2011 International Conference on Life Science and Technology
IPCBEE vol.3 (2011) © (2011) IACSIT Press, Singapore

An Introduction to Principal Component Analysis and Its Importance in Biomedical Signal Processing

Varun Gupta
Dept. of Instrumentation & Control Engg.
National Institute of Technology Jalandhar,
Jalandhar, India
varun_gupta793@indiatimes.com

Ramveer Singh
Dept. of Computer Science & Engg.
Babu Banarasi Das Institute of Technology
Ghaziabad, India
sahil70itian@gmail.com

Gavendra Singh
Dept. of Instrumentation & Control Engg.
National Institute of Technology Jalandhar,
Jalandhar, India
jadaungs@gmail.com

Rajvir Singh
Dept. of Instrumentation & Control Engg.
National Institute of Technology Jalandhar,
Jalandhar, India
dhanoarajvir@gmail.com

Harsimran Singh
Dept. of Instrumentation & Control Engg.
National Institute of Technology Jalandhar,
Jalandhar, India
harsimrans329@gmail.com

***Abstract*-Principal Component Analysis is used where lots of data, all very confusing, too many variables to consider exists, some of them are probably insignificant. PCA was invented in 1901 by Karl Pearson. It has some basic assumptions i.e. Linearity, Large variances, the principal components are orthogonal. In addition, for PCA to work exactly, one should use standardized data so that the mean is zero. Principle Component Analysis (PCA) is commonly used techniques for data classification and dimensionality reduction.**

Index-Terms- Linearity, Large variances, principal components, dimensionality reduction.

I. INTRODUCTION

PCA is a simple, non-parametric method for extracting relevant information from confusing data sets. PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. PCA is a special case of Factor Analysis that is highly useful in the analysis of many time series and the search for patterns of movement common to several series (true factor analysis makes different assumptions about the underlying structure and solves eigenvectors of a slightly different matrix)[1].

$$\text{Variation Explained by Each Principal Component} = \frac{\text{Sum of Eigen Value}}{\text{Number of variables}} \quad (1)$$

II. BENEFIT OF PCA

A primary benefit of PCA arises from quantifying the importance of each dimension for describing the variability of a data set. PCA can also be used to compress the data, i.e. by reducing the number of dimensions, without much loss of information.

III. ACQUISITION OF DATA

Six healthy subjects {all males (2 non-smoker+4 smoker) and with no prior history of cardiovascular disease} aged between 20 and 25 took part in the experiments after giving the informed consent. All of the experiments were performed at the same university laboratory with the room temperature being maintained at about 20 degrees centigrade during the afternoon time (from 2:00 to 4:00 pm). The subjects were required to have a resting period of at least 5 minutes under relaxation laying on bed for acquiring Electrocardiogram (ECG). The Multifunctional physiological data acquisition system MP35 (Biopac System Inc.) was utilized for signal measurement for ECG and Respiratory signals (by module SS2LA). Airflow Transducer SS11LA & Temperature Transducer SS6L are also used for Respiratory Signal Acquisition.The user friendly analysis package Biopac Student Lab 3.7.6 was used for the signal measurement and Biopac Student Lab PRO 3.7.6 for management, including the signal quality pre-screening, data storage and retrieval. The sampling frequency was 500 Hz for ECG & 1000 Hz for Respiratory signal with hanning window.The signals were

verified visually by a well-trained technician. If the signal quality was poor, the signal would be excluded from further analysis and the subject was asked to repeat the experiment once again.

Figure.1 ECG of all six Subjects

IV. ANALYSIS OF ACQUIRED DATA

The analysis was executed after the experiments were finished and approved. The source code for ECG and Respiratory signal Principal component analysis was developed in MATLAB(R2010a) (MathWorksInc.).Firstly we got data from subjects using BSL3.7.6 software with help od MP35. Secondly, we open the files (ECG & Respiratory data) in BSL PRO 3.7.6, then we got the data of ECG and Respiratory in text format. Finally we analyse the text data using the appropriate Mat-lab. source code. We got the Two PC's(principal components)of Respiratory and ECG signal plot for the subject those are participating in the experiment. The analysis of acquired data Table(1) is done through simply the mathematical expressions such as eigen values and eigen vectors.

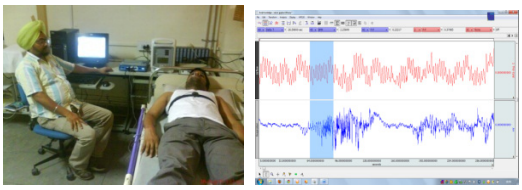

Figure.2 Acquiring Respiratory data of Sinus Patient

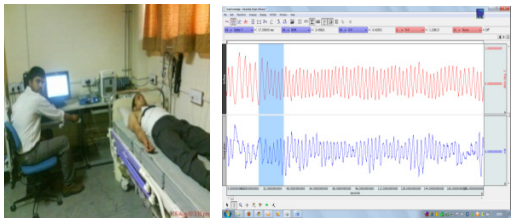

Figure.3 Acquiring Respiratory data of normal person

Table-1

Airflow	ECG
-0.146484	-0.887146
-0.147502	-0.567932
-0.148519	-0.37384
-0.149536	-0.271606
-0.150553	-0.214844
-0.151571	-0.197144
-0.151571	-0.191956
-0.149536	-0.178833
-0.148519	-0.172729
-0.147502	-0.172424
-0.146484	-0.173035
-0.146484	-0.17334
-0.145467	-0.167236
-0.145467	-0.15564
-0.14445	-0.139465
-0.143433	-0.121765
-0.142415	-0.10376
-0.142415	-0.0845337
-0.141398	-0.0686646
-0.140381	-0.0561523
…….	…….

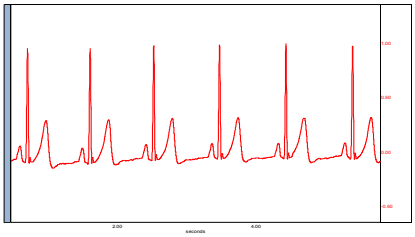

Figure.4 Acquired ECG data of Sinus Patient

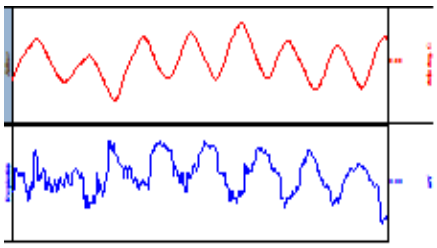

Figure.5 Acquired Respiratory data of Sinus Patient

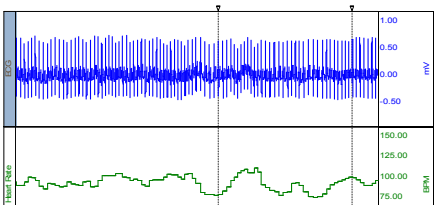

Figure.6 Acquired Respiratory data of normal person

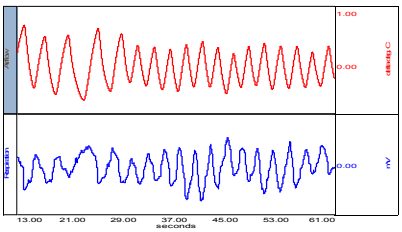

Figure.7 Acquired Respiratory data of normal person

The general processor of analysis is as follows

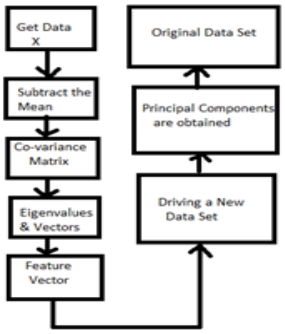

Figure.8 General Processor of Principal Component analysis

A. Acquisition of Data

For signal acquisition we require well established lab setup. In this paper we have done the analysis of two signals (e.g. airflow & ECG) through Principal component analysis. First of all we are acquiring the data through MP36 BIOPAC machinery and after that we are analysing the signal through BIOPAC Acknowledge software.

Figure.9 MP36 BIOPAC SYSTEM

B. Adjust the Data

Adjust the acquired data(signal) simply by subtracting the mean of the particular data from the acquired data .

C. Find the Covariance Matrix

Covariance is always measured between two dimensions. Covariance measures how much the dimensions vary from the mean with respect to one another. If we calculate the covariance between one dimension and itself, we will get the variance of that dimension. The covariance matrix describes all relationships between pairs of measurements in the considered data set.

$$S_x = \frac{1}{n-1} XX^T \qquad (2)$$

The basic formula for Covariance is expressed as

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)} \qquad (3)$$

Where X and Y represents two separate dimensions of data.

D. Find the Eigen Values & Eigen Vectors(Feature vector)

Let, A = n x n matrix. The scalar **λ** is an Eigenvalue of A if there exists a non-zero vector v such that,

$$Av = \lambda v \qquad (4)$$

Where Vector v is called an eigenvector of A corresponding to eigenvalue **λ.**For each eigenvalue **λ**, the set of all vectors v satisfying Av = **λ**v is called eigen space of A corresponding to eigenvalue **λ.**

We can rewrite the condition Av = **λ**v as,

$$(A - \lambda I)\, v = 0 \qquad (5)$$

I : n x n identity matrix.

For a non-zero vector v to satisfy the above eqn., matrix (A- **λ**I) must not be invertible.

Non-invertible → determinant of (A- **λ**I) must be zero.

$$p(\lambda) = \det(A - \lambda I) \qquad (6)$$

is the characteristic polynomial of A.The eigenvalues of A are simply the roots of this characteristic polynomial.

To find eigenvectors, $v = \begin{pmatrix} v_1 \\ v_2 \\ \cdot \\ v_n \end{pmatrix}$

Corresponding to an eigenvalue **λ**, we solve the system of linear equations given by,

$$(A - \lambda I)\, v = 0 \qquad (7)$$

E. Find the Row Feature Vector

We can easily find Row feature vector, it is just the transpose of Eigen vectors matrix.

F. Find the New Data Set

NEW (FINAL) DATA =ROW FEATURE VECTOR * ROW DATA ADJUST

Row data adjust is also the transpose of the estimated values i.e. adjust data values.

V. CALCULATION OF DIFFERENT STEPS

Step-1 Calculate the Mean of the two signals

In this we have considered arbitrary data of Physiological Signals Airflow(X), ECG(Y) for understanding i.e. how to calculate Principal Components.

Input Data X (e.g Airflow)	Input Data Y (e.g ECG)
3.5	3.4
1.5	1.3
2.1	1.2
1.3	0.8
0.7	0.1
0.91	1.6
1.31	1.9
1.56	1.2
2.34	0.78

Sum of X=15.22 **Sum of Y=12.28**

Mean of X=1.6911 **Mean of Y=1.364**

Step-2 Subtract the Mean from the Original Signal (Data)

X-Mean of X	Y-Mean of Y
1.8089	2.0356
-0.1911	-0.0644
0.4089	-0.1644
-0.3911	-0.5644
-0.9911	-1.2644
-0.7811	0.2356
-0.3811	0.5356
-0.1311	-0.1644
0.6489	-0.5844

Step-3 Calculation of covariance

To calculate the cov= $\begin{pmatrix} \textbf{Cov(x,x)} & \textbf{Cov(x,y)} \\ \textbf{Cov(y,x)} & \textbf{Cov(y,y)} \end{pmatrix}$

This is further defined as:

$$\textbf{Cov(x,y)= } \frac{\textbf{summation of (x-mean(x))(y- mean(y))}}{\textbf{(n-1)}}$$

Similarly for cov(x,x) and cov(y,y).

summation of (x-mean(x))(x-mean(x))	summation of (x-mean(x))(y-mean(y))	summation of (y-mean(y))(x-mean(x))	summation of (y-mean(y))(y-mean(y))
3.2721	3.6821	3.6821	4.1436
0.03651	0.01230	0.01230	0.0041473
0.1671	-0.0672	-0.0672	0.02702
0.1529	0.2207	0.2207	0.3185
0.9822	1.2531	1.2531	1.5987
0.6101	-0.1840	-0.1840	0.0555
0.1452	-0.2041	-0.2041	0.2868
0.01718	0.02155	0.02155	0.02702
0.4210	-0.3792	-0.3792	0.3415
summation of (x-mean(x))(x-mean(x))= **5.80429**	summation of (x-mean(x))(y-mean(y))= **4.35525**	summation of (y-mean(y))(x-mean(x))= **4.35525**	summation of (y-mean(y))(y-mean(y))= **6.80278**

Thus cov(x,x)=5.80429/8=0.72553

Cov(x,y)=cov(y,x)=4.35525/8=0.54440

Cov(y,y)=6.80278/8=0.85034

$$\text{Cov}=\begin{pmatrix} 0.72553 & 0.54440 \\ 0.54440 & 0.85034 \end{pmatrix}$$

Step-4 Calculation of Eigen Values & Eigen Vectors from covariance Matrix

$$D=\begin{bmatrix} 1.3359 & 0 \\ 0 & 0.2400 \end{bmatrix}$$

$$V=\begin{bmatrix} 0.6656 & -0.7463 \\ 0.7463 & 0.6656 \end{bmatrix}$$

Step-5 Calculation of new data set

NEW (FINAL) DATA =ROW FEATURE VECTOR * ROW DATA ADJUST

$$NFD=\begin{bmatrix} 0.6656 & 0.7463 \\ -0.7463 & 0.6656 \end{bmatrix}\times\begin{bmatrix} 1.8089 & -0.1911 & \cdots & 0.6489 \\ 2.0356 & -0.0644 & \cdots & -0.5844 \end{bmatrix}$$

Step-6 Principal Components Calculation

PCA_1=[2.7232 -0.1753 0.1495 -0.6815 -1.6033 -0.3441 0.1461 -0.2100 -0.0042]

PCA_2=[0.0049 0.0998 -0.4146 -0.0838 -0.1019 0.7398 0.6409 -0.0116 -0.8733]

VI. RESULTS AND DISCUSSION

Here we used the coefficients of the first principal component to derive the respiratory signal but there were also respiratory related changes in the coefficients of some of the lower principal components. This is illustrated in figure 17 where the coefficients of not only the first but also the second principal component clearly show respiratory related changes.

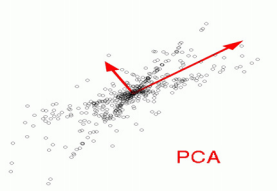

Figure.10 Two Directions of PCA

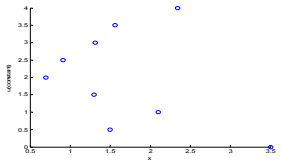

Figure.11 Acquired X Data (e.g. Airflow)

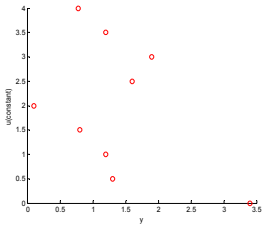

Figure.12 Acquired Y Data (e.g. Heart Rate)

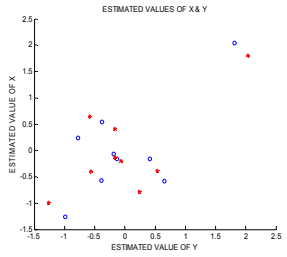

Figure.13 Adjusted values of X(Airflow)&Y(ECG) signals

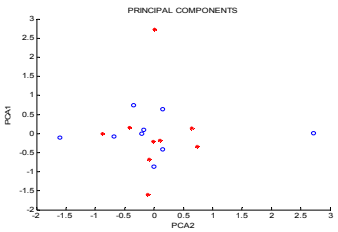

Figure.14 Resulting Principal Components (PCA1 & PCA2)

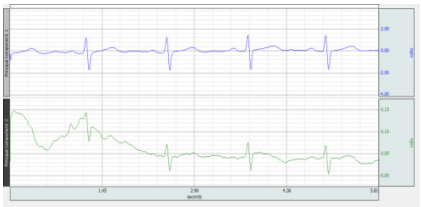

Figure.15 Showing ECG as a respiratory information-carrying signal with first Two PC's(Principal Components)

Eigen values

0.114581 0.000945581

Eigenvector Matrix

$$\begin{bmatrix} 0.0208535 & -0.999783 \\ 0.999783 & 0.0208535 \end{bmatrix}$$

VII. CONCLUSION

It is easy to see that the first principal component is the direction along which the samples show the largest variation. The second principal component is the direction uncorrelated to the first component along which the samples show the largest variation. we have transformed our data so that is expressed in terms of the patterns between them, where the patterns closely describe the relationships between the data. We can define PCA as a meaningful graphical display of model outputs. It has an applicability to both continuous and batch processes. We have shown that in PCA, diagonalization of the covariance matrix results in computation of model parameters directly.

VIII. FUTURE CHALLENGE FOR PCA

PCA to work exactly, PCA should use standardized data so that the mean is zero. But, there is one major drawback to standardization. Standardizing means that PCA results will come out with respect to standardized variables. This makes the further applications of PCA results even more difficult.

REFERENCES

[1] Wold, S., Esbensen, K., Geladi, P. (1987). Principal Components Analysis. Chemometrics and Intelligent Laboratory Systems. 2, 37-55.

[2] Nelson, P, Mac-Gregor, J. F., Taylor, P. A., (1996). Missing Data Methods in PCA and PLS: Score Calculations with Incomplete Observations. Chemometrics and Intelligent Laboratory Systems, 35.

[3] Kourti, T., MacGregor, J.F., (1995). Process Analysis, Monitoring, and Diagnosis using Multivariate Projection Methods - A Tutorial. Chemometrics and Intelligent Laboratory Systems, 28, 3-21.

[4] Sharvan Kumar Pahuja, Varun Gupta, Gavendra Singh,Manish Mittal "Fourier Transform of Untransformable signals using pattern recognition technique"ACT-2010,Indonesia.

[5] M Emdin, A Taddei, M Varanini "Compact Representation of Autonomic Stimulation on Cardiorespiratory Signals by Principal Component Analysis".

[6] Xuan Zhaoyan , Xie Shiman,Sun Qiuyan "The Empirical Mode Decomposition Process of non-stationary signals".

[7] Ruqiang Yan and Robert X. Gao "A Tour of the Hilbert-Huang Transform: An Empirical Tool for Signal Analysis".

[8] Varun Gupta, Manish Mittal , Gavendra Singh, Reena Tyagi" Time-Frequency Description of Signum Function using Principal Component & Linear Discriminant Analysis ", ICCMS-2011,Mumbai.

[9] Zhongchao Huang "A Novel spectral Analysis Method of Atrial Fibrillation Signal Based on Hilbert-Huang Transform".

[10] Eyal Gottlieb, Sean M. Armour "Mitochondrial respiratory control is lost during growth factor deprivation."

[11] Angel Otero " Palliative Performance Status, Heart Rate and Respiratory Rate as Predictive Factors of Survival Time in Terminally Ill Cancer Patients".

[12] Tuan Anh "Principal Components Analysis (PCA)", Financial Pricing Midterm Presentation,2000.