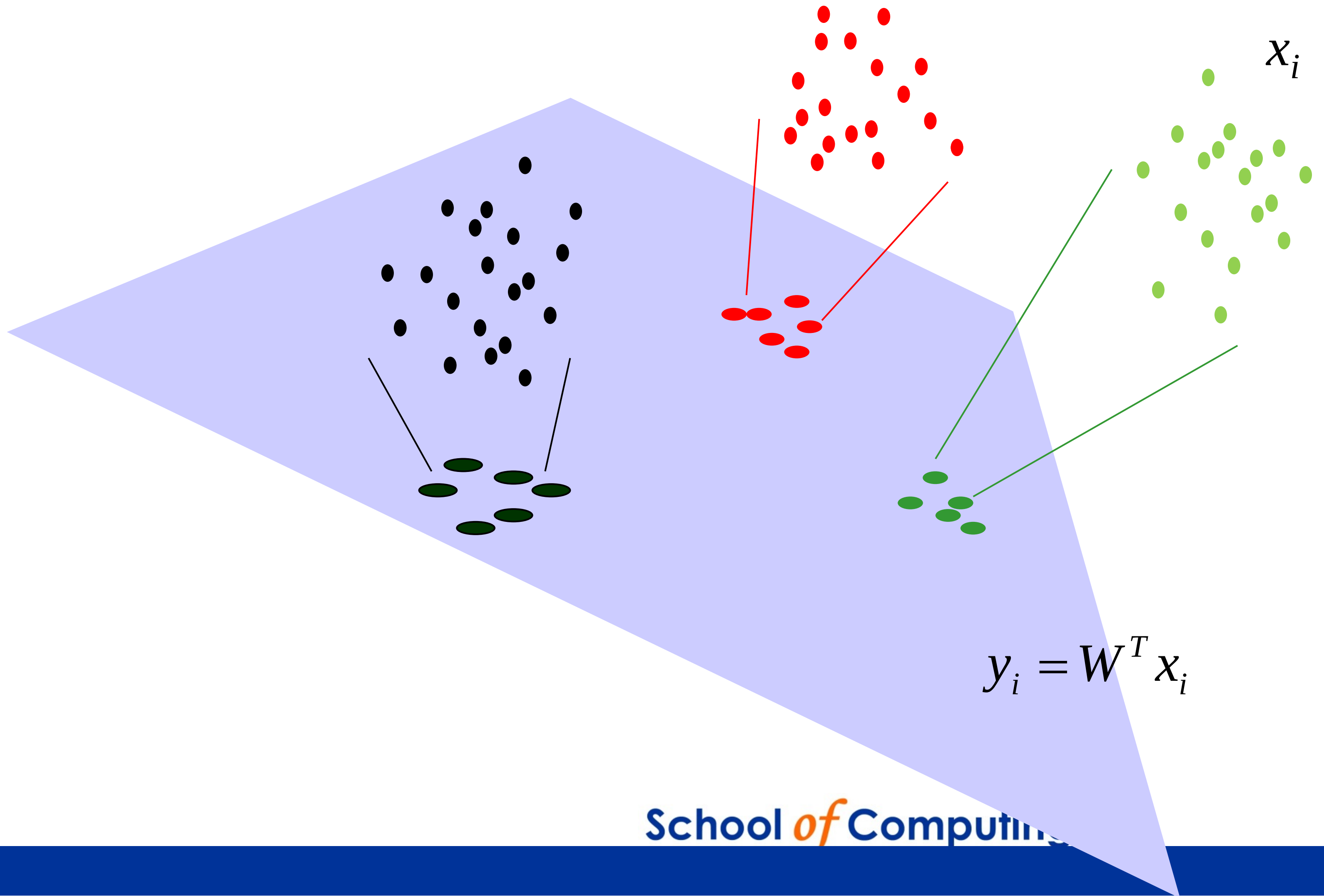
Fisher Linear Discriminant

New and Old

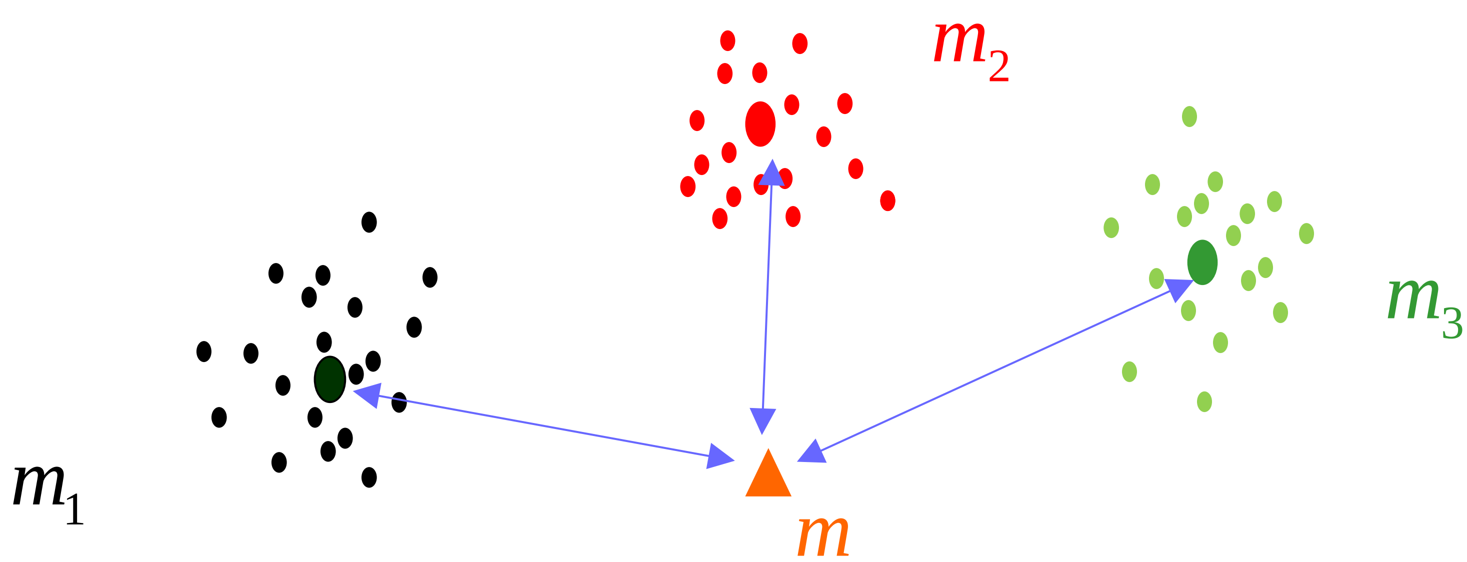
Fisher Linear Discriminant

Goal: find plane (subspace) that "shrinks" each class while separating all classes.



Fisher Linear Discriminant

Class separation: between-class scatter matrix

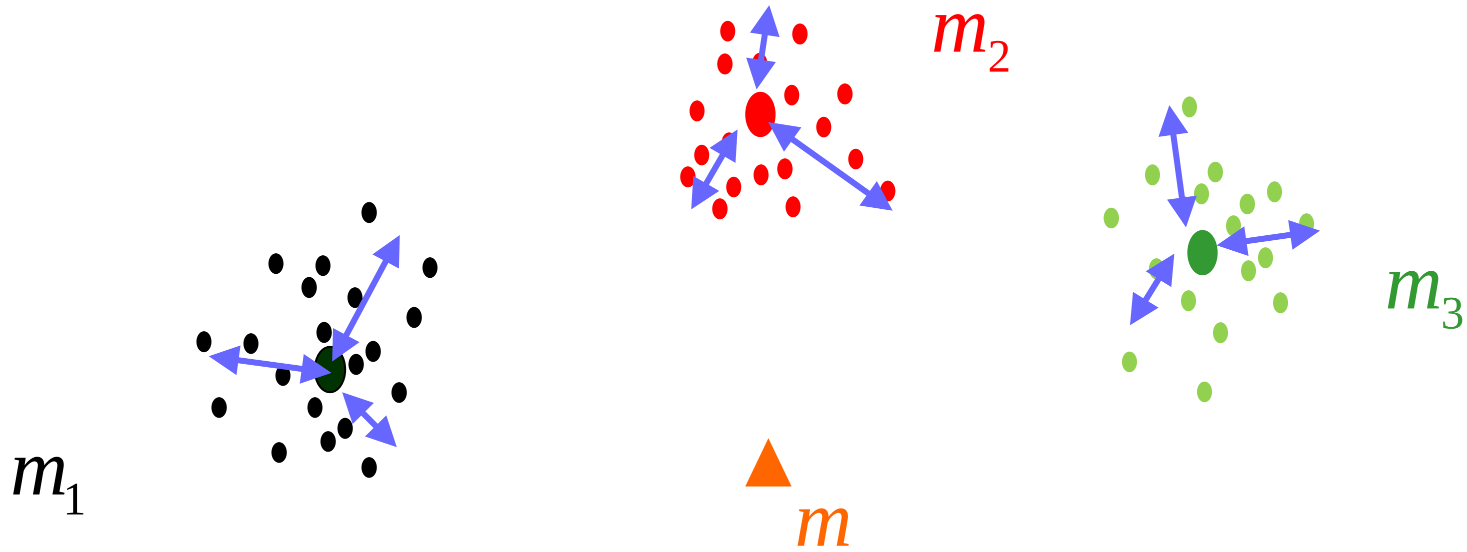


n_i : # samples in class ω_i
m_i : mean of class ω_i
m: global mean

$$S_B = \sum_{i=1}^{C} n_i (m_i - m)(m_i - m)^{\mathrm{T}}$$

Fisher Linear Discriminant or LDA

Class spread: within-class scatter matrix



$$S_W = \sum_{i=1}^{C} \sum_{x \in \omega_i} (x - m_i)(x - m_i)^{\mathrm{T}}$$

Scatter matrices after projection

S_B

S_W

S_T

$x \rightarrow y = W^T x$

$S_B' = W^T S_B W$

$S_W' = W^T S_W W$

$S_T' = W^T S_T W$

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

Total scatter matrix

Fisher's Criterion

$$J_F = \frac{|\mathbf{W}^{\mathrm{T}}\mathbf{S}_B\mathbf{W}|}{|\mathbf{W}^{\mathrm{T}}\mathbf{S}_W\mathbf{W}|}$$

- Numerator is the between-class scatter matrix after projection (i.e. y_{i})
- Denominator is the within-class scatter matrix after projection (i.e. y_{i})
- **W** is the projection matrix.
- Goal: find **W** to maximize Fisher's Criterion
- Recall PCA:

$$J = \frac{|\mathbf{W}^{\mathrm{T}}\mathbf{S}_T\mathbf{W}|}{|\mathbf{W}^{\mathrm{T}}\mathbf{W}|}$$

LDA

- It can be shown that the solution is:

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

- This is the Generalized Eigenvalue Problem
- If Sw is invertible, then

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

- which is the regular eigenvalue problem

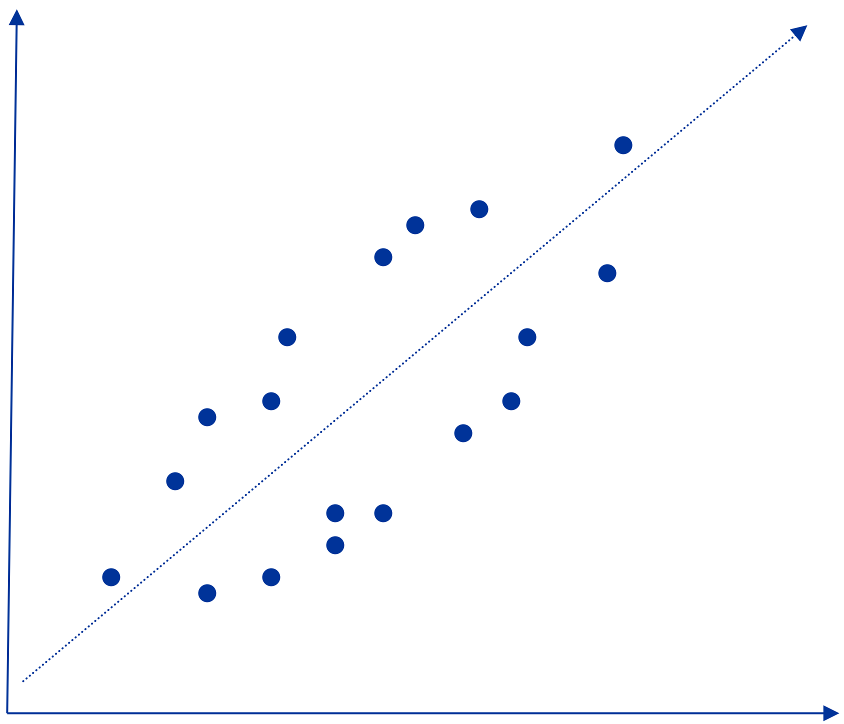
LDA

- There will be up to *C*-1 eigenvectors.
- But these won't be mutually orthogonal
  - Because $\mathbf{S}_W^{-1}\mathbf{S}_B$ may not be symmetric.

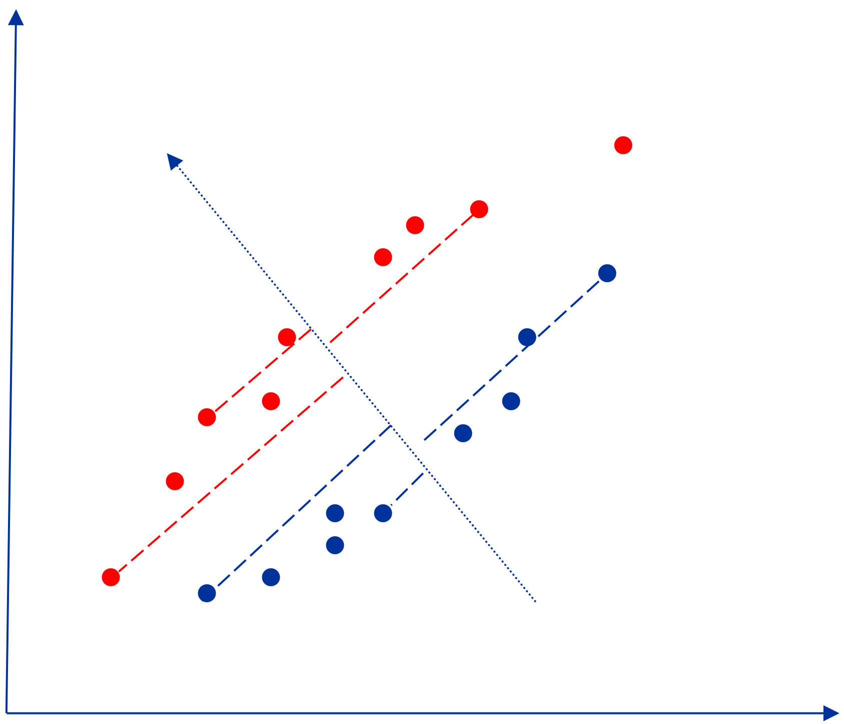
- The eigenvectors form a subspace that maximizes the Fisher Criterion.
- Put them into the matrix **W**, then compute the feature as:

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

PCA & LDA



PCA: max | S_T' |



LDA: max | S_B'| / |S_W' |

Questions about the FLD

- How to compare $\boldsymbol{J_F}$?
  - What is the best possible $\boldsymbol{J_F}$?
  - Answer: $+\infty$
- Does maximizing $\boldsymbol{J_F}$ perfectly separate classes?
  - If not, under what conditions can classes be perfectly separated?
  - Answer: require $D >= N$-1 and linearly independence
- Can these conditions be satisfied in practice?
  - Usually, e.g. face images, but not digit recognition
- How are class patterns (vectors) distributed in the subspace W ?
  - As vertices of a regular simplex

Whitened FLD (WFLD) [Zhang&Sim: [1], also CVPR 06, PAMI 07]

- Whiten the data with **P**:

Whitening matrix

$$\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_w$$
$$\mathbf{S}_T = \mathbf{U}\mathbf{D}^2\mathbf{U}^T$$
$$\mathbf{P} = \mathbf{U}\mathbf{D}^{-1}$$
$$\tilde{\mathbf{x}}_i = \mathbf{P}^T\mathbf{x}_i$$
$$\Rightarrow \tilde{\mathbf{S}}_T = \mathbf{P}^T\mathbf{S}_T\mathbf{P} = \mathbf{I}$$
$$\mathbf{P}^T(\mathbf{S}_B + \mathbf{S}_w)\mathbf{P} = \mathbf{I}$$
$$\tilde{\mathbf{S}}_B + \tilde{\mathbf{S}}_w = \mathbf{I}$$
$$\tilde{\mathbf{S}}_w = \mathbf{I} - \tilde{\mathbf{S}}_B$$

- Let $\mathbf{e}$ be eigenvector of $\tilde{\mathbf{S}}_B$ with eigenvalue λ_B

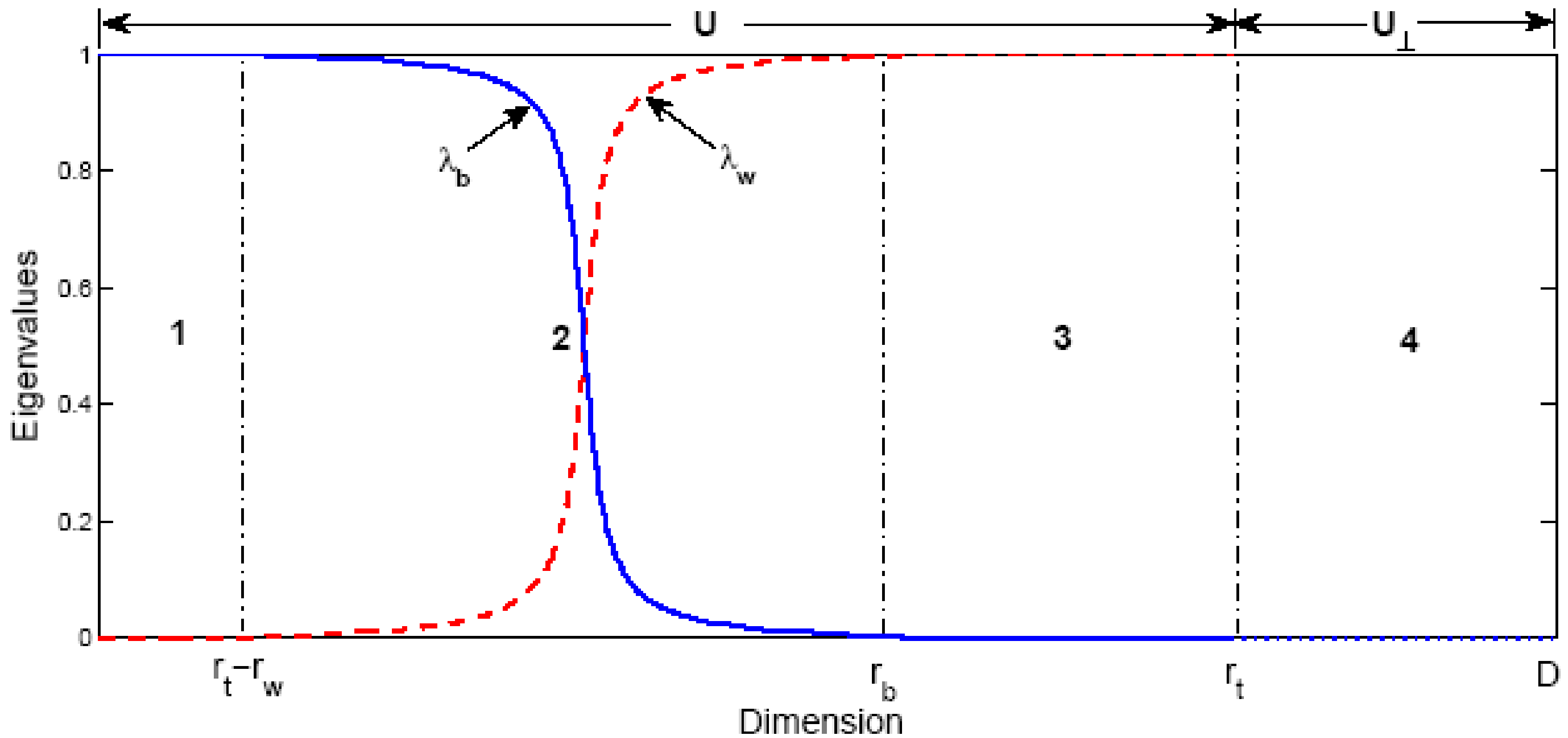
$$\tilde{\mathbf{S}}_W\,\mathbf{e} = (\mathbf{I} - \tilde{\mathbf{S}}_B)\,\mathbf{e}$$
$$= \mathbf{e} - \lambda_B\,\mathbf{e} \quad = (1 - \lambda_B)\,\mathbf{e}$$
$$= \lambda_W\,\mathbf{e}$$

So $\tilde{\mathbf{S}}_B$ and $\tilde{\mathbf{S}}_W$ share the same eigenvectors, and their eigenvalues sum to 1.

Moreover:

$$\frac{\lambda_B}{\lambda_W} = J_F$$

Discriminant Subspace Analysis



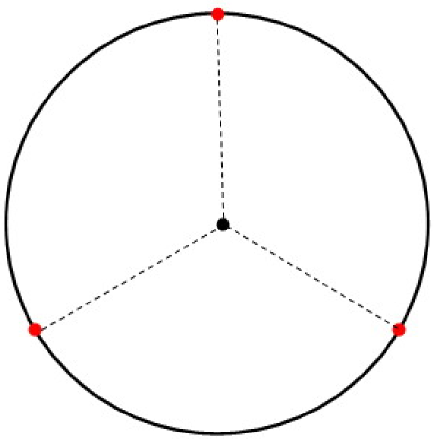
1. Identity Space: $J_F = +\infty$
2. Mixed Space: $0 < J_F < +\infty$
3. Variation Space: $J_F = 0$
4. Null Space: no discrim. info

Identity Space

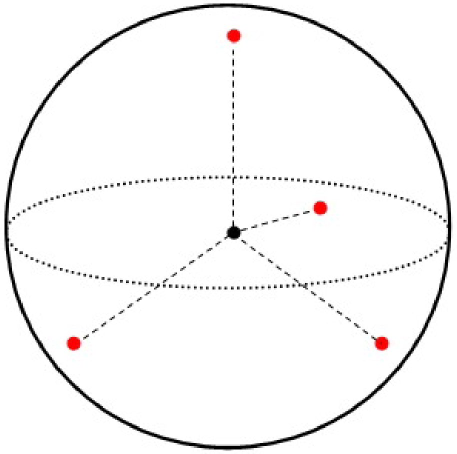
- Definition
  - Subspace 1: eigenvectors with $\lambda_b = 1, \quad \lambda_w = 0$
  - The most discriminant subspace $\mathbf{y}_1 = \mathbf{V}_1^T \mathbf{P}^T \mathbf{x}$
- Properties (Theorem 1 in paper)
  - No within-class variation
    - Samples from one class project onto one point (a.k.a Identity Vector).
  - Classes are cleanly separated.
    - Samples from one class will not project onto another class as long as they have different class means.

Identity Space

- Geometric Structure (Theorem 3 in paper)
  - *C* Identity Vectors are equidistant and distributed on a *C-1* dimensional sphere.
  - The distance between classes has been maximized.



Three Identity Vectors distributed on a 2D circle



Four Identity Vectors distributed on a 3D sphere

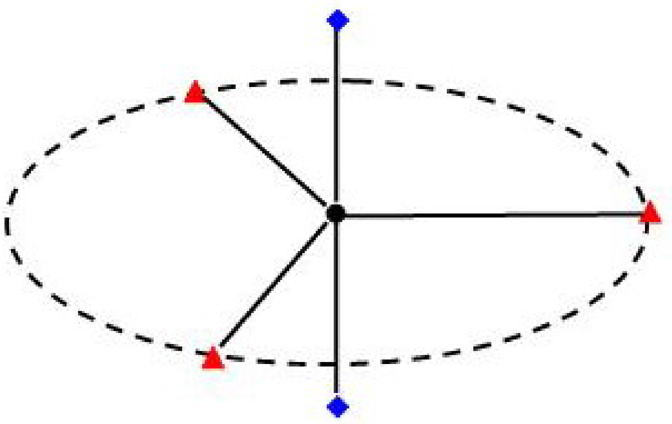
Variation Space

- Definition
  - Subspace 3: eigenvectors with $\lambda_b = 0, \quad \lambda_w = 1$
  - The least discriminant subspace $\mathbf{y}_3 = \mathbf{V}_3^T \mathbf{P}^T \mathbf{x}$

- Property (Theorems 4 & 6 in paper)
  - All class means project to zero.
  - The within-class variation of each class lie in orthogonal subspaces.

Variation Space

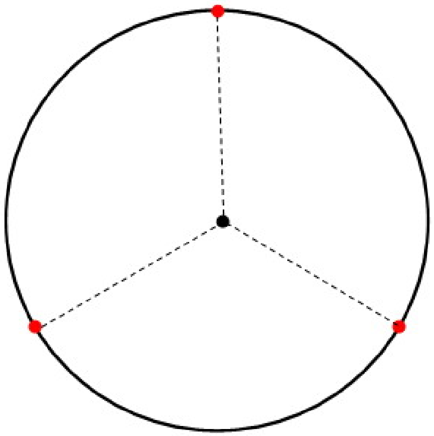
- Geometric Structure
  - After projecting onto Variation Space, any two classes are orthogonal to each other.
  - For each class, samples are equally distributed over a hypersphere (for 1D case, it is a straight line.).



Class One has two samples lying on a
vertical line, and both are orthogonal to Class Two,
which has three samples evenly distributed on a circle.

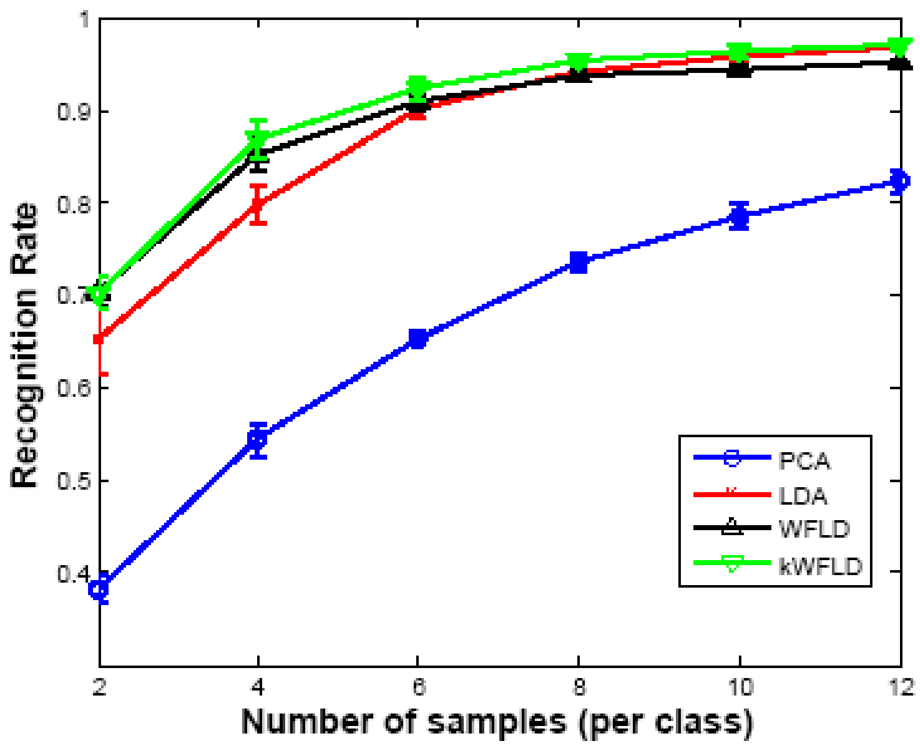
Discriminability of Identity Space

- Project query point onto Identity Space.
- Find nearest class mean
  - Euclidean or cosine distance

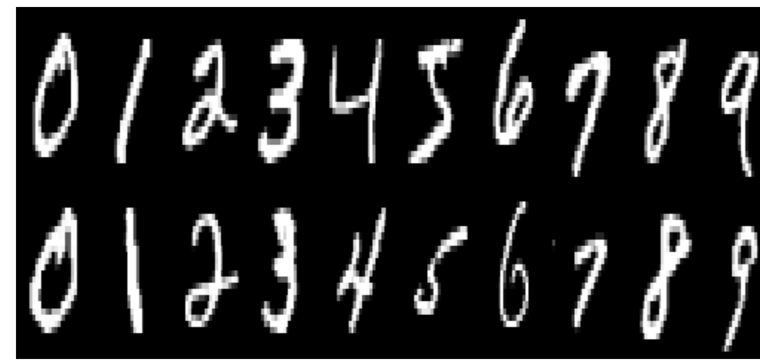


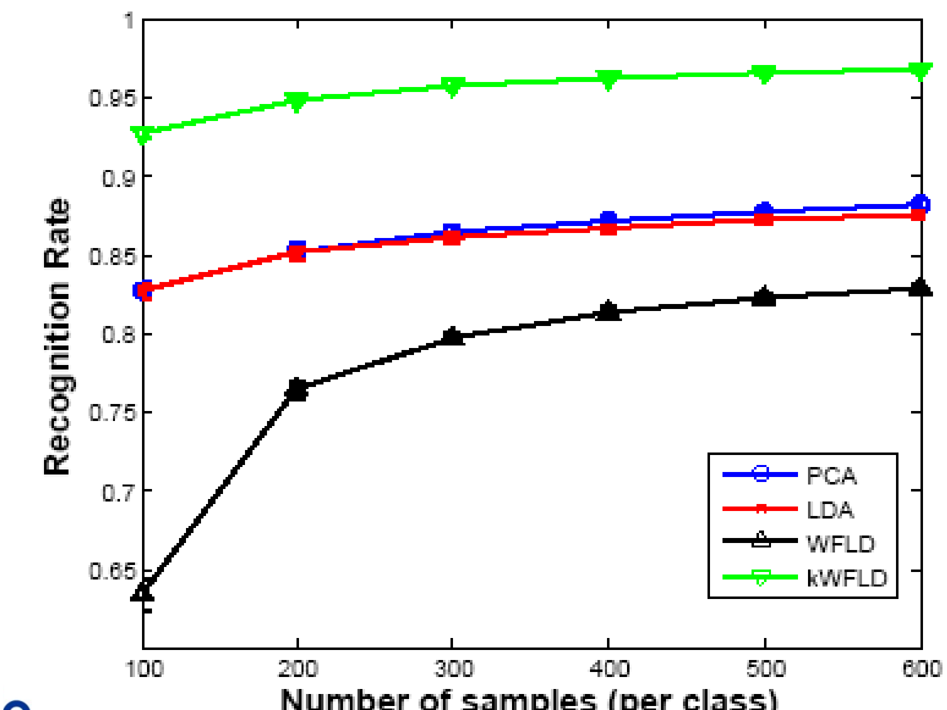
Discriminability of Identity Space

BANCA

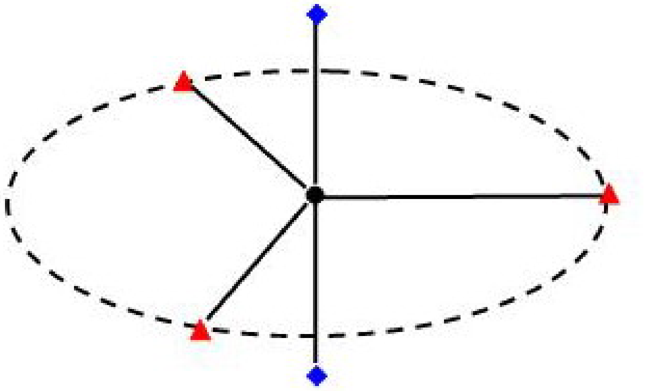


MNIST





Discriminability of Variation Space



- No means for classification!
  - Not according to the Fisher Criterion

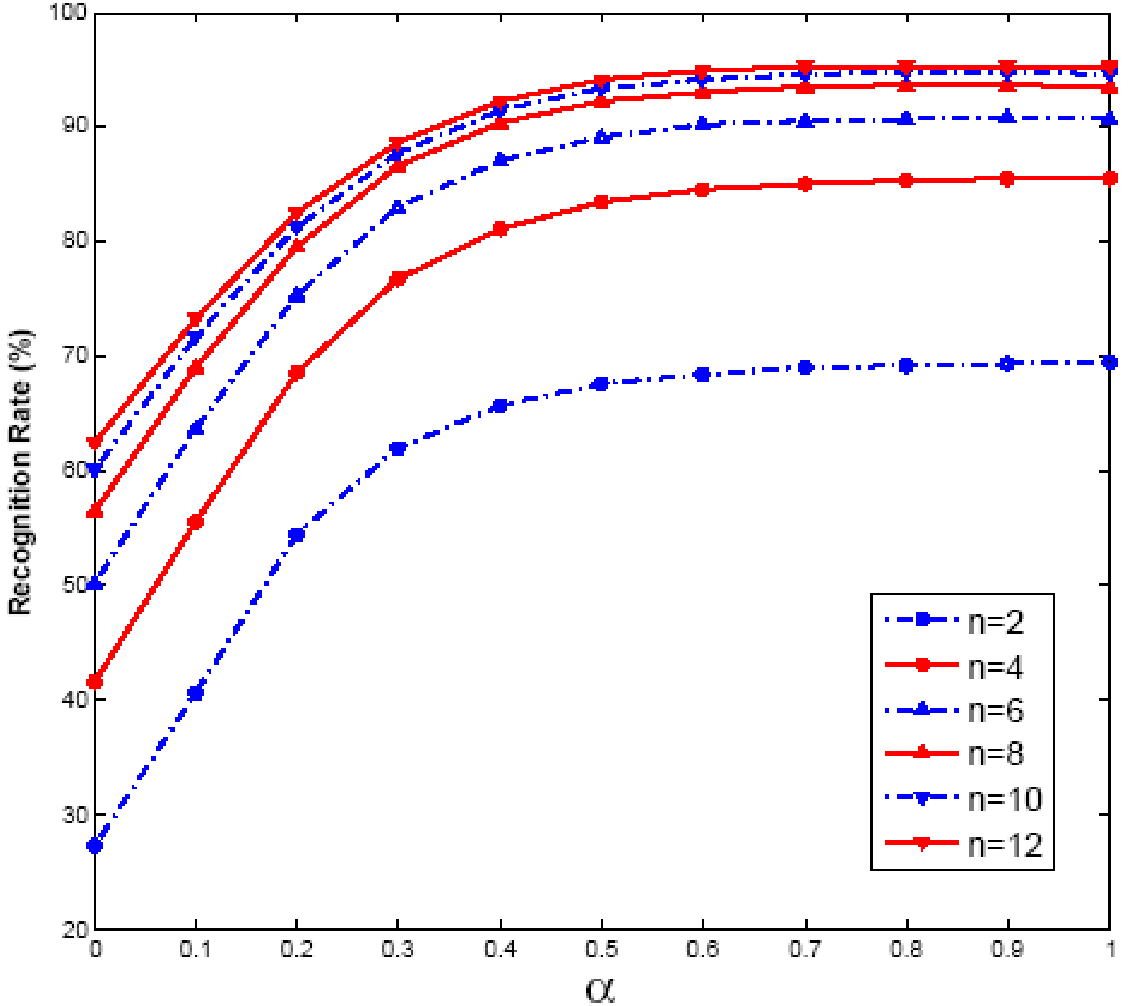
$$J_F = \frac{|\mathbf{W}^{\mathrm{T}}\mathbf{S}_B\mathbf{W}|}{|\mathbf{W}^{\mathrm{T}}\mathbf{S}_W\mathbf{W}|}$$

  - But in WFLD, within-class variations of all classes occupy orthogonal subspaces.
    - We can use this for classification, even if class means are identical!

Discriminability of Variation Space

- Distance is a weighted sum of d_I and d_V.
  - d_I: Euclidean distance to class mean in Identity Space
  - d_V: distance to subspace in Variation Space

$$d = \alpha d_I + (1-\alpha) d_V$$



Other subspace methods

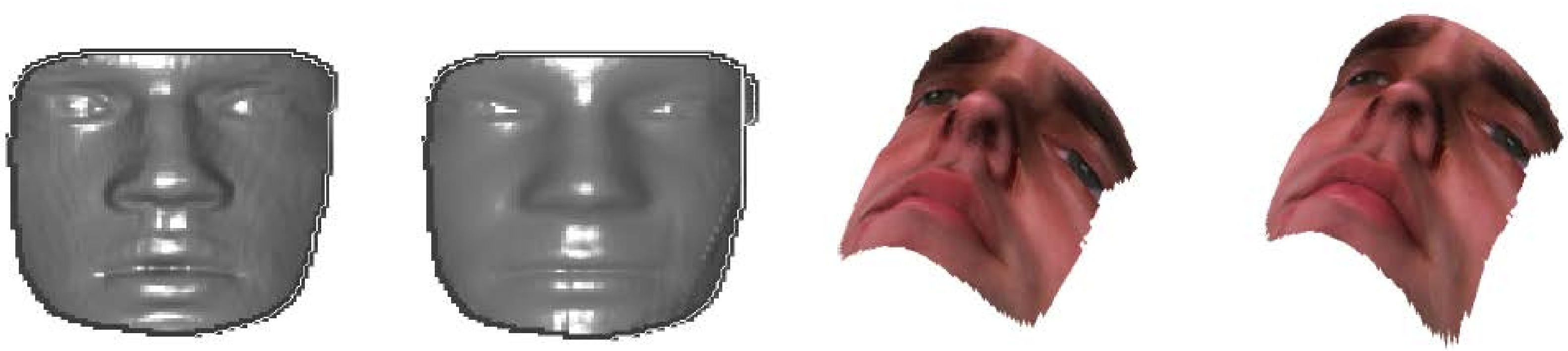
- Independent Components Analysis (ICA)

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

  - Components of **y** are statistically indep., not just uncorrelated.
  - **W** is not orthogonal
  - Algorithm makes lots of approximations, so unclear whether resulting components are indeed indep.
  - Applied in Blind Source Separation problem
    - Cocktail-party problem: separate who said what

Other subspace methods

- Canonical Correlation Analysis (CCA)
  - Give vectors **x**, **y** (possibly different dimensions)
  - Find **w** such that $Y' = w^T y$, $X' = w^T x$ are most correlated.
- Use this to model Y' = RX'
  - Applied in estimating depth map from color image [3]

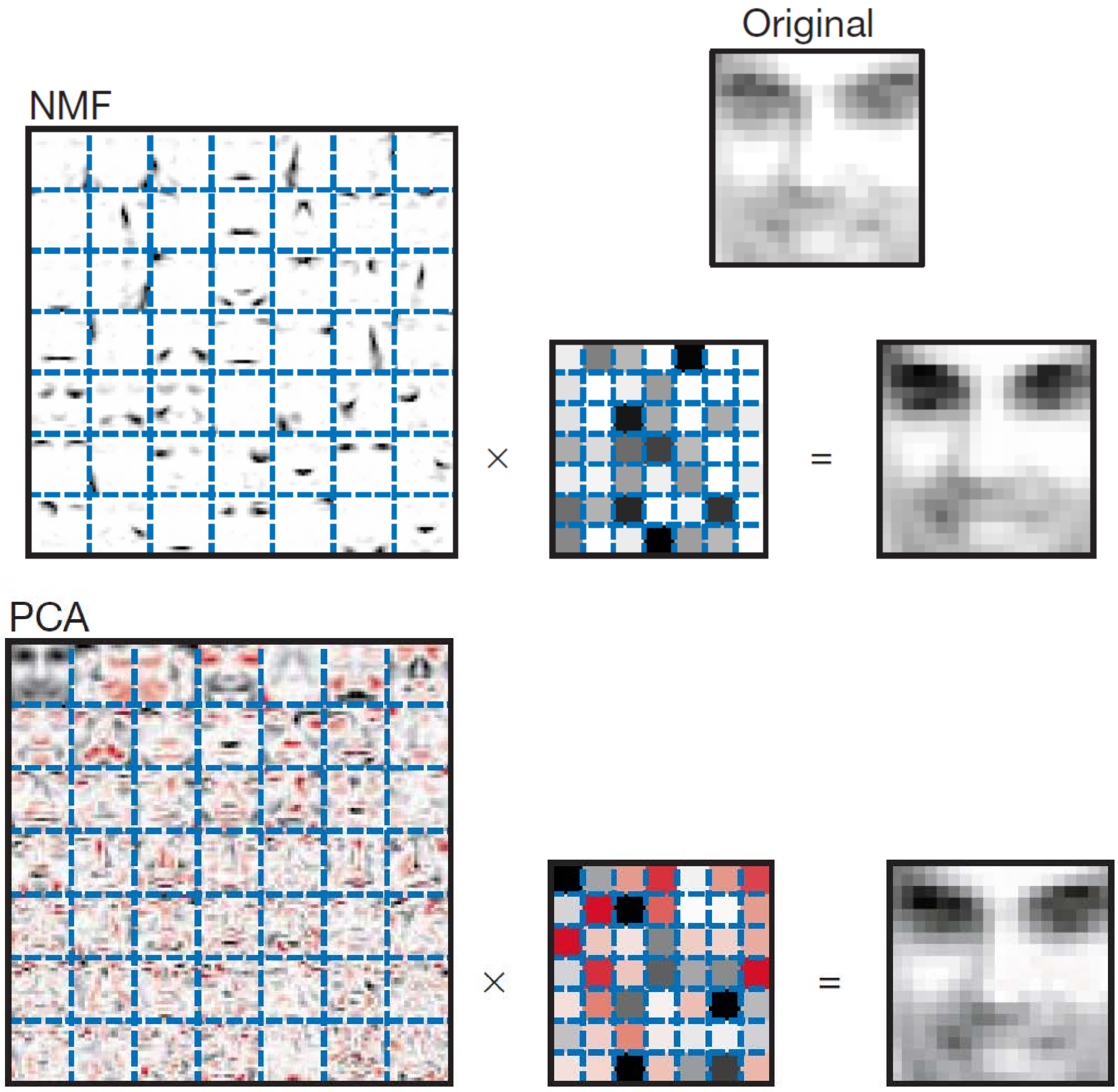


Original surface   Reconstructed surface   Original face patch   Reconstructed face patch

Other subspace methods

- Non-negative Matrix Factorization (NMF) [2]
- In PCA, given data matrix X, find X = WY
  - Entries in W and Y can be negative
- In NMF, find X = WH
  - Entries in W and H must be non-negative
  - Iterative algorithm to find W
  - W is not orthogonal

NMF



References

[1] *Identity and Variation Spaces: Revisiting the Fisher Linear Discriminant.* Sheng Zhang, Terence Sim, Mei-Chen Yeh. *Subspace 2009, at ICCV 2009.*

[2] *Learning the parts of objects by non-negative matrix factorization*. Daniel D. Lee & H. Sebastian Seung. *Nature, vol. 401, Oct. 1999*

[3] *Estimation of Face Depth Maps from Color Textures using Canonical Correlation Analysis*. Michael Reiter, Ren´e Donner, Georg Langs, and Horst Bischof. *Computer Vision Winter Workshop 2006.*

[4] *An Introduction to Independent Component Analysis and Blind Source Separation.* Lucas C. Parra. Tech. Rep., Sarnoff Corporation, 1999.