-- --

Principal Components Analysis (PCA)

• Reading Assignments

S. Gong et al., *Dynamic Vision: From Images to Face Recognition*, Imperial College Press, 2001 (pp. 168-173 and Appendix C: Mathematical Details, hard copy).

K. Kastleman, *Digital Image Processing*, Prentice Hall, (Appendix 3: Mathematical Background, hard copy).

F. Ham and I. Kostanic. *Principles of Neurocomputing for Science and Engineering*, Prentice Hall, (Appendix A: Mathematical Foundation for Neurocomputing, hard copy).

A. Jain, R. Duin, and J. Mao, "Statistical Pattern Recognition: A Review", *IEEE Transactions on Pattern Analysis and Machine Intelligenve*, vol. 22, no. 1, pp. 4-37, 2000 (read pp. 11-13, on-line)

• Case Studies

M. Turk and A. Pentland, "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991 (hard copy)

K. Ohba and K. Ikeuchi, "Detectability, Uniqueness, and Reliability of Eigen Windows for Stable Verification of Partially Occluded Objects", *IEEE Transactions on Pattern Analysis and Machine Intelligenve*, vol. 19, no. 9, pp. 1043-1048, 1997 (on-line)

H. Murase and S. Nayar, "Visual Learning and Recognition of 3D Objects from Appearance", *Interantional Journal of Computer Vision*, vol 14, pp. 5-24, 1995 (hard-copy)

Principal Component Analysis (PCA)

• Pattern recognition in high-dimensional spaces

- Problems arise when performing recognition in a high-dimensional space (e.g., curse of dimensionality).

- Significant improvements can be achieved by first mapping the data into a *lower-dimensionality* space.

$$x = \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_N \end{bmatrix} --> \textit{reduce dimensionality} --> y = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_K \end{bmatrix} \quad (K << N)$$

- The goal of PCA is to reduce the dimensionality of the data while retaining as much as possible of the variation present in the original dataset.

• Dimensionality reduction

- PCA allows us to compute a linear transformation that maps data from a high dimensional space to a lower dimensional space.

$$\begin{gathered} b_1 = t_{11}a_1 + t_{12}a_2 + \ldots + t_{1n}a_N \\ b_2 = t_{21}a_1 + t_{22}a_2 + \ldots + t_{2n}a_N \\ \cdots \\ b_K = t_{K1}a_1 + t_{K2}a_2 + \ldots + t_{KN}a_N \end{gathered}$$

$$\text{or } y = Tx \text{ where } T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1N} \\ t_{21} & t_{22} & \cdots & t_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ t_{K1} & t_{K2} & \cdots & t_{KN} \end{bmatrix}$$

• Lower dimensionality basis

- Approximate the vectors by finding a basis in an appropriate lower dimensional space.

(1) Higher-dimensional space representation:

$$x = a_1 v_1 + a_2 v_2 + \cdots + a_N v_N$$

$v_1, v_2, ..., v_N$ is a basis of the N-dimensional space

(2) Lower-dimensional space representation:

$$\hat{x} = b_1 u_1 + b_2 u_2 + \cdots + b_K u_K$$

$u_1, u_2, ..., u_K$ is a basis of the K-dimensional space

- *Note:* if both bases have the same size ($N = K$), then $x = \hat{x}$)

<u>Example</u>

$$v_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, v_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, v_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \text{ (standard basis)}$$

$$x_v = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = 3v_1 + 3v_2 + 3v_3$$

$$u_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, u_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, u_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \text{ (some other basis)}$$

$$x_u = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = 0u_1 + 0u_2 + 3u_3$$

thus, $x_v = x_u$

• Information loss

- Dimensionality Reduction **implies** Information Loss !!

- Preserve as much information as possible, that is,

$$\textbf{minimize } ||x - \hat{x}|| \text{ (error)}$$

• How to determine the best lower dimensional space?

The best low-dimensional space can be determined by the "best" eigenvectors of the covariance matrix of x (i.e., the eigenvectors corresponding to the "largest" eigenvalues -- also called "principal components").

• Methodology

- Suppose $x_1, x_2, ..., x_M$ are Nx1 vectors

Step 1: $\bar{x} = \frac{1}{M} \sum_{i=1}^{M} x_i$

Step 2: subtract the mean: $\Phi_i = x_i - \bar{x}$

Step 3: form the matrix $A = [\Phi_1 \; \Phi_2 \cdots \Phi_M]$ (NxM matrix), then compute:

$$C = \frac{1}{M} \sum_{n=1}^{M} \Phi_n \Phi_n^T = AA^T$$

(sample **covariance** matrix, NxN, characterizes the *scatter* of the data)

Step 4: compute the eigenvalues of C: $\lambda_1 > \lambda_2 > \cdots > \lambda_N$

Step 5: compute the eigenvectors of C: $u_1, u_2, \ldots, u_N$

- Since C is symmetric, $u_1, u_2, \ldots, u_N$ form a basis, (i.e., any vector x or actually $(x - \bar{x})$, can be written as a linear combination of the eigenvectors):

$$x - \bar{x} = b_1 u_1 + b_2 u_2 + \cdots + b_N u_N = \sum_{i=1}^{N} b_i u_i$$

Step 6: (**dimensionality reduction step**) keep only the terms corresponding to the K largest eigenvalues:

$$\hat{x} - \bar{x} = \sum_{i=1}^{K} b_i u_i \text{ where } K << N$$

- The representation of $\hat{x} - \bar{x}$ into the basis $u_1, u_2, ..., u_K$ is thus

$$\begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_K \end{bmatrix}$$

• Linear tranformation implied by PCA

- The linear tranformation R^N->R^K that performs the dimensionality reduction is:

$$\begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_K \end{bmatrix} = \begin{bmatrix} u_1^T \\ u_2^T \\ \cdots \\ u_K^T \end{bmatrix} (x - \bar{x}) = U^T (x - \bar{x})$$

• An example

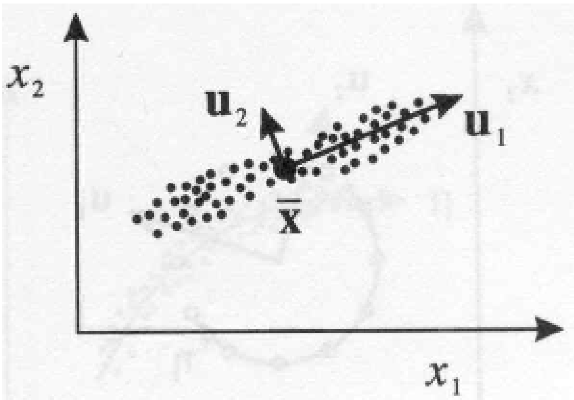
(see Castleman's appendix, pp. 648-649)

• Geometrical interpretation

- PCA projects the data along the directions where the data varies the most.

- These directions are determined by the eigenvectors of the covariance matrix corresponding to the largest eigenvalues.

- The magnitude of the eigenvalues corresponds to the variance of the data along the eigenvector directions.



• Properties and assumptions of PCA

- The new variables (i.e., b_i's) are uncorrelated.

$$\text{the covariance of } b_i\text{'s is: } U^T C U = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ . & . & . \\ . & . & . \\ . & . & . \\ 0 & 0 & \lambda_K \end{bmatrix}$$

- The covariance matrix represents only second order statistics among the vector values.

- Since the new variables are linear combinations of the original variables, it is usally difficult to interpret their meaning.

• **How to choose the principal components?**

- To choose K, use the following criterion:

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} > Threshold \quad \text{(e.g., 0.9 or 0.95)}$$

• **What is the error due to dimensionality reduction?**

- We saw above that an original vector x can be reconstructed using its the principla components:

$$\hat{x} - \bar{x} = \sum_{i=1}^{K} b_i u_i \text{ or } \hat{x} = \sum_{i=1}^{K} b_i u_i + \bar{x}$$

- It can be shown that the low-dimensional basis based on principal components minimizes the reconstruction error:

$$e = ||x - \hat{x}||$$

- It can be shown that the error is equal to:

$$e = 1/2 \sum_{i=K+1}^{N} \lambda_i$$

• **Standardization**

- The principal components are dependent on the *units* used to measure the original variables as well as on the *range* of values they assume.

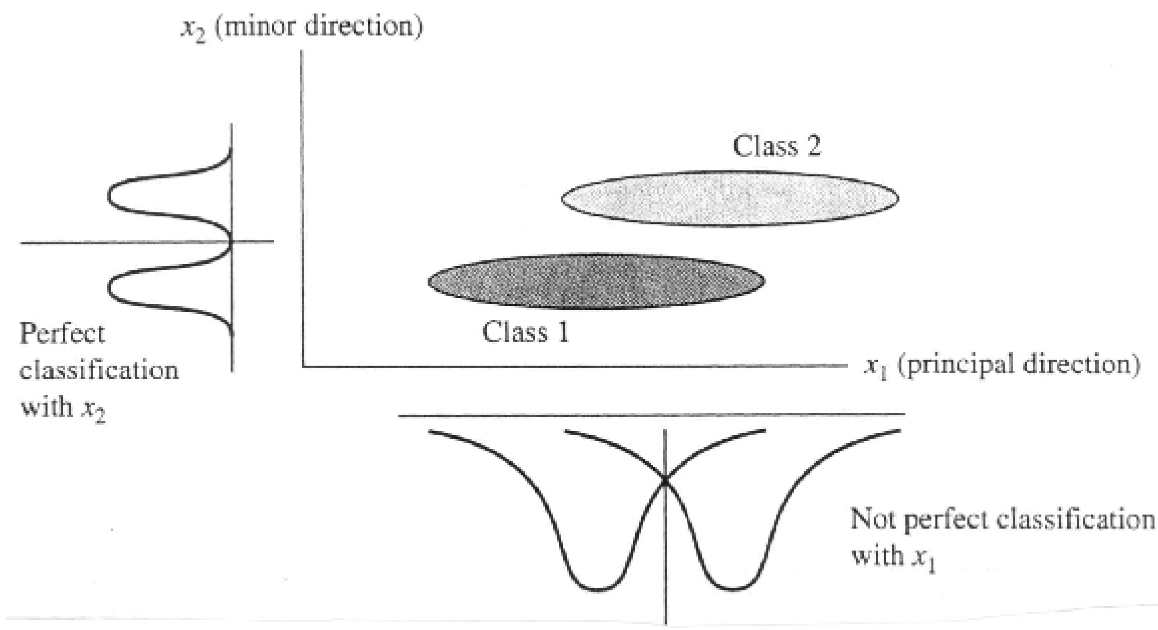
- We should always standardize the data prior to using PCA.

- A common standardization method is to transform all the data to have zero mean and unit standard deviation:

$$\frac{x_i - \mu}{\sigma} \quad (\mu \text{ and } \sigma \text{ are the mean and standard deviation of } x_i\text{'s})$$

• PCA and classification

- PCA is **not** always an optimal dimensionality-reduction procedure for classification purposes:



• Other problems

- How to handle occlusions?

- How to handle different views of a 3D object?