Chapter 3:
PCA and FLD Techniques (part 3)

- Dimensionality Problem
- Component Analysis
- Scatter Matrix
- Principal Component Analysis
- Fisher Linear Discriminant

FEATURE SELECTION

- **Desired properties:**
  - **Patterns belonging to different classes have dissimilar-valued features**. Patterns associated with different classes should have feature values as far apart as possible.
  - **Patterns belonging to the same class have similar-valued features**. Patterns belonging to the same class should be as close as possible to their mean.

- **Dimensionality problem:**
  - Combining features to reduce the dimensionality can be used to mitigate the adverse effects of "curse of dimensionality".
  - **Idea:** represent a set of N d-dimensional samples vectors using a single, p-dimensional vector, where $p < d$ (i.e. reducing the feature space). If $p=1$, all samples patterns are projected on one direction.
  - **Techniques** for reducing excessive dimensionality by combining features through linear transformations:
    - **Component Analysis**: **Principal Component Analysis** (**PCA**) seeks a projection that best represents the data.
    - **Discriminant Analysis**: **Fisher Linear Discriminant** (**FLD**) seeks a projection that best separates (discriminates) the data.

COMPONENT ANALYSIS

- **Goal:** represent a set of N d-dimensional samples $X_1, X_2, \ldots, X_N$, using a single vector X_0 so that the squared distances between X_0 and any X_k are as small as possible.

- Let where m is the **sample mean**:

$$m = \frac{1}{N}\sum_{k=1}^{N} X_k$$

- The **squared-error criterion function** is:

$$J(X_0) = \sum_{k=1}^{N} \left\| X_0 - X_k \right\|^2 = \sum_{k=1}^{N} \left\| (X_0 - m) - (X_k - m) \right\|^2$$

$$J(X_0) = \sum_{k=1}^{N} \left\| X_0 - m \right\|^2 - 2(X_0 - m)^T \sum_{k=1}^{N} (X_k - m) + \sum_{k=1}^{N} \left\| X_k - m \right\|^2 = \sum_{k=1}^{N} \left\| X_0 - m \right\|^2 + \sum_{k=1}^{N} \left\| X_k - m \right\|^2$$

- **Observation:** $J(X_0)$ is minimized by selecting $X_0 = m$.

- This is a zero-dimensional representation of the data set

ONE-DIMENSIONAL PROJECTION

- **Observation:** best 1-D representation of data (minimizing the least-square error) is the projection onto a line through the sample mean.

- **One-dimensional representation:** $\boxed{\widetilde{X}_k = m + a_k e}$
  where *e* is the unit vector of the projection direction, and *a* is a scalar.

- Optimal sets of coefficients a_k are obtained by minimizing the squared-error criterion function:

$$J_1(a,e) = \sum_{k=1}^{N} \left\|(m + a_k e) - X_k\right\|^2 = \sum_{k=1}^{N} \left\|a_k e - (X_k - m)\right\|^2 = \sum_{k=1}^{N} a_k^2 \|e\|^2 - 2\sum_{k=1}^{N} a_k e^T (X_k - m) + \sum_{k=1}^{N} \left\|X_k - m\right\|^2$$

- Since $||e|| = 1$, then: $\boxed{\dfrac{\partial J_1}{\partial a_k} = 0} \longrightarrow \boxed{a_k = e^T\left(X_k - m\right)}$

- The best direction *e* for the projection line can be found by minimizing:

$$J_1(e) = \sum_{k=1}^{N} a_k^2 - 2\sum_{k=1}^{N} a_k^2 + \sum_{k=1}^{N} \left\|X_k - m\right\|^2 = -\sum_{k=1}^{N} \left[e^T (X_k - m)\right]^2 + \sum_{k=1}^{N} \left\|X_k - m\right\|^2$$

SCATTER MATRIX

- Finding the best direction *e* for the projection line involves the **scatter matrix** *S* defined as *(N-1)* times the covariance matrix of the samples:

$$\boxed{S = (N-1)\Sigma = \sum_{k=1}^{N} (X_k - m)(X_k - m)^T}$$

- **Criterion function** (to minimize):

$$J_1(e) = -\sum_{k=1}^{N} \left[e^T (X_k - m)\right]^2 + \sum_{k=1}^{N} \left\| X_k - m \right\|^2 = -\sum_{k=1}^{N} e^T (X_k - m)(X_k - m)^T e + \sum_{k=1}^{N} \left\| X_k - m \right\|^2$$

$$\boxed{J_1(e) = -e^T S e + \sum_{k=1}^{N} \left\| X_k - m \right\|^2}$$

- $J_1(e)$ is minimized when $e^T Se$ is maximized. Applying Lagrange optimization method (with λ an undetermined multiplier):

$$\boxed{u = e^T Se - l(e^T e - 1)} \longrightarrow \boxed{\frac{\partial u}{\partial e} = 2Se - 2\lambda e} \longrightarrow \boxed{Se = \lambda e}$$

- **Conclusion:** since $e^T Se = \lambda e^T e = \lambda$, maximizing $e^T Se$ means selecting the eigenvector corresponding to the largest eigenvalue of the scatter matrix.

MATH REMINDER

- **Eigenvector and Eigenvalue:** given $d \times d$ matrix M and a scalar λ, the d-dimensional vector x satisfying the set of linear equations:
$$Mx = \lambda x$$
is called the *eigenvector* of M corresponding to scalar λ. If I is the identity matrix, then the system of linear equations can be rewritten as:
$$(M - \lambda I)x = 0$$
The solution vector $x = e_k$ and corresponding scalar $\lambda = \lambda_k$ are called, respectively, *eigenvector* and associated *eigenvalue*. Note that any multiple of eigenvector x is also an eigenvector.

- **Lagrange optimization:** suppose we seek the extremum x_0 of function $f(x)$ subject to a constraint expressed in the form $g(x) = 0$. The constrained optimization problem can be solved by employing the *Lagrange undetermined multiplier* λ to form Lagrangian function:
$$L(x, \lambda) = f(x) + \lambda g(x)$$
The position of the extremum is given by the solution of equation:
$$\frac{\partial L(x,\lambda)}{\partial x} = \frac{\partial f(x)}{\partial x} + \lambda \frac{\partial g(x)}{\partial x} = 0$$

PCA

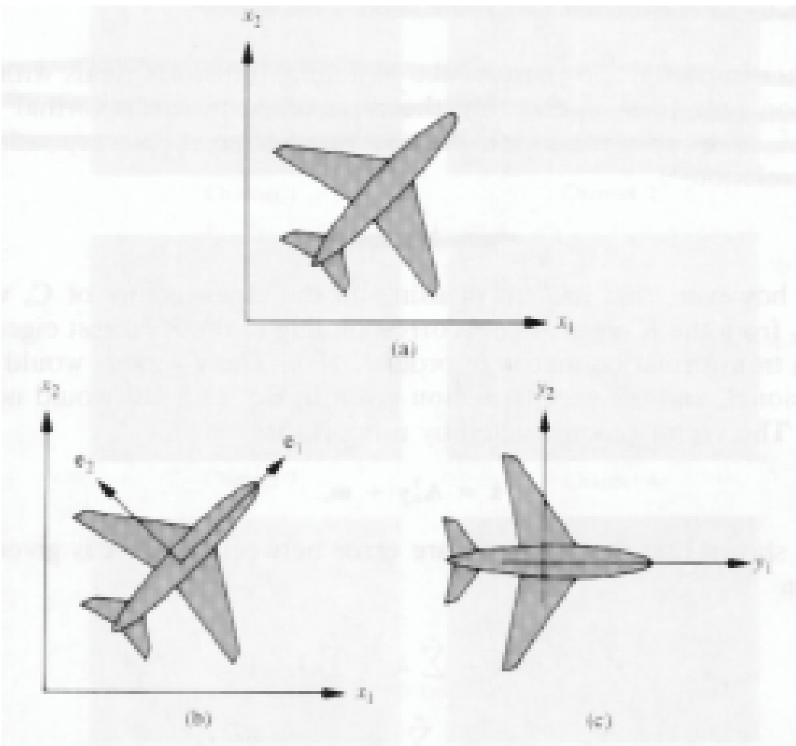
- **Principal Component Analysis** (**PCA**) leads to the projection that best best represents the data in the least-square sense:
  - Optimal projection is onto a line passing through the sample mean.
  - The projection line has the direction of the eigenvector corresponding to the largest eigenvalue of the scatter matrix.
- Coefficients $a_k = e^T(X_k - m)$ are called **principal components**; e^T is the unit vector on the projection direction (and happens to be the eigenvector associated with the largest eigenvalue).

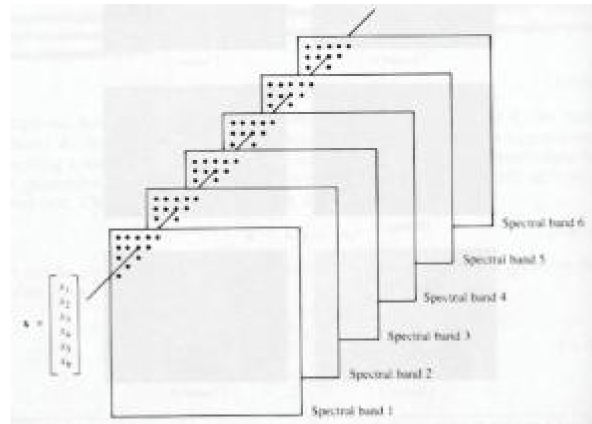
- **General case:** extending projection from one to p dimensions ($p < d$):

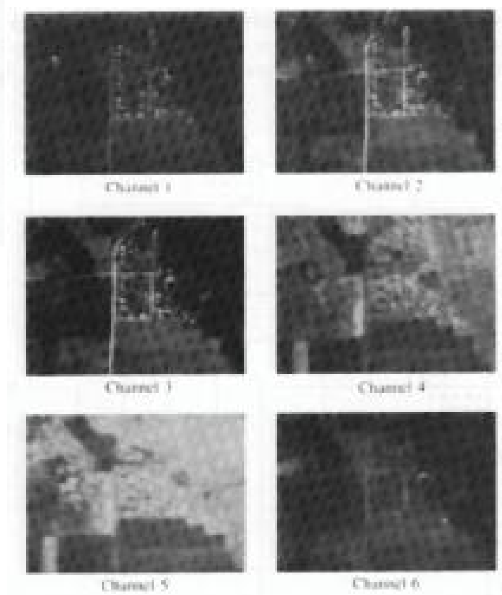
$$\boxed{\tilde{x} = m + \sum_{j=1}^{p} a_j e_j} \longrightarrow \boxed{J_p(a,e) = \sum_{k=1}^{N} \left\| \left(m + \sum_{j=1}^{p} a_j e_j \right) - X_k \right\|^2}$$

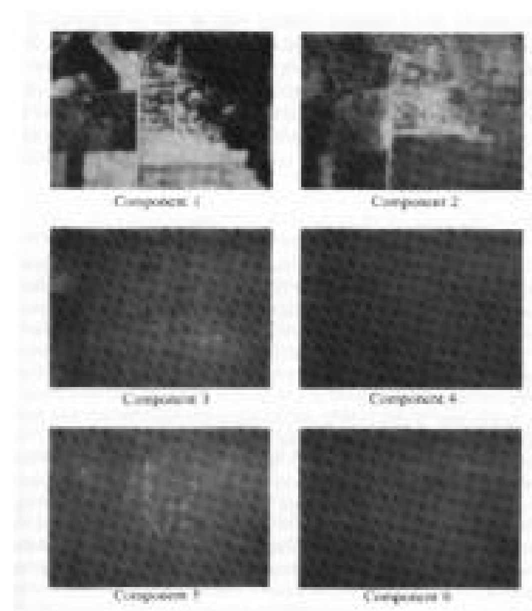
- **PCA optimization:** in *p-dimensions*, the best representation of the sample data is the *p-eigenvectors* of the scatter matrix, corresponding to the largest *p-eigenvalues*.

Examples of PCA's Applications









DISCRIMINANT ANALYSIS

- **PCA** finds the minimum number of components that best represents the data (that is, the best representation is in the least square sense). It does not guarantee any usefulness for classification.

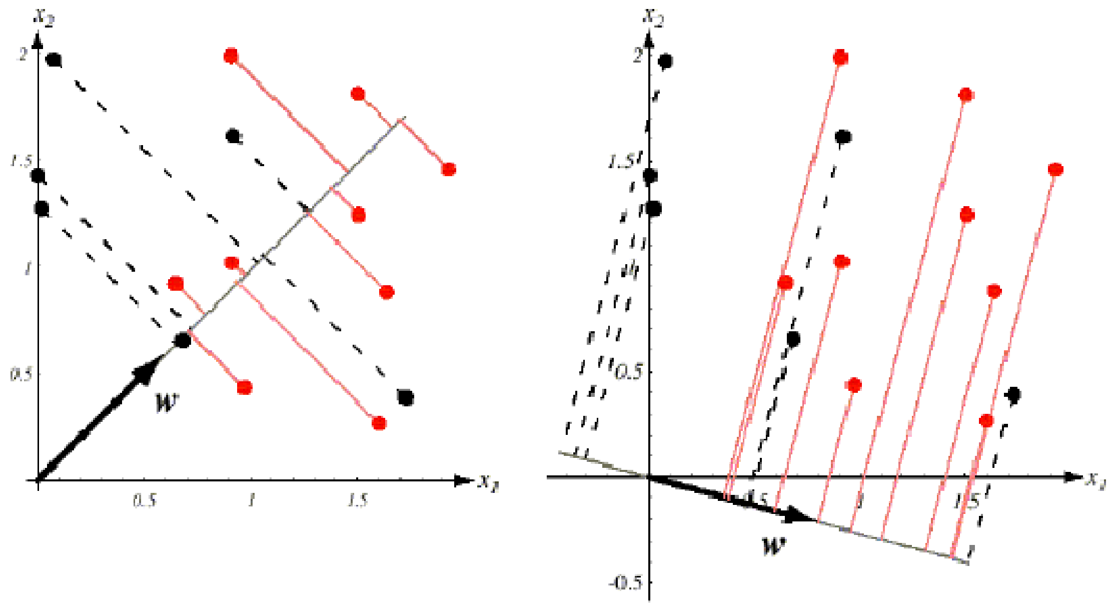
- There is a need to reduce the dimensionality, under some constraint of maximizing the class separation (i.e. discrimination).

- While PCA aims to find the directions most efficient for representing the data, discriminant analysis attempts to identify the directions that are efficient for discriminating data between different classes.

- **Discriminant Analysis:** maximizing the discrimination can be achieved by increasing the inter-cluster distances while reducing the intra-cluster distances. These distances are obtained by employing, respectively, the **between** and **within class scatter matrices**.

OPTIMAL PROJECTION

Projection of the same set of samples onto two different lines in the directions marked by vector *w*. The figure on the right shows greater separation between the red and the black projected points.



If samples are seen as forming a *d*-dimension hyper-ellipsoidally shaped cloud, then the eigenvectors of the scatter matrix are the principal axes of that hyper-ellipsoid.

FISHER LINEAR DISCRIMINANT

- **Data:** a set of N d-dimensional samples $X_1, X_2, \ldots, X_N$ are distributed into c subsets, $D_1, D_2, \ldots, D_c$, subset D_k being associated with class C_k. Let n_k be the number of samples in subset D_k.
- **Fisher Linear Discriminant** (**FLD**) is simply based on a linear transformation such as:

$$y = w^T X$$

  where X is the $[d \times N]$ matrix of the samples and w is a ***[dxp] projection*** matrix
- **Observations:**
  - FLD transforms the set of samples $X_1, X_2, \ldots, X_N$ (via $y_k = w^T X_k$) into a set of sample projections $y_1, y_2, \ldots, y_N$ of reduced dimensionality ($p < d$).
  - Projections are distributed into subsets $Y_1, Y_2, \ldots, Y_c$, with $Y_k = \{y_k \mid X_k \in D_k\}$, corresponding to class-based subsets $D_1, D_2, \ldots, D_c$.
  - If $||w|| = 1$, then y_k is the projection of X_k on a line in the direction w.
  - It can be shown that the the samples mean has a similar projection:

$$\boxed{m_k = \frac{1}{n_k} \sum_{X_k \in D_k} X_k} \longrightarrow \boxed{\widetilde{m}_k = \frac{1}{n_k} \sum_{y_k \in Y_k} y_k = \frac{1}{n_k} \sum_{X_k \in D_k} w^T x_k} \longrightarrow \boxed{\widetilde{m}_k = w^T m_k}$$

TWO-CLASS CRITERIA

- **Definitions:**
  - **Distance** between 2 projected means:
    $$\left|\tilde{m}_1 - \tilde{m}_2\right| = \left|w^T (m_1 - m_2)\right|$$
  - **Scatter** for projected samples:
    $$\tilde{s}_k = \sum_{y \in Y_k} (y - \tilde{m}_k)^2$$
- **Desired Criteria:**
  - **Maximize separation:** separation of projected class means should be as large as possible.
    $$\left|\tilde{m}_1 - \tilde{m}_2\right| = \left|w^T (m_1 - m_2)\right| \to \max$$
  - **Preserve compactness:** sum of the scatter of each of the sets of projected values should be as small as possible.
    $$\tilde{s}_1^2 + \tilde{s}_2^2 \to \min$$
- **FLD Criterion function:**
  - overall measure of FLD "goodness" is given by the separation of the means relative to the compactness.
    $$J(w) = \frac{\left|\tilde{m}_1 - \tilde{m}_2\right|^2}{(\tilde{s}_1^2 + \tilde{s}_2^2)} \to \max$$
  - *w* maximizing *J(w)* leads to the best separation between two projects sets.
    $$\frac{\partial J(w)}{\partial w} = 0$$

SCATTER MATRICES

- **Scatter matrix:** covariance matrix times size of samples population.
  $$S_k = (n_k - 1)\Sigma_k = \sum_{X \in D_k} (X - m_k)(X - m_k)^T$$
- **Within-class scatter matrix**: sum of the scatter matrices for both classes.
  $$\boxed{S_W = S_1 + S_2}$$
- **Between-class scatter matrix**: defined through the class means.
  $$\boxed{S_B = (m_1 - m_2)(m_1 - m_2)^T}$$
- **Scatter** of projected X_k samples:
  $$\tilde{s}_k^2 = \sum_{X \in D_k} (w^T X - w^T m_k)^2 = \sum_{X \in D_k} w^T (X - m_k)(X - m_k)^T w = w^T S_k w$$
- **Within-class scatter** of projected samples (that should be as small as possible):
  $$\boxed{\tilde{s}_1^2 + \tilde{s}_2^2 = w^T S_w w}$$
- **Between-class scatter** of projected samples (that should be as large as possible):
  $$\boxed{|\tilde{m}_1 - \tilde{m}_2|^2 = w^T S_B w}$$

FINDING BEST FEATURES

- **Problem:** identify the matrix *w* for the linear transformation $y = w^T X$ to reduce feature vector dimension from *d* to *p* (with $p < d$). For a *c*-class problem
- **Fisher Linear Discriminant:** set of *w* values that maximizes the criterion function *J(w)*.
- **FLD Criterion:**

$$\boxed{J(w) = \frac{\left|\tilde{m}_1 - \tilde{m}_2\right|^2}{(\tilde{s}_1^2 + \tilde{s}_2^2)} = \frac{w^T S_B w}{w^T S_W w}}$$

- **Methods:**
  - 1. Start with a good guess and iteratively adjust *w*, so that *J(w)* increases with each iteration (i.e. *w* is getting better and better).
  - 2. Solve directly:

$$\boxed{\frac{\partial J(w)}{\partial w} = 0} \longrightarrow \boxed{S_B w = \lambda S_w w}$$

- **Solution:** the eigenvector of $S_W^{-1} S_B$ with the largest absolute eigenvalue.

$$\boxed{S_w^{-1} S_B w = \lambda w} \longrightarrow \boxed{w = S_w^{-1}(m_1 - m_2)}$$

GENERAL CASE

- **Total mean vector:** $$\boxed{m = \frac{1}{N}\sum_{k=1}^{c} X_k = \frac{1}{N}\sum_{k=1}^{c} n_k m_k}$$
- **Total scatter matrix:** $$\boxed{S_T = \sum_{k=1}^{N} (X_k - m)(X_k - m)^T}$$
- **Within-class scatter matrix:** $$\boxed{S_W = \sum_{k=1}^{c} S_k = \sum_{k=1}^{c} \sum_{X \in D_k} (X - m_k)(X - m_k)^T}$$
- **Between-class scatter matrix:** $$\boxed{S_B = \sum_{k=1}^{c} (m_k - m)(m_k - m)^T}$$
- **FLD Criterion function:** $$\boxed{J(w) = \frac{w^T S_B w}{w^T S_W w}}$$

FLD TECHNIQUE

- FLD tries to find the *w* transformation matrix that will maximize the *J(w)* criterion function. It is considered that maximizing this function will increase the inter-cluster distances (through maximizing the between class cluster matrix) and will decrease the intra-cluster distances (through minimizing the within class scatter matrix).

- The columns of the optimal (in least-square sense) *w* matrix are the generalized eigenvectors (of $S_W^{-1}S_B$) corresponding to the largest eigenvalues.

$$\boxed{S_B w_k = \lambda_k S_W w_k}$$

$$\boxed{S_w^{-1} S_B w_k = \lambda_k w_k}$$

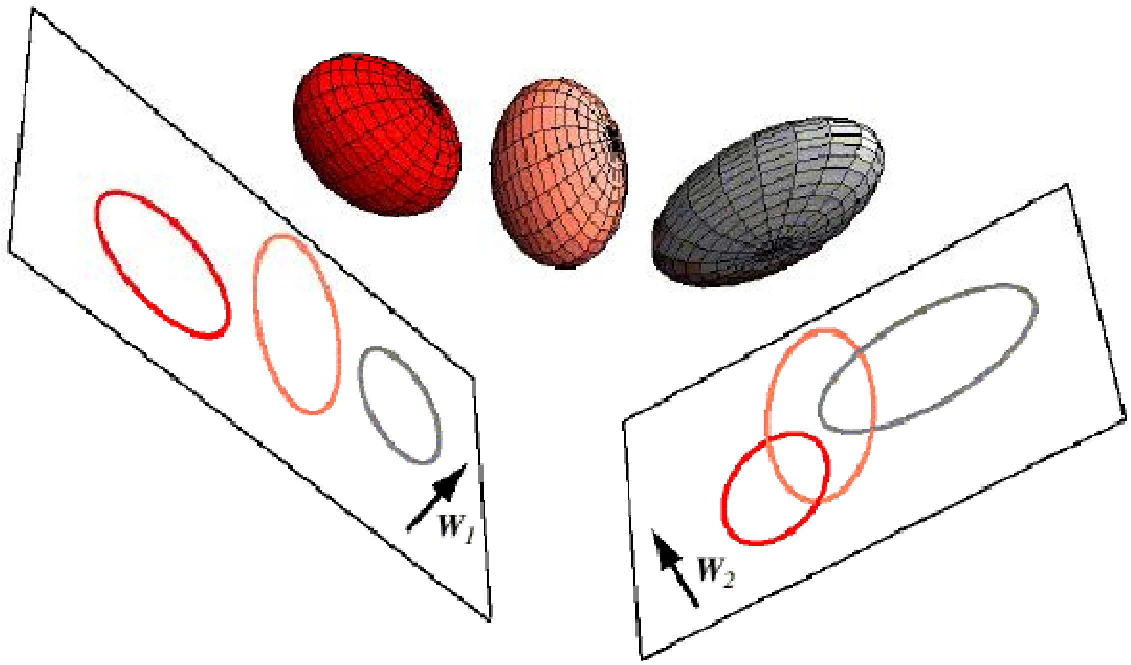
- This generalized eigenvalue problem can be solved by first computing the eigenvalues as the roots of the characteristic polynomial, and then solving the linear set for w_k, the columns of the *w* matrix.

$$\boxed{\left| S_B - \lambda_k S_W \right| = 0}$$

$$\boxed{\left(S_B - \lambda_k S_W \right) w_k = 0}$$

- Note: procedure might generate *d* eigenvalues, and *d* corresponding eigenvectors. However, only p of these eigenvalues should end up being non-zero.

MAXIMIZING SEPARATION



Three three-dimensional distributions are projected onto two-dimensional subspaces, described by a normal vectors w_1 and w_2. Informally, multiple discriminant methods seek the optimum such subspace, that is, the one with the greatest separation of the projected distributions for a given total within-scatter matrix, here as associated with w_1.