

---

# INF 5300

## Linear feature transforms

Anne Solberg (anne@ifi.uio.no)

Today:

- Feature transformation through principal component analysis
- Fisher's linear discriminant function

21.3.12

INF 5300

1

---

## Linear feature transforms

- Feature extraction can be stated as
  - Given a feature space  $x_i \in \mathbb{R}_n$  find an optimal mapping  $y = f(x) : \mathbb{R}_n \rightarrow \mathbb{R}_m$  with  $m < n$ .
  - An optimal mapping in classification :the transformed feature vector  $y$  yield the same classification rate as  $x$ .
- The optimal mapping may be a non-linear function
  - Difficult to generate/optimize non-linear transforms
  - Feature extraction is therefore usually limited to linear transforms  $y = A^T x$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

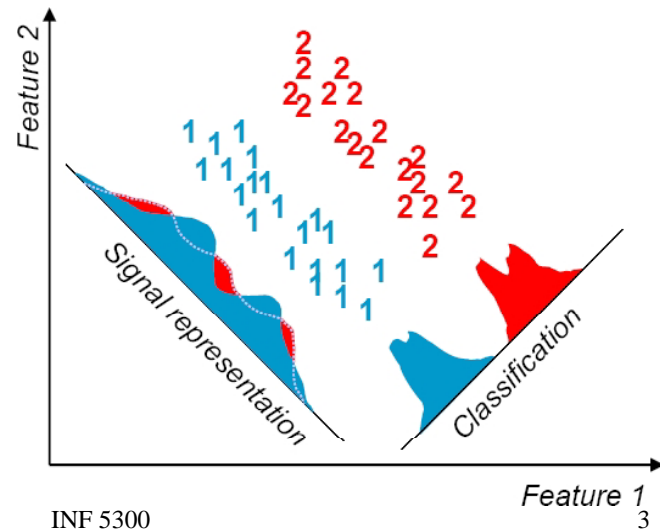
INF 5300

2

# Signal representation vs classification

---

- Principal components analysis (PCA)
  - signal representation, unsupervised
  - Minimize the mean square representation error
- Linear discriminant analysis (LDA)
  - classification, supervised
  - Maximize the distance between the classes



# Correlation matrix vs. covariance matrix

---

- $\Sigma_x$  is the covariance matrix of  $x$

$$\Sigma_x = E[(x - \mu)(x - \mu)^T]$$

- $R_x$  is the correlation matrix of  $x$

$$R_x = E[(x)(x)^T]$$

- $R_x = \Sigma_x$  if  $\mu_x = 0$ .

# Principal component or Karhunen-Loeve transform

---

- Let  $x$  be a feature vector.
- Features are often correlated, which might lead to redundancies.
- We now derive a transform which yields **uncorrelated** features.
- We seek a linear transform  $y=A^T x$ , and the  $y_i$ s should be uncorrelated.
- The  $y_i$ s are uncorrelated if  $E[y(i)y(j)^T]=0$ ,  $i \neq j$ .
- If we can express the information in  $x$  using uncorrelated features, we might need **fewer** coefficients.

---

## Principal component transform

---

- The correlation of  $Y$  is described by the correlation matrix  
 $R_y = E[yy^T] = E[A^T x x^T A] = A^T R_x A$        $R_x$  is the correlation matrix of  $X$   
 $R_x$  is symmetric, thus all eigenvectors are orthogonal.
  - We seek uncorrelated components of  $Y$ , thus  $R_y$  should be diagonal.
- From linear algebra:
- $R_y$  will be diagonal if  $A$  is formed by the orthogonal eigenvectors  $a_i$ ,  $i=0, \dots, N-1$  of  $R_x$ :  $R_y = A^T R_x A = \Lambda$ , where  $\Lambda$  is diagonal with the eigenvalues of  $R_x$ ,  $\lambda_i$ , on the diagonal.
  - We find  $A$  by solving the equation  $A^T R_x A = \Lambda$  (using Singular Value Decomposition (SVD)).
  - $A$  is formed by computing the eigenvectors of  $R_x$ . Each eigenvector will be a column of  $A$ .

# Mean square error approximation

---

- $x$  can be expressed as a combination of all  $N$  basis vectors:

$$x = \sum_{i=0}^{N-1} y(i)a_i, \text{ where } y(i) = a_i^T x$$

- An approximation to  $x$  is found by using only  $m$  of the basis vectors:

$$\hat{x} = \sum_{i=0}^{m-1} y(i)a_i \quad \text{a projection into the } m\text{-dimensional subspace spanned by } m \text{ eigenvectors}$$

- The PC-transform is based on minimizing the mean square error associated with this approximation.
- The mean square error associated with this approximation is

$$E[\|x - \hat{x}\|^2] = E\left[\left\|\sum_{i=m}^{N-1} y(i)a_i\right\|^2\right] = E\left[\sum_i \sum_j (y(i)a_i^T)(y(j)a_j)\right] = \sum_{i=m}^{N-1} E[y^2(i)] = \sum_{i=m}^{N-1} a_i^T E[xx^T] a_i$$

INF 5300

7

- Furthermore, we can find that

$$E[\|x - \hat{x}\|^2] = \sum_{i=m}^{N-1} a_i^T \lambda_i a_i = \sum_{i=m}^{N-1} \lambda_i$$

- The mean square error is thus

$$E[\|x - \hat{x}\|^2] = \sum_{i=1}^{N-1} \lambda_i - \sum_{i=1}^m \lambda_i = \sum_{i=m}^{N-1} \lambda_i$$

- The error is minimized if we select the eigenvectors corresponding to the  $m$  largest eigenvalues of the correlation matrix  $R_x$ .
- The transformed vector  $y$  is called the principal components of  $x$ . The transform is called the principal component transform or Karhunen-Loeve-transform.

## Principal component of the covariance matrix

---

- Alternatively, we can find the principal components of the covariance matrix  $\Sigma_x$ .
- If we have software for computing principal components of  $R_x$ , we can compute principal components from  $\Sigma_x$  by first setting  $z=x-\mu_x$  and compute  $PC(z)$ .
- The principal component transform is not scale invariant, because the eigenvectors are not invariant. Often, normalization to data with zero mean and unit variance is done prior to applying the PC-transform.

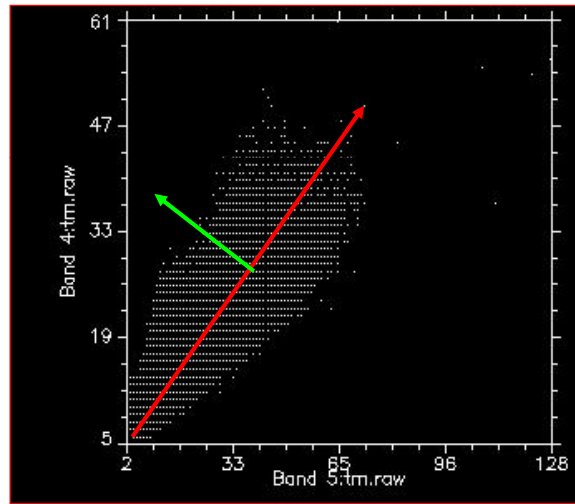
## Principal components and total variance

---

- Assume that  $E[x]=0$ .
- Let  $y=PC(x)$ .
- From  $R_y$  we know that the variance of component  $y_j$  is  $\lambda_j$ .
- The eigenvalues  $\lambda_j$  of the correlation matrix  $R_x$  is thus equal to the variance of the transformed features.
- By selecting the  $m$  eigenvectors with the largest eigenvalues, we select the  $m$  dimensions with the largest variance.
- The first principal component will be along the direction of the input space which has largest variance.

# Geometrical interpretation of principal components

- The eigenvector corresponding to the largest eigenvalue is the direction in n-dimensional space with highest variance. →
- The next principal component is orthogonal to the first, and along the direction with the second largest variance. →

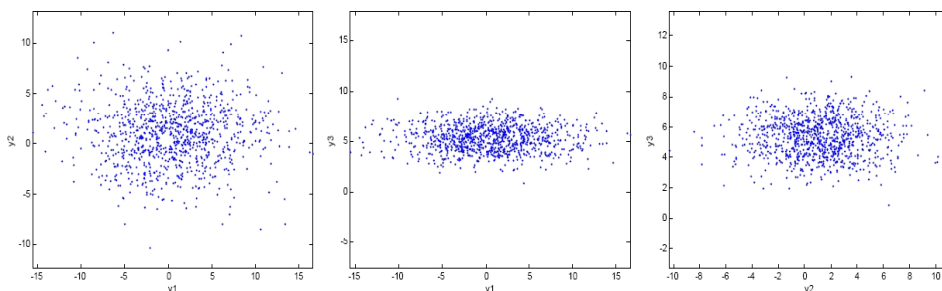
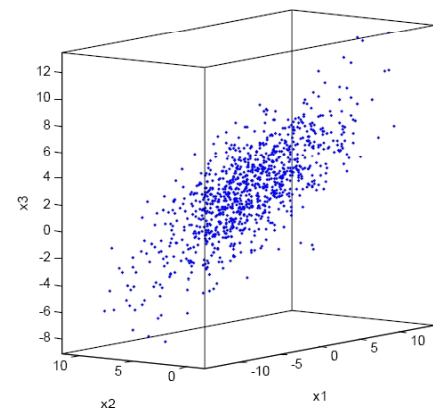


Note that the direction with the highest variance is NOT related to separability between classes.

## PCA example

3d Gaussian with parameters

$$\mu = [0 \ 5 \ 2]^T, \quad \Sigma = \begin{bmatrix} 25 & -1 & 7 \\ -1 & 4 & -4 \\ 7 & -4 & 10 \end{bmatrix}$$



# Principal component images

---

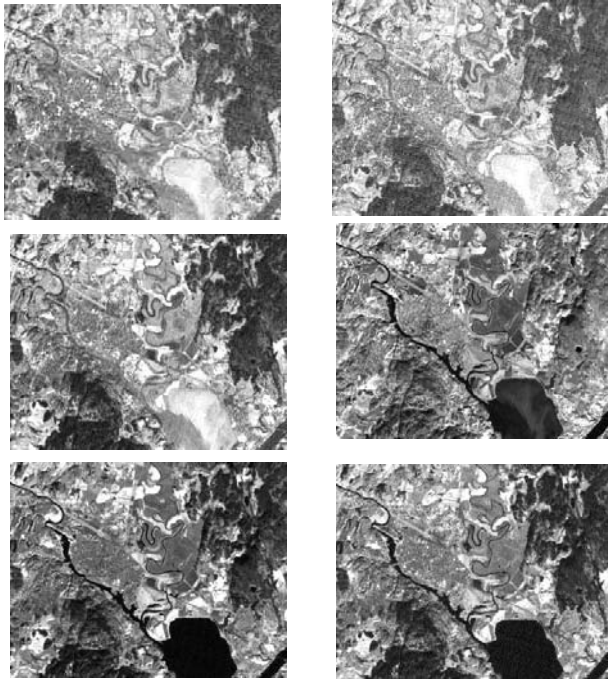
- For an image with  $n$  bands, we can compute the principal component transform of the entire image  $X$ .
- $Y=PC(X)$  will then be a new image with  $n$  bands, but most of the variance is in the bands with the lowest index (corresponding to the largest eigenvalues).

# Principal component images

---

- For an image with  $n$  bands, we can compute the principal component transform of the entire image  $X$ .
- $Y=PC(X)$  will then be a new image with  $n$  bands, but most of the variance is in the bands with the lowest index (corresponding to the largest eigenvalues).

# PCA example – original image



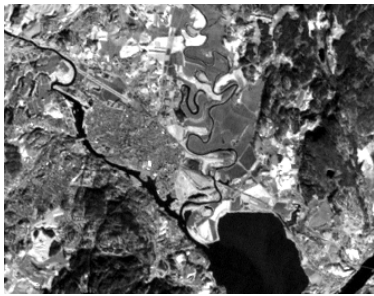
- Satellite image from Kjeller
- 6 spectral bands with different wavelengths

1	Blue	0.45-0.52	Max. penetration of water
2	Green	0.52-0.60	Vegetation and chlorophyll
3	Red	0.63-0.69	Vegetation type
4	Near-IR	0.76-0.90	Biomass
5	Mid-IR	1.55-1.75	Moisture/water content in vegetation/soil
7	Mid-IR	2.08-2.35	Minerals

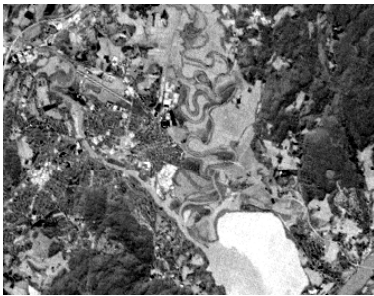
INF 5300

15

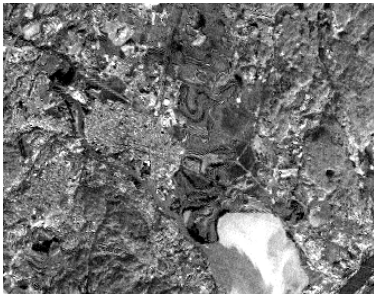
# Principal component images



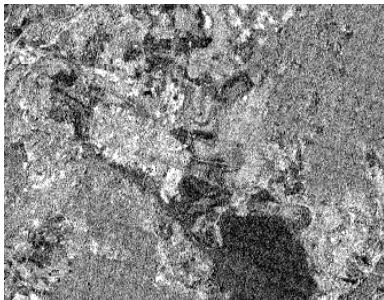
Principal component 1



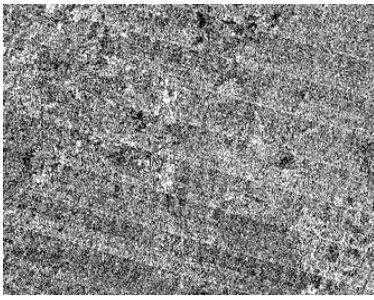
Principal component 2



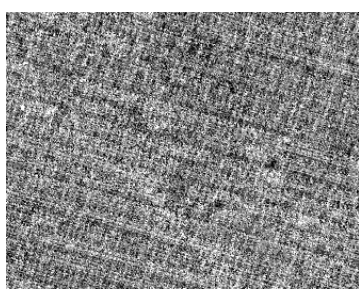
Principal component 3



Principal component 4



Principal component 5



Principal component 6

INF 5300

16

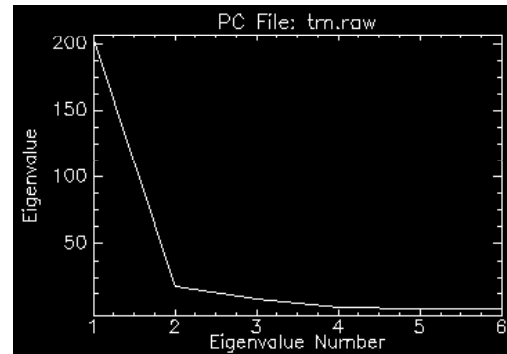


## Example: inspecting the eigenvalues

---

The representation error we get with  $m$  of the  $N$  PCA-components is given as

$$E\left[\|x - \hat{x}\|^2\right] = \sum_{i=1}^{N-1} \lambda_i - \sum_{i=1}^m \lambda_i = \sum_{i=m}^{N-1} \lambda_i$$



Plotting  $\lambda_i$

will give indications on how many features are needed for representation

## PC and compression

---

- PC-transform is optimal transform with respect to preserving the energy in the original image.
- For compression purposes, PC-transform is theoretically optimal with respect to maximizing the entropy (from information theory). Entropy is related to randomness and thus to variance.
- The basis vectors are the eigenvectors and vary from image to image. For transmission, both the transform coefficients and the eigenvectors must be transmitted.
- PC-transform can be reasonably well approximated by the Cosinus-transform or Sinus-transform. These use constant basis vectors and are better suited for transmission, since only the coefficients must be transmitted (or stored).

# PC vs. Fisher's linear discriminant transform

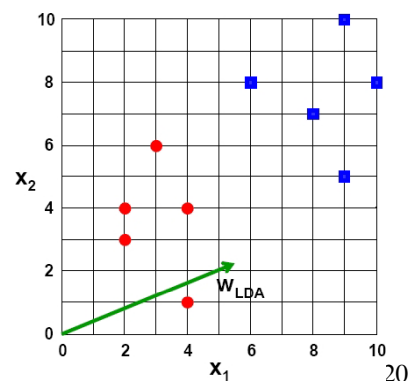
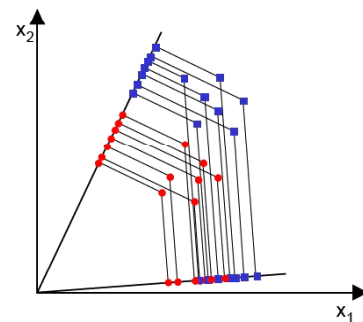
- The principal component transform has no information about the classes in the data.
- The PC-projection might not be helpful to improve class separability.
- From an input vector  $x$  with dimension  $m$ , PC-transform gives us a projection  $y$  with dimensions  $1, \dots, m$  (depending on how many eigenvalues we include).
- A projection with Fishers linear discriminant gives us  $y$  with dimensions  $1, \dots, K-1$ , where  $K$  is the number of classes.
- Fishers linear discriminant find the projection that maximizes the ratio of between-class to within-class scatter.

INF 5300

19

## Fisher's Linear Discriminant

- Goal:
  - Reduce dimension while preserving class discriminatory information
- Strategy (2 classes):
  - We have a set of samples  $x = \{x_1, x_2, \dots, x_n\}$  where  $n_1$  belong to class  $\omega_1$  and the rest  $n_2$  to class  $\omega_2$ . Obtain a scalar value by projecting  $x$  onto a line  $y: y = w^T x$
  - **Challenge: find  $w$  that maximizes the separability of the classes**



INF 5300

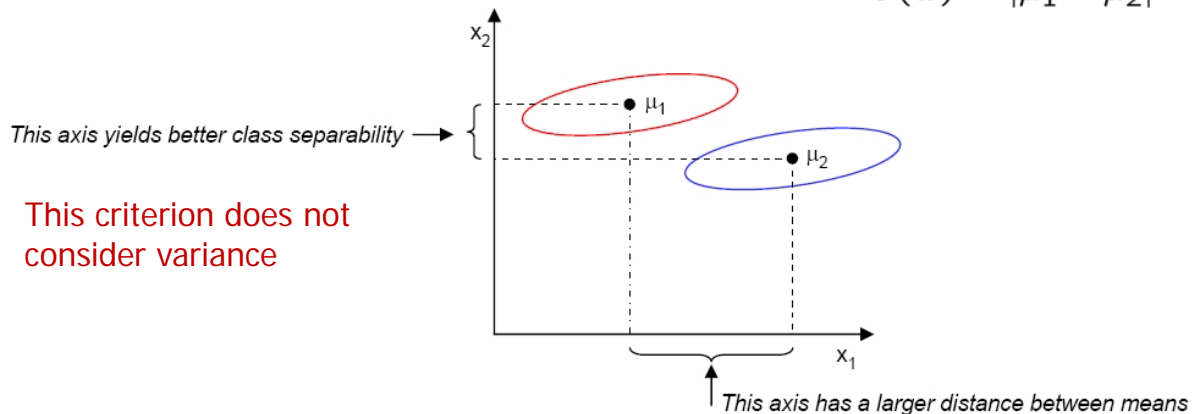
20

## A simple criterion function: 2 features and 2 classes

---

- To find a good projection vector, we need to define a measure of separation between the projections. This will be the criterion function  $J(w)$
- The mean vector of each class in the spaces spanned by  $x$  and  $y$  are
 
$$\mu_i = \frac{1}{n_i} \sum_{x \in \omega_i} x$$

$$\tilde{\mu}_i = \frac{1}{n_i} \sum_{y \in \omega_i} y = \frac{1}{n_i} \sum_{x \in \omega_i} w^T x = w^T \mu_i$$
- A naive choice would be projected mean difference,  $J(w) = |\tilde{\mu}_1 - \tilde{\mu}_2|$



## A criterion function including variance: 2 features and 2 classes

---

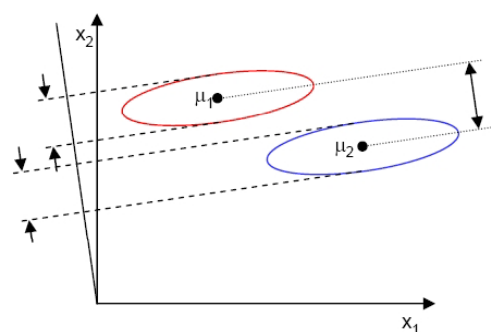
- Fisher's solution: maximize a function that represents the difference between the means, scaled by a measure of the within class scatter
- Define classwise scatter (similar to variance)

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2$$

- $\tilde{s}_1^2 + \tilde{s}_2^2$  is *within class scatter*
- Fisher's criterion is then

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

- We look for a projection where examples from the same class are close to each other, while at the same time projected mean values are as far apart as possible.



# Introducing general scatter matrices

- In M-dimensional space, let us now consider matrices describing the variance:
  - Variance INSIDE each class
  - Variance BETWEEN the classes (how well separated are the classes)
  - The total variance in the data set is constant and independent of any class labels

F1 15.2.06

INF 5300

23

## Scatter matrices – M classes

- Within-class scatter matrix:

$$S_w = \sum_{i=1}^M P(\omega_i) S_i$$

$$S_i = E[(x - \mu_i)(x - \mu_i)^T]$$

Variance within each class

- Between-class scatter matrix:

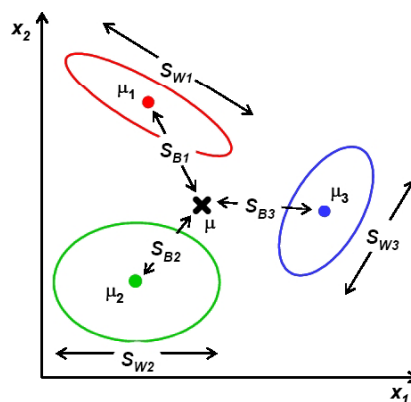
$$S_b = \sum_{i=1}^M P(\omega_i) (\mu_i - \mu_0)(\mu_i - \mu_0)^T$$

$$\mu_0 = \sum_{i=1}^M \mu_i \quad \text{Distance between the classes}$$

- Mixture or total scatter matrix:

$$S_m = E[(x - \mu_0)(x - \mu_0)^T]$$

Variance of feature with  
respect to the global mean



INF 5300

24

# Some matrix algebra

---

- $M$  is a symmetric  $l \times l$  matrix
- $|M|$  is the determinant of  $M$
- $M$  can be expressed in terms of eigenvalues  $\lambda_i$  and eigenvectors  $v_i$ .
- $|M|$  is nonzero only if the matrix has full rank (all eigenvalues are nonzero)
- $\text{trace}(M)$  is equal to the sum of eigenvalues of  $M$ .

## Relation between eigenvalues and the scatter of a matrix

---

- The eigenvalues associated with an eigenvector tells how strong the contribution along this direction is.
- A scalar measure of the scatter matrix  $M$  is its determinant (the product of the eigenvalues). This gives us ONE measure of the scatter in the matrix.
- If  $M$  is a covariance matrix,  $|M|$  is a measure of the  $l$ -dimensional hypervolume of the data. If the data lies in a subspace,  $|M|$  will be zero.
- For a covariance matrix  $M$ ,  $\text{trace}(M)$  is the sum of the eigenvalues and thus a measure of the spread or scatter in  $A$ .

- The total scatter is

$$S_m = S_w + S_b$$

- Consider the criterion function

$$J_1 = \frac{\text{trace}\{S_m\}}{\text{trace}\{S_w\}} \quad \leftarrow \quad \text{Sum of the diagonal elements in } S_m$$

⇕

sum of variance around global mean  
sum of variance inside class

- $J_1$  will be large when the variance among the global mean is large compared to the within-class variance.

## Better scatter criteria functions – J2 and J3

$$J_2 = \frac{|S_m|}{|S_w|} = |S_w^{-1} S_m|$$

$$J_3 = \text{trace}\{S_w^{-1} S_b\}$$

- $J_2$  and  $J_3$  are invariant to linear transformations.

# Fisher's linear discriminant

---

- Fisher's linear discriminant is a transform that uses the information in the training data set to find a linear combination that best separates the classes.
- It is based on the criterion  $J_3$ :

$$J_3 = \text{trace} \left\{ S_w^{-1} S_b \right\}$$

$$S_w = \sum_{i=1}^M P(\omega_i) S_i \text{ -- within - class scatter}$$

$$S_b = \sum_{i=1}^M P(\omega_i) (\mu_i - \mu_0)(\mu_i - \mu_0)^T \text{ -- between - class scatter}$$

- From the feature vector  $x$ , let  $S_{xw}$  and  $S_{xb}$  be the within-class and between-class scatter matrix.
- The scatter matrices for the transformed variable  $y = A^T x$  are:

$$S_{yw} = A^T S_{xw} A \quad S_{yb} = A^T S_{xb} A$$

INF 5300

29

- In subspace  $y$ ,  $J_3$  becomes:

$$J_3 = \text{trace} \left\{ \left( A^T S_w A \right)^{-1} \left( A^T S_b A \right) \right\}$$

- Problem: find  $A$  such that  $J_3$  is maximized.
- Solution: set

$$\frac{\partial J_3(A)}{\partial A} = 0$$

$\Downarrow$

$$\frac{\partial J_3(A)}{\partial A} = -2S_{xw} A \left( A^T S_{xw} A \right)^{-1} \left( A^T S_{xb} A \right) \left( A^T S_{xw} A \right)^{-1} + 2S_{xb} A \left( A^T S_{xw} A \right)^{-1} = 0$$

$\Downarrow$

$$S_{xw}^{-1} S_{xb} A = A \left( S_{yw}^{-1} S_{yb} \right)$$

INF 5300

30

- 
- The scatter matrices  $S_{yw}$  and  $S_{yb}$  are symmetric, and can thus be diagonalized by the linear transform (appendix B)

$$B^T S_{yw} B = I \text{ and } B^T S_{yb} B = D$$

- $B$  is a  $l \times l$  matrix,  $I$  the identity matrix, and  $D$  an  $l \times l$  diagonal matrix.  $I$  and  $D$  are the scatter matrices of the transformed vector

$$\hat{y} = B^T y = B^T A^T x$$

- 
- $J_3$  is invariant under linear transformations and:

$$\begin{aligned} J_3(\hat{y}) &= \text{trace} \left\{ \left( B^T S_{yw} B \right)^{-1} \left( B^T S_{yb} B \right) \right\} = \text{trace} \left\{ B^{-1} S_{yw}^{-1} S_{yb} B \right\} \\ &= \text{trace} \left\{ S_{yw}^{-1} S_{yb} B B^{-1} \right\} = J_3(y) \end{aligned}$$

- Furthermore,

$$\left( S_{xw}^{-1} S_{xb} \right) C = C D, \text{ where } C = A B$$

- Because  $D$  is diagonal, this is an eigenvalue-problem,
- $D$  must have the eigenvalues of  $\left( S_{xw}^{-1} S_{xb} \right)$  on the diagonal
- $C$  must have the corresponding eigenvectors of  $\left( S_{xw}^{-1} S_{xb} \right)$  as columns...



- 
- Note that  $S_{xb} = \sum_{i=1}^M P(\omega_i)(\mu_i - \mu_0)(\mu_i - \mu_0)^T$

is a sum of M (M=no. of classes) matrices of rank 1, only m-1 of these elements are independent, meaning that the rank of  $S_{xb}$  is M-1 or less (no more than M-1 eigenvalues are nonzero).

- This means also that  $S_{xw}^{-1}S_{xb}$  has rank M-1 or less.
- Fisher's discriminant transform can give us a l-dimensional projection, where  $l \leq M-1$ .
  - Note: with 30 features (m=30) and 5 classes (M=5) this gives us a projection with dimension 4 or less.

---

## Computing Fishers linear discriminant

---

- For  $l=M-1$ :
  - Form a matrix C such that its columns are the M-1 eigenvectors of  $S_{xw}^{-1}S_{xb}$
  - Set  $\hat{y} = C^T x$
  - This gives us the maximum  $J_3$  value.
  - This means that we can reduce the dimension from m to M-1 without loss in class separability power (but only if  $J_3$  is a correct measure of class separability.)
  - Alternative view: with a Bayesian model we compute the probabilities  $P(\omega_i|x)$  for each class ( $i=1, \dots, M$ ). Once M-1 probabilities are found, the remaining  $P(\omega_M|x)$  is given because the  $P(\omega_i|x)$ 's sum to one.

## Computation: Case 2: $I < M-1$

---

- Form C by selecting the eigenvectors corresponding to the  $I$  largest eigenvalues of

$$S_{xw}^{-1} S_{xb}$$

- We now have a loss of discriminating power since

$$J_{3,\hat{y}} < J_{3,x}$$

---

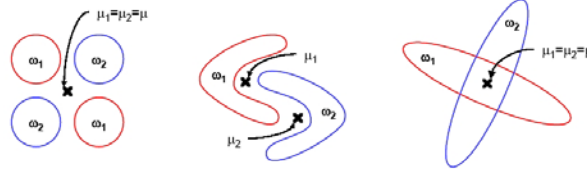
## Comments on Fishers discriminant rule

---

- In general, projection of the original feature vector to a lower dimensional space is associated with some loss of information.
- Although the projection is optimal with respect to  $J_3$ ,  $J_3$  might not be a good criterion to optimize for a given data set. (Note that  $J_3$  is a kind of sum of a product of between-class and within-class scatter, where the sum is over all classes)
- Minimizing  $J_3$  is not equivalent to minimizing the classification error.

# Limitations of Fisher's discriminant

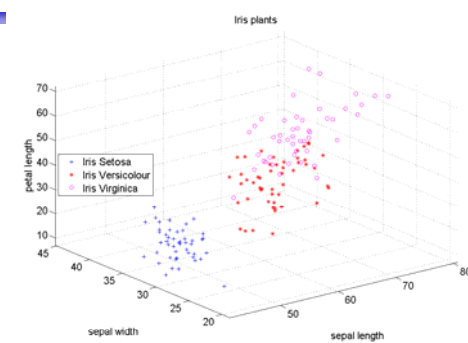
- It produces at most  $C-1$  feature projections
- It is parametric, since it assumes unimodal gaussian likelihoods
- It will fail when the discriminatory information is not in the mean but in the variance of the data



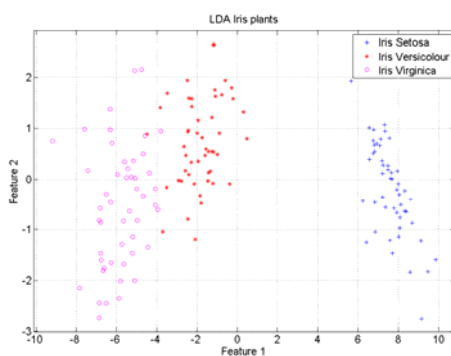
INF 5300

37

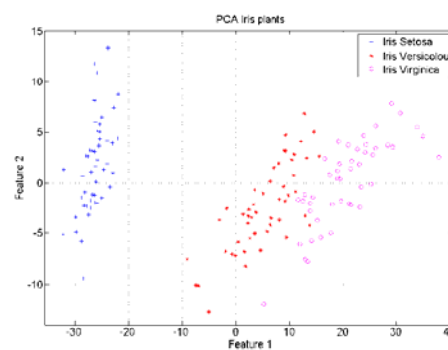
# Fisher's discriminant example



Original data



Best 2 Fisher's



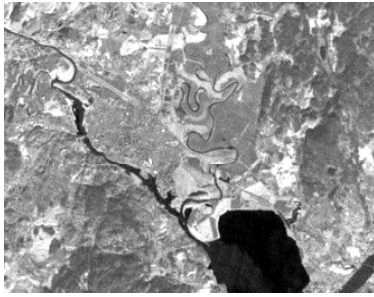
Best 2 PCA

INF 5300

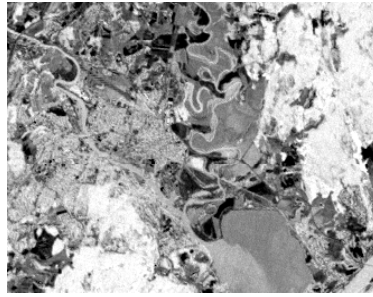
38

# Fisher's linear discriminant on the Landsat image

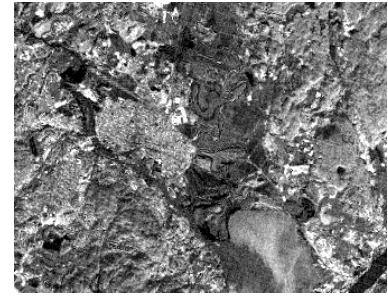
---



1. Fisher feature



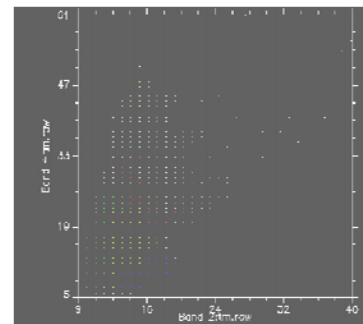
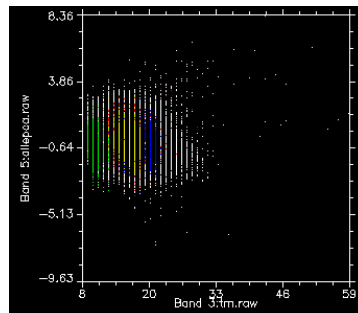
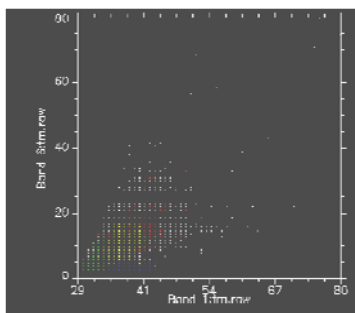
2. Fisher feature



3. Fisher feature

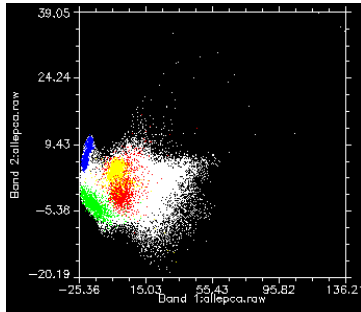
# Scatter plots for the example – Original spectral bands

---

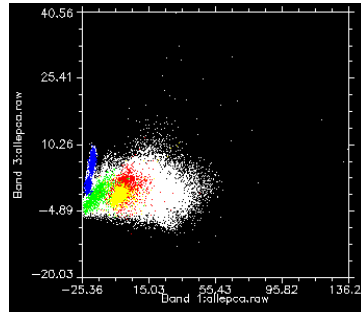


## Scatter plots for the example – PCA-components

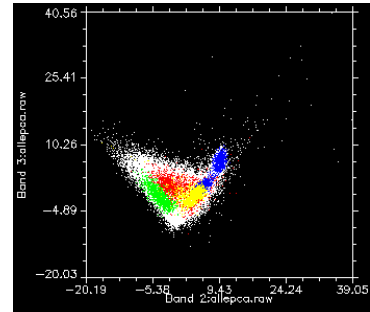
---



PCA 1 and 2



PCA 1 and 3



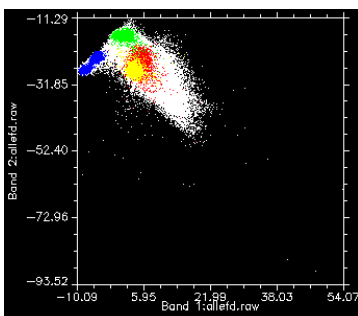
PCA 2 and 3

INF 5300

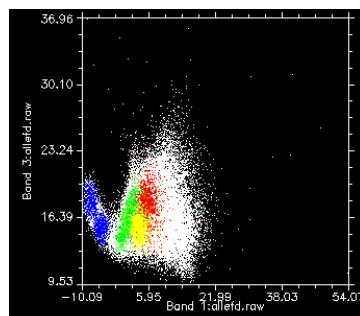
41

## Scatter plots for the example – Fisher-components

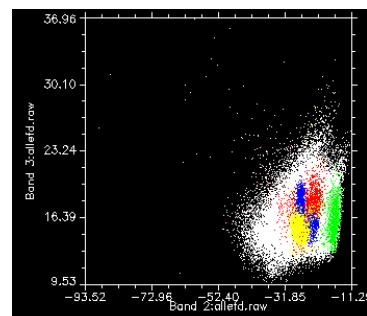
---



Fisher 1 and 2



Fisher 1 and 3



Fisher 2 and 3

INF 5300

42

# Comparison of overall classification accuracy

---

- All 6 original spectral bands: 91.9% correct classification
- PCA components 1-3: 90.8% correct
- Fisher components 1-3: 91.5% correct

# Literature on pattern recognition

---

- Updated review and statistical pattern recognition:
  - A. Jain, R. Duin and J. Mao: Statistical pattern recognition: a review, IEEE Trans. Pattern analysis and Machine Intelligence, vol. 22, no. 1, January 2001, pp. 4--
- Classical PR-books
  - R. Duda, P. Hart and D. Stork, Pattern Classification, 2. ed. Wiley, 2001
  - B. Ripley, Pattern Recognition and Neural Networks, Cambridge Press, 1996.
  - S. Theodoridis and K. Koutroumbas, Pattern Recognition, Academic Press, 2006.