Vol.5. Num.2 Año 2010

Universidad de Puerto Rico. Recinto de Río Piedras http://paideia.uprrp.edu

Álgebra matricial de las técnicas factoriales para usuarios no matemáticos

Juan P. Vázquez Pérez

María P. Galindo Villardón

Resumen

Este artículo presenta un resumen del Álgebra matricial aplicada en la estadística multivariada, específicamente, en las técnicas factoriales. El propósito del mismo es reforzar el conocimiento de usuarios no matemáticos en dicha área, dada la necesidad de su comprensión. Se incluye una explicación de las operaciones básicas y su aplicación en las operaciones superiores, como el cálculo de los valores y vectores propios. Por otro lado, se menciona la utilidad de estas operaciones en la reducción de la dimensionalidad de la información. El tipo de técnica multivariada que se seleccione para dicha reducción depende de la información que contenga la matriz de datos y el propósito de la investigación. Finalmente, se presenta la aplicación del Álgebra matricial en el procedimiento para realizar un Análisis de Componentes Principales. Como parte de la explicación, se utilizan ejemplos reales para facilitar el aprendizaje en torno a la realización del análisis y la interpretación de los resultados.

Descriptores: álgebra matricial; estadísticas multivariantes; reducción de información estadística; análisis de componentes principales

Abstract

This article presents the summary of Matrix Algebra applied in multivariate statistic, in specific, factor techniques. Its purpose is to reinforce knowledge in this area to non-mathematician users. It includes an explanation of basic operations and their application in higher operations, as the calculation of eigenvalues and eigenvectors. In the other hand, it mentions the utility of these operations in the reduction of information dimensionally. The selection of the multivariate technique type depends of the information contained in the data matrix and the purpose of statistic analysis. Finally, it presents the matrix algebra application in the procedure to conduct a Principal Components Analysis. As part of the explanation, real examples are used to facilitate the learning on how to perform the analysis and interpretation of the results.

Descriptors: matrix algebra; multivariate statistics; statistical information reduction; principals components

Al estudiarse las estadísticas multivariadas se parte de medidas múltiples de los individuos u objetos bajo investigación. De acuerdo con Hair, Anderson, Tatham & Black (1999), las estadísticas multivariadas son extensiones del análisis univariado (i.e., análisis de las distribuciones de una sola variable) y del análisis bivariado (i.e., análisis conjunto de las medidas de dos variables). Según estos autores, las estadísticas multivariadas surgen de la necesidad de entender el concepto básico del análisis multivariado, que es el valor teórico. El mismo se define como una combinación de variables con una ponderación que se determina de manera empírica. Esto ayuda a que el conjunto de las variables se represente de la mejor manera posible y se adapte al objetivo de cada análisis multivariado.

Los datos de individuos, que a su vez correspondan a unas variables específicas, se ubican en matrices para organizarlos. Estas matrices permiten la visualización de la información detalladamente. Sin embargo, para realizar operaciones con las mismas se requiere un conocimiento específico. El mismo puede carecerse si la preparación académica de los interesados no es, específicamente, en Matemáticas o en áreas que se relacionen con las mismas (i.e., usuarios no matemáticos) o si se ha estudiado pero hace mucho tiempo que no se practica. También, se utilizan programas estadísticos para realizar el análisis de datos, lo que puede propiciar que el manejo manual de las operaciones con matrices se olvide.

Conocer el Álgebra matricial básica que se aplica al analizar datos, en este caso datos multivariados, puede facilitar el entendimiento de los resultados y cómo se obtienen los mismos. Por esta razón, este artículo tiene como propósito explicar, de una manera simple y detallada, las operaciones con matrices que necesitan conocerse para entender la información teórica del análisis multivariado. Para esto, Álgebra matricial de técnicas factoriales se abordan, en primer lugar, los conceptos básicos que se relacionan con la estructura y la nomenclatura de las matrices para centrar el conocimiento previo en el texto. Luego, se presentan y explican los conceptos pertinentes a las operaciones que se aplican en el análisis multivariado y su utilidad en la reducción de la información estadística. De manera complementaria, se ejemplifica la aplicación del Álgebra matricial en el análisis multivariado, en específico, en el Análisis de Componentes Principales.

Conceptos básicos del álgebra matricial

Por definición, una matriz es un grupo de elementos organizados de forma vertical y horizontal dentro de corchetes (Munem & Tschirhart, 1988; Bronson, 1989; Hall, 1994; Hair, Anderson, Tatham & Black, 1999; Smith, Charles, Dossey & Bittinger, 2001; Pintos & Urdaneta, 2009). Es decir, organización de filas y columnas de una tabla cuyos valores se conocen como elementos.

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{1j} \\ x_{21} & x_{22} & x_{2j} \\ x_{i1} & x_{i2} & x_{ij} \end{bmatrix}$$

La matriz se identifica con una letra mayúscula, en este caso X. Para referirse a las filas de la matriz se utiliza la letra i minúscula y, para referirse a las columnas se utiliza la letra j minúscula. Los elementos que componen la matriz se representan utilizando la misma letra con que se nombró a la matriz, pero en minúscula seguido por un suscrito de una doble numeración. El dígito de la izquierda del suscrito representa el número de la fila en que se encuentra el elemento y el dígito de la derecha representa el número de la columna. Por ejemplo, el elemento x12 de X se refiere al elemento que se encuentra en la primera fila y en la segunda columna de la matriz. De igual forma, el elemento xij de X se refiere al elemento que se encuentra en la fila i-ésima y j-ésima de la matriz. Cabe señalar que las filas comienzan a enumerarse de arriba hacia abajo y las columnas de izquierda a derecha.

Definiciones

Para efectos de este artículo, se utilizarán una serie de conceptos de los cuales se ofrecerá una breve definición. Así, el lector se relacionará con los conceptos a los que se hará referencia en el desarrollo de la exposición. De acuerdo con la forma de las matrices, pueden clasificarse de diferentes maneras (Bronson, 1989; Barrios, 2006).

1. Matriz cuadrada – Es una matriz que tiene el mismo número de filas como de columnas.

2. Matriz rectangular – Es una matriz cuyo número de filas no concuerda con el número de columnas.

3. Matriz identidad - Es una matriz cuadrada que contiene sólo unos (1) en su diagonal principal y ceros (0) en todos los demás elementos, sin importar su posición.

a. Propiedades

1. La suma de los elementos de la diagonal (i.e., los unos) es lo que se conoce como la traza de dicha matriz identidad.

2. Cualquier matriz que se multiplique por una matriz identidad será igual a la matriz identidad.

Operaciones con matrices

Suma y resta de matrices

De acuerdo con Munem & Tschirhart (1988), Bronson (1989), Hall (1994), Mayer (2000), Smith, Charles, Dossey & Bittinger (2001), Barrios (2006) y Pintos & Urdaneta (2009), para poder sumar o restar matrices (i.e., sus elementos), éstas deben tener el mismo número de filas y de columnas. Estos autores indican que para ambas operaciones se suman o se restan los elementos que ocupan un mismo lugar en las matrices. Por ejemplo, si una matriz es de orden 3 x 2 y otra de 3 x 3, no se pueden sumar ni restar sus elementos. Sin embargo, las matrices no necesariamente tienen que ser cuadradas.

$$\text{Sean } A = \begin{bmatrix} 3 & 1 & 2 \\ 0 & 5 & -3 \\ 7 & 0 & 4 \end{bmatrix} \text{ y } B = \begin{bmatrix} -1 & 2 & 4 \\ 2 & 5 & 8 \\ 0 & 1 & -2 \end{bmatrix}$$

$$\boldsymbol{A+B} = \begin{bmatrix} 3 & 1 & 2 \\ 0 & 5 & -3 \\ 7 & 0 & 4 \end{bmatrix} + \begin{bmatrix} -1 & 2 & 4 \\ 2 & 5 & 8 \\ 0 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 6 \\ 2 & 10 & 5 \\ 7 & 1 & 6 \end{bmatrix}$$

$$\boldsymbol{A-B} = \begin{bmatrix} 3 & 1 & 2 \\ 0 & 5 & -3 \\ 7 & 0 & 4 \end{bmatrix} - \begin{bmatrix} -1 & 2 & 4 \\ 2 & 5 & 8 \\ 0 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 4 & -1 & -2 \\ -2 & 0 & -11 \\ 7 & -1 & 2 \end{bmatrix}$$

Producto de matrices

Para poder multiplicar dos matrices, la primera de éstas debe tener el mismo número de columnas que filas de la segunda. La matriz resultante del producto quedará con el mismo número de filas de la primera y con el mismo número de columnas de la segunda (Munem & Tschirhart, 1988; Bronson, 1989; Hall, 1994; Mayer, 2000; Smith, Charles, Dossey & Bittinger, 2001; Barrios, 2006; Pintos & Urdaneta, 2009). Es decir, si tenemos una matriz 2 x 3 y la multiplicamos por una de orden 3 x 5, la matriz resultante será de orden 2 x 5. De acuerdo con esto, puede observarse que el producto de matrices no es conmutativo. Si se quisiera multiplicar las matrices anteriores en distinto orden, no podría hacerse ya que el número de columnas de la primera matriz no es igual al número de filas de la segunda (i.e., (matriz 3 x 5) x (matriz 2 x 3)).

El producto de la matriz resultante se obtiene multiplicando cada fila de la primera matriz por cada columna de la segunda matriz utilizando el siguiente algoritmo:

$$\begin{bmatrix} s & p \\ c & w \end{bmatrix} \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{bmatrix} =$$

$$\begin{bmatrix} sa_1 + pb_1 & sa_2 + pb_2 & sa_3 + pb_3 \\ ca_1 + wb_1 & ca_2 + wb_2 & ca_3 + wb_3 \end{bmatrix}$$

© Revista Paideia Puertorriqueña
Universidad de Puerto Rico. Recinto de Río Piedras http://paideia.uprrp.edu

Si se presume que A y B son matrices tales que el número de columnas de A coincide con el número de filas de B, entonces

Sean $A = \begin{bmatrix} 2 & 1 \\ 5 & 4 \\ 3 & -2 \end{bmatrix}$ $B = \begin{bmatrix} 0 & -3 & -2 \\ -1 & 2 & 4 \end{bmatrix}$ Entonces, ***AB*** =

$$\begin{bmatrix} (2)(0)+(1)(-1) & (2)(-3)+(1)(2) & (2)(-2)+(1)(4) \\ (5)(0)+(4)(-1) & (5)(-3)+(4)(2) & (5)(-2)+(4)(4) \\ (3)(0)+(-2)(-1) & (3)(-3)+(-2)(2) & (3)(-2)+(-2)(4) \end{bmatrix} =$$

$$\begin{bmatrix} 0+-1 & -6+2 & -4+4 \\ 0+-4 & -15+8 & -10+16 \\ 0+2 & -9+-4 & -6+-8 \end{bmatrix} =$$

$$\begin{bmatrix} -1 & -4 & 0 \\ -4 & -7 & 6 \\ 2 & -13 & -14 \end{bmatrix}$$

Determinantes

Bronson (1989), Mayer (2000) y Pintos & Urdaneta (2009) definen a los determinantes como un escalar particular que se asigna a cada matriz n-cuadrada X. Al mismo se le denomina como determinante de X y se escribe de la siguiente manera: det (A) o det | A |. En el caso en que se tenga una tabla n x n de escalares entre dos líneas verticales (| |), a la que se llama determinante de orden n, no se considera como matriz.

Determinantes de orden uno y dos. El determinante de una matriz de orden uno es el mismo escalar, ya que es el único del sistema. Por el contrario, para el determinante de una matriz de orden dos se multiplican los escalares de manera cruzada y se restan. El determinante de orden uno se define como sigue:

$$\textit{det}\ |x_{11}| = x_{11}$$

El determinante de orden dos se define como sigue:

$$\textit{det} \begin{vmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{vmatrix} = (x_{11})(x_{22}) - (x_{12})(x_{21})$$

Al aplicarse el algoritmo en cada caso se observa lo siguiente:

$$\textit{det}\ |24| = 24 \qquad \textit{det}\ |-3| = -3$$

$$\textit{det}\begin{vmatrix} -7 & 6 \\ 3 & 2 \end{vmatrix} = (-7)(2) - (6)(3) = (-14) - 18 = -32$$

Determinantes de orden tres. Para calcular el determinante de una matriz cuadrada de orden tres debe utilizarse el siguiente algoritmo:

$$\textit{det}\ (X) = \begin{vmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{vmatrix}$$

$$= (x_{11})(x_{22})(x_{33}) + (x_{12})(x_{23})(x_{31}) + (x_{21})(x_{32})(x_{13}) - (x_{13})(x_{22})(x_{31}) - (x_{12})(x_{21})(x_{33}) - (x_{32})(x_{23})(x_{11})$$

Al aplicar el algoritmo a la matriz A se obtiene:

$$A = \begin{bmatrix} 3 & 1 & 2 \\ 0 & 5 & -3 \\ 7 & 0 & 4 \end{bmatrix}$$

$$(3)(5)(4) + (1)(-3)(7) + (0)(0)(2) - (2)(5)(7) - (1)(0)(4) - (0)(-3)(3) =$$

$$60 + (-21) + 0 - 70 - 0 - 0 =$$

$$39 - 70 = -31$$

Traspuesta de una matriz

La traspuesta o la trasposición de una matriz, consiste en ubicar las filas como columnas y las columnas como filas (Bronson, 1989; Mayer, 2000; Pintos & Urdaneta, 2009). Para denominar una matriz traspuesta se utiliza la letra mayúscula, con la que se nombró la matriz de partida u original, con una t mayúscula (T) o un apóstrofo (’) como exponente (X^T o X’)[1]. Estos autores afirman que si la matriz de partida no es cuadrada, su orden cambiará. Es decir, que el orden de la matriz traspuesta será el contrario de la matriz de partida (e.g., si el orden de A era 2 x 3, entonces el orden de A^t será 3 x 2). Por su parte, en el caso de una matriz cuadrada, al transponerse, los elementos de la diagonal principal de la misma no cambian su posición.

$$A = \begin{bmatrix} 3 & 1 \\ 2 & 7 \end{bmatrix} \qquad A^T = \begin{bmatrix} 3 & 2 \\ 1 & 7 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 3 \\ 4 & -9 \\ -8 & 5 \end{bmatrix} \quad X^T = \begin{bmatrix} 1 & 4 & -8 \\ 3 & -9 & 5 \end{bmatrix}$$

Vectores

Bronson (1989) y Mayer (2000) definen un vector como una matriz que sólo tiene una fila o columna. El número de elementos en un vector fila o vector columna es igual a su orden o dimensión. A los elementos de un vector se llaman componentes. Por ejemplo, si es un vector fila con tres componentes sería un vector de orden 1 x 3. Si es un vector columna de tres componentes, su orden sería 3 x 1. Por otro lado, cuando se traspone un vector fila se obtiene un vector columna y viceversa.

$$A = \begin{bmatrix} 2 & -4 & 1 & 5 \end{bmatrix} \qquad B = \begin{bmatrix} 3 \\ -1 \\ 7 \\ 6 \end{bmatrix}$$

$$A^T = \begin{bmatrix} 2 \\ -4 \\ 1 \\ 5 \end{bmatrix} \qquad B^T = \begin{bmatrix} 3 & -1 & 7 & 6 \end{bmatrix}$$

Valores propios y vectores propios

De acuerdo con Bronson (1989), Mayer (2000), López Gómez (2005) y Pintos & Urdaneta (2009), un vector columna X, de valor distinto de cero, es un vector propio de una matriz cuadrada A si existe un escalar λ (Lambda) tal que AX = λX. Entonces λ es un valor propio de A. Estos autores indican que los valores propios pueden tener valor de cero. Sin embargo, un vector propio no puede ser un vector cero. Por su parte, Vázquez & Ramírez (2000), afirman que λ es un valor propio de una matriz A cuadrada si existe un vector X, distinto de cero, que pertenezca a un cuerpo K, que satisfaga la ecuación AX = λX. En este caso, el vector X se denominará autovector de la A y el mismo estará asociado con el autovalor λ. Cabe señalar, que la ecuación matricial AX = λX no es lineal porque contiene un producto de dos valores desconocidos (i.e., λ y X) (Vázquez & Ramírez, 2000). Estos autores resaltan que una condición necesaria para que λ sea un valor propio de A es que, el valor absoluto de la matriz A restándole el escalar λ multiplicad por la matriz identidad sea igual a cero (A –λI) = 0).

De acuerdo con Bronson (1989), Mayer (2000) y Luzardo & Peña (2006) cada matriz cuadrada satisface su propia ecuación característica. Es decir, que si A es una matriz cuadrada, I es la matriz identidad de igual orden que A y p(x) = det (A –λI), entonces p(A) = 0. Como esta ecuación puede considerarse un polinomio, a la misma se le conoce como ecuación polinomial característica. Al resolver dicha ecuación característica para los λ, se obtienen como resultado los valores propios de una matriz. Una vez que se determinan los valores propios, pueden sustituirse en la ecuación característica y entonces, esa ecuación puede resolverse para los correspondientes vectores propios. Esta ecuación, es el determinante de la matriz de partida menos λ multiplicado por la matriz identidad (i.e., determinante de (A –λI) = 0).

$$\textit{Si } A = \begin{bmatrix} 3 & 5 \\ -2 & -4 \end{bmatrix}. \textit{ Entonces,}$$

$$A - \lambda I = \begin{bmatrix} 3 & 5 \\ -2 & -4 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 3-\lambda & 5 \\ -2 & (-4)-\lambda \end{bmatrix}$$

De ahí,

$$\begin{aligned} &\det(A - \lambda I) = (3-\lambda)(-4-\lambda) - (5)(-2) \\ &= -12 - 3\lambda + 4\lambda + \lambda^2 + 10 \\ &= \lambda^2 - 3\lambda + 4\lambda - 12 + 10 \\ &= \lambda^2 + \lambda - 2 \end{aligned}$$

La ecuación polinomial característica de A sería $\lambda^2 2 + \lambda - 2 = 0$. Al resolver la misma para λ (por factorización) se obtienen dos valores: λ = 1 y λ = -2

Propiedades de los valores propios y vectores propios

Bronson (1989) presenta unas propiedades para los valores propios y vectores propios, a saber:

1. La suma de los valores propios de una matriz son iguales a su traza, la cual se define como la suma de los elementos de la diagonal principal de la matriz.

2. Los vectores propios correspondientes a diferentes valores propios son linealmente independientes. Esto significa que si se multiplican por constantes de valor cero (i.e., c1 = c2 = … = cn = 0) y se suman los vectores resultantes, se obtendrá como resultado un vector cero.

3. Una matriz es singular si y sólo si tiene un valor propio igual a cero.

4. Una matriz (A) y su traspuesta (AT) tienen los mismos valores propios.

5. Los valores propios de una matriz triangular, sea ésta superior o inferior, son los elementos que están en su diagonal principal.

6. El producto de valores propios de una matriz (i.e., multiplicaciones continuas)es igual al determinante de la matriz.

Aplicaciones del Álgebra matricial en técnicas multivariadas

Cuadras (2007) define el análisis multivariado como la parte de la estadística y del análisis de datos en la que se estudian, se analizan, se representan y se interpretan los datos que se obtienen de la observación de un número de p variables estadísticas, mayor que uno (i.e., $p > 1$), sobre una muestra de n individuos. Este autor afirma, que las variables observables son homogéneas y se correlacionan, sin que alguna de ellas predomine sobre las otras. También, añade que la información estadística que se utiliza en el análisis multivariado es de carácter multidimensional. Por lo tanto, la geometría, el cálculo matricial y las distribuciones multivariadas tienen importancia en el mismo.

En la realidad, cuando se realizan estudios se toman en consideración numerosas variables. Esto hace que las matrices de datos sean de un orden mucho mayor a los que se han discutido en los apartados iniciales de este artículo. También, no necesariamente se cuenta con la misma cantidad de individuos y de variables para conformar matrices cuadradas. Sin embargo, lo que se pretende es que pueda entenderse, a menor escala, lo que se hace en el análisis multivariado a mayor escala. Por otro lado, se observará que las matrices no se presentan en la forma en que se planteó en el apartado de Conceptos Básicos. Las mismas, se presentan en forma de tablas donde, a su vez, se visualizan igualmente las filas y columnas. Este tipo de presentación permite situar al lector en el contexto de una matriz de datos en un programa estadístico (i.e., pantalla que muestra la hoja de datos) (Martín, Cabero & de Paz, 2008).

Cuando se desean analizar estadísticamente grandes matrices de datos, pueden utilizarse técnicas de análisis multivariado que se basen en la reducción de la dimensionalidad de la información. Éstas permiten la proyección de los datos originales (i.e., información de la matriz de partida) sobre un subespacio cuyo ajuste sea óptimo, de tal forma que se mantengan los patrones que hacen referencia a la variación conjunta de los individuos y las variables (Cárdenas, Galindo & Vicente, 2007). Entre las técnicas multivariadas que permiten la reducción de la información pueden mencionarse el Análisis de Componentes Principales, el Análisis Factorial, el Análisis de correspondencias y los Métodos Biplots. La selección de una técnica u otra dependerá del tipo de información que contenga la matriz de datos y del propósito de la investigación y del análisis estadístico (Cárdenas, Galindo & Vicente, 2007).

En esta parte se abordará, de manera general, el Análisis de Componentes Principales, como técnica de análisis multivariado. Esto, se hará enfatizando en parte del procedimiento de los conceptos del Álgebra matricial. Entre los aspectos del Álgebra matricial que se observarán se encuentran: la nomenclatura, los conceptos que se relacionan con las operaciones, la estructura y la elaboración de la matriz de datos, la aplicación de las operaciones y la obtención de valores y vectores propios y sus utilidades.

Cuando se recopila información acerca de un conjunto de individuos, es recomendable que se tome el mayor número posible de variables. No obstante, al seleccionarse muchas variables, hay que detenerse a pensar en que la cantidad de coeficientes de correlación será alta y, por consiguiente, difícil de visualizar. Esto puede suceder cuando no se conocen demasiado los datos o sólo se están explorando los mismos. Cabe señalar, que el hecho que el concepto de mayor información se relaciona con el de mayor variabilidad o varianza. Esto quiere decir, que a medida que la variabilidad de los datos sea mayor, puede considerarse la existencia de mayor información. A su vez, esto amerita la reducción de la información cuantitativa.

Utilización del Álgebra matricial en el Análisis de Componentes Principales

De acuerdo con Cuadras (2007) y Martín, Cabero & de Paz (2008), el Análisis de Componentes Principales (ACP) permite transformar un conjunto de variables intercorrelacionadas en un nuevo conjunto de variables no correlacionadas (i.e., componentes o variables latentes). En específico, Martín, Cabero & de Paz (2008) indican que lo que se hace en este tipo de análisis es aglomerar la matriz de correlaciones en unos componentes principales (i.e., variables latentes) de la variabilidad o varianza [2] total. De esta forma, el primer componente principal que se extrae resume a información contenida en la matriz de partida. Igualmente, el segundo componente que se extrae explica la mayor cantidad de varianza que el primer componente no explicó. Así, los componentes principales subsiguientes siguen contribuyendo, en menor proporción, a la varianza total.

En otras palabras, el ACP es una técnica estadística multivariante que se utiliza con el propósito de hacer una síntesis de la información o reducción de la dimensión (i.e., número de variables) de una matriz de datos (Peña, 2002). Este autor añade que, al reducirse la cantidad de variables en una cantidad menor, debe perderse la menor información posible. Para reducir la cantidad de información, se necesita que las variables de la matriz de partida se relacionen entre sí. El conjunto de aquellas variables que se relacionen pasarán a formar una nueva variable a la que se denomina como componente (Figura 1). Dependiendo de la cantidad de variables de la matriz de partida y de la relación que exista entre ellas, se obtiene una mayor o menor cantidad de componentes.

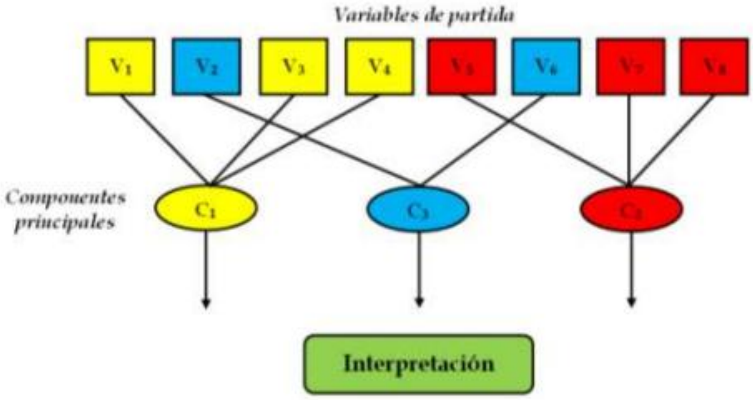

Figura 1. Modelo del Análisis de Componentes Principales

Tabachnick & Fidell (2001), Martín, Cabero & de Paz (2008) y Cuadras (2007) exponen que, por el P rincipio de Parsimonia3, el número de componentes debe reducirse lo más posible y ser apropiados para una interpretación sustantiva. De esta forma, estos autores afirman que una buena solución del ACP será aquella que sea sencilla e interpretable.

La interpretación del los componentes no viene dada en un principio. A partir de las variables que conformen cada componente, se procede a explicarlas. Cabe resaltar que, como los componentes se forman de un conjunto de variables que se relacionan entre sí, hay que encontrar una explicación que resuma lo que se mide con cada una de las variables y que se visualice en el contexto del problema o tema bajo estudio (Tabachnick & Fidell, 2001; Martín, Cabero & de Paz, 2008).

Representación de una matriz de datos

Para representar una matriz de datos multivariados n x p, en un espacio de dimensión reducida, por ejemplo de dimensión dos, se necesita introducir una distancia que permita la proyección de los elementos de la matriz de partida. Para esto puede utilizarse la distancia euclídea [4.] . Si X = [x, …, xp], entonces la distancia euclídea entre los elementos de X sería:

$$x_i = (x_{i1}, \ldots, x_{ip}), \ \ x_j = (x_{j1}, \ldots, x_{jp}) \ \text{donde,}$$

$$\delta_{ij}^2 = (x_i - x_j)^T (x_i - x_j) = \sum_{h=1}^{p} (x_{ih} - x_{jh})^2$$

La matriz delta, Δ = (δij), es la matriz n x n de las distancias entre las filas y las columnas de la matriz X. Con ésta se obtendrá una nueva matriz de distancias que permitirá la representación de los elementos de la matriz de partida como puntos en un espacio R^p (Figura 2). Los mismos estarán distanciados de acuerdo con la métrica de δij.

Universidad de Puerto Rico. Recinto de Río Piedras http://paideia.uprrp.edu

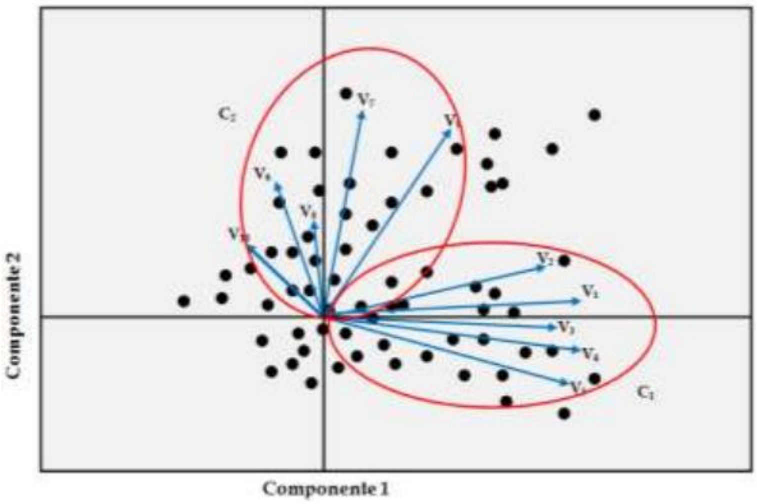

Figura 2. Representación de una matriz en el espacio R^p

A pesar de que tanto el ACP como el Análisis Factorial son técnicas que se utilizan para la reducción de la información, el interés del ACP está en la información acerca de los individuos, a diferencia del Análisis Factorial en el que el interés está en el estudio de la estructura de las correlaciones entre las variables. En específico, en el ACP se intenta describir los valores de los individuos a través de un número reducido de las variables, que sean una combinación de las variables de la matriz de partida.

Cálculo de los componentes principales

En este apartado se presenta, de manera resumida, el procedimiento lógico para calcular los componentes principales. En el mismo se enfatizará en cuatro aspectos principales: 1) Análisis de la matriz de correlaciones, 2) Obtención de los componentes principales, 3) Análisis de la matriz factorial e 4) Interpretación de los componentes. Cabe mencionar, que el cálculo puede hacerse utilizando un programa estadístico, en este caso se utilizó SPSS. Este programa considera el ACP como una solución de un problema de Análisis Factorial, porque el mismo es una técnica descriptiva para la reducción de la información.

Análisis de la matriz de correlaciones. Esta matriz se obtiene calculando la correlación entre las variables bajo estudio. Esto quiere decir que, en la matriz de correlaciones se observará tanto en las filas como en las columnas de la matriz las variables bajo estudio, formando así una matriz cuadrada. Los elementos de la matriz corresponderán a la relación lineal que existe entre cada una de las variables (Tabla 1). Este paso permite conocer si el ACP tiene sentido. Hair, Anderson, Tatham Black (1999), Tabachnick & Fidell (2001) y Cuadras (2007) indican que si los elementos de la matriz (i.e., las correlaciones entre cada una de las variables) tienen valores muy pequeños ($r_{ij} < 0.30$), hay poca relación entre las variables y, por consiguiente no podrían establecerse los componentes principales. Por el contrario, si los valores de los elementos son muy altos ($r_{ij} > 0.90$) se considera

© Revista Paideia Puertorriqueña
Universidad de Puerto Rico. Recinto de Río Piedras http://paideia.uprrp.edu

que hay información redundante, lo que permitiría que pocos factores (i.e., componentes) expliquen gran parte de la variabilidad total. Los programas estadísticos facilitan el cálculo de esta matriz de correlaciones.

Tabla 1.

Matriz de correlaciones en el Análisis de Componentes Principales

Tabla 1.

Matriz de correlaciones en el Análisis de Componentes Principales

	Coeficientes de correlación entre las variables									
	V_1	**V_2**	**V_3**	**V_4**	**V_5**	**V_6**	**V_7**	**V_8**	**V_9**	**V_{10}**
V_1 - Seguridad durante el horario escolar	1.000	.984	.966	.972	.977	-.073	.092	.169	-.039	-.216
V_2 - Seguridad cuando los estudiantes llegan y se van de la escuela		1.000	.966	.969	.975	-.036	.108	.231	.008	-.166
V_3 - Seguridad en algunas actividades de la escuela			1.000	.969	.959	-.036	.066	.153	-.068	-.240
V_4 - Mantenimiento de la disciplina escolar				1.000	.976	-.050	.083	.139	-.022	-.191
V_5 - Coordinación con la policía local					1.000	-.043	.136	.146	-.007	-.179
V_6 – Guías para docentes acerca del manejo de la sala de clases						1.000	.363	.349	.535	.273
V_7 - Guías para docentes acerca de las políticas disciplinarias							1.000	.042	.633	.393
V_8 - Guías para docentes acerca de procedimientos de seguridad								1.000	.309	-.054
V_9 - Guías para docentes acerca de la identificación temprana de problemas de comportamiento									1.000	.674
V_{10} - Guías para docentes acerca del abuso de drogas y alcohol										1.000

Desde el punto de vista algebraico, la descomposición en vectores y valores propios se define por

$$S_{(p \times p)} = X^T X$$

$$S_{(p \times p)} = X^T X = X^T_{(n \times p)} X_{(p \times n)} = V_{(p \times r)} \Lambda_{(r \times r)} V^T_{(r \times p)}$$

La matriz V es la matriz de los vectores propios. Éstos se representan en las columnas de V cuyos valores son iguales a los cosenos de los ángulos que se forman entre los componentes principales y las variables de la matriz de partida X. Por su parte, la matriz diagonal Λ contiene los valores propios. Éstos representan las varianzas de los componentes. Los mismos se ordenan de acuerdo con su magnitud (i.e., orden decreciente) y su suma corresponde a la traza de la matriz S, que a su vez, es igual a la variabilidad total de los datos. Al graficarse los coeficientes de la matriz de correlación Λ, la varianza de la nube de puntos que explica cada componente coincide con el valor propio que se asocia a la misma. De esta forma, se conoce la cantidad de varianza que cada componente recoge. Además, indica la bondad de ajuste de la representación en el espacio de baja dimensión.

Junto al análisis de la matriz de correlaciones se realizan otras pruebas para determinar el grado de asociación entre las variables y así, tener más evidencia estadística al respecto (Hair, Anderson, Tatham & Black, 1999; Martín, Cabero & de Paz, 2008). Entre estas pruebas se encuentran, la Prueba de contraste de esfericidad de Bartlett y la Medida de adecuación de la muestra (KMO, por sus siglas en Inglés de Kaiser-Meyer-Oklin) (Tabla 2).

La prueba de Contraste de esfericidad de Bartlett contrasta que la matriz de correlaciones sea una matriz identidad. La hipótesis nula en la prueba de Bartlett indica que no existe asociación entre las variables bajo estudio. Esto significa que no existe relación entre las variables, por lo que el ACP es inadecuado. Según Hair, Anderson, Tatham & Black (1999) y Martín, Cabero & de Paz (2008), con valores altos de χ 2en la prueba de Contraste de esfericidad, se rechaza la hipótesis nula a un nivel de significación de 0.0001. Esto significa que las variables están correlacionadas, por lo que puede aplicarse el ACP.

Por otro lado, KMO es un índice para calcular la magnitud de los coeficientes de correlación observados con la magnitud de los coeficientes de correlación parcial. En esta prueba, se contrasta que las correlaciones parciales son pequeñas. Hair, Anderson, Tatham & Black (1999) y Martín, Cabero & de Paz (2008) indican que, si la suma de los coeficientes de correlación parcial al cuadrado entre todos los posibles pares de variables es pequeña, al compararse con los coeficientes de correlación al cuadrado, el índice de KMO se aproximará a uno, en una escala de cero a uno. Estos autores recomiendan que con valores bajos del KMO no se aplique el ACP. Los valores en la prueba del KMO deben ser mayores que 0.75. Esto significaría que las correlaciones entre los pares de variables podrían explicarse a partir de otras variables.

Tabla 2.

Medidas de asociación entre las variables: Medida de adecuación de la muestra y Prueba de Bartlett

Prueba		**Resultado**
Medida de adecuación de la muestra (KMO)		.80
Prueba de esfericidad de Bartlett	Valor del Chi-cuadrado χ^2	698.0
	Significancia	.000

Obtención de los componentes principales. La selección de los componentes principales se realiza de forma tal que, en el primer componente se recoja la mayor cantidad posible de la variabilidad total, siendo la que mejor contribuye para explicar la varianza total. El segundo componente debe recoger la máxima cantidad de la variabilidad posible que no se recogió en el primer componente. De igual manera, los componentes principales subsiguientes siguen contribuyendo, en menor proporción, a la varianza total. Del total de componentes, se elegirán aquéllas que recojan el porcentaje de variabilidad que se considere suficiente. A éstas se les denominará componentes principales. Las coordenadas de los individuos sobre los componentes, se obtienen al proyectar la matriz X sobre el subespacio de dimensión reducida. Si el mismo es de dimensión dos, entonces, se retienen los dos primeros vectores propios (i.e., componentes), y la matriz que contiene las coordenadas de individuos será: $A(n \times 2) = X(n \times p)\, V(p \times 2)$.

Para esto, se observa la matriz de varianza total explicada por los componentes (Tabla 3). En la misma se incluyen los valores propios de cada uno de los componentes posibles que surgen del ACP. Para determinar los componentes que se retendrán, algunos autores (Hair, Anderson, Tatham & Black, 1999; Tabachnick & Fidell, 2001; Cuadras, 2007) recomiendan el uso de la conocida Regla de Kaiser. En la misma se observa el valor propio de cada componente. Las que tengan un valor propio mayor que uno ($\lambda > 1.0$) serán las que se retengan. También, puede utilizarse el criterio de la cantidad de varianza explicada por los componentes. Hair, Anderson, Tatham & Black (1999) recomiendan que en un estudio en el área de las Ciencias Naturales se retengan los componentes necesarias hasta que se obtenga un 95% de varianza explicada y en las Ciencias Sociales un 60%. Por su parte, Cuadras (2007) recomienda un 80% de varianza explicada, mientras que Tabachnick & Fidell (2001) recomienza un 70%. En la Tabla 3 puede observarse que con las primeras dos componentes se explica un 75% de la varianza total. Por ser los datos de este ejemplo un caso de Ciencias Sociales, permite que el investigador retenga sólo las dos primeras componentes. Además, el tercer componente en adelante se explica menos cantidad de varianza, en comparación con los dos primeros componentes.

Tabla 3.

Matriz de la varianza total explicada por los componentes

Componentes	Autovalores	Porcentaje de varianza	Porcentaje de varianza acumulada
1	4.984	49.835	49.835
2	2.536	25.361	75.196
3	1.118	11.182	86.379
4	.598	5.978	92.357
5	.497	4.967	97.323

Por otro lado, podría utilizarse el criterio del codo o bastón roto (Hair, Anderson, Tatham & Black, 1999; Tabachnick & Fidell, 2001; Cuadras, 2007). Este criterio consiste en la representación gráfica de los valores propios, gráfico de sedimentación (conocido en Inglés como Scree Plot). En el mismo, puede trazarse una línea recta por lo puntos que representan a los valores propios bajos hasta el eje de y, el eje que indica el autovalor de los valores propios. De esta forma, se retienen los componentes que se corresponden a los autovalores que tengan un valor mayor al que indicó el punto de corte de la línea trazada en el eje de y (Figura 3). En el caso de que el investigador conozca de antemano la cantidad de componentes que se obtienen o que desee retener una cantidad específica de componentes, puede hacerlo fijándolo en el programa estadístico.

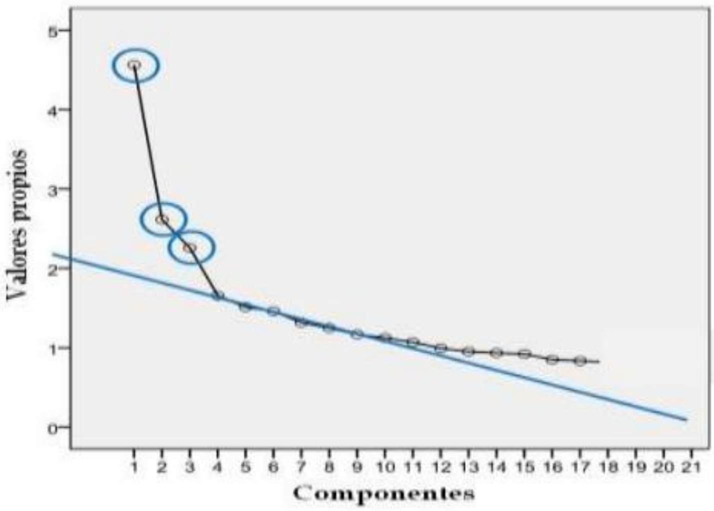

Figura 3. Ejemplo del procedimiento a seguir al utilizarse el criterio del Scree Plot

Análisis de la matriz factorial. Luego de que se seleccionen los componentes principales, se representan en forma de matriz. Los elementos de esta matriz representan a cada coeficiente factorial de las variables (i.e., los coeficientes de correlación entre las variables y los componentes principales). Los mismos pueden nombrarse como pesos o saturaciones. La cantidad de columnas de la matriz factorial dependerá de la cantidad de componentes principales y la cantidad de filas dependerá de la cantidad de individuos, que en este caso serán las variables de partida.

Para determinar las variables que forman parte de cada componente, se observan los elementos de la matriz factorial (Tabla 4). Cada variable debe tener un peso alto con sólo uno de los componentes y bajo con los demás. La variable de partida se ubicará en el componente donde tenga el valor absoluto mayor de cada peso (Hair, Anderson, Tatham & Black, 1999; Tabachnick & Fidell, 2001; Cuadras, 2007). A partir de las variables que conforman cada uno de los componentes, se procede a explicarlos. En el ejemplo que se presenta abajo (Tabla 4) puede observarse que el primer componente lo conforman las variables V1, V2, V3, V4 y V5. El segundo componente la conforma las variables V6, V7, V8, V9 y V10. Cabe resaltar, que en la matriz factorial no siempre se ubican en los componentes las variables juntas; pueden quedar entremezcladas. Esto depende del orden en que se entren en el programa estadístico que se utilice y las correlaciones entre ellas.

Tabla 4.

Matriz factorial del Análisis de Componentes Principales

Variables	Componentes 1	Componentes 2
V_1 - Seguridad durante el horario escolar	.991*	.009
V_2 - Seguridad cuando los estudiantes llegan y se van de la escuela	.989*	.064
V_3 - Seguridad en algunas actividades de la escuela	.983*	-.009
V_4 - Mantenimiento de la disciplina escolar	.985*	.021
V_5 - Coordinación con la policía local	.985*	.049
V_6 - Guías para docentes acerca del manejo de la sala de clases	-.058	.702*
V_7 - Guías para docentes acerca de las políticas disciplinarias	.084	.747*
V_8 - Guías para docentes acerca de procedimientos de seguridad	.203	.359*
V_9 - Guías para docentes acerca de la identificación temprana de problemas de comportamiento	-.053	.930*
V_{10} - Guías para docentes acerca del abuso de drogas y alcohol	-.253	.696*

Nota. * Valor absoluto mayor.

Interpretación de los componentes. Como los componentes se forman de un conjunto de variables que se relacionan entre sí, hay que encontrar una explicación que resuma lo que se mide con cada una de las variables y que a su vez, se visualiza en el contexto del problema o tema bajo estudio. Para que un componente principal sea fácilmente interpretable, debe tener las siguientes características:

- Los pesos o saturaciones deben ser próximos a uno (1).
- Cada variable debe tener pesos elevados en sólo un componente.
- No deben existir variables con pesos similares en distintos componentes.

A partir de las variables que conforman cada uno de los componentes, se procede a explicar los mismos. Cabe señalar, que como los componentes se forman a partir de un conjunto de variables, hay que encontrar una explicación que resuma lo que se mide con cada una de las variables. A su vez, ésta debe hacerse en el contexto del problema o tema bajo estudio (Hair, Anderson, Tatham & Black, 1999; Tabachnick & Fidell, 2001; Cuadras, 2007; Martín, Cabero & de Paz, 2008).

Interpretación final de los resultados del análisis

En la Tabla 4 pudo observarse que las variables que conformaban el primer componente se relacionaban con aspectos de seguridad en la escuela. Por su parte, las variables del segundo componente se relacionaban con guías que se ofrecen al personal docente de la escuela. De acuerdo con esa información, se procede a asignar un nombre a cada componente que resuma la información que aportan las variables que las conforman. Entonces, los nombres para los componentes podrían ser: 1) Servicios para promover la seguridad escolar y 2) Guías para el mejoramiento profesional del personal docente. Así, cuando se procede a interpretar los resultados del análisis, se hace enfatizando en cómo los individuos se proyectan en torno a los componentes principales (Figuras 4 y 5).

Como se mencionó anteriormente, el énfasis del ACP es describir los valores que representan los individuos a través de un número reducido de variables latentes que a su vez, sean producto de la combinación de las variables de partida. Si se proyectan los individuos con respecto a cada componente, puede conocerse la posición que cada individuo toma con respecto a la nueva información que se obtuvo del ACP. Esto permite visualizar cómo cada individuo responde a los componentes de acuerdo con el valor que se obtiene de la combinación de las variables que conforman cada componente (Figuras 4 y 5).

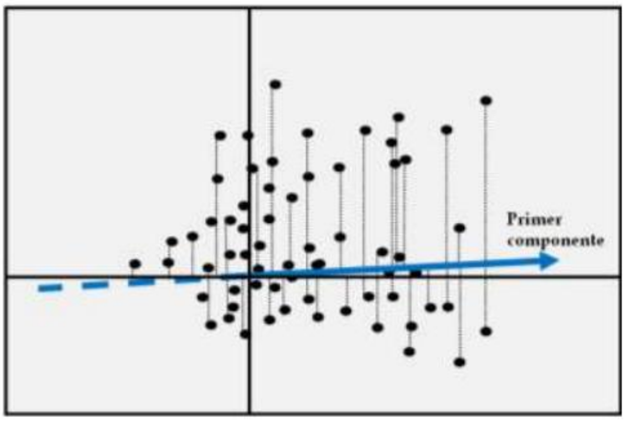

Figura 4. Proyección de los individuos con respecto al primer componente

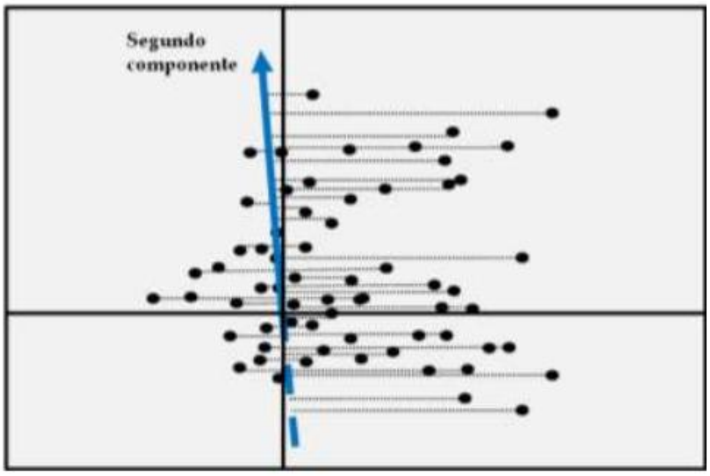

Figura 5. Proyección de los individuos con respecto al segundo componente

En este caso, el ejemplo que se utilizó contenía datos acerca de las prácticas de seguridad que se utilizaban en las escuelas públicas de los Estados Unidos. En los mismos se indicaba una puntuación con respecto a la utilización de cada una de las prácticas (i.e., variables de partida). Si se suman las puntuaciones de las variables que conforman cada uno de los componentes se obtienen dos nuevas puntuaciones. A su vez, las mismas permitirán que se determinen las nuevas puntuaciones de los individuos (i.e., las escuelas) para cada uno de los componentes. A partir de los mismos se presentan los resultados. En este caso sería a partir de los componentes Servicios para promover la seguridad escolar y Guías para el mejoramiento profesional del personal docente.

Este tipo de análisis estadístico también puede utilizarse como una preparación previa de los datos para realizar nuevos análisis con los componentes que se obtengan. La información se reduce y facilita su uso en un nuevo contexto estadístico. De esta forma, se cumple con el Principio de Parsimonia que indican Tabachnick & Fidell (2001), Martín, Cabero & de Paz (2008) y Cuadras (2007), lo que contribuye a que se trabaje con una información cuantitativa reducida pero lo más completa posible para seguir indagando acerca del problema o tema de investigación.

Referencias

Barrios, J. A. (2006). Algebra matricial para economía y empresa. Madrid, España: Delta Publicaciones.

Bronson, R. (1989). Matrix operations. New York, NY: McGraw-Hill.

Cárdenas, O., Galindo, P. & Vicente, J. L. (2007). Los métodos biplots: Evolución y aplicaciones. Revista Venezolana de Análisis de Coyuntura, 13(001), 279-303.

Cuadras, C. M. (2007). Nuevos métodos de análisis multivariado. Barcelona, España: CMC Editions.

Hair, J. F., Anderson, R. E., Tatham, R. L. & Black, W. C. (1999). Análisis multivariado (5a. ed.). Madrid, España: Prentice Hall Iberia.

Hall, J. W. (1994). Algebra for college students (2a. ed.). Boston, MA: PWS Publishing Company.

López Gómez, M. (2005). Reducción por semejanza de una matriz: Diagonalización.

Recuperado de http://www.scribd.com/doc/3900131/reduccion-pordiagonalizacion

Luzardo, D. & Peña, A. J. (2006). Historia del Álgebra Lineal hasta los albores del siglo XX. Divulgaciones matemáticas, 14(2), 153-170.

Martín, Q., Cabero, M. T. & de Paz, Y. (2008). Tratamiento estadístico de datos con SPSS: Prácticas resueltas y comentadas. Madrid, España: Thomson Editores Spain.

Mayer, C.D. (2000). Matrix analysis & applied linear algebra. Philadelphia,PA: SIAM.

Munem, M. A. & Tschirhart, W. (1988). Algebra for college students. New York, NY:

Worth Publishers, Inc.

Peña, D. (2002). Análisis de datos multivariados. Madrid, España: Mc Graw-Hill.

Pintos, S. & Urdaneta, G. (2009). Breves de álgebra lineal. Recuperado de

http://www.scribd.com/doc/17225553/Algebra-Lineal-Breve

Smith, S. A., Charles, R. I., Dossey, J. A. & Bittinger, M. L. (2001). Algebra with trigonometry. Upper Saddle River, NJ: Prentice Hall.

Tabachnick, B. G. & Fidell, L. S. (2001). Using multivariate statistics. Boston, MA: Allyn & Bacon.

Vázquez, M. & Ramírez, G. (2000). Aspectos teóricos del álgebra matricial con aplicaciones estadísticas. Caracas, Venezuela: Unidad de Publicaciones, Universidad Central de Venezuela. Revista Paideia Puertorriqueña Vol. 5 Núm. 2

Notas

1 En este artículo se utilizó la T.

2 Medida de dispersión que presenta la posición de los datos de una variable respecto a su media.

3 Principio que afirma que los fenómenos deben explicarse con el menor número posible de pasos.

4 Medida de similitud que indica la distancia de la línea recta entre dos objetos.

5 En el ACP se excluye toda relación que no sea lineal.