The usage of Principal Components applied to Bio – Data Compression

Carlos Alvarez Picaza, Magister[1], Ángel Esteban Piacenza, Especialista[2], Julián Ignacio Veglia, Magister[1], y Alberto Daniel Valdéz, Magister[1]
[1]Facultad de Ciencias Exactas y Naturales y Agrimensura, Universidad Nacional del Nordeste, Argentina, cpicaza@gmail.com, jveglia@exa.unne.edu.ar, advdanieladv@gmail.com
[2]Facultad de Medicina, Universidad Nacional del Nordeste, Argentina, aepiacenza@yahoo.com.ar

***Abstract*– The use of tools provided by biostatistics will allow in the present paper to address the analogical-digital processing of bio-potentials. Optimization in information processing is essential to acquire relevant conclusions. Subjects with heart conditions will be of interest in the development of this work. By using digital techniques, the conditioning of the different variables will be sought for the identification and classification of patterns. The current development will aim at the treatment and analysis of electrical potentials from bio-signals for their consequent processing through biostatistics, using data compression tools such as Principal Component Analysis (PCA). The PCA was initially used in psychology, the social and natural sciences. However, for some years now its application has been extended to the physical sciences, engineering, economics, pattern recognition and data compression. As a result, we acquired reductions of more than 50% in the number of variables. From a total of thirteen (13) original variables, it was possible to concentrate more than 94 % of the information in only six (6) principal components.**

***Keywords*—PCA, Correlation, Variance.**

El uso de Componentes Principales aplicado a la Compresión de Bio – Datos

Carlos Alvarez Picaza, Magister[1], Ángel Esteban Piacenza, Especialista[2], Julián Ignacio Veglia, Magister[1], y Alberto Daniel Valdéz, Magister[1]
[1]Facultad de Ciencias Exactas y Naturales y Agrimensura, Universidad Nacional del Nordeste, Argentina, cpicaza@gmail.com, jveglia@exa.unne.edu.ar, advdanieladv@gmail.com
[2]Facultad de Medicina, Universidad Nacional del Nordeste, Argentina, aepiacenza@yahoo.com.ar

Resumen– ***El uso de las herramientas provistas por la bioestadística, permitirán en el presente trabajo abordar el procesamiento analógico-digital de biopotenciales. La optimización en el procesado de la información es fundamental para la obtención de conclusiones relevantes. Sujetos con afecciones cardíacas serán de interés en el desarrollo de este trabajo. Mediante el uso de técnicas digitales se buscará el acondicionamiento de las diferentes variables para la identificación y clasificación de patrones. El actual desarrollo tendrá como objetivo el tratamiento y análisis de potenciales eléctricos provenientes de bioseñales para su consecuente procesamiento mediante bioestadística, utilizando herramientas de compresión de datos como el Análisis de Componentes Principales (PCA). El PCA se empleó inicialmente en la psicología, las ciencias sociales y naturales. Sin embargo, desde hace ya algunos años se ha extendido su aplicación a las ciencias físicas, la ingeniería, la economía, el reconocimiento de patrones, la compresión de datos, etc. Como resultado obtuvimos, reducciones de más del 50 % en el número de variables. De un total de trece (13) variables originales, se logró concentrar más del 94 % de la información en sólo seis (6) componentes principales.***

Palabras clave—PCA, Correlación, Varianza.

I. Introducción

La Bioingeniería es una rama de la Ingeniería que estudia entre otras cosas, la cuantificación de los fenómenos biológicos, como ser, la conductividad de sangre y tejidos, la respuesta mecánica a un estímulo eléctrico o el estudio de fenómenos bioeléctricos. Dentro de éstos últimos se encuadra el análisis de la señal cardíaca.

Cuando se recoge la información de una muestra de datos, lo más frecuente es tomar el mayor número posible de variables. Sin embargo, si tomamos demasiadas de ellas sobre un conjunto de objetos, tendremos que considerar muchos posibles coeficientes de correlación, y va aumentando, si consideramos un número aún mayor de variables.

Otro problema que frecuentemente aparece es la fuerte correlación que muchas veces se presenta entre las variables, si tomamos demasiadas (cosa que en general sucede cuando no se sabe demasiado sobre los datos o sólo se tiene ánimo exploratorio), lo normal es que estén relacionadas o que midan lo mismo bajo distintos puntos de vista. Por ejemplo, en estudios médicos, las presiones sanguíneas a la salida del corazón y a la salida de los pulmones están fuertemente relacionadas. Se hace necesario, pues, reducir el número de variables.

Es importante resaltar el hecho de que el concepto de mayor información se relaciona con el de mayor variabilidad o varianza. Cuanto mayor sea la variabilidad de los datos se considera que existe mayor información. El Análisis de Componentes Principales o PCA, por sus siglas en inglés ,Principal Component Analysis, es una técnica estadística de síntesis de la información o reducción de la dimensión (número de variables). En bancos de datos de muchas variables, la técnica de PCA permite reducir el número de ellas, sin perder información substancial.

Los nuevos factores o componentes serán una combinación lineal de las variables originales, e independientes entre sí [1]. Un aspecto clave en el Análisis de Componentes Principales es la interpretación de los factores, que no viene dada a priori, sino que se deduce tras observar la relación de los resultados con las variables iniciales. El propósito fundamental de la técnica consiste en la reducción de la dimensión de los datos con el fin de simplificar el problema de estudio.

Últimamente PCA se está utilizando para mejorar la precisión en los dignósticos medicos [2], concentrando la compresión de datos en una matriz y midiendo la distancia entre los datos de PCA y la secuencia de datos de referencia.

Yalin et al. [3] solucionó mediante el Análisis de Componentes Principales las limitaciones de los métodos tradicionales de detección de estado continuo en el tratamiento de potenciales eléctricos biológicos, inclusive llevando éste procedimiento a aplicaciones industriales y de generación de energía. Existen programas computacionales específicos que ayudan a simplificar el desarrollo del trabajo (XLSTAT, HOMER, MATLAB).

II. Metodología

El Análisis de Componentes Principales es un método que reduce la dimensión de los datos realizando un análisis de covarianza entre factores [4].

En muchas aplicaciones, un conjunto de n objetos se representan a través de una colección de m descriptores, índices o parámetros. En algunos casos m es un número muy grande, lo que dificulta el análisis del conjunto de datos en toda su dimensión, es decir que se pueden considerar los n objetos como n puntos ubicados en un espacio de m dimensiones. El objetivo, es el de clasificar esos objetos y

representarlos en un espacio de dimensión menor p ($p < m$), de tal manera que la proyección en ese espacio sea óptima.

Conceptos tales como, desviación estándar, covarianza, autovectores y autovalores, explicados en un trabajo previo [5], son fundamentales para una descripción detallada del funcionamiento de PCA.

III. Desarrollo

El primer tratamiento numérico que debe hacerse es el de escalar las columnas de descriptores de la matriz **A**. Esto es así porque cada columna (cada variable) puede estar especificada en un sistema de unidades distinto. De hecho, cada variable no tiene porqué ser de la misma naturaleza que las otras. Hay varias posibilidades de escalado. La más común consiste en obtener vectores columnas centrados y normalizados adimensionales, así pues de cada columna a_j de la matriz **A**,

$$\mathbf{A} = (\mathbf{a}_1 \;\; \mathbf{a}_2 \ldots \mathbf{a}_m) . \tag{1}$$

se calcula su media

$$\bar{a}_j = \frac{1}{n}\sum_{1}^{n} a_{ij} . \tag{2}$$

y las desviaciones estándar multiplicadas por n

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

obteniendo la matriz de variables adimensionales siguiente:

$$\mathbf{A} \rightarrow \mathbf{Z} = (\mathbf{z}_1 \;\; \mathbf{z}_2 \ldots \mathbf{z}_m) . \tag{4}$$

donde cada vector columna $\mathbf{z}_j$ se define a partir de la transformación

$$a_j \rightarrow z_j = \frac{a_j - \bar{a}_j}{s_j} . \tag{5}$$

La matriz de variables homogeneizadas adimensionales permite calcular la matriz de los coeficientes de correlación entre cada par de columnas de datos:

$$\mathbf{R} = \mathbf{Z}^{\mathrm{T}} \mathbf{Z} . \tag{6}$$

esta matriz es de dimensión mxm.

$$\mathbf{RX} = \mathbf{X\Delta} . \tag{7}$$

donde

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

Todos los valores propios son no negativos. Precisamente los valores propios de esta matriz son los parámetros que indican qué fracción de la varianza total original retiene cada nuevo componente principal (CP).

$$f_i = 100 \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j} \% . \tag{9}$$

Por ello, el ordenamiento, de mayor a menor, de los valores propios induce un orden de preferencia de los CP. A partir de ahora supondremos que

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m . \tag{10}$$

Ahora

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

donde $\mathbf{x}_1$ es el vector propio asociado a λ_1, $\mathbf{x}_2$ a λ_2 y así sucesivamente hasta m. El primer Componente Principal representa la mayor cantidad de varianza de los datos originales, retiene la segunda mayor varianza, y así hasta m. A los coeficientes de cada vector propio $\mathbf{x}_j$ se les llama pesos (loadings) e indican qué combinaciones lineales de las variables originales se deben construir para definir las nuevas coordenadas adimensionales [6].

IV. Resultados, Avances/Discusión

En las siguientes tablas podemos observar algunas caracteríscas propias referentes al funcionamiento de la pared cardíaca a saber Presión Sistólica (PS) y Velocicidad de Onda de Pulso (VOP) de la Tabla I y Compliancia (Cm) y la Medida Interna Carótida (ImrCa) de la Tabla II entre otras.

Vemos un total de trece variables (columnas) entre ambas tablas que intervienen en dicho funcionamiento.

TABLA I
Actividad Cardíaca de Catorce Pacientes Hipertensos (HTA)

Nº de paciente	Edad (1) años	Peso (2) kg	Altura (3) m	PS (4) mmHg	PD (5) mmHg	PM (6) mmHg	VOP (7) m/s
1	55	73	1,63	146	96	113	15,21
2	63	79	1,72	106	84	91	14,58
3	42	75	1,76	116	65	82	10,07
4	50	83	1,84	157	89	112	11,15
5	60	84	1,86	166	98	121	14,25
6	59	80	1,72	164	92	116	16,26
7	50	106	1,82	127	82	97	11,44
8	64	83	1,76	155	92	113	17,16
9	38	58	1,63	155	70	98	10,83
10	40	81	1,72	139	100	113	11,28
11	54	93	1,74	134	84	101	10,27
12	45	83	1,78	114	75	88	11,31
13	54	72	1,65	117	78	91	14,07
14	62	80	1,72	125	83	97	14,78

TABLA II
ACTIVIDAD CARDÍACA DE CATORCE PACIENTES HIPERTENSOS (HTA)

Nº de	Cm (8)	DS (9)	DD (10)	DM (11)	Etha (12)	ImtCa (13)
paciente	e-4 cm/mmHg	mm	mm	mm	mmHg s/mm	mm
1	2,00	7,70	7,32	7,50	5,07	0,73
2	2,21	7,42	7,03	7,25	2,92	1,14
3	3,49	5,72	5,32	5,53	4,81	0,75
4	3,76	7,54	7,22	7,37	5,83	0,9
5	2,68	8,75	8,36	8,55	3,99	0,9
6	2,22	9,33	8,85	9,11	3,73	0,85
7	3,76	7,92	7,50	7,71	4,66	0,71
8	2,48	7,83	7,60	7,72	9,17	0,82
9	3,70	7,11	6,69	6,89	4,52	1,08
10	4,04	8,15	7,95	8,05	9,28	0,79
11	4,82	8,13	7,81	7,97	5,17	0,76
12	3,12	6,63	6,10	6,30	2,18	0,89
13	2,31	7,44	6,94	7,19	2,39	0,98
14	1,68	5,81	5,60	5,71	8,42	1,04

La Figura 1 muestra los autovalores del sistema original de la matriz de catorce (14) filas por trece (13) columnas [7].

A partir de aquí ocuparemos la técnica de Análisis de Componentes Principales para realizar la compression de datos correspondiente y trabajar con un porcentaje de información relevante.

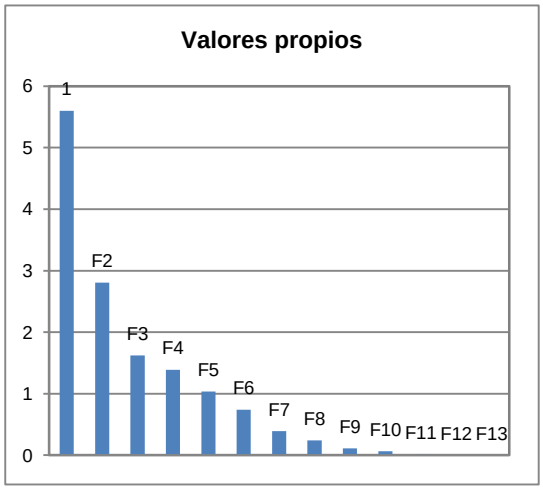

Fig. 1 Orden de preferencia de los Componentes Principales.

Los valores indicativos de la mayor varianza nos lo da el valor de la desviación standard σ, ya que la varianza se define como $var = \sigma^2$.

TABLA III
MEDIA Y DESVIACIÓN STANDARD DE LAS VARIABLES (COLUMNAS)

Variables	Media	σ
Edad	52,571	8,398
Peso	80,714	10,368
Altura	1,739	0,069
PS	137,307	19,360
PD	84,857	10,063
PM	102,341	11,598
VOP	13,047	2,292
CM	3,019	0,890
DS	7,533	0,963
DD	7,165	0,960
DM	7,348	0,964
Etha	5,152	1,735
ImtCa	0,881	0,131

Las Tablas IV y V nos indican el orden de los loadings (pesos) de las combinaciones lineales de cada una de las variables originales.

TABLA IV
COMPONENTES PRINCIPALES DE LOS PACIENTES CARDÍACOS

	F1	F2	F3	F4	F5	F6
Valor propio	5,598	2,806	1,619	1,386	1,034	0,739
% varianza	39,985	20,045	11,568	9,899	7,386	5,279
% acumulado	39,985	60,030	71,598	81,497	88,883	94,162

TABLA V
COMPONENTES PRINCIPALES DE LOS PACIENTES CARDÍACOS

F7	F8	F9	F10	F11	F12	F13
0,391	0,242	0,109	0,064	0,011	0,000	0,000
2,792	1,728	0,780	0,460	0,078	0,000	0,000
96,954	98,681	99,462	99,922	100,000	100,000	100,000

La Figura 2 muestra el diagrama de Pareto obtenido en función de los Componentes Principales (factores) seleccionados. El análisis de Pareto es una comparación ordenada de factores relativos a un problema. Esta comparación ayuda a identificar y enfocar los pocos factores vitales diferenciándolos de los muchos factores útiles. La aplicación del mismo permite exhibir visualmente en orden de importancia, la contribución de cada elemento en el efecto total. En el gráfico sólo de visualizan los 6 primeros Componentes Principales, PS – PM – Peso – Edad – VOP – Etha,

Los pesos de los 7 restantes son insignificantes respecto de los primeros, en los cuales se concentra más del 94 % de la información de la matriz original.

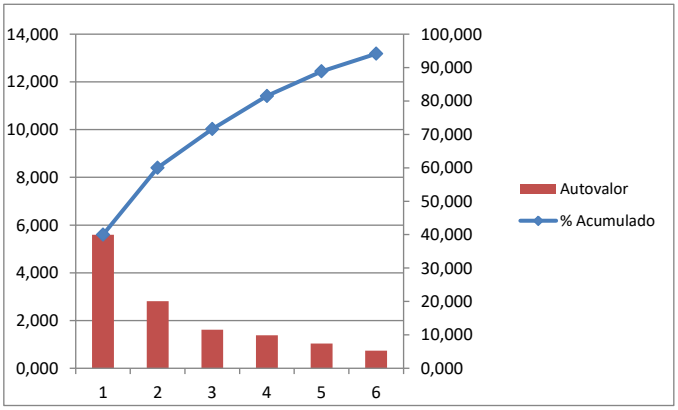

Fig. 2 Diagrama de Pareto.

A. *Trabajos futuros*

La utilización del método de Análisis de Componentes Principales abarca un abanico de aplicaciones que utilizaremos en nuevos trabajos. Como ser la clasificación de anomalías de los latidos cardíacos, los cuales pueden realizarse a partir de estudios temporales y morfológicos provenientes de la señal electrocardiográfica; además tenemos diagramado emplearlo en el estudio de optimización de energía de servomotores que accionan las prótesis mioeléctricas.

B. *Alcances del PCA en Ingeniería*

El PCA como herramienta estadística ya está siendo aprovechada en la Ingeniería Eléctrica y Electrónica en la compilación de las variables necesarias para proyectar el diseño de equipamiento y sistemas relacionados con el control de generación y distribución de energía eléctrica. Además, de su aplicación en programas computacionales utilizados en energías renovables (eólica, solar, hidráulica, geotermia, etc.), destacando los parámetros más importantes a la hora de realizar observaciones, comparaciones e identificaciones.

Como se explicó a lo largo del trabajo presentado, esta herramienta matemática es aplicable a cualquier situación donde existe una gran cantidad de datos recabados.

V. Conclusiones

El método de Análisis de Componentes Principales (PCA) es efectivo, y permitió cumplir con los objetivos mencionados anteriormente. En el caso de los datos electrocardiográficos, se logró reducir una matriz de 13 variables a solo 6, recabando el 94% de la información contenida en la matriz original. Con esta herramienta se eliminó la redundancia de datos para agilizar los tiempos computacionales, lo que constituye un objetivo primordial en el procesamiento de información.

Reconocimiento

Este trabajo fue realizado dentro del marco del Proyecto de Investigación 22F019 financiado por la Secretaría General de Ciencia y Técnica de la Universidad Nacional del Nordeste.

Referencias

[1] M.I. Pisarello, C. Alvarez Picaza, J.E. Monzón, "Análisis de Componentes Principales para la Compresión del ECG". *Proceedings de la 5ta Conferencia Iberoamericana, Cibernética e Informática* (CISCI 2006). Orlando - Florida - EE.UU, 2006.

[2] K. Tusongjiang, G. Wensheng, "Power transformer fault diagnosis using FCM and improved PCA". *The Journal of Engineering. IET Electrical Engineering Academic Forum*, 2017.

[3] W. Yalin, S. Kenan, Y. Xiaofeng, C. Yue, L. Ling, N.K. Heikki, "A Novel Sliding Window PCA-IPF Based Steady-State Detection Framework and Its Industrial App.". *IEEE Access.Magazine.DigitalObjectIdentifier* 10.1109/ACCESS.2018.2825451, 2018.

[4] C. Alvarez Picaza, "Modelado - Aplicación de Técnicas de Control Moderno – Utilización de PCA (Análisis de Componentes Principales) para Sistemas de Energías Renovables". *Trabajo Final de Maestría* – pp. 1-82 Biblioteca (UNSa – Argentina.), 2014.

[5] C. Alvarez Picaza, M.I. Pisarello, J.E. Monzón, "Análisis de Componentes Principales desarrollado en Energías Renovables. Aplicación a Sistemas Dinámicos y Biomédicos". *Proceedings del III Congreso Argentino de Ingeniería*, Chaco, Argentina, 2016.

[6] A.J. González, R.C. Castrillón, E.C. Quispe, "Energy efficiency improvement in the cement industry through energy management". *IEEE-IAS/PCA 54th Cement Industry Technical Conference*, 2013.

[7] XLSTAT, Toolbox User´s Guide. Addinsoft, Version 7.5.3.