



Desarrollo Comparado de los Algoritmos Análisis de Componentes Principales (ACP) y Mínimos Cuadrados Alternos (MCA). Basado en un Enfoque Matemático

(Comparative Development of Principal Component Analysis (ACP) and Alternate Least Squares (MCA) Algorithms. Based on a Mathematical Approach)

Recibido: 05/05/2020 Aprobado: 30/09/2021

Eddy Jackeline Rodríguez.

Centro de investigación de Matemática Aplicada (CIMA). Facultad de Ingeniería. Universidad del Zulia. Maracaibo. Venezuela. eddyjackeline@gmail.com.

Código ORCID: 0000-0003-3558-2189

Línea de investigación: Programación lineal y no lineal.

Ana Isolina Prieto de Osorio.

Centro de investigación de Matemática Aplicada (CIMA). Facultad de Ingeniería. Universidad del Zulia. Maracaibo. Venezuela. aisolinap@hotmail.com.

Josefina del Carmen Matera Acosta.

Centro de investigación de Matemática Aplicada (CIMA). Facultad de Ingeniería. Universidad del Zulia. Maracaibo. Venezuela. pinamatera@hotmail.com

Resumen

El presente artículo muestra los resultados de la investigación llevada a cabo, cuyo objetivo fue realizar un desarrollo comparado de los algoritmos análisis de componentes principales (ACP) y mínimos cuadrados alternos (MCA) bajo un enfoque matemático-optimal, apoyada en autores especialistas como Cuadras (1991), Esteban, Ariño, Díaz-Cruz, Díaz-Cruz y Tauler (2000) y De Juan, Vander Heyden, Tauler & Massart (1997), entre otros. La metodología empleada fue teórica-práctica. La investigación fue del tipo aplicada, partiendo de una revisión bibliográfica de libros y artículos publicados sobre el tema en estudio, seguido de la descripción, análisis y formulación de los algoritmos de algunas técnicas multivariantes aplicadas al modelado de problemas con un diseño documental no experimental. La población y muestra de la investigación estuvo conformada por las diferentes fuentes bibliográficas o documentales y literatura sobre el tema de investigaciones consultadas. Como técnicas de recolección de datos se empleó la observación directa y como instrumentos se utilizaron la lista de cotejo y el cuaderno de notas. Los resultados demostraron que: (a) Los algoritmos descomponen de manera satisfactoria las matrices arrojando errores pequeños basados en la norma de Frobenius; (b) Ambos algoritmos reducen significativamente el número de componentes de la matriz en estudio, asegurando que las componentes principales son independientes; (c) El algoritmo mínimos cuadrados alternos determina número componentes en la fase inicial; (d) El algoritmo de análisis de componentes principales determina el número de componentes en la fase de desarrollo y (e) La matriz se puede representar con dos componentes principales conservando una varianza no menor a un 90%. Es decir, los resultados demostraron la eficiencia de los métodos estudiados, tanto para la obtención de componentes principales como en la descomposición bilineal de matrices.



Palabras claves: Técnicas Multivariantes, Algoritmo, Componentes, Mínimos Cuadrados Alternos, Desarrollo Matemático-Optimal.

Abstract

This article shows the results of the research carried out, the objective of which was to carry out a comparative development of the principal component analysis (ACP) and alternate least squares (ALC) algorithms under a mathematical-optimal approach, supported by specialist authors such as Cuadras (1991), Esteban, Ariño, Diaz-Cruz, Diaz-Cruz and Tauler (2000) and De Juan, Vander Heyden, Tauler & Massart (1997), among others. The methodology used was theoretical-practical. The research was of the applied type, starting from a bibliographic review of books and articles published on the subject under study, followed by the description, analysis and formulation of the algorithms of some multivariate techniques applied to the modeling of problems with a non-experimental documentary design. The research population and sample was made up of the different bibliographic or documentary sources and literature on the topic of the researches consulted. Direct observation was used as data collection techniques and the checklist and notebook were used as instruments. The results showed that: (a) The algorithms satisfactorily decompose the matrices, yielding small errors based on the Frobenius norm; (b) Both algorithms significantly reduce the number of components of the matrix under study, ensuring that the principal components are independent; (c) The alternate least squares algorithm determines number of components in the initial phase; (d) The principal components analysis algorithm determines the number of components in the development phase and (e) The matrix can be represented with two principal components, keeping a variance of no less than 90%. That is, the results demonstrated the efficiency of the studied methods, both for obtaining main components and in the bilinear decomposition of matrices.

Key words: Multivariate Techniques, Algorithm, Components, Alternate Least Squares, Mathematical-Optimal Development.

Introducción

En diferentes áreas de la industria se presentan situaciones que requieren control, predicción y modelado, por tanto, crear modelos matemáticos para su control, haciéndose necesario el estudio de las características de las variables participantes en los problemas, con el fin de facilitar la regulación del mismo. En la actualidad existen diferentes métodos para alcanzar este fin, tales como: el método de mínimos cuadrados alternos (MCA) y el método de análisis de componentes principales (ACP), los cuales forman el objeto de estudio de este trabajo. Se han desarrollado áreas específicas en la investigación como son: la quimiometría, la econometría, entre otras; las cuales analizan el desarrollo de técnicas multivariantes para dar respuesta a estos inconvenientes.

La creación de algoritmos para extraer la información relevante cuando se resuelve un determinado problema o la representación de un conjunto de datos dividido en grupos con características de un problema son, por ejemplo, situaciones donde se aplican las técnicas multivariantes, análisis de componentes principales y mínimos cuadrados alternos, algoritmos que constituyen el objeto de este estudio.

Al estudiar una matriz de datos, es posible encontrar correlaciones altas (en valor absoluto) entre varias variables, el caso más extremo sucede cuando hay colinealidad entre variables. También pueden existir casos donde el número de variables sea tan grande que dificulta su análisis global y, por tanto, es necesario reducirlas a un conjunto de menor dimensión,



mediante una matriz de datos cuyo contenido es la selección de un subconjunto de variables originales o combinaciones lineales de ellas.

Estos algoritmos, ACP y MCA, son usados para resolver las contribuciones de los componentes en un conjunto de datos con estructura bilineal, pudiendo ser aplicados a problemas con o sin restricciones (Ximénez y Martín, 2013). La solución es alcanzada mediante procedimiento de optimización restringida o no restringida, según sea el caso. El principio de los MCA está condicionado por la estimación de parámetros, sucesivamente se van alternando entre las iteraciones realizadas hasta alcanzar la optimización de los datos observados (Lahoz-Beltrá, 2004). Se define una función de pérdida que frecuentemente es la suma del cuadrado de los residuos u otra medida de disparidad entre las observaciones y las estimaciones.

El método de ACP se fundamenta en la rotación de las variables originales en las direcciones de máxima varianza, donde se define una función y se establece el objetivo a maximizar y la restricción de ortogonalidad (Johnson, Miller y Freund, 2013). El proceso obtiene las componentes una a una. En esta investigación se presenta brevemente la descripción de los métodos mínimos cuadrados alternos y análisis de componentes principales, realizando un desarrollo comparativo de sus algoritmos basados en su estructura matemática, de optimización y estadística, facilitando su comprensión e implementación en diferentes áreas de investigación.

Objetivo General

Comparar los algoritmos análisis de componentes principales (ACP) y mínimos cuadrados alternos (MCA) bajo un enfoque matemático.

Objetivos específicos

Describir las técnicas de análisis multivariantes en el análisis de componentes principales (PCA) y mínimos cuadrados alternos (MCA).

Analizar los algoritmos por fase de las técnicas de análisis multivariantes en el análisis de componentes principales (PCA) y mínimos cuadrados alternos (MCA).

Evaluar el comportamiento de los algoritmos por fase de las técnicas de análisis multivariantes.

Metodología

La investigación fue del tipo aplicada, partiendo de una revisión bibliográfica de libros y artículos publicados sobre el tema en estudio, seguido de la descripción, análisis y formulación de los algoritmos de algunas técnicas multivariantes aplicadas al modelado de problemas (Cuadras, 1991), con un diseño documental no experimental ya que se realizó sin manipular deliberadamente la variable. En este sentido Hurtado (2015) señala, la investigación documental es un proceso basado en la búsqueda, recuperación, análisis, críticas e interpretación de datos secundarios, es decir, los obtenidos y registrados por investigadores en fuentes documentales, impresas, audiovisuales o electrónicas.

La población y muestra de la investigación estuvo conformada por la confrontación documental de las diferentes fuentes bibliográficas o documentales y literatura sobre el tema de investigaciones consultadas. Como técnicas de recolección de datos se elaboraron instrumentos, revisión de textos, trabajos y publicaciones científicas, así como la observación directa, lista de cotejo y el cuaderno de notas (Hurtado, 2015), utilizados para el análisis documental.

Durante la investigación se comparó la estructura matemática de estas técnicas, alcanzando el desarrollo de los objetivos específicos y general, bajo una metodología teórica-práctica mediante la consulta bibliográfica de diferentes autores, quienes plantean la formulación de los algoritmos de distintas técnicas multivariantes analizando la obtención de las componentes principales, aplicando funciones objetivas de costo y evaluando el algoritmo mediante parámetros de calidad de ajuste.

Fundamentación Teórica

Análisis de componentes principales

Este modelo se describe con la siguiente ecuación:

$$X_{(n,p)} = C_{(n,k)}P_{(k,p)}^T + E_{(n,p)} \quad (1)$$

Donde X es la matriz experimental, C es la matriz de *scores* la cual describe las muestras en el espacio de componentes principales (posee información sobre las filas de la matriz X), P es la matriz de loadings que describe las variables en los ejes de componentes principales (contiene información sobre las columnas de la matriz X), E es la matriz de los residuales, en ella se encuentra la porción de la información no recogida en el producto de CP^T , n es el número de muestras, p es el número de variables y k es el número de componentes principales seleccionadas (Enrique, 2006).

Este método permite reducir el espacio original manteniendo la información significativa de los datos, los cuales son descritos en nuevos ejes de coordenadas llamados componentes principales (Tauler, Maeder y De Juan, 2009), como se muestra de forma particular para el plano en la figura 1, donde el objetivo es encontrar una matriz ortogonal P que gira los ejes originales (X_1, X_2) para alinearlos con las extensiones naturales de la nube de puntos de la matriz de datos encerradas en una elipse, estas nuevas variables (C_1, C_2) son ortogonales, asegurando la no correlación entre las mismas.

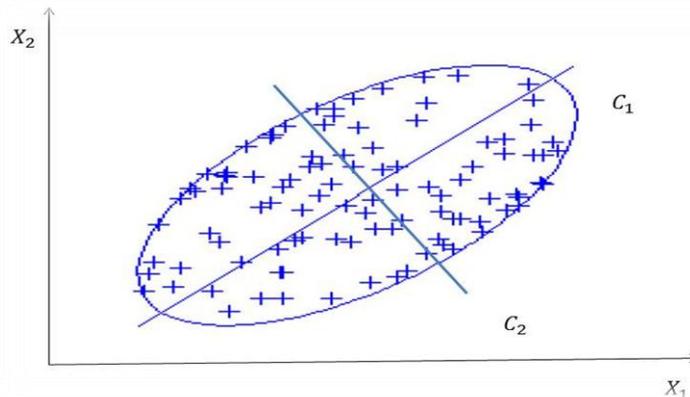


Figura 1. Representación de las componentes principales en el plano
Fuente: Elaboración Propia (2018)

Fundamento matemático, estadístico y optimal de las componentes principales

Sea X una matriz escalada por columnas de orden (n, p) , donde la matriz de correlación R , es simétrica y semidefinida positiva:

$$R = \frac{X^T X}{(n-1)} \quad (2)$$

Con n igual al número de filas de la matriz X .

La matriz R , puede descomponerse espectralmente como:

$$R = \Gamma \Lambda \Gamma^T \quad (3)$$

Donde $\Gamma = [\gamma_1, \dots, \gamma_p]$ es una matriz ortogonal de orden (p, p) , formado por los auto-vectores de norma uno. Λ es una matriz diagonal de orden (p, p) que contiene los auto-valores. Por ortogonalidad de la matriz Γ , la ecuación (3) se puede escribir como:

$$\Gamma^T R \Gamma = \Lambda \quad (4)$$

$$\begin{bmatrix} \gamma_1^T \\ \vdots \\ \gamma_p^T \end{bmatrix} R [\gamma_1 \quad \dots \quad \gamma_p] = \begin{bmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_p \end{bmatrix}$$

De lo anterior se tiene que para todo $i, j = 1, \dots, p$:

$$\gamma_i^T R \gamma_j = \begin{cases} \lambda_i & i = j \\ 0 & i \neq j \end{cases} \quad (5)$$

La matriz de componentes principales C de orden (n, p) , se obtiene transformando la matriz X , de la siguiente forma:

$$C = X \Gamma = [X \gamma_1 \quad \dots \quad X \gamma_p] \quad (6)$$

De las ecuaciones (5) y (6) se concluye que las componentes principales son ortogonales. Este método se basa en maximizar la varianza de las componentes principales C , con la condición de tener auto-vectores de norma uno ($\gamma_i^T \gamma_i = 1$), $i = 1, \dots, p$. Defínase a F , como la función a maximizar mediante el proceso de multiplicadores de Lagrange:

$$F(\gamma, \lambda) = \text{Var}(C) - \lambda(\gamma^T \gamma - 1) \quad (7)$$

Como:

$$\text{Var}(C) = \gamma^T \text{Var}(X) \gamma = \gamma^T \frac{X^T X}{(n-1)} \gamma = \gamma^T R \gamma \quad (8)$$

Entonces sustituyendo (8) en (7), queda:

$$F(\gamma, \lambda) = \gamma^T R \gamma - \lambda(\gamma^T \gamma - 1)$$

Luego:

$$\frac{\partial F}{\partial \gamma} = 2R\gamma - 2\lambda\gamma = 0$$

$$\frac{\partial F}{\partial \lambda} = \gamma^T \gamma - 1 = 0$$

$$R\gamma = \lambda\gamma \quad \gamma^T \gamma = 1 \quad (9)$$

De las ecuaciones (9), queda:

$$\gamma^T R\gamma = \lambda \quad (10)$$

De esta manera se establece que el proceso óptimo de maximizar la varianza de las componentes es equivalente a encontrar los auto-vectores y auto-valores de la matriz X (Cuadras, 1991). Por lo tanto la varianza de la primera componente corresponde a λ_1 con el vector γ_1 , y así sucesivamente, la varianza de la última componente es λ_p con el vector γ_p .

Mínimos cuadrados alternos

El conjunto de datos se representa como una matriz que se descompone en forma bilineal, esto por trabaja con dos variables dependientes (Curtis, Baum, Cook, Edgerton y Rabidoux, 2011; Tauler et al., 2009).

$$D = CS^T + E \quad (11)$$

Donde D es la matriz de datos de tamaño (n, p) , C y S^T forman la descomposición bilineal de tamaños (n, k) y (k, p) respectivamente y, E es la matriz de residuos no ajustados por el modelo y es de tamaño (n, p) . Este método es un proceso iterativo que calcula en cada ciclo nuevas estimaciones de las matrices C y S^T , con el objetivo de minimizar la matriz de residuos E .

Fundamento matemático, estadístico y óptimo de los mínimos cuadrados alternos

El estudio de la ecuación (11), se basa en obtener los vectores expresados en las matrices C , S^T , se plantea la optimización de las siguientes funciones de costo (De Juan, Vander Heyden, Tauler & Massart, 1997)

$$F(c_j) = \|D_j - c_j S^T\| + \lambda \|c_j\| \quad (12)$$

$$F(s_i) = \|D_i - C s_i\| + \lambda \|s_i\| \quad (13)$$

Con $j = 1, \dots, k$ $i = 1, \dots, k$ y λ es el término de regularización para las funciones de costo. En este método las variables de las matrices C , S^T son desconocidas, el proceso para resolver las ecuaciones (12) y (13) se fundamenta en la regresión lineal (Geladi y Kowalski, 1986) por mínimos cuadrados ordinarios, una vez se fija alguna de las dos matrices: Si la matriz fijada es C :

$$\|D - C\beta\| \quad (14)$$

Donde la solución es:

$$\beta = (C^T C)^{-1} C^T D \quad (15)$$

Donde el valor de s_i corresponde al β calculado en (15). Si la matriz que se fija es S :

$$\|D^T - S\beta\| \quad (16)$$

La solución es:

$$\beta = (S^T S)^{-1} S^T D^T \quad (17)$$

El valor de β representa el de c_j . Las matrices C , S^T pueden obtenerse también de forma completa y no por vectores, si se asume en las ecuaciones (12) y (13) a los vectores c_j , s_i como matrices C_j , S_i , donde i, j expresa el número de iteración.

Fases de los algoritmos

El desarrollo matemático del algoritmo del método de análisis de componentes principales y del algoritmo del método de mínimos cuadrados alternos puede plantearse a través de las siguientes fases:

Fase inicial

Algoritmo ACP. Cuando las variables están correlacionadas el elipsoide formado por la nube de puntos no está orientado de forma paralela a alguno de los ejes de las variables X , y se debe rotar los ejes de forma que correspondan a los ejes del elipsoide (. Primeramente se hace la traslación del origen de las variables de X a \bar{x} (vector media de X), de donde se hará la rotación (Johnson et al., 2013). La traslación consiste en escalar por columna a la matriz de estudio X , calculando la media, por columna, de la matriz:

$$\bar{x} = \frac{1}{n} \sum_{r=1}^n x_r = \frac{1}{n} X^T \mathbf{1} \quad (18)$$

Donde x_r representa los vectores columnas de X , $\mathbf{1}$ representa un vector columnas de unos, y \bar{x} representa la media de X . Luego la matriz escalada es:

$$X = X - \mathbf{1}\bar{x}^T \quad (19)$$

Quedando, para el posterior desarrollo, que la matriz X viene dada por (19).

Algoritmo MCA. La aplicación del método MCA requiere saber cuál es el número de componentes que se usaran del conjunto dado en la matriz de datos experimentales D , esto se determina por conocimiento técnico de los datos o por estudio matemático:

$$D_{(n,p)} = C_{(n,k)} P_{(k,p)}^T \quad (20)$$

$$D_{(n,p)} = U_{(n,n)} S_{(n,p)} V_{(p,p)}^T \quad (21)$$

En la ecuación (20) este número es determinado usando análisis de componentes principales, donde el número de columnas de la matriz P , suministra este valor. En la ecuación (21) el número de componentes es obtenido mediante descomposición de valores singulares, queda determinado por la escogencia del número de valores singulares significativos de la

matriz S . La matriz D con la que se trabaja será la obtenida luego de la reducción al número de componentes y su tamaño es de (n, k) .

Una estimación inicial de la matriz C o S^T , en la ecuación (11), estas estimaciones pueden ser obtenidas de manera aleatoria o a partir de técnicas conocidas como: análisis de factores evolutivos (EFA), simple-to-use Iterative self-modeling mixture analysis (SIMPLISMA) o descomposición de valores singulares (Bogomolov, Hachey y Williams, 2005).

Para aplicar EFA, de acuerdo con Esteban, Ariño, Díaz-Cruz, Díaz-Cruz y Tauler (2000), se toma una sub-matriz que solo contiene la información de las dos primeras columnas de la matriz D y se realiza una descomposición de valores singulares, una vez obtenidos los valores singulares, se agrega la tercera columna a la sub-matriz y luego se vuelve a aplicar la descomposición de valores singulares, repitiendo el proceso sobre toda la matriz D .

Este análisis se puede hacer de manera sucesiva desde el inicio hasta el final (*forward*) o viceversa (*backward*). Las estimaciones iniciales corresponden a aquellas que tengan los valores singulares más grandes en cada proceso. Sea D la matriz en estudio, el proceso general para encontrar la estimación inicial para la matriz C , de la ecuación (11), consiste en:

Se aplica descomposición de valores singulares de forma sucesiva a la matriz D , llamada D_i :

$$D_i = [D(:, 1) D(:, 2) \dots D(:, i)] \text{ con } i = 1, \dots, k \quad (22)$$

$$D_i = U_i S_i V_i^T \quad (23)$$

Donde U_i, V_i son ortogonales entre sí y son los vectores propios por derecha y por izquierda de la matriz D_i . Los valores singulares están en la matriz S_i . Se seleccionan los vectores de la matriz U_i que corresponden a los mayores valores singulares en cada matriz, para formar los vectores columnas de la matriz C .

Para encontrar la estimación inicial de S^T se procede de igual forma como se indicó para obtener C , trabajando con la matriz traspuesta de D . Para encontrar la estimación inicial aplicando la técnica SIMPLISMA (Rodríguez, Matera y Pérez, 2016), en primer lugar, se halla la columna con la mayor desviación estándar relativa, que va a corresponder a la primera variable pura, luego se procede a la normalización de dicha columna.

La segunda variable de mayor pureza, debe poseer la mayor desviación y, la mínima correlación con la primera variable pura encontrada, para esto debe calcularse un factor de ponderación, el cual corresponde al determinante de la matriz compuesta por las variables puras encontradas y cada columna de la matriz en estudio que aún no ha sido seleccionada. Se escoge la de mayor valor. Así sucesivamente hasta seleccionar todas las variables puras.

Considerando la matriz de estudio D , las variables puras a encontrar formaran la estimación inicial, siguiendo con Rodríguez et al. (2016), el procedimiento para la matriz inicial C consiste en: Calcular los vectores media y desviación estándar de D :

$$\mu = \frac{1}{n} D^T \mathbf{1} \quad (24)$$

$$\sigma = \sqrt{\frac{1}{n} \left(D^T D - \frac{1}{n} D^T \mathbf{1} \mathbf{1}^T D \right)} \quad (25)$$

Calcular el valor de α :

$$\alpha = 0.01 * \mu_{max} \quad (26)$$

μ_{max} : Es el mayor valor de μ

Calcular el vector de la desviación estándar relativa:

$$p = \frac{\sigma}{(\mu + \alpha)} \quad (27)$$

El máximo valor de p , corresponde a la primera variable pura. Calcular la matriz de correlación y factor de ponderación con:

$$Q = \frac{D}{\sqrt{\sigma^2 + (\mu + \alpha)^2}} \quad (28)$$

Se calcula la matriz de correlación M_c :

$$M_c = \frac{1}{n} Q^T Q \quad (29)$$

Se calcula el factor de peso:

$$w_i = \det(M_{c_i}^T M_{c_i}) \quad (30)$$

Donde M_{c_i} es la matriz formada por las variables puras encontradas y cada i -ésima columna de la matriz de datos que aún no ha sido seleccionada. Para las siguientes variables puras, se calcula la desviación estándar relativa modificada:

$$p_i = w_i \left(\frac{\sigma_i}{(\mu_i + \alpha)} \right) \quad (31)$$

Donde el subíndice i , representa al número de la columna correspondiente a la variable. Se selecciona la que tenga mayor p . Se detiene el proceso cuando se ha encontrado k variables puras, las cuales formaran la estimación de la matriz C . Para estimar la matriz S^T , se procede de igual manera en la búsqueda de las variables puras usando a la matriz traspuesta de D .

Para aplicar descomposición en valores singulares, según Mardia, Kent y Bibby (1979), en la búsqueda de la matriz inicial, se factoriza la matriz experimental D en tres matrices:

$$D = USV^T \quad (32)$$

Las matrices U y V son unitarias y ortogonales y están formadas por los vectores propios por la derecha y por la izquierda, respectivamente, D la matriz S es diagonal y contiene los valores singulares. Para la estimación de S^T , en la ecuación (11), se parte de la ecuación (32) quedando:

$$DV = US \quad (33)$$

En esta ecuación se observa que la matriz V no causa modificación en la variabilidad de la matriz D , solo produce una rotación de ejes, por tanto, la matriz S puede considerarse como una estimación inicial para S^T . Tomando solo las primeras k columnas. Para estimar a la matriz C , en la ecuación (11), se trabaja de forma igual con la matriz traspuesta de D .

En esta fase inicial se hace una traslación de la matriz original, sin causar modificación en la variabilidad de los elementos. En los mínimos cuadrados alternos se estima la matriz C o S^T a partir de la matriz original (Esteban et al., 2000), estas estimaciones se caracterizan por tener

vectores no correlacionados, ya sea que se realice por métodos matemáticos o de forma aleatoria. En este caso no se altera la variabilidad de la matriz experimental.

Fase de desarrollo

Algoritmo ACP. Este método tiene por objetivo reducir la dimensión de las variables preservando en lo posible la estructura de la varianza de la matriz original. Obtención de la primera componente principal:

$$c_1 = Xp_1 \quad (34)$$

La varianza de c_1 :

$$\sigma_{c_1}^2 = p_1^T S p_1 \quad (35)$$

Donde $S = \frac{1}{n-1} X^T X$, es la matriz de covarianza de X . Los vectores p y c indican las correspondientes filas de las matrices P y C . Se maximiza la varianza de c_1 , resolviendo el siguiente problema de optimización con restricciones:

$$\text{Maximizar: } p_1^T S p_1$$

$$\text{sujeto a: } \|p_1\| = 1 \quad (36)$$

Aplicando los multiplicadores de Lagrange:

$$F(\lambda, p_1) = p_1^T S p_1 - \lambda(p_1^T p_1 - 1)$$

Derivando respecto a las variables e igualando a cero:

$$\frac{\partial F}{\partial p_1} = 2S p_1 - 2\lambda p_1 = 0$$

$$\frac{\partial F}{\partial \lambda} = p_1^T p_1 - 1 = 0$$

La solución cumple que:

$$S p_1 = \lambda p_1 \quad (37)$$

El vector p_1 , define la primera componente principal, es un vector propio de la matriz de covarianza S , corresponde al mayor valor propio de la matriz de covarianza, como se demostró en la ecuación (10). Se calcula la matriz residual E , que corresponderá a la nueva matriz X :

$$E = X - C P^T \quad (38)$$

$$X = E \quad (39)$$

La obtención de la componente principal k se encuentra a partir de:

$$c_k = Xp_k \quad (40)$$

La varianza de c_k :

$$\sigma_{c_k}^2 = p_k^T S p_k \quad (41)$$

El problema consiste en:

$$\begin{aligned} & \text{Maximizar: } p_k^T S p_k \\ & \|p_k\| = 1 \\ \text{sujeto a: } & \begin{cases} p_i^T p_k = 0 \text{ (ortogonalidad) con } i = 1, 2, \dots, k-1 \end{cases} \end{aligned} \quad (42)$$

Se resuelve aplicando los multiplicadores de Lagrange:

$$F(p_k, \lambda_1, \lambda_i) = p_k^T S p_k - \lambda_1 (p_k^T p_k - 1) - \lambda_2 p_1^T p_k - \dots - \lambda_k p_{k-1}^T p_k$$

Derivando respecto a las variables e igualando a cero:

$$\begin{aligned} \frac{\partial F}{\partial p_k} &= 2S p_k - 2\lambda_1 p_k - \lambda_2 p_1 - \dots - \lambda_k p_{k-1} = 0 \\ \frac{\partial F}{\partial \lambda_1} &= p_k^T p_k - 1 = 0 \\ \frac{\partial F}{\partial \lambda_2} &= p_1^T p_k = 0 \\ \frac{\partial F}{\partial \lambda_k} &= p_{k-1}^T p_k = 0 \end{aligned}$$

Pre-multiplicando la $\frac{\partial F}{\partial p_k}$ por p_k^T :

$$2p_k^T S p_k - 2\lambda_1 p_k^T p_k - \lambda_2 p_k^T p_1 - \dots - \lambda_k p_k^T p_{k-1} = 0$$

Queda:

$$S p_k = \lambda_1 p_k \quad (43)$$

Entonces, el vector p_k , que define la componente k principal, es el vector propio asociado al k mayor valor propio de la matriz de covarianza S . La magnitud de los valores propios indica la cantidad de variabilidad o información retenida por cada uno de las componentes principales. Se calcula la matriz residual E y la nueva matriz X a partir de las ecuaciones (38) y (39).

Algoritmo MCA. En esta fase se determinan las matrices C, S^T , formando la descomposición bilineal de la matriz D .

Si en la fase inicial se calculó a C , se obtiene a:

$$S^T = C^+ D \quad (44)$$

Si en la fase inicial se obtuvo a S^T , entonces se determina:

$$C = D(S^T)^+ \quad (45)$$

Donde C^+ , $(S^T)^+$ corresponden a las pseudo-inversa de las matrices C, S^T :

$$C^+ = (C^T C)^{-1} C^T \quad (46)$$

$$(S^T)^+ = S(S^T S)^{-1} \quad (47)$$

Estas matrices son encontradas a partir del teorema de Gauss-Markov que da respuesta a las soluciones óptimas de las ecuaciones formadas por el método de mínimos cuadrados ordinarios en las ecuaciones (14) y (16). Con la nueva matriz estimada mediante las ecuaciones (44) o (45) y con la matriz D , se determina la nueva matriz S^T o C la cual sustituirá a la matriz obtenida como estimación inicial:

$$C = D \quad (48)$$

En el caso de que en la fase inicial se haya obtenido a C .

$$S^T = C^+ D \quad (49)$$

En el caso de que se haya calculado en la fase inicial a S^T . Luego se calcula la matriz de residuo E :

$$E = D - CS^T \quad (51)$$

Este proceso se repite tantas veces hasta alcanzar el valor mínimo deseado en la matriz de residuos. En esta fase de desarrollo se observa que, en el algoritmo de análisis de componentes principales los vectores que formaran las matrices C y P son obtenidos uno a uno basado en el concepto de maximizar la varianza del vector c , durante lo cual se forma un problema de optimización con restricciones, cuyo objetivo es encontrar el vector p , en cada iteración se redefine la matriz de datos X , mediante la resta de la información obtenida en CP^T .

Mientras en el algoritmo de mínimos cuadrados alternos el objetivo es encontrar las matrices C, S^T mediante la aplicación del método de mínimos cuadrados ordinarios a través de las pseudo-inversa de las matrices y la matriz de datos originales D ; por último se calcula la matriz residual, restándole a la matriz D la información aportada en CS^T .

Fase final

Esta fase puede ser también llamada fase de convergencia debido a que en ambos algoritmos se busca la descomposición bilineal de la matriz original, la convergencia ocurre cuando:

$$CP^T \rightarrow X$$

En el caso de componentes principales.

$$CS^T \rightarrow D$$

En el caso de mínimos cuadrados alternos, los procedimientos para detener los algoritmos iterativos de los métodos ACP y MCA consisten en colocar un valor máximo de iteraciones, y evaluar sucesivamente la matriz residual, deteniendo el algoritmo cuando se alcanza el valor mínimo residual o al llegar al número máximo de iteraciones, lo que suceda primero.

Evaluación de la matriz residual

La determinación del valor mínimo para la matriz residual se puede calcular mediante la norma euclídea

$$\|E\| = \left(\sum_{i=1}^m \sum_{j=1}^n |e_{ij}|^2 \right)^{1/2} = \|E\|_F \quad (52)$$

Donde e_i ($i = 1, \dots, n$) son los vectores columnas de la matriz residual. También se puede usar la raíz del cuadrado medio del error (*RMS*, del inglés *root mean square error*): definido como:

$$RMS = \sqrt{\frac{\sum_i \sum_j (e_{ij})^2}{n}} \quad (53)$$

El algoritmo se detiene cuando los valores de la norma y RMS se acercan a cero. Comparación de los algoritmos análisis de componentes principales y mínimos cuadrados alternos. Al presentar los algoritmos dividiendo sus funciones por fases, como se hizo anteriormente, se puede deducir un conjunto de semejanzas y diferencias entre ellos.

Entre las semejanzas se puede observar:

- Ambos algoritmos tienen por objetivo la descomposición bilineal de la matriz de datos originales.
- En el proceso final de los algoritmos, los parámetros usados para detenerlos son los mismos.

Las diferencias se observan en las fases de inicio y de desarrollo:

a) El algoritmo de MCA requiere de la determinación del número de componentes la cual regirá el tamaño de las matrices, esto asegura que las componentes a usar en la matriz D sean independientes, desde una visión matemática, de esta forma la variabilidad de ruido presente sea mínima, desde un enfoque estadístico. En el algoritmo ACP el número de componentes se determinan durante el proceso.

b) El algoritmo MCA necesita, para comenzar, de un algoritmo generador de la matriz de inicio, los cuales pueden ser: EFA, SVD o SIMPLISMA, la matriz generada tiene la característica de estar formada por vectores independientes. Mientras que en el algoritmo de ACP solo requiere para su inicio, centrar la matriz de datos, la cual se hace restando la media a las columnas.

c) En el algoritmo ACP, se obtienen las matrices de la descomposición mediante el concepto de varianza con la restricción de que una nueva componente debe ser ortogonal a las obtenidas previamente, para esto se plantea un problema de optimización con restricciones cuya solución se obtiene al aplicar multiplicadores de Lagrange. Por otro lado, el algoritmo MCA obtiene las matrices de la descomposición basada en el método de mínimos cuadrados ordinarios, cuya solución puede plantearse como un problema de optimización sin restricciones aplicando el teorema de Gauss-Markov.

d) En el algoritmo de ACP, la matriz X es renovada en cada iteración al restarle la información suministrada por el producto de CP^T . En el algoritmo MCA, la matriz D permanece constante durante la fase de desarrollo.

e) Además, el algoritmo MCA, en la fase de desarrollo, exige trabajar con matrices inversas, lo que no representa obstáculo por los requisitos alcanzados en la fase inicial donde se selecciona el número de componentes y se estima una matriz inicial, formada con vectores independientes.

f) Por último, en el algoritmo ACP, las matrices que forman la descomposición son únicas, en cambio, en el algoritmo MCA, las matrices no son únicas, el proceso para obtenerlas es libre de restricción rotacional, es decir:

Considere:

$$S_{nueva}^T = ZS^T$$

$$C_{nueva} = CZ^{-1}$$

Donde Z es una matriz diagonal que tiene inversa. Entonces:

$$D = C_{nueva}S_{nueva}^T = CZ^{-1}ZS^T = CS^T$$

De esta manera, las matrices C , S pueden asumir diferentes valores, sin alterar la matriz D . Un resumen comparado de lo explicado se muestra en la Cuadro 1.

Cuadro 1

Comparación de los algoritmos ACP y MCA

| Algoritmos | ACP | MCA |
|---------------------|---|--|
| Fase Inicial | Se escala la matriz de datos: $X = X - 1\bar{x}^T$ Con \bar{x} como la media de X y 1 como un vector columna de unos. | Se determina el número de componentes por: ACP: $D_{(n,p)} = C_{(n,k)}P_{(k,p)}^T$ Donde el número de columnas de P indica el número de componentes. SVD: $D_{(n,p)} = U_{(n,n)}S_{(n,p)}V_{(p,p)}^T$ Donde el número de valores singulares significativos indican el número de componentes. |

Cuadro 1. *Cont.*

| Algoritmos | ACP | MCA |
|---------------------------|--|--|
| Fase Inicial | | La matriz D luego de la reducción será de tamaño (n, k) . Se calcula la matriz inicial de: C o S^T de forma aleatoria o por alguna técnica tal como: EFA o SIMPLISMA o SVD. |
| Fase de desarrollo | Se resuelve el problema de optimización con restricciones: <i>Maximizar:</i> $p_k^T S p_k$ <i>sujeto a:</i> $\begin{cases} \ p_k\ = 1 \\ p_i^T p_k = 0 \text{ con } i = 1, 2, \dots, k \end{cases}$ Se calcula: $c_k = X p_k$ Se calcula la matriz residual, que será la nueva matriz X : $E = X - C P^T$ $X = E$ | Se aplica mínimos cuadrados ordinarios: $\min_C \ D - C S^T\ $ $\min_{S^T} \ D - C S^T\ $ Para obtener a: $S^T = C^+ D$ O $C = D(S^T)^+$ Dependiendo de cuál fue la matriz inicial estimada. Se procede a obtener la matriz C o S^T que sustituirá a la matriz inicial estimada. Se calcula la matriz residual: $E = D - C S^T$ |
| Fase final | Se busca la convergencia de: $C P^T \rightarrow X$ Usando la matriz residual E , se calcula: $\ E\ _F = \sqrt{\text{tr}(E^T E)}$ $RMS = \sqrt{\frac{\sum_i \sum_j (e_{ij})^2}{n}}$ Cuando estos valores son $< \varepsilon$ con $\varepsilon = 10^{-4}$, se detiene el algoritmo. Otra manera de detener el algoritmo es considerar un máximo de iteraciones. | Se busca la convergencia de: $C S^T \rightarrow D$ Usando la matriz residual E , se calcula: $\ E\ _F = \sqrt{\text{tr}(E^T E)}$ $RMS = \sqrt{\frac{\sum_i \sum_j (e_{ij})^2}{n}}$ Cuando estos valores son $< \varepsilon$ con $\varepsilon = 10^{-4}$, se detiene el algoritmo. Otra manera de detener el algoritmo es considerar un máximo de iteraciones. |

Fuente: Elaboración propia (2018)

Resultados

Para medir la eficiencia de los algoritmos desarrollados en esta investigación se realizó el estudio de dos ejemplos diseñados para evaluar los beneficios de cada algoritmo. La evaluación consiste en analizar la matriz de entrada de cada ejemplo, buscando la descomposición bilineal de las mismas y verificando que la variabilidad de las componentes de la matriz original no es alterada.

Los parámetros de evaluación aplicados son: Mínimo valor propio (*mvp*) se puede seleccionar las componentes asociadas a valores propios superior a un valor prefijado, por ejemplo la varianza media.

$$var_m = \sum_{i=1}^k \frac{\lambda_i}{k}$$

Donde λ_i con $(i = 1, \dots, k)$ son los valores propios.

Representación gráfica: los valores propios, se pueden representar en el plano graficando el número de componentes versus los valores propios (λ_i), y buscar un punto a partir del cual los valores propios son relativamente pequeños. Se indican las variables de entrada y la variable de salida o respuesta, esto con el objeto de mostrar la aplicación de las variables en cada ejemplo.

El primer ejemplo fue seleccionado para demostrar la robustez del algoritmo en identificar la colinealidad de los datos. El segundo ejemplo fue seleccionado para evidenciar la eficiencia de los algoritmos ante problemas de conjuntos grandes de datos, donde el número de variables supera significativamente al número de muestras.

Ejemplo 1: Datos de cemento Portland

Ingredientes (%): (variables independientes)

Columna1: $3CaO \cdot Al_2O_3$ (aluminato tricálcico); Columna 2: $3CaO \cdot SiO_2$ (silicato tricálcico)

Columna3: $4CaO \cdot Al_2O_3 \cdot Fe_2O_3$ (aluminoferrita tetracálcico); Columna 4: $2CaO \cdot SiO_2$ (silicato beta - dicálcico) Calor (cal / gm): (variable respuesta)

Calor de endurecimiento después de 180 días

Este ejemplo contiene una matriz de entrada formada por los vectores de ingredientes y un vector de salida llamado calor (Woods, Steinour y Starke, 1932). El número de observaciones es de trece por vector.

| | | | |
|----|----|----|----|
| 7 | 26 | 6 | 60 |
| 1 | 29 | 15 | 52 |
| 11 | 56 | 8 | 20 |
| 11 | 31 | 8 | 47 |
| 7 | 52 | 6 | 33 |
| 11 | 55 | 9 | 22 |
| 3 | 71 | 17 | 6 |
| 1 | 31 | 22 | 44 |
| 2 | 54 | 18 | 22 |
| 21 | 47 | 4 | 26 |
| 1 | 40 | 23 | 34 |
| 11 | 66 | 9 | 12 |
| 10 | 68 | 8 | 12 |

Figura 1. Matriz de entrada.

Fuente: Elaboración propia (2018)

Los datos en estudio será la matriz de entrada original más la repetición de las columnas 2 y 4 en las posiciones 4 y 5 quedando una matriz de tamaño trece por seis, se llamará matriz de estudio (X).

Descomposición mediante el algoritmo ACP

Al aplicar el algoritmo de análisis de componentes principales se obtienen las matrices C de tamaño $(13, 4)$ y la matriz P de tamaño $(6, 4)$, esto indica que el número de componentes principales son cuatro, como era de esperarse debido a su correspondencia con el número de vectores independientes. La descomposición es alcanzada en cuatro iteraciones con un error de $6,5244e-14$, calculado con la norma de Frobenius. La representación de $X = CP^T$ es:

| | | | | | |
|---------|---------|---------|---------|---------|---------|
| 7.0000 | 26.0000 | 6.0000 | 26.0000 | 60.0000 | 60.0000 |
| 1.0000 | 29.0000 | 15.0000 | 29.0000 | 52.0000 | 52.0000 |
| 11.0000 | 56.0000 | 8.0000 | 56.0000 | 20.0000 | 20.0000 |
| 11.0000 | 31.0000 | 8.0000 | 31.0000 | 47.0000 | 47.0000 |
| 7.0000 | 52.0000 | 6.0000 | 52.0000 | 33.0000 | 33.0000 |
| 11.0000 | 55.0000 | 9.0000 | 55.0000 | 22.0000 | 22.0000 |
| 3.0000 | 71.0000 | 17.0000 | 71.0000 | 6.0000 | 6.0000 |
| 1.0000 | 31.0000 | 22.0000 | 31.0000 | 44.0000 | 44.0000 |
| 2.0000 | 54.0000 | 18.0000 | 54.0000 | 22.0000 | 22.0000 |
| 21.0000 | 47.0000 | 4.0000 | 47.0000 | 26.0000 | 26.0000 |
| 1.0000 | 40.0000 | 23.0000 | 40.0000 | 34.0000 | 34.0000 |
| 11.0000 | 66.0000 | 9.0000 | 66.0000 | 12.0000 | 12.0000 |
| 10.0000 | 68.0000 | 8.0000 | 68.0000 | 12.0000 | 12.0000 |

Figura 2. Matriz $X = CP^T$
Fuente: Elaboración propia 2018)

Los valores propios de la matriz X y los de la matriz obtenida por el producto de CP^T , coinciden, con lo cual se verifica que la variabilidad de las componentes no son alteradas por la descomposición, estos valores, ordenados de mayor a menor, son:

$$[294.6894 \ 109.0479 \ 28.5406 \ 10.4002 \ 0.0000 \ 0.0000]$$

El número de valores significativos son cuatro, coincidiendo con el número de componentes. Queda comprobado que las matrices obtenidas a través de este algoritmo ACP descomponen de manera satisfactoria a la matriz X . Aplicando los parámetros de evaluación, el mínimo valor propio, indica como las dos primeras componentes son suficientes para representar las componentes principales:

$$var_m = 73,7797$$

Los valores propios mayores a var_m , son:

$$[294.6894 \ 109.0479]$$

La figura 3 representa el número de componentes versus los valores propios, representando en la parte superior los datos de la matriz original de entrada, y en la parte inferior se representan los datos del producto de la descomposición $X = CP^T$, se puede observar que a partir de la segunda componente los valores propios tienden a ser relativamente pequeños.

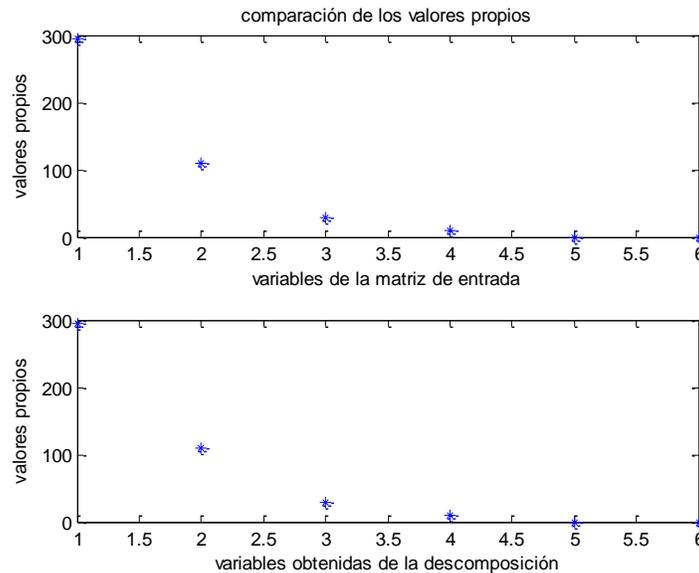


Figura 3. Ejemplo 1. Representación de los valores propios. Algoritmo ACP. Fuente: Elaboración propia (2018)

La tabla 1 muestra el error calculado por la norma de Frobenius, observándose una mejor aproximación cuando se trabaja con las cuatro componentes, en la cual no se altera la varianza total, esto es aplicado cuando se requiere de la descomposición simple de una matriz. Mientras, si se trabaja en un problema donde es aceptable la pérdida de varianza total a favor de disminuir el número de componentes, se puede trabajar, en este ejemplo, con las dos primeras componentes como se demostró con los parámetros de evaluación.

Tabla 1

Ejemplo 1. Relación de número de componentes con la varianza total

| Número de | Error | Varianza total (%) |
|-----------|------------|--------------------|
| 1 | 32,3676 | 66,57 |
| 2 | 15,1300 | 91,20 |
| 3 | 1,9318 | 97,65 |
| 4 | 6,5244e-14 | 100 |
| 5 | 1,2018e-14 | 100 |
| 6 | 6,0934e-15 | 100 |

Fuente: Elaboración propia (2018)

Descomposición a través del algoritmo MCA

Al aplicar el algoritmo de mínimos cuadrados alternos a la matriz de estudio se obtienen en una iteración y con un error de $1,1017e-13$ (norma de Frobenius) las matrices C con tamaño 13 por 2 y la matriz S de tamaño 6 por 2. El número de columnas indica que el algoritmo depura las variables, quedando de esta manera, solo dos vectores. La representación de la matriz $D = CS^T$ es:

| | | | | | |
|------|------|------|------|------|------|
| 5.39 | 25.5 | 13.2 | 25.5 | 59.4 | 59.4 |
| 5.58 | 28.8 | 12.5 | 28.8 | 42.1 | 42.1 |
| 8.10 | 55.9 | 11.1 | 55.9 | 19.7 | 19.7 |
| 5.70 | 31.0 | 11.9 | 31.0 | 46.7 | 46.7 |
| 7.93 | 51.4 | 12.5 | 51.4 | 32.4 | 32.4 |
| 8.05 | 55.0 | 11.3 | 55.0 | 21.8 | 21.8 |
| 9.65 | 70.9 | 11.0 | 70.9 | 6.37 | 6.37 |
| 5.67 | 31.4 | 11.6 | 31.4 | 44.7 | 44.7 |
| 7.94 | 54.0 | 11.2 | 54.0 | 22.4 | 22.4 |
| 7.17 | 47.4 | 10.8 | 47.4 | 25.6 | 25.6 |
| 6.54 | 40.4 | 11.3 | 40.4 | 34.8 | 34.8 |
| 9.17 | 65.9 | 11.2 | 65.9 | 11.8 | 11.8 |
| 9.41 | 67.7 | 11.4 | 67.7 | 11.7 | 11.7 |

Figura 4. Matriz $D = CS^T$
Fuente: Elaboración propia (2018)

Los valores propios de la matriz de estudio X y los de la matriz obtenida por el producto de CS^T , son los mismos, comprobándose que la variabilidad de los vectores principales no son alterados por la descomposición, estos valores, ordenados de forma decreciente, son:

Matriz X :

[294.6894 109.0479 28.5406 10.4002 0.0000 0.0000]

Matriz CS^T :

[294.6894 109.0479 0.0000 0.0000 0.0000 0.0000]

Queda verificado que las matrices obtenidas a través de este algoritmo MCA descomponen de manera satisfactoria a la matriz X . El estudio de los parámetros de evaluación coincide con el estudio realizado a través del algoritmo ACP, obteniendo los mismos valores propios. Reduciendo el número de componentes a las dos primeras, las cuales corresponden a los dos valores propios mayores a:

$$var_m = 73,7797$$

En la figura 5 se representa en la parte superior el número de variables versus los valores propios de la matriz de entrada, y en la parte inferior el número de variables contra los valores propios del producto obtenido de las matrices de descomposición.

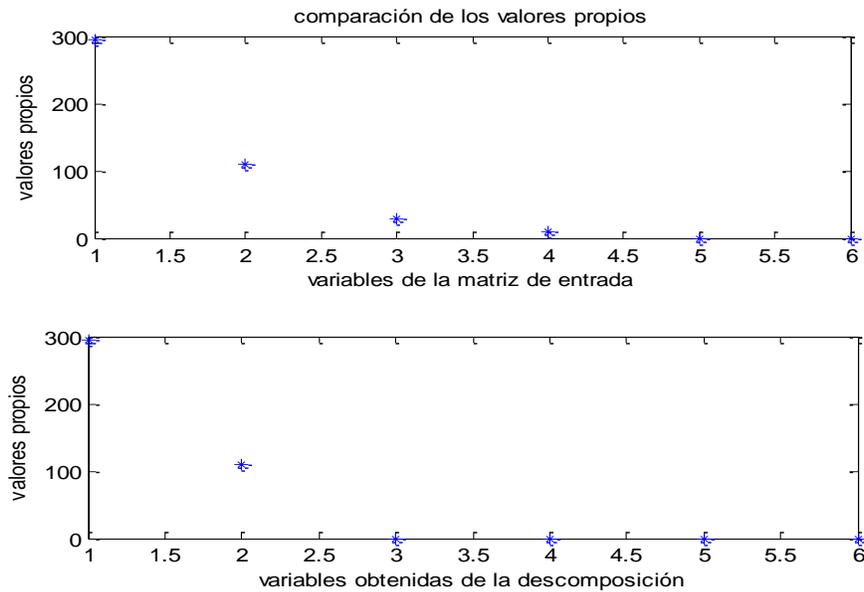


Figura 5. Ejemplo 1: Representación de los valores propios. Algoritmo MCA.
Fuente: Elaboración propia (2018)

En el algoritmo MCA, el número de componentes fue determinado en la fase inicial mediante la descomposición de valores singulares, quedando que la matriz de estudio era de tamaño 13 por 2. Para este ejemplo la matriz estimada inicial fue C , con un tamaño 13 por 2.

Ejemplo 2: Datos espectrales y de octanaje de la gasolina

NIR: espectros NIR, medidos en intervalos de 2 nm de 900 nm a 1700 nm
Octano: octanaje

En este ejemplo la matriz de entrada está representada por NIR, es de tamaño 60 filas (muestras) por 401 columnas (variables), la matriz de respuesta o salida está representada por octano y es un vector de 60 muestras.

Descomposición a través del algoritmo ACP

Al aplica a los datos NIR el algoritmo de análisis de componentes principales se obtienen las matrices C de tamaño 60 por 59 y la matriz P de tamaño 401 por 59, esto indica que hay 59 componentes principales. La descomposición es alcanzada en cincuenta y nueve iteraciones con un error de $4,8351e-15$, calculado con la norma de Frobenius. Los valores propios de la matriz de entrada (NIR) y de la matriz obtenida por la descomposición CP^T , coinciden:

| | | | | | | | | | |
|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 44.6814 | 1.5311 | 0.4997 | 0.4699 | 0.3659 | 0.2107 | 0.1615 | 0.1309 | 0.1103 | 0.0979 |
| 0.0866 | 0.0778 | 0.0647 | 0.0613 | 0.0571 | 0.0504 | 0.0456 | 0.0427 | 0.0366 | 0.0305 |
| 0.0284 | 0.0244 | 0.0230 | 0.0202 | 0.0178 | 0.0170 | 0.0161 | 0.0146 | 0.0136 | 0.0115 |
| 0.0100 | 0.0093 | 0.0087 | 0.0082 | 0.0080 | 0.0074 | 0.0067 | 0.0066 | 0.0059 | 0.0057 |
| 0.0053 | 0.0049 | 0.0046 | 0.0045 | 0.0043 | 0.0041 | 0.0039 | 0.0038 | 0.0035 | 0.0034 |
| 0.0034 | 0.0033 | 0.0031 | 0.0030 | 0.0029 | 0.0028 | 0.0026 | 0.0024 | 0.0023 | 0.0020 |

Queda comprobado que las matrices obtenidas a través del algoritmo ACP descomponen de manera satisfactoria a la matriz X . Para este ejemplo no hay valores propios con valor cero, debido a la no existencia de vectores repetidos. Aplicando los parámetros de evaluación queda que, con el mínimo valor propio al calcularse una varianza media de 0,8191 se obtienen las dos primeras componentes como suficientes para representar la matriz de componentes principales, con valores propios:

$$[44.6814 \quad 1.5311]$$

En la figura 6 se muestra la relación del número de componentes con los valores propios, evidenciando que los dos primeros valores propios aportan la mayor variabilidad, el resto tienden a valores cercanos a cero. En la parte superior de la figura se muestran los valores correspondientes a la matriz de entrada, y en la parte inferior se muestran los valores obtenidos de la descomposición de la matriz.

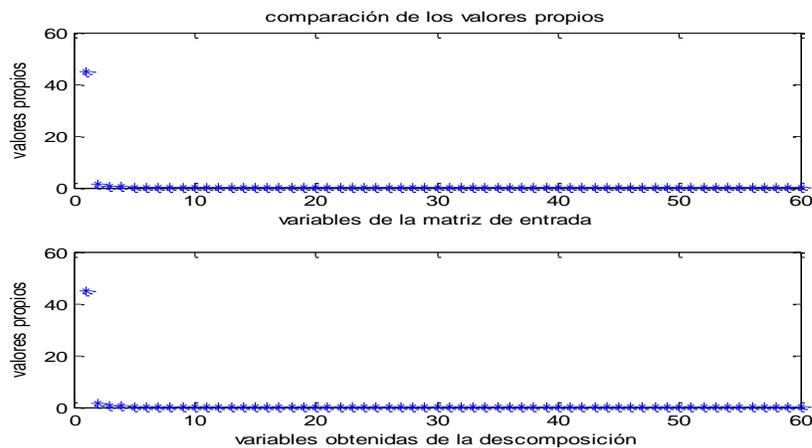


Figura 6. Ejemplo 2: Representación de los valores propios. Algoritmo ACP
Fuente: Elaboración propia (2018)

En la tabla 2 se muestra el error calculado a través de la norma de Frobenius. Se representaron los primeros 10 componentes y la correspondiente al componente 59. Se observó que se obtiene una mejor aproximación cuando se trabaja con las 59 componentes, en la cual no se altera la varianza total, esto es aplicado cuando se requiere de la descomposición simple de una matriz. Si se trabaja en un problema donde es aceptable la pérdida de varianza total con el objeto de reducir el número de componentes, solo son necesarias las dos primeras componentes para este ejemplo, como se demostró con los parámetros de evaluación.

Tabla 2
Ejemplo 2. Relación de número de componentes con la varianza total

| Número de componentes | Error | Varianza total (%) |
|------------------------------|--------------|---------------------------|
| 1 | 0,9924 | 90,92 |
| 2 | 0,7602 | 94,03 |
| 3 | 0,5729 | 95,05 |
| 4 | 0,4038 | 96,00 |
| 5 | 0,3443 | 96,75 |
| 6 | 0,2896 | 97,18 |

Tabla 2. Cont.

| Número de componentes | Error | Varianza total (%) |
|-----------------------|------------|--------------------|
| 7 | 0,2571 | 97,51 |
| 8 | 0,2304 | 97,77 |
| 9 | 0,2029 | 98,00 |
| 10 | 0,1812 | 98,20 |
| 59 | 4,8351e-15 | 100,00 |

Fuente: Elaboración propia (2018)

Descomposición mediante el algoritmo MCA

En la descomposición de la matriz de estudio del ejemplo 2 mediante mínimos cuadrados alternos se obtuvieron las matrices C , S de tamaños 60 por 2 y 401 por 2 respectivamente en una iteración con un error, calculado por la norma de Frobenius, de $9,9371e-15$. Los valores propios de la matriz de estudio y los de la matriz obtenida por el producto de CS^T , coinciden para las dos primeras componentes, verificándose que la variabilidad de los vectores principales no son alterados por la descomposición, estos valores, ordenados de forma decreciente, son:

Matriz de estudio:

| | | | | | | | | | | |
|---|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| [| 44.6814 | 1.5311 | 0.4997 | 0.4699 | 0.3659 | 0.2107 | 0.1615 | 0.1309 | 0.1103 | 0.0979 |
| | 0.0866 | 0.0778 | 0.0647 | 0.0613 | 0.0571 | 0.0504 | 0.0456 | 0.0427 | 0.0366 | 0.0305 |
| | 0.0284 | 0.0244 | 0.0230 | 0.0202 | 0.0178 | 0.0170 | 0.0161 | 0.0146 | 0.0136 | 0.0115 |
| | 0.0100 | 0.0093 | 0.0087 | 0.0082 | 0.0080 | 0.0074 | 0.0067 | 0.0066 | 0.0059 | 0.0057 |
| | 0.0053 | 0.0049 | 0.0046 | 0.0045 | 0.0043 | 0.0041 | 0.0039 | 0.0038 | 0.0035 | 0.0034 |
| | 0.0034 | 0.0033 | 0.0031 | 0.0030 | 0.0029 | 0.0028 | 0.0026 | 0.0024 | 0.0023 | 0.0020 |
|] | | | | | | | | | | |

Matriz CS^T :

| | | | | | | | |
|---|---------|--------|--------|--------|--------|-----|--------|
| [| 44.6814 | 1.5311 | 0.0000 | 0.0000 | 0.0000 | ... | 0.0000 |
|---|---------|--------|--------|--------|--------|-----|--------|

Queda verificado que las matrices obtenidas a través del algoritmo MCA descomponen de manera satisfactoria a la matriz de estudio. El estudio de los parámetros de evaluación coincide con el realizado a través del algoritmo ACP, obteniéndose los mismos valores propios. De esta forma el número de componentes se puede reducir a las dos primeras, las cuales corresponden a los dos valores propios mayores a:

$$var_m = 0,8191$$

En la figura 7 se representan los valores propios de las matrices, en la parte superior se grafica el número de variables versus los valores propios de la matriz de entrada, y en la parte inferior se grafica el número de variables contra los valores propios del producto obtenido de las matrices de descomposición.

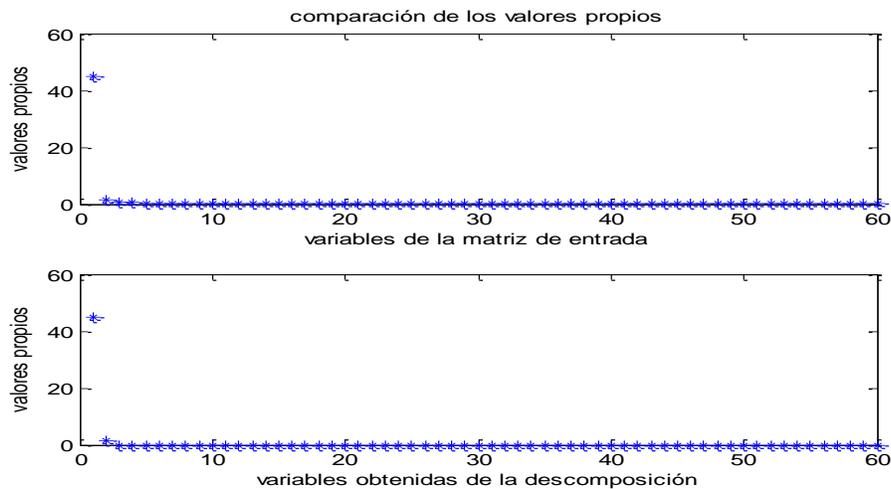


Figura 7 Ejemplo 2: Representación de los valores propios. Algoritmo MCA
Fuente: Elaboración propia 2018)

En el algoritmo MCA, el número de componentes fue determinado en la fase inicial usando la descomposición de valores singulares quedando la matriz de estudio de tamaño 401 por 2. Para este ejemplo la matriz estimada inicial fue S , con un tamaño 60 por 2. Mediante un análisis comparativo de los resultados de los ejemplos se puede concluir:

- a) Los algoritmos descomponen de manera satisfactoria las matrices arrojando errores pequeños basados en la norma de Frobenius.
- b) Ambos algoritmos reducen significativamente el número de componentes de la matriz en estudio, asegurando que las componentes principales son independientes.
- c) El algoritmo mínimos cuadrados alternos determina número componentes en la fase inicial.
- d) El algoritmo de análisis de componentes principales determina el número de componentes en la fase de desarrollo.
- e) Se pudo observar que la matriz se puede representar con dos componentes principales conservando una varianza no menor a un 90%.

Conclusiones

Los métodos estudiados en la investigación permitieron a los datos originales sometidos a transformaciones, conservar las características de variabilidad. La determinación del número de componentes se basó en la optimización de la varianza, ya sea que el proceso se lleve a cabo de manera directa, como en el algoritmo de análisis de componentes principales o de forma indirecta, como en el algoritmo de mínimos cuadrados alternos.

Los objetivos son alcanzados mediante la solución de problemas de optimización que buscan la convergencia entre los datos originales y los calculados. Para aplicar el algoritmo MCA a la matriz de estudio, debe estar formada por vectores independientes y el número de componentes seleccionarse en la fase inicial, previo al desarrollo óptimo del algoritmo.

El método ACP trabaja con toda la matriz de datos originales escogiendo las componentes una a una. Ambos algoritmos ACP y MCA pueden trabajar con matrices de cualquier tamaño, como se comprobó mediante los ejemplos. Al aplicar los algoritmos, la matriz C contiene la



información sobre las filas de la matriz de datos originales, y la matriz P (caso de ACP) o la matriz S (caso de MCA) contienen información sobre las columnas de la matriz de datos.

En estos algoritmos al igual que en otras técnicas de modelado de problemas se debe tener presente el uso del menor número de variables principales para asegurar la efectividad y robustez del modelo.

Referencias Bibliográficas

- Bogomolov, A; Hachey, M; y Williams, A (2005). *Software for Interactive Curve Resolution Using SIMPLISMA*. Progress in Chemometrics Research. New York, Estados Unidos: Editorial Nova Science.
- Cuadras, C. (1991) *Métodos de Análisis Multivariante*. Barcelona, España: Promociones y Publicaciones Universitarias, S. A., Eunivar.
- Curtis, M., Baum, J., Cook, C., Edgerton, J. y Rabidou, S. (2011). *Statistical Properties of Alternating Least Squares Estimators of a Collaborative Filtering Model*. Recuperado de: <http://userpages.umbc.edu/~gobbert/papers/CurtisUMBCReview2011.pdf>
- De Juan, A., Vander Heyden, Y., Tauler, R. & Massart, D. (1997). Assessment of new constraints applied to the alternating least squares method. *Analytica Chimica, Acta*. p. 307-318. Recuperado de: [https://www.researchgate.net/profile/Anna_De_Juan2/publication/223092318_Assessment_of_new_constraints_applied_to_the_alternating_least_squares_\(ALS\)_method/links/09e4151026e6761430000000.pdf](https://www.researchgate.net/profile/Anna_De_Juan2/publication/223092318_Assessment_of_new_constraints_applied_to_the_alternating_least_squares_(ALS)_method/links/09e4151026e6761430000000.pdf) .
- Esteban, M., Ariño, C., Diaz-Cruz, J., Diaz-Cruz, M., Tauler, R. (2000). Multivariate curve resolution with alternating least squares optimisation: a soft-modelling approach to metal complexation studies by voltammetric techniques. *TrAC-Trend Anal Chemical*, 19, 49-61.
- Enrique, M. (2006). *Seguimiento Cuantitativo de Reacciones de Resinas Epoxi mediante Espectroscopia de Infrarrojo Cercano y Métodos de Resolución de Curvas*. Tesis Doctoral. Universitat Rovira I Virgili. Cataluña España.
- Geladi P. y Kowalski B. (1986). Partial Least-Squares Regression: A Tutorial. *Analytica Chimica, Acta* 185, p.1-17. Indirect Observations. Elsevier.
- Hurtado, J. (2015). *Metodología de la Investigación Holística*. Caracas, Venezuela: Ediciones SYPAL
- Johnson R., Miller I. & Freund J. (2013). *Probability and Statistics for Engineers*. Pearson New International. Edition Pearson Education Limited
- Lahoz-Beltrá, R. (2004). *Bioinformática: Simulación, vida artificial e inteligencia artificial*. Madrid: Ediciones Díaz de Santos, S.A.
- Mardia K., Kent J. y Bibby J. (1979). *Multivariate analysis*. Academic Press. San Diego, United States: Academic Press. INC. Recuperado de: <https://statisticalsupportandresearch.files.wordpress.com/2017/06/k-v-mardia-j-t-kent-j-m-bibby-multivariate-analysis-probability-and-mathematical-statistics-academic-press-inc-1979.pdf>



- Rodríguez E., Matera J. Pérez M. (2016). *Determinación de Variables en Problemas Multivariantes. Método: SIMPLISMA. Impacto Científico*, 11(2), 137-154.
- Tauler R, Maeder M, De Juan A. (2009). *Multiset Data Analysis. Extended Multivariate Curve Resolution. Elsevier*. Recuperado de:
<http://www.iasbs.ac.ir/chemistry/chemometrics/history/11th/multiset%20data%20analysis.pdf>
- Woods, H., H. Steinour, H. Starke. (1932). Effect of Composition of Portland Cement on Heat Evolved during Hardening. *Industrial and Engineering Chemistry*, 24(11), 1207-1214.
- Ximénez G., C. y San Martín C. R. (2013). *Fundamentos de las Técnicas Multivariantes*. Madrid, España. Editorial Universidad Nacional de Educación a Distancia (UNED).