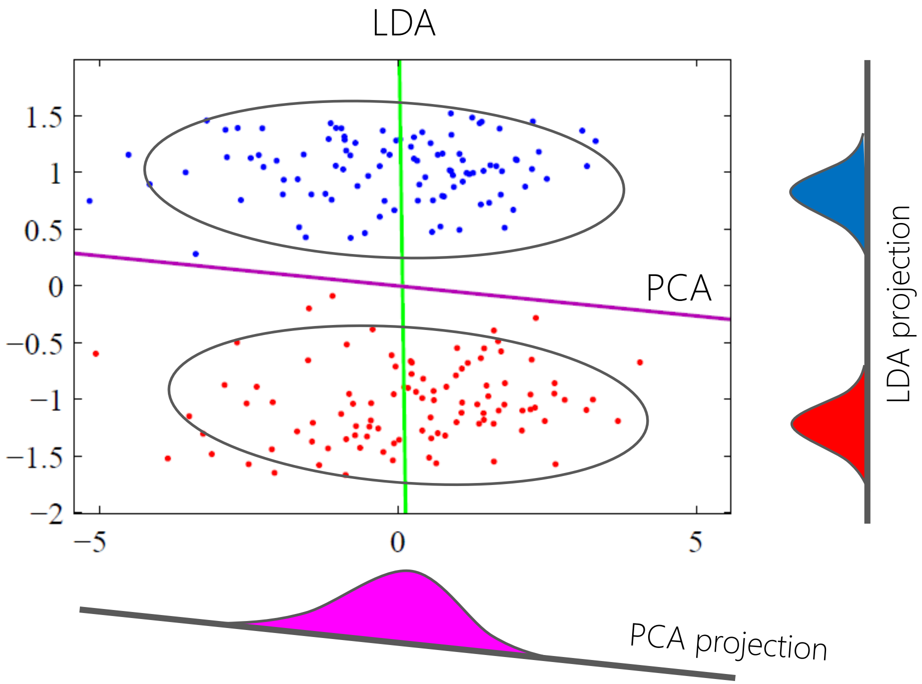
- *Max-Mahalanobis Linear Discriminant Analysis Networks,* Tianyu Pang · Chao Du · Jun Zhu
- *Discovering Interpretable Representations for Both Deep Generative and Discriminative Models,* Tameem Adel · Zoubin Ghahramani · Adrian Weller
- *Mixed batches and symmetric discriminators for GAN training,* Thomas LUCAS · Corentin Tallec · Yann Ollivier · Jakob Verbeek
- *Batch Bayesian Optimization via Multi-objective Acquisition Ensemble for Automated Analog Circuit Design,* Wenlong Lyu · Fan Yang · Changhao Yan · Dian Zhou · Xuan Zeng
- *High-Quality Prediction Intervals for Deep Learning: A Distribution-Free, Ensembled Approach,* Tim Pearce · Alexandra Brintrup · Mohamed Zaki · Andy Neely
- *Generalized Robust Bayesian Committee Machine for Large-scale Gaussian Process Regression,* Haitao Liu · Jianfei Cai · Yi Wang · Yew Soon ONG

NIPS 2018

- *Discrimination-aware Channel Pruning for Deep Neural Networks,* Zhuangwei Zhuang · Mingkui Tan · Bohan Zhuang · Jing Liu · Yong Guo · Qingyao Wu · Junzhou Huang · Jinhui Zhu
- *Hunting for Discriminatory Proxies in Linear Regression Models,* Samuel Yeom · Anupam Datta · Matt Fredrikson
- *Virtual Class Enhanced Discriminative Embedding Learning,* Binghui Chen · Weihong Deng · Haifeng Shen
- *Power-law efficient neural codes provide general link between perceptual bias and discriminability,* Michael Morais · Jonathan W Pillow
- *Unsupervised Text Style Transfer using Language Models as Discriminators,* Zichao Yang · Zhiting Hu · Chris Dyer · Eric Xing · Taylor Berg-Kirkpatrick
- *Why Is My Classifier Discriminatory?,* Irene Chen · Fredrik Johansson · David Sontag
- *Learning from discriminative feature feedback,* Sanjoy Dasgupta · Sivan Sabato · Nicholas Roberts · Akansha Dey
- *On preserving non-discrimination when combining expert advice,* Avrim Blum · Suriya Gunasekar · Thodoris Lykouris · Nati Srebro
- *Sample-Efficient Reinforcement Learning with Stochastic Ensemble Value Expansion,* Jacob Buckman · Danijar Hafner · George Tucker · Eugene Brevdo · Honglak Lee
- *Diverse Ensemble Evolution: Curriculum based Data-Model Marriage,* Tianyi Zhou · Shengjie Wang · Jeff Bilmes
- *Knowledge Distillation by On-the-Fly Native Ensemble,* Xu lan · Xiatian Zhu · Shaogang Gong
- *Using Large Ensembles of Control Variates for Variational Inference,* Tomas Geffner · Justin Domke

Motivation

- Projection that best separates the data in a least-squares sense:
  - PCA finds components that are useful for representing data.
  - Pooling (or projecting) data may discard essential information for discriminating between data in different classes.
  - PCA finds the direction for maximum data variance (unsupervised/generative).
  - LDA (Linear Discriminant Analysis) or MDA (Multiple Discriminant Analysis) finds the direction that optimally separates data of different classes (supervised/discriminative).



Fisher Linear Discriminant (FLD)

- We first consider *2-class problem* i.e. binary-classification.
- Data are projected from *D* dimensions onto a line, i.e. one-dimensional subspace.

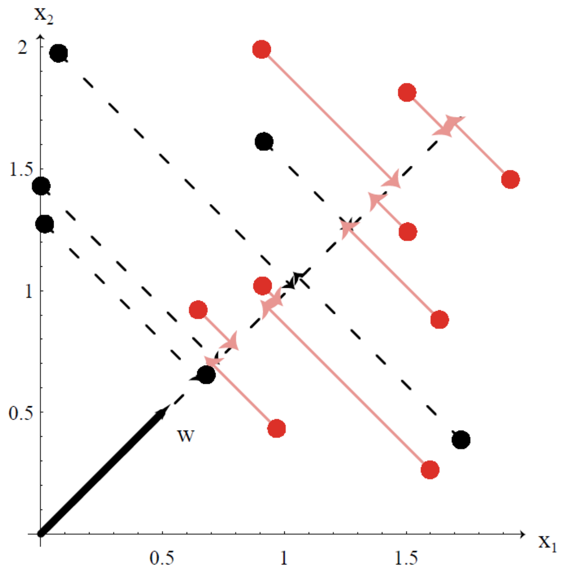
- Given a set of N *D*-dimensional samples $\mathbf{x_1}, \ldots, \mathbf{x}_N$,
  where $N_\mathbf{1}$ samples belong to the class $c_\mathbf{1}$ and $N_\mathbf{2}$ to the class $c_\mathbf{2}$.

- We wish to form a linear combination of the components of $\mathbf{x}$ as
  $y = \mathbf{w}^T\mathbf{x}$
  and a corresponding set of N samples $y_\mathbf{1}, \ldots, y_N$.

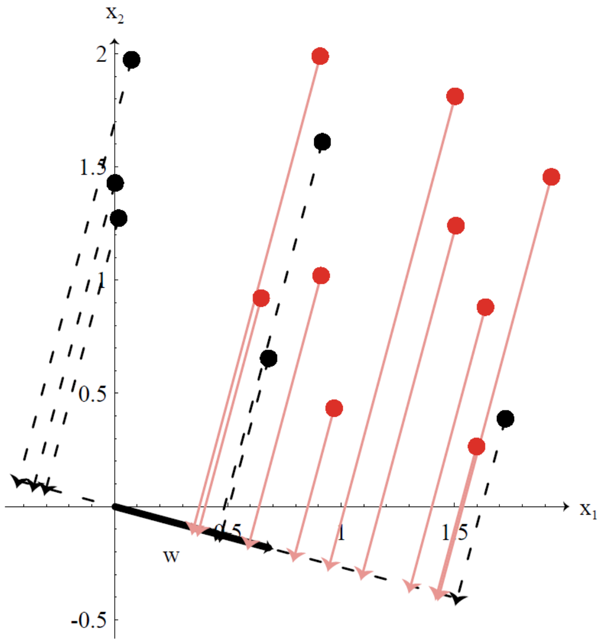
FLD: two-dimensional example

- 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 class means is

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

Criterion for Fisher Linear Discriminant

- Define the scatter of the projected samples as

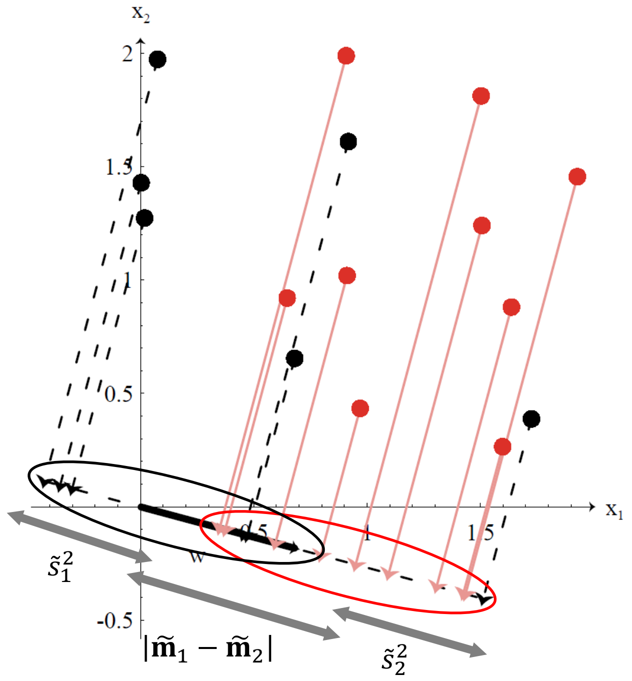
$$\tilde{s}_i^2 = \sum_{y \in c_i} (\mathbf{y} - \tilde{\mathbf{m}}_i)^2$$

- Thus $(1/N)(\tilde{s}_1^2 + \tilde{s}_2^2)$ is 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

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

is maximum and independent of ||w||.



Scatter matrices

- To obtain *J(·)* as an explicit function of 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

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

$(w^T x - w^T m)(w^T x - w^T m)$

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

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

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

- The within-class scatter matrix $\mathbf{S}_W \in R^{DxD}$ is symmetric and positive semidefinite, and is nonsingular if N>D.

Scatter matrices

- Similarly, the separation of the projected class means is

$$|\widetilde{\mathbf{m}}_1 - \widetilde{\mathbf{m}}_2|^2 = (\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2)^2$$
$$= \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w}$$
$$=\mathbf{w}^T\mathbf{S}_B\mathbf{w}$$

Where Between-class scatter matrix $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$.

- The between-class scatter matrix $\mathbf{S}_B$ 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 and optimisation

- 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.

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

- 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

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

- 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 $\mathbf{S}_W^{-1}\mathbf{S}_B$.

Multiple Discriminant Analysis

- Generalization of Fisher's Linear Discriminant, for multiple *c* classes, involves M discriminant functions $\mathbf{w_i}, \ \ \mathrm{i} = 1, \dots, \mathrm{M}$.
- Projection is from a D-dimensional space to a M-dimensional subspace.

- The Within-class and Between-class scatter matrices are defined as

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

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

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

- The desired projections are found as generalised eigenvectors:

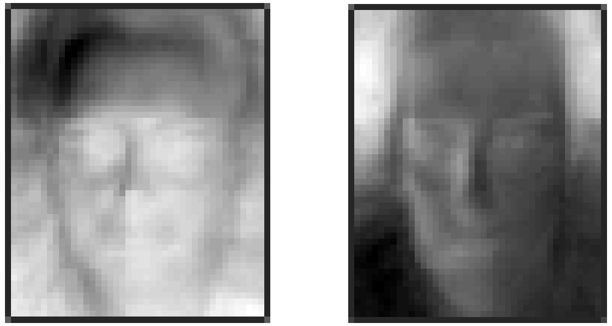
$$\mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{S}_W \mathbf{w}_i, \quad i = 1, \dots, \mathrm{M}$$

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.

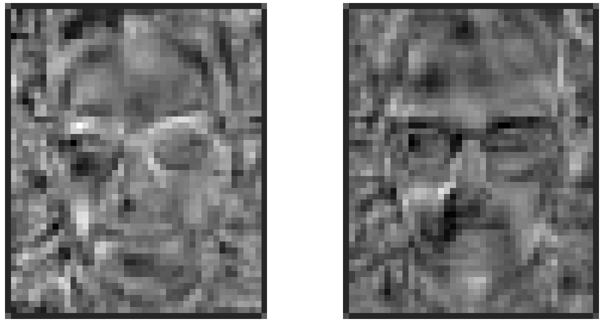
Fisherfaces

Eigenfaces



vs

Fisherfaces



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

Recognition using class specific linear projection

- Let us consider N sample images $\{x_n\}$, n = 1,…,N and $x_n \in R^D$ in an D-dimensional image space, and assume that each image belongs to one of c classes $\{\boldsymbol{c_i}\}$, 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 $y_n \in R^M$ are defined by the following linear transformation:

$$y_n = \mathbf{W}^T x_n$$

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

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 W_{opt} is chosen to maximize the determinant of the total scatter matrix of the projected samples, i.e.,

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

where $\mathbf{w_i}$, i = 1,...,M is the set of D-dimensional eigenvectors of $\mathbf{S}_T$ corresponding to the M largest eigenvalues.

- A drawback of this approach is that both the between-class and within-class scatter are maximized, since $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$.

Recognition using class specific linear projection

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

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

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

- If $\boldsymbol{S}_W$ is nonsingular, the optimal projection 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.,

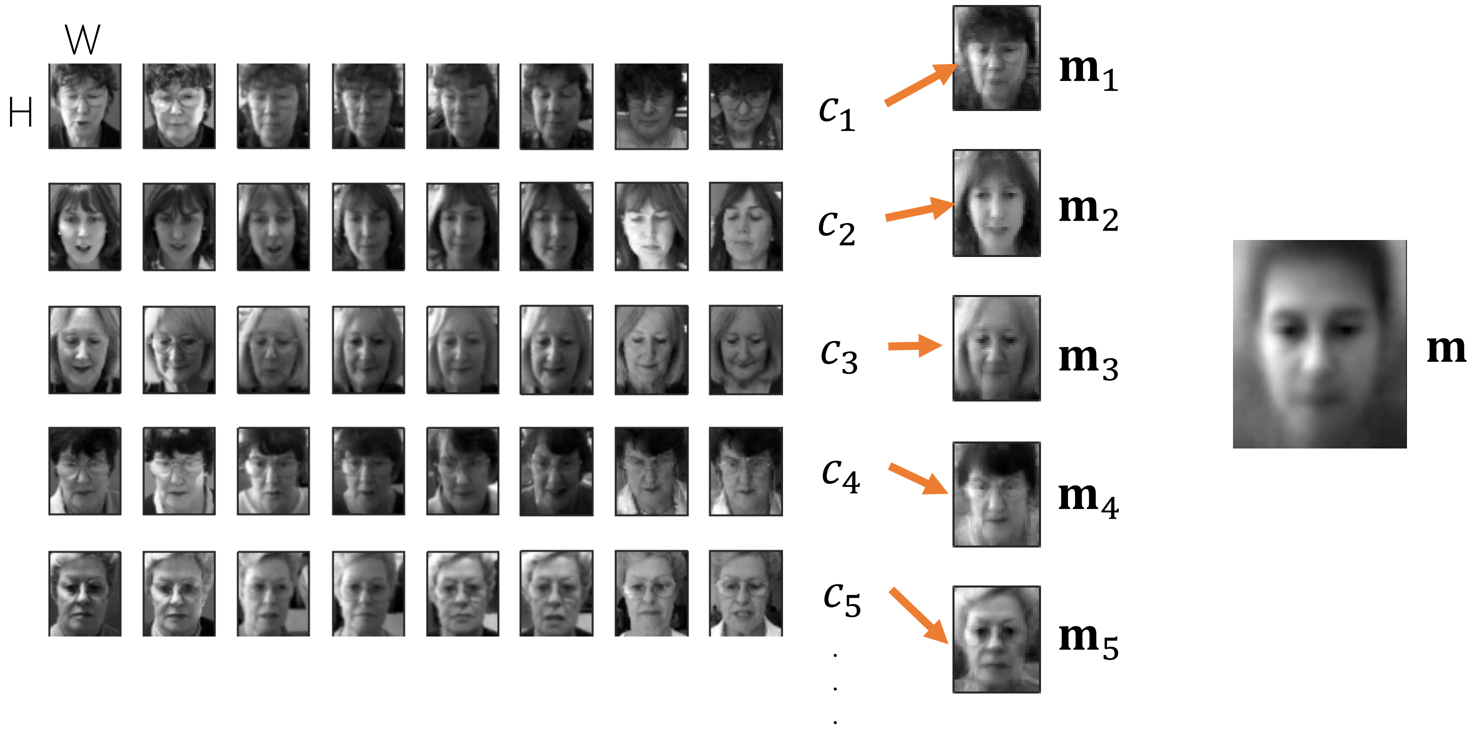
$$\mathbf{W}_{\text{opt}} = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^T \boldsymbol{S}_B \mathbf{W}|}{|\mathbf{W}^T \boldsymbol{S}_W \mathbf{W}|} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{\mathrm{M}}]$$

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

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

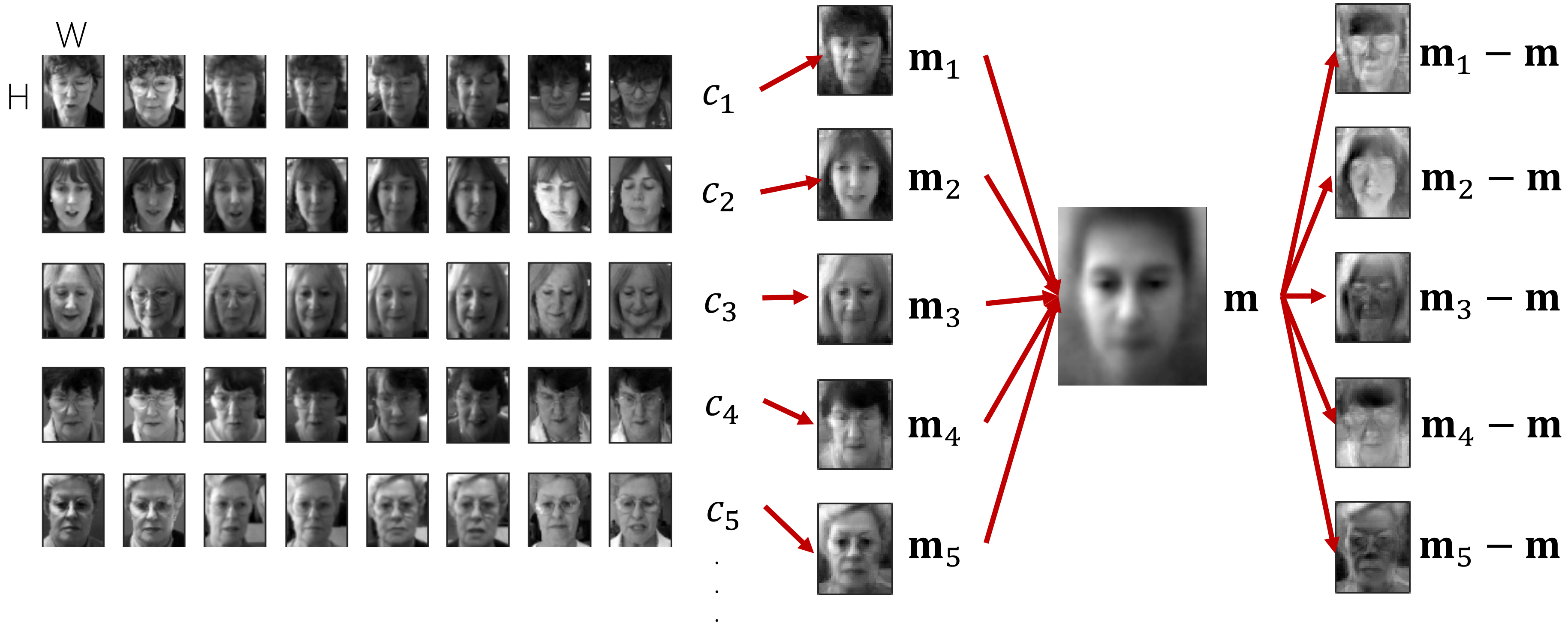
Procedures: Fisherfaces

- 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 $\mathbf{m}$



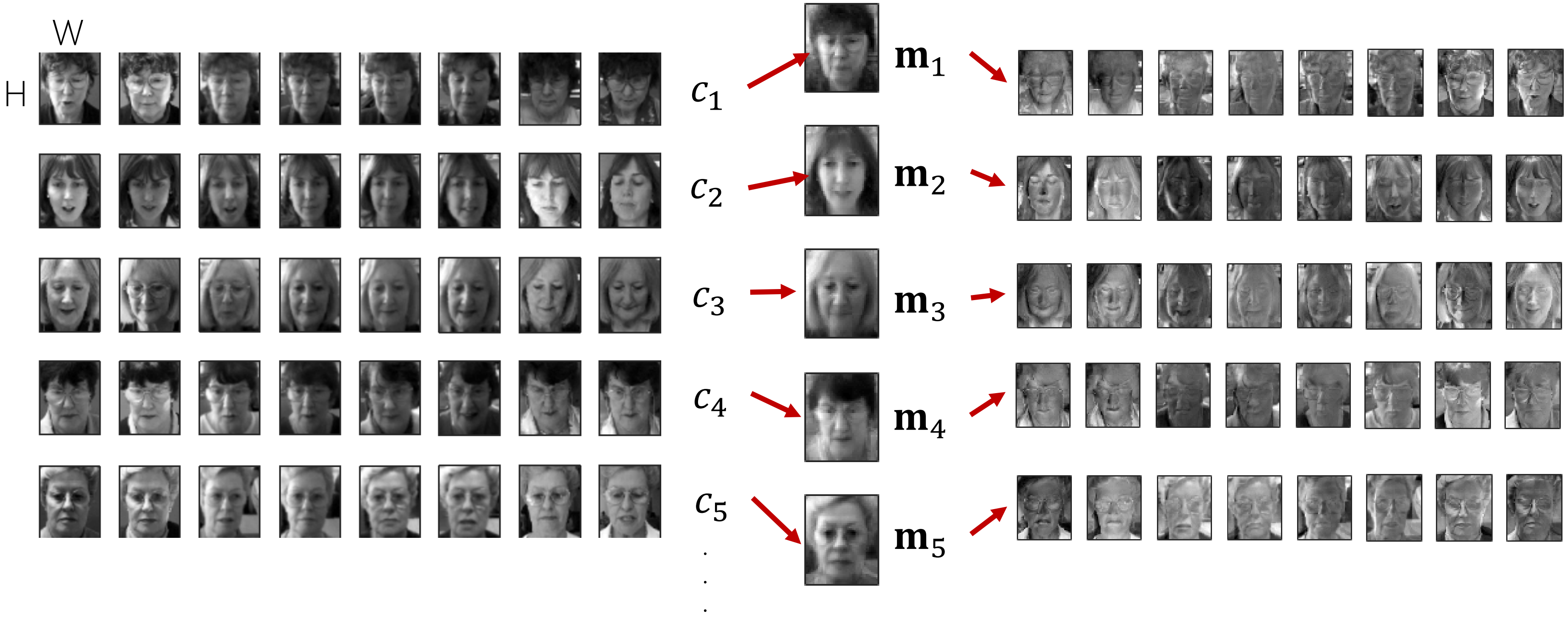
Procedures: Fisherfaces

- Compute $\mathbf{m}_i - \mathbf{m}$, and $\mathbf{S}_B$, where rank($\mathbf{S}_B$) = c -1.



Procedures: Fisherfaces

- Compute $\mathbf{x} - \mathbf{m}_i$, and $\mathbf{S}_W$, where rank($\mathbf{S}_W$) is N – c.



Fisherfaces

- Given the generalized eigenvalue/vector problem of $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$:

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

- Note that there are at most c - 1 nonzero generalized eigenvalues i.e. the rank of $\boldsymbol{S}_B$, and so an upper bound on M is c - 1.
- The within-class scatter matrix $\boldsymbol{S}_W \in \mathrm{R}^{D x D}$ is often singular, since the rank of $\boldsymbol{S}_W$ is at most N - c, and, in general, N is smaller than D.

Fisherfaces

- In order to overcome the singular $\mathbf{S}_W$, 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).
- Formally, W_{opt} is given by

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

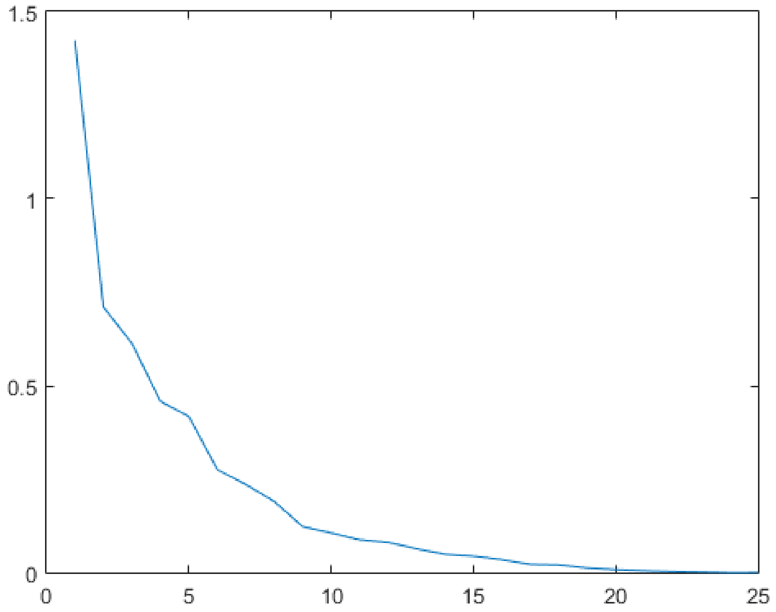
$$\mathbf{W}_{\mathrm{pca}} = \arg\max_{\mathbf{W}} |\mathbf{W}^T \mathbf{S}_T \mathbf{W}|$$

$$\mathbf{W}_{\mathrm{lda}} = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^T \mathbf{W}_{\mathrm{pca}}^T \mathbf{S}_B \mathbf{W}_{\mathrm{pca}} \mathbf{W}|}{|\mathbf{W}^T \mathbf{W}_{\mathrm{pca}}^T \mathbf{S}_W \mathbf{W}_{\mathrm{pca}} \mathbf{W}|}$$

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

Procedures: Fisherfaces

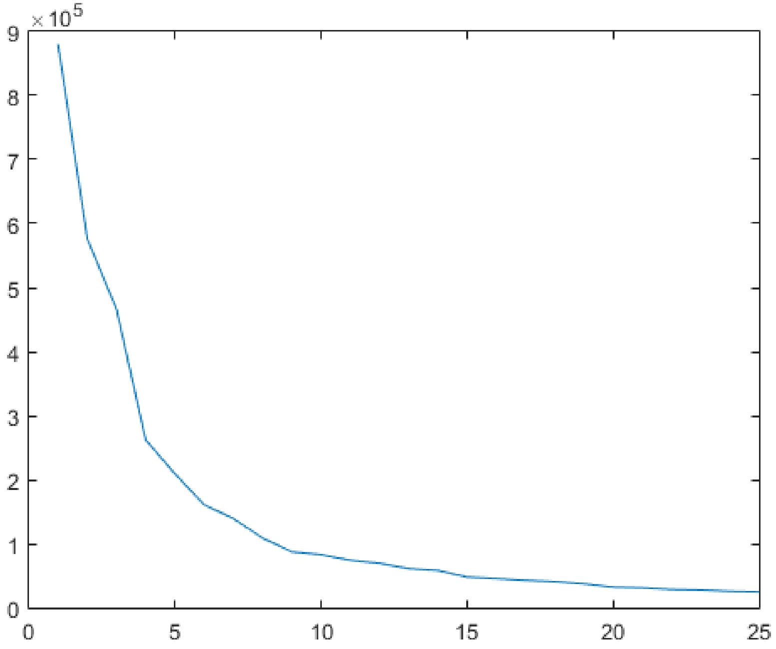
- rank(Sw) = 182 (=N – c), rank(Sb) = 25 (=c – 1)
- Perform PCA to get $\mathbf{W}_{\mathbf{pca}}$ (Mpca=25), and compute $\mathbf{W}_{\mathbf{pca}}^T \mathbf{S}_B \mathbf{W}_{\mathbf{pca}}$ and $\mathbf{W}_{\mathbf{pca}}^T \mathbf{S}_W \mathbf{W}_{\mathbf{pca}}$.
- Get the generalized eigenvectors of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



Generalized eigenvalues


Top 25 generalized eigenvectors

Comparison to Eigenfaces

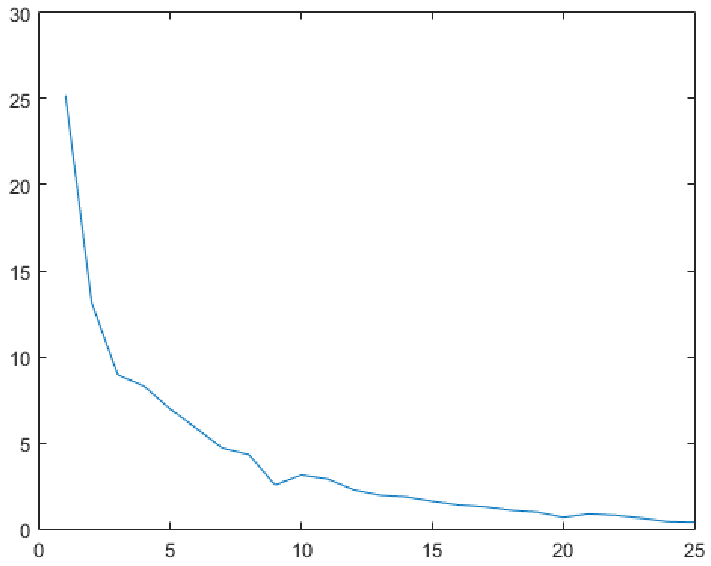


Eigenvalues

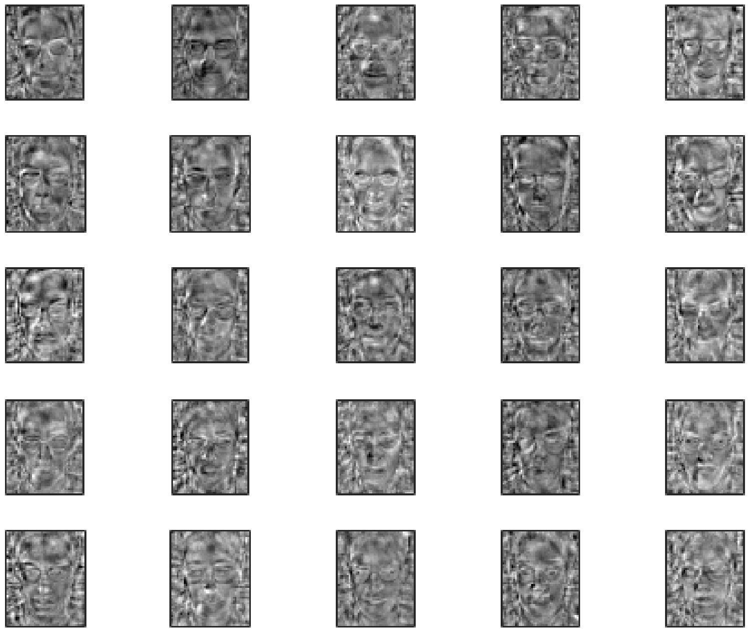
Top 25 eigenvectors

Procedures: Fisherfaces

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

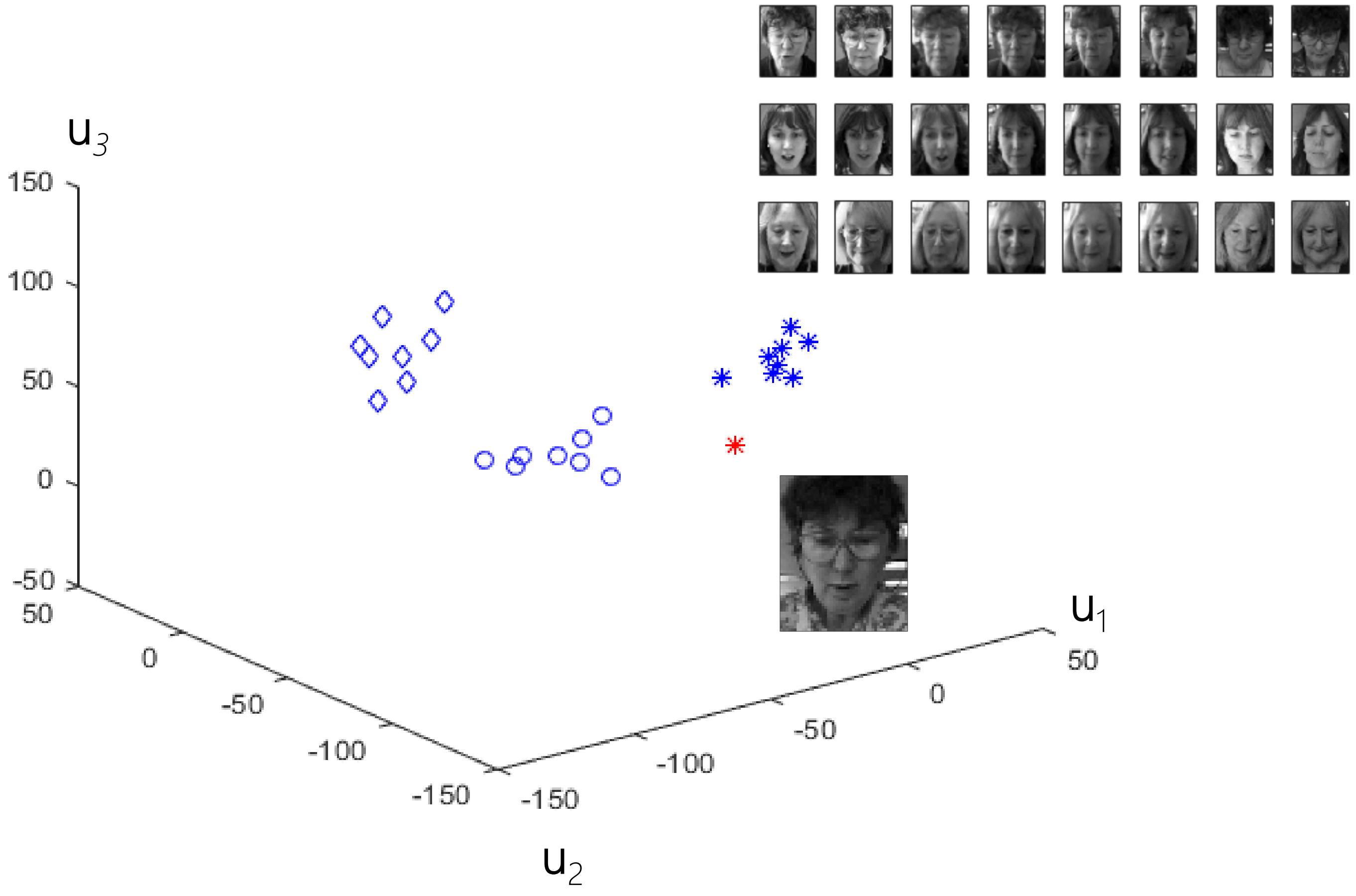


Generalized eigenvalues



Top 25 generalized eigenvectors

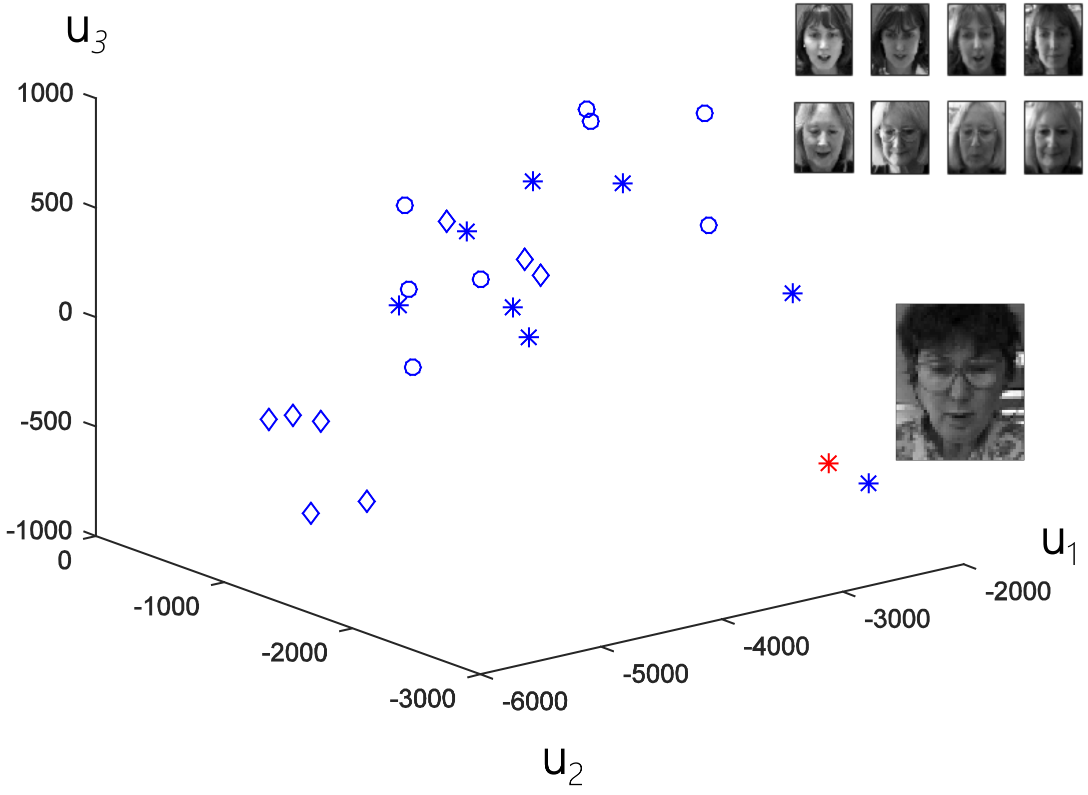
Procedures: Fisherfaces



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



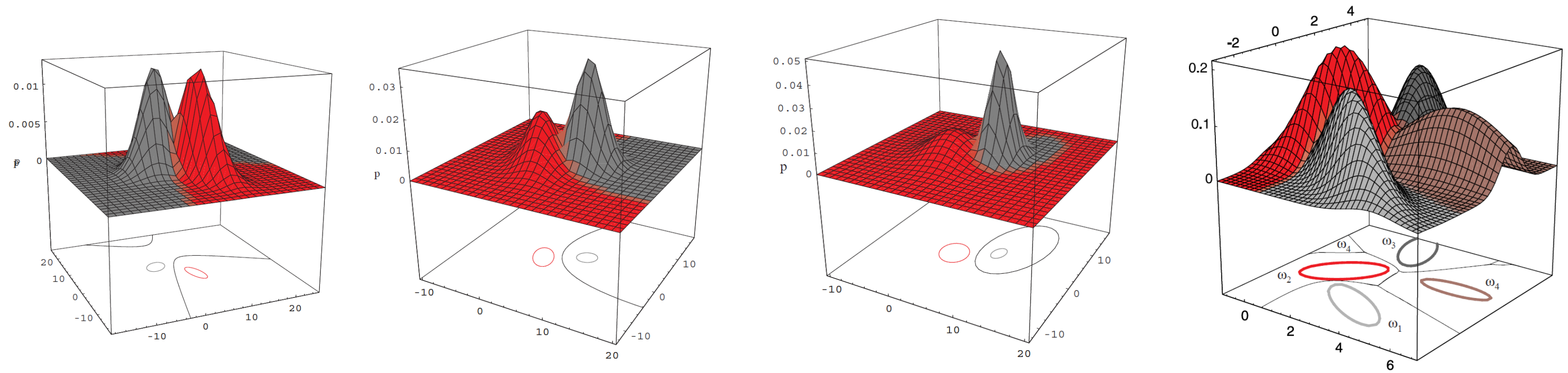
Face images in 3-dimensional eigen-subspace

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 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



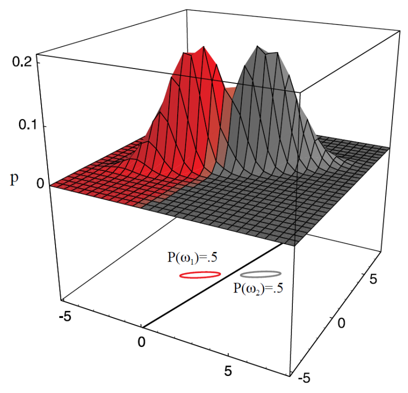
http://web.stanford.edu/class/stats202/content/lec9-cond.pdf

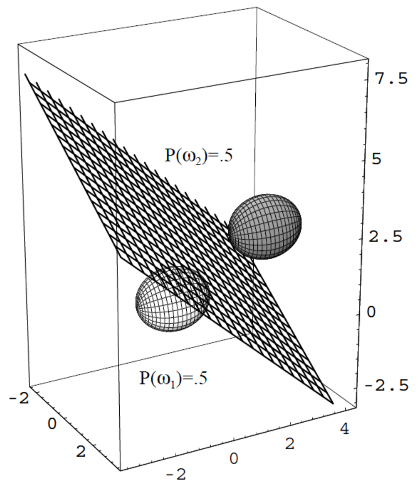
Relation to optimal Bayesian decision theory

- $\Sigma_i = \Sigma$
  - 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.





Optimisation for Fisher Discriminant

- In our particular case i.e. 2 class problem, using the definition of $\boldsymbol{S_B}$

$$\mathbf{S}_W^{-1}(\mathbf{m_1} - \mathbf{m_2})(\mathbf{m_1} - \mathbf{m_2})^T\mathbf{w} = \lambda\mathbf{w}$$

- 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 w

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m_1} - \mathbf{m_2})$$

Relation to optimal Bayesian decision theory

- If $\Sigma_{\mathbf{1}} = \Sigma_{\mathbf{2}}$, 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.
  - The extension from two-classes to multiple classes in LDA is ad-hoc.

T-K Kim, PhD dissertation: Discriminant Analysis of Patterns in Images, Image Ensembles, and Videos, Univ. of Cambridge, 2008.