Advanced Pattern Recognition -- Model-Based Classifiers

Xiaojun Qi

Outline

- Principal Component Analysis (PCA)
- FYI: Linear Discriminant Analysis (LDA)

Principal Component Analysis (PCA): Introduction

- Principal Component Analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

PCA

- This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.
- The resulting vectors are an **uncorrelated** orthogonal basis set.

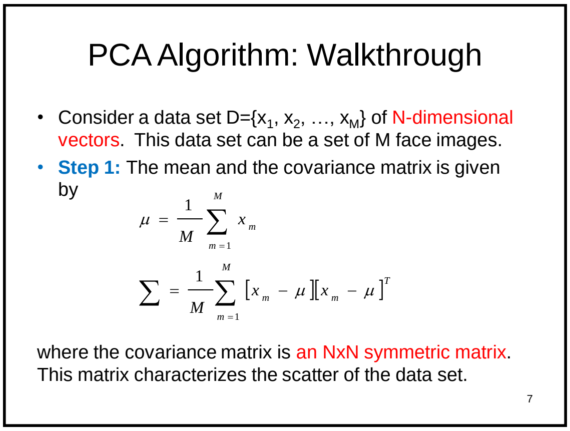
Possible Use of PCA

- Dimensionality reduction
- The determination of linear combinations of variables
- Feature selection: the choice of the most useful variables.
- Visualization of multidimensional data
- Identification of underlying variables.
- Identification of groups of objects or of outliers

PCA Algorithm

- Goal: Represent a data set D={x_1, x_2, …, x_M} of N-dimensional vectors using K-dimensional vectors, where K<<N.

1. Compute mean and covariance matrix.
2. Calculate the eigenvalues and eigenvectors of the covariance matrix.
3. Construct the transformation matrix using the eigenvectors corresponding to the first K largest eigenvalues.
4. Transform each observation into a K-dimensional vector for another representation.



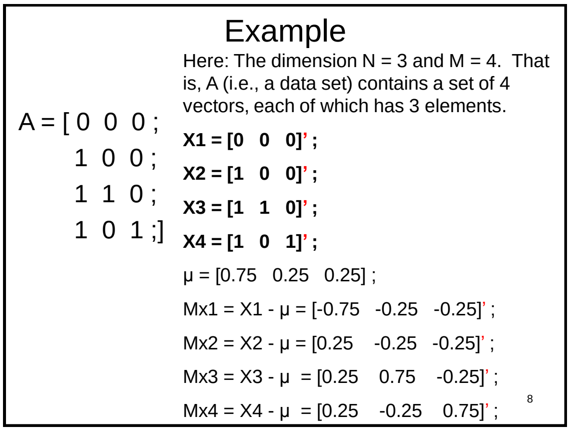
PCA Algorithm: Walkthrough

- Consider a data set D={x_1, x_2, …, x_M} of N-dimensional vectors. This data set can be a set of M face images.
- **Step 1:** The mean and the covariance matrix is given by

$$\mu = \frac{1}{M}\sum_{m=1}^{M} x_m$$

$$\Sigma = \frac{1}{M}\sum_{m=1}^{M} [x_m - \mu][x_m - \mu]^T$$

where the covariance matrix is an NxN symmetric matrix. This matrix characterizes the scatter of the data set.



Example

Here: The dimension N = 3 and M = 4. That is, A (i.e., a data set) contains a set of 4 vectors, each of which has 3 elements.

A = [0 0 0 ;
1 0 0 ;
1 1 0 ;
1 0 1 ;]

X1 = [0 0 0]' ;

X2 = [1 0 0]' ;

X3 = [1 1 0]' ;

X4 = [1 0 1]' ;

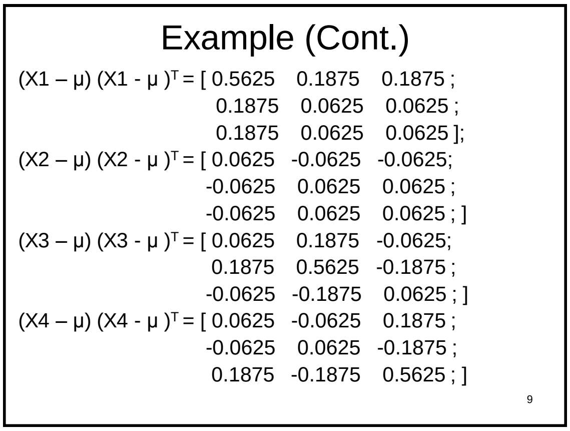
μ = [0.75 0.25 0.25] ;

Mx1 = X1 - μ = [-0.75 -0.25 -0.25]' ;

Mx2 = X2 - μ = [0.25 -0.25 -0.25]' ;

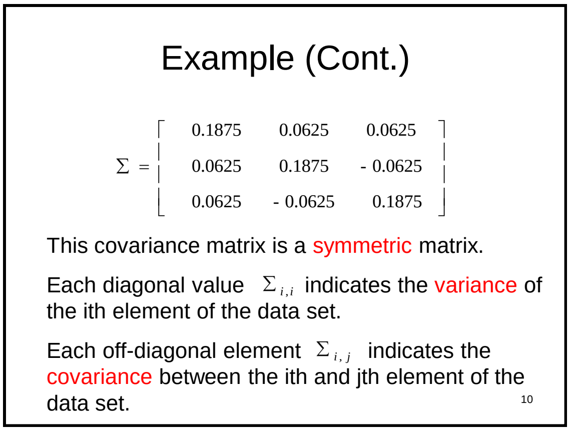
Mx3 = X3 - μ = [0.25 0.75 -0.25]' ;

Mx4 = X4 - μ = [0.25 -0.25 0.75]' ;



Example (Cont.)

(X1 – μ) (X1 - μ)^T = [0.5625 0.1875 0.1875 ;
0.1875 0.0625 0.0625 ;
0.1875 0.0625 0.0625];
(X2 – μ) (X2 - μ)^T = [0.0625 -0.0625 -0.0625;
-0.0625 0.0625 0.0625 ;
-0.0625 0.0625 0.0625 ;]
(X3 – μ) (X3 - μ)^T = [0.0625 0.1875 -0.0625;
0.1875 0.5625 -0.1875 ;
-0.0625 -0.1875 0.0625 ;]
(X4 – μ) (X4 - μ)^T = [0.0625 -0.0625 0.1875 ;
-0.0625 0.0625 -0.1875 ;
0.1875 -0.1875 0.5625 ;]



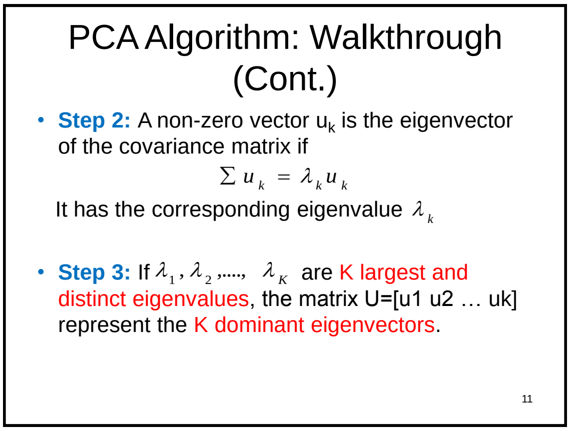
Example (Cont.)

$$\Sigma = \begin{bmatrix} 0.1875 & 0.0625 & 0.0625 \\ 0.0625 & 0.1875 & -0.0625 \\ 0.0625 & -0.0625 & 0.1875 \end{bmatrix}$$

This covariance matrix is a symmetric matrix.

Each diagonal value $\Sigma_{i,i}$ indicates the variance of the ith element of the data set.

Each off-diagonal element $\Sigma_{i,j}$ indicates the covariance between the ith and jth element of the data set.



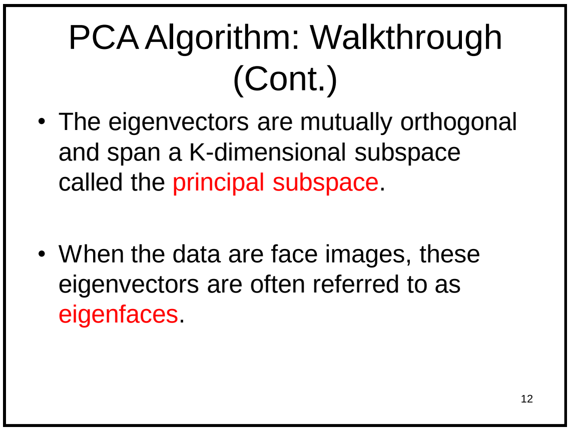
PCA Algorithm: Walkthrough (Cont.)

- **Step 2:** A non-zero vector u_k is the eigenvector of the covariance matrix if

$$\Sigma u_k = \lambda_k u_k$$

It has the corresponding eigenvalue λ_k

- **Step 3:** If $\lambda_1, \lambda_2, \ldots, \lambda_K$ are K largest and distinct eigenvalues, the matrix U=[u1 u2 … uk] represent the K dominant eigenvectors.



PCA Algorithm: Walkthrough (Cont.)

- The eigenvectors are mutually orthogonal and span a K-dimensional subspace called the principal subspace.
- When the data are face images, these eigenvectors are often referred to as eigenfaces.

PCA Algorithm: Walkthrough (Cont.)

- **Step 4:** If U is the matrix of dominant eigenvectors, an N-dimensional input x can be linearly transformed into a K-dimensional vector α by:

$$\alpha = U^{T}(x - \mu)$$

- After applying the linear transform U^T, the set of transformed vectors $\{\alpha_1, \alpha_2, \dots \alpha_M\}$ has scatter

$$U^{T} \Sigma U = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \dots\dots & \\ & & & \lambda_k \end{bmatrix}$$

- PCA chooses U so as to maximize the determinant of this scatter matrix.

PCA Reconstruction

- An original vector x can be approximately constructed from its transformed α as:

$$\tilde{x} = \sum_{k=1}^{K} \alpha_k u_k + \mu$$

- In fact, PCA enables the training data to be reconstructed in a way that minimizes the squared reconstruction error over the data set. This error is:

$$\varepsilon = \frac{1}{2} \sum_{m=1}^{M} \left\| x_m - \tilde{x}_m \right\|^2$$

PCA Eigenvector Calculations

- Geometrically, PCA consists of projection onto K orthonormal axes.
- These principal axes maximize the retained variance of the data after projection.
- In practice, the covariance matrix is often singular, particularly, if M<N.
- However, the K<M principal eigenvectors can still be estimated using Singular Value Decomposition (SVD) or Simultaneous Diagonalization.

```
[pc, newdata, variance, t2] = princomp(A)
pc =
    0.8165         0    0.5774
    0.4082   -0.7071   -0.5774
    0.4082    0.7071   -0.5774
newdata =
   -0.8165   -0.0000   -0.1443
    0.0000   -0.0000    0.4330
    0.4082   -0.7071   -0.1443
    0.4082    0.7071   -0.1443
variance =
    0.3333
    0.3333
    0.0833
t2 =
    2.2500
    2.2500
    2.2500
    2.2500
```

PCA Summary

- The PCA method generates a new set of variables, called principal components.
- Each principal component is a linear combination of the original variables.
- All the principal components are orthogonal to each other so there is no redundant information.
- The principal components as a whole form an orthogonal basis for the space of the data.

The First Principal Component

- The first principal component is a single axis in space. When you project each observation on that axis, the resulting values form a new variable. And the variance of this variable is the maximum among all possible choices of the first axis.

The Second Principal Component

- The second principal component is another axis in space, perpendicular to the first. Projecting the observations on this axis generates another new variable. The variance of this variable is the maximum among all possible choices of this second axis.

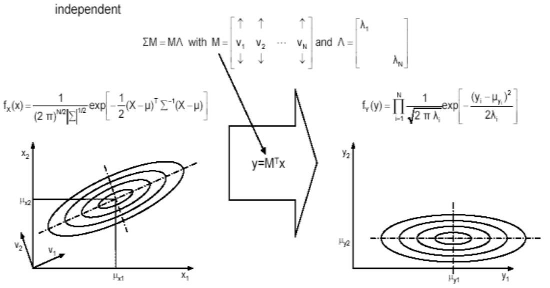
Principal Components

- The full set of principal components is as large as the original set of variables.

- But it is commonplace for the sum of the variances of the first few principal components to exceed 80% of the total variance of the original data. By examining plots of these few new variables, researchers often develop a deeper understanding of the driving forces that generated the original data.

PCA Illustration

- If the distribution happens to be Gaussian, then the transformed vectors will be statistically independent



$$\Sigma M = M\Lambda \text{ with } M = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ v_1 & v_2 & \cdots & v_N \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \text{ and } \Lambda = \begin{bmatrix} \lambda_1 & & \\ & & \\ & & \lambda_N \end{bmatrix}$$

$$f_X(x) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)\right]$$

$$f_Y(y) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left[-\frac{(y_i-\mu_{y_i})^2}{2\lambda_i}\right]$$

PCA Illustration Explanation

- Given an nxn matrix that does have eigenvectors, there are n of them.
- Scale the vector by some amount before multiplying it, the same multiple of it will be obtained.
- All the eigenvectors of a matrix are perpendicular, i.e., at right angles to each other, no matter how many dimensions you have.➔ You can express the data in terms of these perpendicular eigenvectors, instead of expressing them in terms of the x and y axes.

Linear Discriminant Analysis (LDA) -- Face Identification

- Suppose a data set X exists, which might be face images, each of which is labeled with an identity. All data points with the same identity form a class. In total, there are C classes. That is:

  X = {X1, X2, …., Xc}

LDA Algorithm

1. Compute the within-classes scatter matrix W and the between-classes scatter matrix B
2. Calculate the eigenvalues and eigenvectors of $W^{-1}B$.
3. Construct the transformation matrix using the eigenvectors corresponding to the first K largest eigenvalues.
4. Transform each observation into a K-dimensional vector for classification.

LDA

- The sample covariance matrix for the entire data set is then a NxN symmetric matrix

$$\Sigma = \frac{1}{M}\sum_{x \in X}[x - \mu][x - \mu]^T$$

where M is the total number of faces.

- This matrix characterize the scatter of the entire data set, irrespective of class-membership.

LDA

- However, a within-classes scatter matrix, W, and a between-classes scatter matrix, B are also estimated.

$$W = \frac{1}{C}\sum_{c=1}^{C}\left\{\frac{1}{M_c}\sum_{x \in X_c}[x - \mu_c][x - \mu_c]^T\right\}$$

$$B = \frac{1}{C}\sum_{c=1}^{C}[\mu_c - \mu][\mu_c - \mu]^T$$

- Where Mc is the number of samples of class c, μ_c is the sample mean for class c, and µ is the sample mean for the entire data set X.

LDA

- The goal is to find a linear transform U which maximizes the between-class scatter while minimizing the within-class scatter.
- Such a transformation should retain class separability while reducing the variation due to sources other than identity, for example, illumination and facial expression.

LDA

- An appropriate transformation is given by the matrix U = [u1 u2 … u_K] whose columns are the eigenvectors of $W^{-1}B$.
- In other words, the generalized eigenvectors corresponding to the K largest eigenvalues

$$BU_k = \lambda_k WU_k$$

- There are at most C-1 non-zero generalized eigenvalues, so K<C

LDA

- The data are transformed as follows:

$$\alpha = U^T(x - \mu)$$

- After this transformation, the data has between-class scatter matrix U^TBU and within-class scatter matrix U^TWU.
- The matrix U is such that the determinant of the new between-class scatter is maximized while the determinant of the within-class scatter is minimized.
- This implies that the following ratio is to be maximized:

$$|U^TBU| / |U^TWU|$$

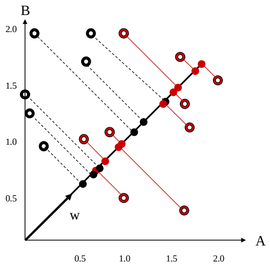
LDA

- In practice, the within-class scatter matrix (W) is often singular. This is nearly always the case when the data are image vectors with large dimensionality since the size of the data set is usually small in comparison (M<N).
- For this reason, PCA is first applied to the data set to reduce its dimensionality. The discriminant transformation is then applied to further reduce the dimensionality to C-1.

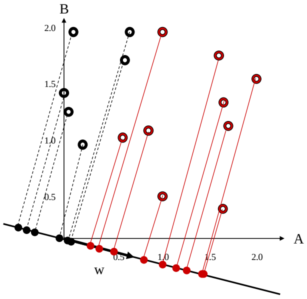
PCA vs. LDA

- PCA seeks directions that are efficient for representation;
- LDA seeks directions that are efficient for discrimination.
- To obtain good separation of the projected data, we really want the difference between the means to be large relative to some measure of the standard deviation for each class.

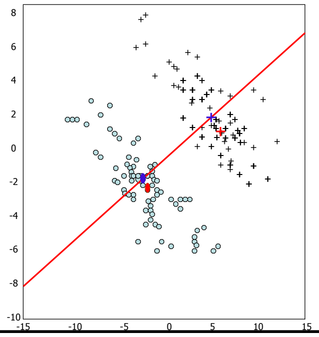
LDA Illustration
-- Bad Separation



LDA Illustration
-- Good Separation



LDA Illustration
-- 2-class case


8
6
4
2
0
-2
-4
-6
-8
-10
-15
-10
-5
0
5
10
15