Encuentro Internacional de Educación en Ingeniería ACOFI
2º Congreso Latinoamericano de Ingeniería
2019 10 al 13 de septiembre - Cartagena de Indias, Colombia
RETOS EN LA FORMACIÓN DE INGENIEROS EN LA ERA DIGITAL

COMPRESIÓN DE DATOS APLICADO A SISTEMAS DE ENERGÍAS RENOVABLES. ENFOQUE ASOCIADO A BIO-INFORMACIÓN

Alvarez Picaza C, Veglia JI, Piacenza AE y García Roth JC

Universidad Nacional del Nordeste
Corrientes, Argentina

Resumen

Al realizar el estudio de un sistema de generación compuesto por módulos fotovoltaicos y aerogeneradores, las variables a considerar que entran en juego para su análisis de rendimiento energético son numerosas, a saber, potencia involucrada en los paneles solares, velocidad del viento, carga de baterías del inversor, etc. Es inevitable la acumulación de gran cantidad de datos que no siempre resultan imprescindibles para obtener resultados certeros. El método de Análisis de Componentes Principales (PCA) tiene por objeto transformar un conjunto de variables, a las que se denomina originales, en un nuevo conjunto de variables denominadas componentes principales. Estas últimas se caracterizan por estar correlacionadas entre sí y, además, pueden ordenarse de acuerdo con la información que llevan incorporada. La obtención de datos normalizados y procesos más eficientes agilizan los tiempos computacionales y además economizan en espacio de almacenamiento. 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. A partir del análisis de nuestra central de estudio, conformamos una matriz dinámica, a la cual aplicamos las técnicas de PCA. Como resultado obtuvimos, del estudio de ciertos días determinados, reducciones de hasta el sesenta por ciento (60%) en el número de variables. De un total de diez (10) variables originales, se logró concentrar poco más del noventa y seis por ciento (96%) de la información en tan sólo cuatro (4) componentes principales. El diagrama de Pareto permite visualizar en forma gráfica la ponderación de cada uno de estos componentes. Además, su aplicación es compatible a cualquier clase de sistemas, inclusive los biomédicos. A fin de identificar y clasificar patrones anómalos en las distintas patologías cardíacas, se intenta encontrar una metodología que acerque precisión a la hora de determinar diagnósticos más certeros.

Palabras clave: PCA; compresión de datos; correlación

Abstract

When studying a photovoltaic module - wind turbine generation system, there are many variables to be considered for its energy efficiency analysis, like solar panels power, wind speed, electric inverter, battery charge, etc. It cannot be avoided large amounts data accumulation, they are not essentials to get accurate results. The Principal Component Analysis (PCA) method proposes to transform a set of variables, called originals, into a new set of variables, called principal components. These ones are characterized for being correlated between each other and, and also, they can be ordered according to their built-in information. Getting standardized data and more efficient processes speeds up computational times and economizes on storage space. PCA was first used in psychology, social and natural sciences. However, for some years now, its application has been extended to the physical sciences, engineering, economics, pattern recognition, data compression, etc. From the analysis of our electric central, we create a dynamic matrix, to which we apply the PCA techniques. As a result, it becomes in reductions up to sixty percent (60%) in the variable number on particular days. From ten (10) original variables, it was possible to concentrate just over ninety six percent (96%) of the information in only four (4) principal components. Pareto's diagram allows to visualize in a graphic way, each component loading. Its application is compatible with many kinds of systems, including biomedical systems. In order to identify and classify anomalous patterns in different cardiac pathologies, we try to find out a methodology that brings precision when determining accurate diagnoses.

Keywords*: PCA; data compression; correlation*

1. Introducción

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 se presenta 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 (varianza) se considera que existe mayor información. El Análisis de Componentes Principales o PCA (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 tales, sin perder información substancial. Como lo expresaron Pisarello et al (2006), los nuevos factores o componentes serán una combinación lineal de las variables originales, e independientes entre sí. 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.

Una de las complicaciones que acarrea la energía eólica es que su disponibilidad es variable y, por lo tanto, necesita ser respaldada por otras fuentes de potencia. Los sistemas fotovoltaicos tienen la ventaja adicional de ser estáticos y de casi no requerir mantenimiento ni reparaciones. Sin embargo, la potencia fotovoltaica es típicamente cinco veces más cara que la potencia eólica. Actualmente, existen investigaciones y esfuerzos por desarrollar paneles fotovoltaicos de bajo costo para aplicaciones generales. La eficiencia de la conversión de la potencia solar es típicamente de 37.8% para paneles solares sin concentrador (marca Spectrolab). El conjunto de ambos sistemas integra una vasta cantidad de variables a considerar, lo cual representa un problema de desarrollo y síntesis.

Últimamente PCA se está utilizando para mejorar la precisión en los dignósticos médicos, de acuerdo a trabajos presentados por Tusongjiang y Wensheng (2017) 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. (2018) 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 este procedimiento a aplicaciones industriales y generación de energía.

Existen programas computacionales específicos que ayudan a simplificar el desarrollo del trabajo (XLSTAT, HOMER, MATLAB).

2. Materiales y Métodos

El Análisis de Componentes Principales es un método que reduce la dimensión de los datos realizando un análisis de covarianza entre factores (Alvarez Picaza, 2014).

A. Técnicas de PCA

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 (Alvarez Picaza et al. 2016), son fundamentales para una descripción detallada del funcionamiento de PCA.

Este trabajo está basado en la configuración de una central eléctrica de energías renovables compuesta por paneles fotovoltaicos y aerogeneradores. En la metodología de PCA se ordenan los descriptores en una matriz **A** de dimensión *n x m*. El criterio matemático utilizado para conseguir la reducción de la dimensión tal que, para un valor prefijado de *p*, se retenga en ese subespacio la máxima varianza estadística total de los datos originales. Esto conduce a especificar una nueva serie de ejes ortogonales entre sí, los componentes principales (CP). Cada CP es una combinación lineal de las variables o descriptores originales.

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, cada columna a_j de la matriz **A**,

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

se le calcula su media

$$\overline{a}_j = \frac{1}{n}\sum_{1}^{n} a_{ij} \quad (2)$$

y las desviaciones estándar multiplicadas por *n*

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

obteniendo la matriz de variables adimensionales siguiente:

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

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

$$a_j \rightarrow z_j = \frac{a_j - \overline{a}_j}{s_j}, \quad (5)$$

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

$$\mathbf{R} = \mathbf{Z}^{\mathrm{T}}\,\mathbf{Z}, \quad (6)$$

esta matriz es de dimensión *mxm*.
Los CP vienen dados por los vectores propios de de la matriz **R**:

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

donde

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

Todos los valores propios son no negativos (recordemos que la matriz **Z** se obtiene de tal manera que es definida no negativa). Precisamente los valores propios de esta matriz son los parámetros que indican qué fracción de la varianza total original retiene cada nuevo CP.

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

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

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \quad (10)$$

Ahora

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

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, $\mathbf{x}_1$ representa la mayor cantidad de varianza de los datos originales, $\mathbf{x}_2$ retiene la segunda mayor varianza, y así hasta *m*. El conjunto de los *m* Componentes Principales genera una nueva matriz de coordenadas. 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, como lo explica González et al (2013). Lo más usual, al reducir la dimensionalidad del problema.

B. XLSTAT

Cabe destacar que para nuestro análisis se tomaron las 24 horas de actividad de un día determinado de la central eléctrica. El objetivo es analizar las correlaciones entre las variables e identificar características que se distinguen fuertemente de los demás (del Manual XLSTAT). El análisis de Componentes Principales (PCA) es un método muy eficaz para el análisis de datos cuantitativos (continuos o discretos) que se presentan bajo la forma de cuadros de M observaciones / N variables. Los límites del PCA vienen del hecho que es un método de proyección, y que la pérdida de información inducida por la proyección puede provocar interpretaciones erróneas. El buen discernimiento permite, sin embargo, evitar estos inconvenientes (Hernández, 1998).

3. Resultados

Datos Central No Convencional

Cabe destacar que el HOMER en este trabajo solamente se utilizó para la captura de datos (4 de enero) para el estudio de los mismos a través del Análisis de Componentes Principales.

Hour	Global Solar kW/m2	Incident Solar kW/m2	Wind Speed m/s	DC Primary Load kW	PV Power kW	SW AIR X (4) kW	Excess Electricity kW	Battery Power kW	Battery State of Charge %	Battery Energy Cost $/kWh
73	0	0	4,216	0,002	0	0,023	0	0,021	96,195	0,98
74	0	0	5,895	0,002	0	0,056	0,005	0,049	97,809	0,97
.										
.										
.										
96	0	0	2,730	0,001	0	0,006	0	0,005	99,373	0,98

Tabla 1. Datos totales correspondientes a la actividad de la central el día 4 de enero

.

Valores propios:

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
Valor propio	4,631	2,677	1,694	0,604	0,204	0,140	0,034	0,014	0,001	0,000
% varianza	46,310	26,767	16,945	6,044	2,040	1,398	0,343	0,138	0,014	0,001
% acumulado	46,310	73,077	90,022	96,066	98,106	99,504	99,847	99,985	99,999	100,000

Tabla 2. Componentes Principales de la central

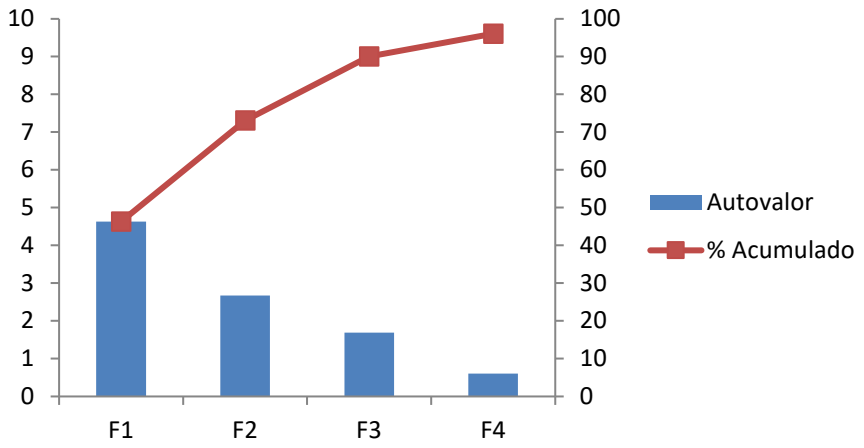

Figura 1. Diagrama de Pareto

La Figura 1 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 4 primeros Componentes Principales debido a que los pesos de los 6 restantes son insignificantes respecto de los primeros, en los que se concentra más del 96 % de la información de la matriz original.

Datos cardiológicos

Datos obtenidos de catorce pacientes hipertensos (HTA).

N° de paciente	Edad	Peso Kg	Altura m	PS mmHg	PD mmHg	PM mmHg	Vop m7s	Cm e-4 cm/mmHg	DS mm	DD mm	DM mm	etha mmHg s/mm	ImtCa 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
.													
.													
.													
14	62	80	1,72	125	83	97	14,78	1,68	5,81	5,60	5,71	8,42	1,04

Tabla 3. Datos totales correspondientes a la actividad cardíaca de catorce (14) pacientes

Valores ropios:

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
Valor propio	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
% varianza	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
% acumulado	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

Tabla 4. Componentes Principales de los pacientes cardíacos

Podemos observar que de los seis (6) primeros factores es posible recopilar más del 94% de la información brindada por los datos en forma individual.

4. Conclusiones

El método de Análisis de Componentes Principales (PCA) es efectivo, y permitió cumplir con los objetivos mencionados anteriormente. Se logró reducir una matriz de 10 variables (central eléctrica) a una matriz de 4 variables en las que se concentra más del 96% de la información. En el caso de los datos electrocardiográficos, de 13 a 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.

5. Referencias

Artículos de revistas

- Hernandez, M. (1998). Temas de Análisis Estadístico Multivariado. Editorial Universidad de Costa Rica.
- Tusongjiang, K., Wensheng, G. (2017). Power transformer fault diagnosis using FCM and improved PCA. The Journal of Engineering. IET Electrical Engineering Academic Forum.

Libros

- Alvarez Picaza, C. (2014). 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 – Arg.).
- XLSTAT. Toolbox User's Guide by Addinsoft. 2018.

Memorias de congresos

- Alvarez Picaza, C., Pisarello, M.I., Monzón, J.E. (2016). Análisis de Componentes Principales desarrollado en Energías Renovables. Aplicación a Sistemas Dinámicos y Biomédicos. Proceedings del III Congreso Argentino de Ingeniería – Chaco - Arg.
- González, A.J., Castrillón, R.C., Enrique C. Quispe, E.C. (2013). Energy efficiency improvement in the cement industry through energy management. IEEE-IAS/PCA 54th Cement Industry Technical Conference.
- Pisarello, M.I., Álvarez Picaza, C., Monzón, J.E. (2006). Análisis de Componentes Principales para la Compresión del ECG. Proceedings de la 5ta Conferencia Iberoamericana, Cibernética e Informática (CISCI 2006). Orlando - Florida - EE.UU.

Fuentes electrónicas

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

Sobre los autores

- **Carlos Álvarez Picaza**: Ingeniero Electricista O.I. Msc en Energías Renovables – UNSa. Profesor Responsable Electrónica Industrial y Teoría de Control – UNNE. cpicaza@gmail.com
- **Julián Ignacio Veglia**: Ingeniero Electricista O.I. JTP Electrónica Industrial y Teoría de Control – UNNE. julianv04@gmail.com
- **Ángel Esteban Piacenza**: Médico Cardiólogo. Profesor Adjunto Medicina I – Universidad Nacional del Nordeste. aepiacenza@yahoo.com.ar

- **Juan Carlos García Roth**: Médico Esp. Terapia Intensiva. JTP Fisiología – Universidad Nacional del Nordeste. juancarlosroth@hotmail.com

---