Acta Scientific MEDICAL SCIENCES (ISSN: 2582-0931)

Volume 9 Issue 6 June 2025

Review Article

Biostatistical Analysis of Heart Diseases: Compression of Bio-Data using Principal Component Analysis

Carlos Alvarez Picaza*, Alberto Daniel Valdéz, Paola Luciana Schlesinger and Juan Ángel Chiozza

Universidad Nacional del Nordeste, Argentina

***Corresponding Author:** Carlos Alvarez Picaza, Universidad Nacional del Nordeste, Argentina.

DOI: 10.31080/ASMS.2025.09.2094

Received: April 01, 2025

Published: May 20, 2025

Abstract

Bioengineering is a branch of engineering that studies, among other things, the quantification of biological phenomena, such as the conductivity of blood and tissues, the mechanical response to an electrical stimulus, or the study of bioelectrical phenomena. Among these, the analysis of the cardiac signal, the electrocardiogram (ECG), is included. The electrocardiographic signal is the most studied biological signal in the world; despite this, there is no automated method that allows for the classification of the wave or signal to identify a normal heartbeat from an abnormal one. In order to identify and classify anomalous patterns, the routine analysis of the clinical ECG almost exclusively depends on visual inspection. Currently, efforts are being made to find a methodology that brings precision in determining normal beats from aberrant ones. The use of tools provided by biostatistics will allow this work to address the analog-digital processing of biopotentials. The optimization in information processing is fundamental for obtaining relevant conclusions. Using digital techniques, the conditioning of signals will be sought for the identification and classification of patterns. The current paper will aim at the treatment and analysis of electrical potentials from biosignals for their subsequent processing through biostatistics, using data compression tools such as Principal Component Analysis (PCA). In order to identify and classify anomalous patterns in various cardiac pathologies, an attempt is made to find a methodology that brings precision when determining more accurate diagnoses.

Keywords: PCA; Correlation; Variance

Introduction

When collecting information from a data sample, it is most common to take as many variables as possible. However, if we take too many of them over a set of objects, we will have to consider many possible correlation coefficients, and it increases if we consider an even greater number of variables.

Another problem that arises is the strong correlation that often exists between the variables. If we take too many (which generally happens when we don't know much about the data or are just being exploratory), it is normal for them to be related or to measure the same thing from different perspectives. For example, in medical studies, the blood pressures at the exit of the heart and at the exit of the lungs are strongly related.

It is therefore necessary to reduce the number of variables. It is important to highlight the fact that the concept of greater information is related to that of greater variability or variance. The greater the variability of the data (variance), the more information is considered to exist. Principal Component Analysis or PCA (Principal Component Analysis) is a statistical technique for information synthesis or dimensionality reduction (number of variables). In databases with many variables, the PCA technique allows for reducing the number of such variables without losing substantial information.

It is well known that random (or trivial) PCs can be estimated from datasets without any correlational structure between the original variables due to sampling error, especially when the number of observations is low in relation to the number of variables [1].

As expressed by M. Sudharsan and G. Thailambal [2] in their study on predicting Alzheimer's disease, the most representative measures of the details are preserved, while the smaller ones are omitted. PCA produces new properties that are a linear composition of the preliminary features and vectors, in a d-dimensional domain. A key aspect of Principal Component Analysis is the interpretation of the factors, which is not given a priori, but is deduced after observing the relationship between the results and the initial variables. The fundamental purpose of the technique is to reduce the dimensionality of the data in order to simplify the problem under study.

Recently, PCA has been used to improve the accuracy of medical diagnoses, focusing data compression into a matrix and measuring the distance between PCA data and the reference data sequence [3].

Yalin., *et al*. solved the limitations of traditional continuous state detection methods in the treatment of biological electric potentials using Principal Component Analysis, even extending this procedure to industrial applications and energy generation [4].

There are specific software programs that help simplify the development of work (XLSTAT, HOMER, MATLAB).

Methodology

Principal Component Analysis is a method that reduces the dimensionality of data by performing a covariance analysis between factors [5].

In many applications, a set of n objects is represented through a collection of m descriptors, indices, or parameters. In some cases, m is a very large number, which makes it difficult to analyze the dataset in its entirety, meaning that the n objects can be considered as n points located in an m-dimensional space. The objective is to classify those objects and represent them in a lower-dimensional space p (p < m), in such a way that the projection in that space is optimal.

Concepts such as standard deviation, covariance, eigenvectors, and eigenvalues are fundamental for a detailed description of PCA's functioning [6].

Development

The first numerical treatment that must be done is to scale the descriptor columns of matrix A. This is because each column (each variable) may be specified in a different system of units. In fact, each variable does not have to be of the same nature as the others. There are several scaling possibilities. The most common one consists of obtaining centered and dimensionless normalized column vectors, thus each column aj of the matrix A,

$$\mathbf{A} = (\mathbf{a}_1 \quad \mathbf{a}_2 \ldots \mathbf{a}_m) \text{ -------- (1)}$$

Its average is calculated

$$\bar{a}_j = \frac{1}{n}\sum_{1}^{n} a_{ij} \text{ ----------- (2)}$$

y standard deviations multiplied by *n*

$$s_j = \sqrt{\sum_{i=1}^{n}(a_{ij} - \bar{a}_j)^2} \text{ ----------- (3)}$$

Obtaining the following matrix of dimensionless variables:

$$\mathbf{A} \to \mathbf{Z} = (\mathbf{z}_1 \quad \mathbf{z}_2 \ldots \mathbf{z}_m) \text{ ------------- (4)}$$

Where each vector column $\mathbf{z}_j$ It is defined from the transformation

$$a_j \to z_j = \frac{a_j - \bar{a}_j}{s_j} \text{ -------------- (5)}$$

The matrix of dimensionless homogenized variables allows you to calculate the matrix of the correlation coefficients between each pair of data columns:

$$\mathbf{R} = \mathbf{Z}^{\mathrm{T}} \ \mathbf{Z} \text{ ----------- (6)}$$

This matrix is of dimension *mxm*.

$$\mathbf{RX} = \mathbf{X\Delta} \quad \text{---------} \ (7)$$

Where

$$\mathbf{X} = (\mathbf{x}_1 \ \ \mathbf{x}_2 \ldots \mathbf{x}_m) \ ; \ \ \mathbf{\Delta} = Diag(\lambda_1 \ \lambda_2 \ \ldots \lambda_m) \text{-----------} \ (8)$$

All eigenvalues are non-negative. Precisely the eigenvalues of this matrix are the parameters that indicate what fraction of the original total variance each new one retains CP.

$$f_i = 100 \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j} \% \ \text{----------------} \ (9)$$

Therefore, the ordering, from highest to lowest, of eigenvalues (eigenvalues) induces an order of preference of the PCs. From now on we will assume that

$$\lambda_1 \geq \lambda_2 \geq \ \ldots \ \geq \lambda_m \ \text{---------} \ (10)$$

Now

$$\mathbf{R} \rightarrow \mathbf{X} = (\mathbf{x}_1 \ \ \mathbf{x}_2 \ldots \ \mathbf{x}_m) \ \_\_\_\_\_\_\_\_\_ \ (11)$$

Where $\mathbf{x}_1$ is the associated self-vector a λ_1, $\mathbf{x}_2$ a λ_2 y and so on until *m*. The First Main Component $\mathbf{x}_1$ represents the highest amount of variance from the original data, $\mathbf{x}_2$ retains the second highest variance, y so on until *m*. To the coefficients of each eigenvector $\mathbf{x}_j$ they are called weights (loadings) and indicate which linear combinations of the original variables must be constructed to define the new dimensionless coordinates [7]. Most usually, by reducing the dimensionality of the problem.

Results, Progress/Discussion

Data obtained from fourteen hypertensive patients (HTN).

Nº of Patients	Age	Weight	Height	PS	PD	PM	Vop	Cm	DS	DD	DM	etha	ImtCa
	Years	Kg	m	mmHg	mmHg	mmHg	m/s	e-4 cm/ mmHg	mm	mm	mm	mmHg s/mm	mm
1	55	73	1,63	146	96	113	15,21	2,00	7,70	7,32	7,50	5,07	0,73
2	63	79	1,72	106	84	91	14,58	2,21	7,42	7,03	7,25	2,92	1,14
3	42	75	1,76	116	65	82	10,07	3,49	5,72	5,32	5,53	4,81	0,75
4	50	83	1,84	157	89	112	11,15	3,76	7,54	7,22	7,37	5,83	0,9
5	60	84	1,86	166	98	121	14,25	2,68	8,75	8,36	8,55	3,99	0,9
6	59	80	1,72	164	92	116	16,26	2,22	9,33	8,85	9,11	3,73	0,85
7	50	106	1,82	127	82	97	11,44	3,76	7,92	7,50	7,71	4,66	0,71
8	64	83	1,76	155	92	113	17,16	2,48	7,83	7,60	7,72	9,17	0,82
9	38	58	1,63	155	70	98	10,83	3,70	7,11	6,69	6,89	4,52	1,08
10	40	81	1,72	139	100	113	11,28	4,04	8,15	7,95	8,05	9,28	0,79
11	54	93	1,74	134	84	101	10,27	4,82	8,13	7,81	7,97	5,17	0,76
12	45	83	1,78	114	75	88	11,31	3,12	6,63	6,10	6,30	2,18	0,89
13	54	72	1,65	117	78	91	14,07	2,31	7,44	6,94	7,19	2,39	0,98
14	62	80	1,72	125	83	97	14,78	1,68	5,81	5,60	5,71	8,42	1,04

Table 1: Total data corresponding to the cardiac activity of fourteen (14) patients.

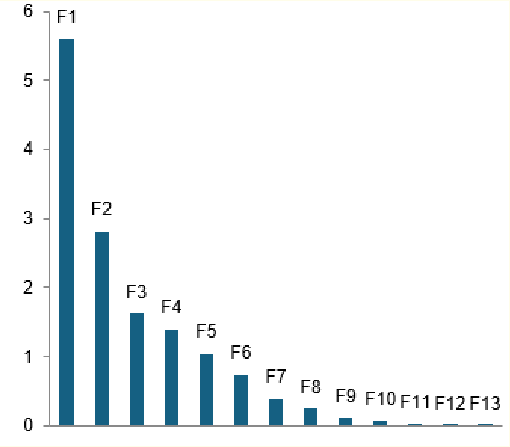

Figure 1: Influence of Eigenvalues.

	M. Value	S. Deviation
Nº of patient	7,500	4,031
Age	52,571	8,398
Weight Kg	80,714	10,368
Height m	1,739	0,069
PS mmHg	137,307	19,360
PD mmHg	84,857	10,063
PM mmHg	102,341	11,598
Vop m7s	13,047	2,292
Cm e-4 cm/mmHg	3,019	0,890
DS mm	7,533	0,963
DD mm	7,165	0,960
DM mm	7,348	0,964
etha mmHg s/mm	5,152	2,235
ImtCa mm	0,881	0,131

Table 2: Mean Value and Standard Deviation of the columns.

Eigenvalues	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
Eigenvalue	5,598	2,806	1,619	1,386	1,034	0,739	0,391	0,242	0,109	0,064	0,011	0,000	0,000
% variance	39,985	20,045	11,568	9,899	7,386	5,279	2,792	1,728	0,780	0,460	0,078	0,000	0,000
% accumulated	39,985	60,030	71,598	81,497	88,883	94,162	96,954	98,681	99,462	99,922	100,000	100,000	100,000

Table 3: Principal Components of Heart Patients.

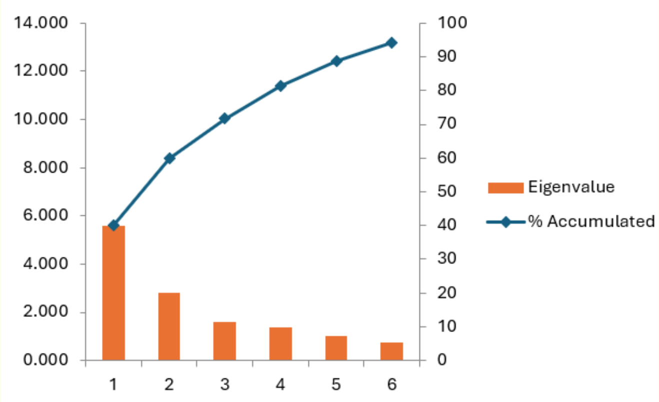

Figure 2: Pareto´s Diagram.

Figure 2 shows the Pareto´s diagram obtained based on the selected Principal Components (factors). The Pareto analysis is an ordered comparison of factors related to a problem. This comparison helps to identify and focus on the few vital factors, distinguishing them from the many useful factors. The application of it allows visually displaying in order of importance, the contribution of each element to the total effect. In the graph, only the first 6 Principal Components are displayed because the weights of the remaining 7 are insignificant compared to the first ones, in which more than 94% of the information from the original matrix is concentrated.

Conclusion

The Principal Component Analysis (PCA) method is effective and allowed us to meet the aforementioned objectives. In the case of the electrocardiographic data, it was possible to reduce a matrix of 13 variables to just 6, capturing 94% of the information contained in the original matrix. With this tool, data redundancy was eliminated to speed up computational times, which constitutes a primary objective in information processing.

Bibliography

1. Camargo A. “PCA test: testing the statistical significance of Principal Component Analysis in R”. *Peer Journal* 10 (2022): e12967.

2. M Sudharsan and G Thailambal. “Alzheimer’s disease prediction using machine learning techniques and principal component analysis (PCA)”. *Proceedings of IVCSM 2K20 (International Virtual Conference on Sustainable Materials)* 81.2 (2023): 87-1176.

3. Tusongjiang K and Wensheng G. “Power transformer fault diagnosis using FCM and improved PCA”. The Journal of Engineering. IET Electrical Engineering Academic Forum (2017).

4. Yalin W., *et al*. “A Novel Sliding Window PCA-IPF Based Steady-State Detection Framework and Its Industrial App”. IEEE Access. Magazine. Digital Object Identifier.

5. Alvarez Picaza C., *et al*. “Compresión de Datos aplicado a Sistemas de Energías Renovables”. Enfoque asociado a Bio-Información. Proceedings del II Congreso Latinoamericano de Ingeniería. Cartagena, Colombia (2019).

6. Alvarez Picaza C., *et al*. “Análisis de Componentes Principales desarrollado en Energías Renovables”. Aplicación a Sistemas Dinámicos y Biomédicos. Proceedings del III Congreso Argentino de Ingeniería – Chaco – Argentina (2016).

7. González AJ., *et al*. “Energy efficiency improvement in the cement industry through energy management”. IEEE-IAS/PCA 54th Cement Industry Technical Conference (2023).