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).


• Case Studies


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.

\[ x = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} \rightarrow \text{reduce dimensionality} \rightarrow y = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix} \quad (K << N) \]

- 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.

\[
\begin{align*}
    b_1 &= t_{11}a_1 + t_{12}a_2 + \ldots + t_{1N}a_N \\
    b_2 &= t_{21}a_1 + t_{22}a_2 + \ldots + t_{2N}a_N \\
    &\vdots \\
    b_K &= t_{K1}a_1 + t_{K2}a_2 + \ldots + t_{KN}a_N
\end{align*}
\]

or \( y = Tx \) where \( T = \begin{bmatrix} t_{11} & t_{12} & \ldots & t_{1N} \\ t_{21} & t_{22} & \ldots & t_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ t_{K1} & t_{K2} & \ldots & t_{KN} \end{bmatrix} \)
• **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 \(x = \hat{x}\)

### Example

\[
\begin{align*}
v_1 &= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad v_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \text{(standard basis)} \\

x_v &= \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = 3v_1 + 3v_2 + 3v_3
\end{align*}
\]

\[
\begin{align*}
u_1 &= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad u_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad u_3 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \quad \text{(some other basis)} \\

x_u &= \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = 0u_1 + 0u_2 + 3u_3
\end{align*}
\]

thus, \(x_v = x_u\)
• **Information loss**

  - Dimensionality Reduction **implies** Information Loss !!
  
  - Preserve as much information as possible, that is,

    \[
    \text{minimize } \|x - \hat{x}\| \text{ (error)}
    \]

• **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 \(N\times1\) vectors

  \[
  \text{Step 1: } \bar{x} = \frac{1}{M} \sum_{i=1}^{M} x_i
  \]

  \[
  \text{Step 2: subtract the mean: } \Phi_i = x_i - \bar{x}
  \]

  \[
  \text{Step 3: form the matrix } A = [\Phi_1 \Phi_2 \cdots \Phi_M] \quad (N\times M \text{ matrix}), \text{ then compute:}
  \]

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

  (sample covariance matrix, \(N\times N\), characterizes the scatter of the data)

  \[
  \text{Step 4: compute the eigenvalues of } C: \lambda_1 > \lambda_2 > \cdots > \lambda_N
  \]

  \[
  \text{Step 5: compute the eigenvectors of } C: u_1, u_2, \ldots, u_N
  \]

  - 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

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

• **Linear transformation implied by PCA**

- The linear transformation $R^N \rightarrow R^K$ that performs the dimensionality reduction is:

$$\begin{bmatrix} b_1 \\ b_2 \\ ... \\ b_K \end{bmatrix} \begin{bmatrix} u_{1}^T \\ u_{2}^T \\ ... \\ u_{K}^T \end{bmatrix} (x - \bar{x}) = U^T (x - \bar{x})$$

• **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.

![Diagram of PCA projection](image)

• **Properties and assumptions of PCA**

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

  \[
  \begin{bmatrix}
  \lambda_1 & 0 & 0 \\
  0 & \lambda_2 & 0 \\
  \cdot & \cdot & \cdot \\
  \cdot & \cdot & \cdot \\
  0 & 0 & \lambda_K
  \end{bmatrix}
  \]

  The covariance of $b_i$’s is: $U^T CU$

- 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 usually difficult to interpret their meaning.
• **How to choose the principal components?**

  - To choose $K$, use the following criterion:

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

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

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

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

• **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:

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

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

    ![Diagram showing PCA and classification](image)

• **Other problems**

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