SIGNATURE VERIFICATION ENTAINING PRINCIPAL COMPONENT ANALYSIS AS A FEATURE EXTRACTOR

Er. Mandeep Kaur Randhawa, G.N.Khalsa College, Karnal, Haryana; Dr. A.K Sharma, N.D.R.I (deemed University), Karnal, Haryana

Abstract

In analysis of biometrics data, a unique challenge arises from the high dimensionality of measurements. Principal component analysis (PCA) is a classic dimension reduction approach. It constructs linear combinations of feature extractors, called principal components (PCs). PCA is computationally simple and can be realized using many existing software packages like MATLAB. In this article, signature verification is used as a representative example of biometric. In this article, firstly, we review the standard PCA technique and its applications in biometrics data analysis like in signature verification. Then, we introduce several recently proposed PCA-based techniques, including the supervised PCA, sparse PCA and functional PCA. The goal of this article is to make biometrics researchers aware of the PCA technique.

Keywords— PCA, Biometric, signature verification, MATLAB.

Introduction

The Principal Component Analysis (PCA) is one of the most successful techniques that have been used in image recognition and compression. It is a statistical method which comes under the broad title of factor analysis. The purpose of PCA is to reduce the large dimensionality of the data space (observed variables) to the smaller intrinsic dimensionality of feature space (independent variables), which are needed to describe the data economically. This is the case when there is a strong correlation between observed variables. Principal component analysis is appropriate when we have obtained measures on a number of observed variables and wish to develop a smaller number of artificial variables (called principal components) that will account for most of the variance in the observed variables. The principal components may then be used as predictor or criterion variables in subsequent analyses.

PCA is a useful statistical technique that has found application in fields such as signature verification and image compression, and is a common technique for finding patterns in data of high dimension. The jobs which PCA can do are prediction, redundancy removal, feature extraction, data compression, etc. In this paper, we will discuss about the PCA; its advanced techniques like supervised PCA, functional PCA and sparse PCA; advantages; limitations. An applications of PCA in biometrics will also be discussed like signature verification. The details are described in the following sections. Section 2 describes the introduction about PCA. Section 3 describes the mathematical formulation of PCA and Matlab code for the same. Section 4 gives describes the applications of PCA in signature recognition. Finally, some limitations of it are discussed.

PCA

PCA was invented in 1901 by Karl Pearson [1]. It is mostly used as a tool in exploratory data analysis and for making predictive models. PCA can be done by eigen value decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after mean centering the data for each attribute. The results of a PCA are usually discussed in terms of component scores (the transformed variable values corresponding to a particular case in the data) and loadings (the weight by which each standardize original variable should be multiplied to get the component score) [2]. PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analysing data. The other main advantage of PCA is that once you have found these patterns in the data, and you compress the data, i.e., by reducing the number of dimensions, without much loss of information. This technique used in image compression, as we will see in a later section. The success of PCA is due to the following two important optimal properties:
a) Principal components sequentially capture the maximum variability among X, thus guaranteeing, minimal information loss;

b) Principal components are uncorrelated, so we can talk about one principal component without referring to others. Following are the steps to compute PCA by co-variance method:

- Organize the data set
- Calculate the empirical mean
- Calculate the deviations from the mean
- Find the covariance matrix
- Find the eigenvectors and eigenvalues of the covariance matrix
- Rearrange the eigenvectors and eigenvalues
- Compute the cumulative energy content for each eigenvector
- Select a subset of the eigenvectors as basis vectors
- Convert the source data to z-scores
- Project the z-scores of the data onto the new basis

Mathematical Background of PCA

PCA is very useful method to pattern recognition in image processing which is based on statistics and matrix algebra. In all cases, we suppose discrete model of computing. In brief, we can summarize mathematical background of PCA to following steps:

- transform images to vector form (input) and compute the mean
- compute co-variances among vectors and construct a covariance matrix
- From covariance matrix get eigenspace – eigen values and eigenvectors
- assess an optimal threshold \( T \) for choosing the K largest determining components

Generally PCA is a transform from correlated data to other uncorrelated data and dimensionality reduction. At the first step, we must express each image as a vector of equal size. Each 2-D image is accordingly represented like a 1-D long vector pixel by pixel of brightness values. Images from ultrasound are naturally in grayscale (R=G=B) and we will get vector of brightness values. It is a general input to PCA. Each vector has the following form:

\[ V_i = (x_{i1}, x_{i2}, ..., x_{im}; m \times n) \]  

where index \( i \) is i-th images in set and \( m \times n \) is a resolution of image. Number of vectors is number of images in collection. Second step is the computing centered data, from each vector is subtracted the arithmetic mean of each vector (dataset). We will need it to compute covariances. Formally we express:

\[ x = \frac{1}{n} \sum_{i=1}^{n} x_i \]  

Next processing will be focused on computing of covariances and from these covariance’s we construct a covariant matrix. This step is critical for next parts. From definition of covariance follows that covariant matrix is real and symmetric. In probability theory and statistics, covariance is a measure of how much two variables change together. Covariance is a kind of variance for 2 or more datasets (vectors). Variance is only for 1 dataset and covariance is simply extension for 2 or more sets. Covariance’s are symmetric too, exactly expressed by \( \text{cov}(X, Y) = \text{cov}(Y, X) \) for datasets X and Y. Thus covariance for \( n \) finite number of datasets is denoted by

\[ \text{cov}(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) - (x_{i1} - \bar{x}_1) \]  

where \( \bar{x}_1 \) is arithmetic mean of Xi dataset, \( n \) is number of images. \( \bar{x}_1 \) is the arithmetic mean (Equation (2)) of Xi dataset. The covariance is also used for a correlation coefficient in neural networks to compute MSE (Mean Square Error). Correlation coefficient is the one of the most important indicators in statistics. Important fact is that covariance matrix has identical dimension like input vectors. Zero covariance would indicate that the two variables are independent of each other. PCA is based on computing of eigen space that is eigen values and corresponding eigenvectors. We constructed covariant matrix and now we can compute the eigen space. Because the covariant matrix is real and symmetric, we can simply compute an eigen space. Number of eigen values is equal to number of input vectors. In practice, if we have 20 input vectors (images), then covariance matrix has a dimension 20x20 and so 20 eigen values with their corresponding eigenvectors. Computed eigen values are in descending order, \( \lambda_1 > \lambda_2 > .. > \lambda_n \)

PCA is the simplest of the true eigenvector-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way which best explains the variance in the data. If a multivariate dataset is visualised as a set of coordinates in a high-dimensional data
space (1 axis per variable), PCA can supply the user with a lower-dimensional picture, a "shadow" of this object (Fig.1) when viewed from it's (in some sense) most informative viewpoint. This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced.

**Fig.1 PCA of multivariate Gaussian distribution centered [adopted from [3]].**

**MATLAB code for PCA [4]**

Function [signals, PC, V] = pca1 (data)

% PCA1: Perform PCA using covariance.
% data - M×N matrix of input data
% (M dimensions, N trials)
% signals - M×N matrix of projected data
% PC - each column is a PC
% V - M×1 matrix of variances

[M, N] = size (data);
% subtract off the mean for each dimension
mn = mean(data,2);
data = data - repmat(mn,1,N);
% calculate the covariance matrix
covariance = 1 / (N-1) * data * data';
% finds the eigenvectors and eigenvalues
[PC, V] = eig (covariance);
% extract diagonal of matrix as vector
V = diag (V);
% sort the variances in decreasing order
[Junk, rindices] = sort (-1*V);
V = V (rindices); PC = PC (:, rindices);
  a. % project the original data set
  b. signals = PC' * data;

**Advantages of PCA**

1. Smaller representation of database because we only store the training images in the form of their projections on the reduced basis.

2. Noise is reduced because we choose the maximum variation basis and hence features like background with small variation are automatically ignored.

   1. **PCA based techniques**

   Following are the PCA based techniques:

   - **Supervised PCA**

     Components of supervised Principal are:

     a. Compute (univariate) standard regression coefficients for each feature.

     b. Form a reduced data matrix consisting of only those features whose univariate.

     c. Co-efficient exceeds a threshold in absolute value (is estimated by cross-validation).

     d. Compute the first (or first few) principal components of the reduced data matrix.

     e. Use these principal component(s) in a regression model to predict the outcome [5].

   - **Sparse PCA**

     Sparse PCA (SPCA) is built on the fact that PCA can be written as a regression-type optimization problem, thus the lasso (elastic net) can be directly integrated into the regression criterion such that the resulting modified PCA produces sparse loadings [6]. Though wildly used in many fields of applications (medical engineering, statistical signal processing, financial data modelling, psychology) the choice of the (ideally "small") number of principal components to include into the description of the data without losing too much information is somewhat arbitrary and mostly only based on empirical criteria. The same problem arises in a related question: how many and which of the original components of the multivariate data vector should be taken into account when trying to identify the most relevant contributions to the above-mentioned linear combination (making up the first PCA, the second, and so on)?

   - **Functional PCA**

     In Functional Principal Components Analysis (FPCA) the dimension reduction can be achieved via the same route: Find ortho-normal weight functions γ1, γ2,..., γ n such that the variance of the linear transformation is maximum.
Application in Signature verification

Signature has been a distinguishing feature for person identification through ages. Signature verification is a well established of research. Handwritten signatures are socially and legally readily accepted as a convenient means of document authentication. Although handwritten signatures are not the most reliable means of personal identification, signature verification systems are inexpensive and non-intrusive. Signature verification systems are categorized into online systems and offline systems. In online case, a special pen is used on an electronic surface. A signature is captured dynamically and then stored as a function of time. In offline systems, a signature is digitized using a flat bed scanner and then stored as an image. These images are called static or offline signatures. Off-line data is a 2-D image of the signature. Offline systems are of interest in scenarios where only hard copies of signature are available, e.g., where a large number of documents need to be authenticated [7]. A novel online signature verification method is also presented [8] that use PCA for dimensional-reduction of signature snapshot. The resulting vectors from PCA are submitted to a multilayer perceptron (MLP) neural network with EBP and sigmoid activation function. In the other hand, Dynamic features such as x, y coordinates, pressure, velocity, acceleration, pen down time, distance, altitude, azimuth and inclination angles, etc. are processed statistically. During enrollment, five reference signatures are captured from each user. One-way ANOVA is used to analyze relative X-Coordinates in 6 groups (5 reference group, 1 testing group). ANOVA test will be repeated for relative Y-Coordinates, pressure value, azimuth and inclination angles. Thus, the algorithm will fill up a vector of five distances (F-scores) between all the possible pairs of testing and reference vectors. The resulting vector is compared to a threshold vector. The database includes 130 genuine signatures and 170 forgery signatures. The verification system has achieved a false acceptance rate (FAR) of 2% and a false rejection rate (FRR) of 5%. The critical analysis of the data glove-based signature identification and forgery detection system emphasizes the essentiality of noise-free signals for input. Lucid inputs are expected for the accuracy enhancement and performance. The raw signals that are captured using 14- and 5-electrode data gloves for this purpose have a noisy and voluminous nature. Reduction of electrodes may reduce the volume but it may also reduce the efficiency of the system. The principal component analysis (PCA) technique has been used for this purpose to condense the volume and enrich the operational data by noise reduction without affecting the efficiency. The advantage of increased discernment in between the original and forged signatures using 14-electrode glove over 5-electrode glove has been discussed here and proved by experiments with many subjects. Calculation of the sum of mean squares of Euclidean distance has been used to project the advantage of our proposed method. 3.1% and 7.5% of equal error rates for 14 and 5 channels further reiterate the effectiveness of this technique [9]. The performance of MLP (multilayer perceptron) network acting as a classifier in signature recognition system for this purpose, a database of 100 signature images of 10 people (10 signatures for each person) is used for training and evaluation. The principal components of images have been used as inputs for purposes network and to decrease the vector’s dimension and to extract eigen pictures(fig.2). The experimental results revealed that MLP network exhibits appropriate performance with recognition accuracy of 98.5%[10].

Limitations of PCA

- The directions with largest variance are assumed to be of most interest.
- We only consider orthogonal transformations (rotations) of the original variables. (Kernel PCA is an extension of PCA that allows non-linear mappings).
- PCA is based only on the mean vector and the covariance matrix of the data. Some distributions (e.g. multivariate normal) are completely characterized by this, but others are not.
- Dimension reduction can only be achieved if the original variables were correlated. If the original variables were uncorrelated, PCA does nothing, except for ordering them according to their variance.
- PCA is not scale invariant.
Conclusion

A useful statistical technique called PCA is described in this paper that has found application in fields such as image compression. This is an effective dimension reduction technique that can be used in biometric data analysis. In this paper, signature verification is used as a representative example of biometric and the analysis techniques discussed are also applicable to other types of biometrics studies. This technique is a step towards improving the current situation in the biometric applications.

References


Biographies

Er. Mandeep Kaur Randhawa received the M.Sc. degree in Computer Science from the Kurukshetra University, Kurukshetra, Haryana, in 2009 and M.Tech degree in 2012.. Currently, she is working in Dept. of Computer Science and Applications as Asst. Professor at Guru Nanak Khalsa College, Karnal, Haryana. Her teaching and research areas include pattern recognition, signature verification, neural network. She has authored more than 11 research papers in international journals and conferences. She is the member of IACSIT (IACSIT is a registered international Association of Computer Science and Information Technology).She has authored two textbooks: Basic Computer Education (for B.A) and Practical file of BA, B.Sc, PGDCA, B.Com. She may be reached at mandeeprandhawa85@yahoo.com.

Dr. A.K Sharma received the PhD degree in Computer Science from Thapar University, Patiala, Punjab, in 2006, respectively. Currently, He is Senior Scientist and Center Head of Computer Science Department, at National Dairy Research Institute (NDRI)-deemed university. He has authored more than 20 research papers in international journals and conferences. He remained the Key note speaker of many international conferences He has authored four textbooks. He is the fellow member of reputed international journal. He may be reached at adesh_kumar.sharma@gmail.com.