Análisis de Datos Multivariados: Análisis General

Jessica Fernández-Garza* y José Antonio Díaz-García

Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro. Buenavista. 25315. Saltillo, Coah., México. Tel.: (844) 4 11 03 33, 4 11 03 34 Fax:(844) 4 11 02 28. E-mail: jadiaz@uaaan.mx (*Autor responsable).

Abstract

The objective of this assay was to study the ideas that serve as a basis for AG, and also serve to obtain both, the ACP and the ACS, like particular cases of it. Here are presented the algebraic ideas on which the General Analysis (AG) is based, since this conforms the theoretical base of some multivariate methods. It is mentioned under which circumstances the Principal Components Analysis (ACP) and the Simple Correspondence Analysis (ACS) are obtained as particular cases of the AG. An example is given for the case of the ACS.

Key words: Multivariate Analysis, Principal Components Analysis, Normalized Principal Components Analysis, Simple Correspondence Analysis, Singular value decomposition.

Resumen

El objetivo del presente ensayo fue estudiar las ideas en las que se fundamenta el AG y obtener el ACP y el ACS como casos particulares del mismo. Se presentan las ideas algebraicas en las que se fundamenta el Análisis General (AG), el cual conforma la base teórica de algunos métodos multivariados. Se menciona bajo que circunstancias el Análisis de Componentes Principales (ACP) y el Análisis de Correspondencias Simple (ACS) se obtienen como casos particulares del AG. Se ejemplifica para el caso del ACS.

Palabras clave: Análisis Multivariado, Análisis de Componentes Principales, Análisis de Componentes Principales Normado, Análisis de Correspondencias Simple, Descomposición en valores singulares.

Introducción

Una de las líneas de investigación de mayor desarrollo del Análisis Multivariado, surgida en las últimas décadas, es el Análisis de Datos Multivariados. Sus métodos contribuyen a describir grandes y complicados conjuntos de datos multivariados, permitiendo obtener representaciones simplificadas que facilitan la interpretación.

En la actualidad existe una gran cantidad de métodos multivariados, cada uno de los cuales es estudiado por separado y en la mayoría de los casos sin una conexión entre ellos. Varios de estos métodos multivariados comparten un fundamento matemático en común que permite estudiarlos partiendo desde un mismo punto y no de forma aislada, como se ha hecho hasta ahora. Esta base teórica en común es conocida como el Análisis General y utiliza criterios de álgebra matricial que se basan en las ideas establecidas por Eckart y Young (1936), la descomposición en valores singulares de una matriz (SVD) y la aproximación por mínimos cuadrados de una matriz por otra de menor rango. Una detallada demostración sobre el teorema establecido por Eckart y Young puede ser revisada en Johnson (1963).

En particular el Análisis de Componentes Principales y el Análisis de Correspondencias Simple pueden ser obtenidos a partir del AG. Una aplicación del AG para obtener el ACP, enfatizando el aspecto geométrico es presentada en Lebart (1984). Una excelente presentación del ACS, utilizando la SVD, es dada en Greenacre (1994).

El objetivo del presente ensayo fue estudiar las ideas en las que se fundamenta el AG y obtener el ACP y el ACS como casos particulares del mismo. Además se presenta como ejemplo la obtención del ACS, utilizando la clasificación de los 300 municipios más representativos de México, de acuerdo a su región geográfica y a su porcentaje destinado a obra pública, este último obtenido de las Finanzas públicas estatales y municipales de México (2001), INEGI.

Descomposición en valores singulares

El teorema de Eckart y Young establece que para cualquier matriz $\mathbf{A} \in R^{m \times n}$ se pueden encontrar dos matrices ortogonales, $\mathbf{U} \in R^{m \times m}$ y $\mathbf{V} \in R^{n \times n}$, tal que

$$\mathbf{U}^{\mathrm{t}}\mathbf{A}\mathbf{V} = \begin{bmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{bmatrix} = \mathbf{\Delta}, \tag{1}$$

donde $\mathbf{D} = \mathrm{diag}(\alpha_1, \alpha_2, ..., \alpha_k)$, tal que $\alpha_1 \geq \alpha_2 \geq ... \geq .\alpha_k > 0$.

Considerando la expresión (1), la matriz puede ser factorizada como el producto de tres matrices, esto es: $\mathbf{A} = \mathbf{U}\Delta\mathbf{V}^t$.(2)

Esta factorización es conocida en la literatura como la descomposición en valores singulares, y se denota como SVD(**A**).

Teorema 1. *Sea* $\mathbf{A} \in \mathrm{R}^{\mathrm{m\,x\,n}}$ *de rango* $k \leq \min(m, n)$*, entonces*

$$\mathrm{SVD}(\mathbf{A}) = \sum_{i=1}^{k} \alpha_i u_i v_i^{\,t} = \mathbf{U}_1 \mathbf{D} \mathbf{V}_1^{\,t}, \tag{3}$$

donde $\mathbf{U}_1 \in \mathrm{R}^{\mathrm{m\,x\,k}}$ *y* $\mathbf{V}_1 \in \mathrm{R}^{\mathrm{n\,x\,k}}$*, tales que* $\mathbf{U}_1\mathbf{U}_1^{\,t} = \mathbf{I}_k = \mathbf{V}_1\mathbf{V}_1^{\,t}$*, esto es, las columnas de* $\mathbf{U}_1$ *y* $\mathbf{V}_1$ *son ortonormales. Esta factorización es conocida como la parte no singular de la SVD.*

Demostración. Sean u_i los eigenvectores ortonormales correspondientes a los eigenvalores no nulos $\alpha_i^{\,2}$, para $i = 1,2,...,k$, de $\mathbf{A}\mathbf{A}^{\mathrm{t}}$, recuerde que $r(\mathbf{A}) = r(\mathbf{A}\mathbf{A}^{\mathrm{t}})$. Además, sea $v_i = \alpha_i^{-1}\mathbf{A}^t u_i$. Entonces los v_i son ortonormales, pues:

$$v_i^{\,t} v_i = \left(\alpha_i^{-1}\mathbf{A}^t u_i\right)^t \alpha_i^{-1}\mathbf{A}^t u_i = \alpha_i^{-2} u_i^{\,t}\mathbf{A}\mathbf{A}^t u_i = \alpha_i^{-2}\alpha_i^{\,2} = 1 \tag{4}$$

y

$$v_i^{\,t} v_j = \left(\alpha_i^{-1}\mathbf{A}^t u_i\right)^t \alpha_j^{-1}\mathbf{A}^t u_j = \alpha_i^{-1}\alpha_j^{-1} u_i^{\,t}\mathbf{A}\mathbf{A}^t u_j = \alpha_i^{-1}\alpha_j^{-1}\alpha_j^{\,2} u_i^{\,t} u_j = 0 \tag{5}$$

Además, defina $\mathbf{U}_2 = u_{k+1},...,u_m$ tal que $\mathbf{U} = u_1,...,u_m$ es un conjunto completo de vectores ortonormales, esto es

$\mathbf{U}\mathbf{U}^t = u_1 u_1^{\,t} + ... + u_m u_m^{\,t} = \mathbf{I}_m$, entonces

$$\mathbf{A} = (u_1 u_1^{\,t} + ... + u_m u_m^{\,t})\mathbf{A} \tag{6}$$

$= (u_1 u_1^{\,t} + ... + u_k u_k^{\,t})\mathbf{A}$ pues $u_i^{\,t}\mathbf{A} = 0$ para i > k,

$= u_1 u_1^{\,t}\mathbf{A} + ... + u_k u_k^{\,t}\mathbf{A} = u_1(\mathbf{A}^t u_1)^t + ... + u_k(\mathbf{A}^t u_k)^t$

$= u_1(\alpha_1 v_1)^t + ... + u_k(\alpha_k v_k)^t = \sum_{i=1}^{k} \alpha_i u_i v_i^{\,t} = \mathbf{U}_1\mathbf{D}\mathbf{V}_1^{\,t}$

tomando $\mathbf{V}_1 = (v_1, v_2, ..., v_k)$, $\mathbf{U}_1 = (u_1, u_2, ..., u_k)$ y $\mathbf{D} = \mathrm{diag}(\alpha_1, \alpha_2, ..., \alpha_k)$.

Observación 1

i) *La expresión (2), se obtiene considerando los siguientes conjuntos de vectores ortonormales* $\mathbf{U}_2 = (u_{k+1}, ..., u_m)$ *y* $\mathbf{V}_2 = (v_{k+1}, ..., v_n)$ *tales que las matrices* $\mathbf{V} = (v_1, v_2, ..., v_n)$ *y* $\mathbf{U} = (u_1, u_2, ..., u_m)$*, son ortogonales.*

ii) *A los vectores* u_i *y* v_i *se les conoce como vectores singulares por la izquierda y por la derecha, respectivamente. Los elementos de la diagonal principal de la matriz* $\mathbf{D}$ *son conocidos como valores singulares de* $\mathbf{A}$.

iii) *Otro resultado, que se puede comprobar fácilmente, es que los vectores singulares por la izquierda son los eigenvectores de* $\mathbf{A}\mathbf{A}^{\mathrm{t}}$*, mientras que los*

eigenvectores de $\mathbf{A}^t\mathbf{A}$ *son los vectores singulares por la derecha de* **A**, *los valores singulares no nulos de* **A** *son las raíces cuadradas positivas de los eigenvalores no nulos de ambas matrices,* $\mathbf{AA}^t$ y $\mathbf{A}^t\mathbf{A}$.

iv) *Note que las coordenadas de los renglones de* **A** *son los renglones de* $\mathbf{H} = \mathbf{U}_1\mathbf{D}$, *con respecto a la base* $\mathbf{V}_1$ *y las coordenadas de las columnas son las filas de la matriz* $\mathbf{C} = \mathbf{V}_1\mathbf{D}$, *con respecto a la base* $\mathbf{U}_1$.

Teorema 2. *Sea como en el teorema 1 y sean* $\Omega \in R^{m\ x\ m}$ *y* $\phi \in R^{n\ x\ n}$ *matrices definidas positivas, entonces* **A** *se puede factorizar como*

$$\mathbf{A} = \mathbf{NDM}^t = \sum_{i=1}^{k} \alpha_i n_i m_i^t, \qquad (7)$$

donde $\mathbf{N} \in R^{m\ x\ k}$ *y* $\mathbf{M} \in R^{n\ x\ k}$, *sus columnas están ortornormalizadas con respecto a* Ω *y* ϕ, *respectivamente, esto es,* $\mathbf{N}^t\Omega\mathbf{N} = \mathbf{M}^t\phi\mathbf{M} = \mathbf{I}_k$.

Demostración. Esta es inmediata a partir del teorema 1, sólo defina $\mathbf{N} = \Omega^{-1/2}\mathbf{U}_1$ y $\mathbf{M} = \phi^{-1/2}\mathbf{V}_1$ y considere la SVD de la matriz $\Omega^{1/2}\mathbf{A}\phi^{1/2}$. •

En este caso, los elementos de la diagonal principal de **D** son conocidos como los valores singulares generalizados de **A**, a las columnas de la matriz **N** se les llaman vectores singulares generalizados por la izquierda y son las bases ortonormales para las columnas de **A**, de forma análoga, las columnas de **M** se conocen como vectores singulares generalizados por la derecha y conforman una base ortonormal para las filas de **A**.

Aproximación de una matriz por otra de menor rango

Si los últimos valores singulares de la matriz **D**, (*k*-*r*), son muy pequeños, de tal forma que puedan ser omitidos, entonces se obtiene la aproximación a la matriz **A** con otra matriz, $\mathbf{A}_{[r]} = \mathbf{U}_r\mathbf{D}_r\mathbf{V}_r^t$, de menor rango, ($r < k$), dicha aproximación es de mínimos cuadrados, ya que se minimizan las distancias al cuadrado entre los elementos de las dos matrices, como lo muestra el siguiente teorema:

Teorema 3. *Sea* $\mathbf{A}_{[r]}$ *de orden m x n de rango r, tal que* $\mathbf{A}_{[r]} = \mathbf{U}_r\mathbf{D}_r\mathbf{V}_r^t = \sum_{i=1}^{r} \alpha_i u_i v_i^t$. *Entonces* $\mathbf{A}_{[r]}$ *es la matriz de aproximación de mínimos cuadrados que minimiza la siguiente expresión.*

$$\sum_{i=1}^{m}\sum_{j=1}^{n} (a_{ij} - x_{ij})^2 = \mathrm{tr}(\mathbf{A} - \mathbf{X})(\mathbf{A} - \mathbf{X})^t = \| \mathbf{A} - \mathbf{X} \|^2 \qquad (8)$$

Demostración

$$\| \mathbf{A} - \mathbf{X} \|^2 = \mathrm{tr}[\mathbf{UU}^t(\mathbf{A} - \mathbf{X})\mathbf{VV}^t(\mathbf{A} - \mathbf{X})^t] = \mathrm{tr}(\Delta - \mathbf{G})(\Delta - \mathbf{G})^t \qquad (9)$$

$$= \sum_{i=1}^{k} (\alpha_i - g_{ii})^2 + \sum_{1 \le j \le n, i \ne j} g_{ij}^2$$

Donde $\mathbf{G} = \mathbf{U}^t\mathbf{XV}$. La sumatoria anterior será mínima si la matriz **G** es diagonal. Note que $r(\mathbf{G}) = r(\mathbf{U}^t\mathbf{XV}) = r(\mathbf{X}) = r$, entonces (*k*-*r*) elementos de la diagonal principal de la matriz son ceros. Además si se cumple que para $i = 1,2,...,r$ la expresión se minimiza. Luego la matriz de aproximación es:

$\mathbf{X} = \mathbf{UGV}^t = \sum_{i=1}^{r} \alpha_i u_i v_i^t = \mathbf{A}_{[r]}$. •

Observe que una medida de la calidad de la aproximación de **A** por $\mathbf{A}_{[r]}$ se puede definir como:

$$\tau_r = \left[\frac{\sum_{i=1}^{r} \alpha_i^2}{\sum_{i=1}^{k} \alpha_i^2} \right] 100 \qquad (10)$$

Finalmente se presenta una extensión del teorema 3, bajo las condiciones del teorema 2.

Teorema 4. *La matriz de aproximación por mínimos cuadrados generalizados o ponderados es*

$$\tilde{\mathbf{A}}_{[r]} = \sum_{i=1}^{r} \alpha_i n_i m_i^t = \mathbf{N}_r\mathbf{D}_r\mathbf{M}_r^t \qquad (11)$$

Demostración. Similar a la demostración anterior, sólo defina y observe que .

$\| \Omega^{-1/2}(\mathbf{A} - \mathbf{X})\varphi^{-1/2} \|^2 = \| \Delta - \tilde{\mathbf{G}} \|^2$. •

Aplicación al ACP

Sea $\mathbf{X} = [x_{ij}]$ la matriz donde los renglones representan *n* individuos y las columnas son variables observadas, $j = 1, 2,...,m$. Ahora defina la matriz $\mathbf{Y} = \left[\frac{x_{ij} - \bar{x}}{\sqrt{n}} \right]$, donde $\bar{x} = \sum_{i=1}^{n} \frac{x_{ij}}{n}$. Observe que $\mathbf{Y}^t\mathbf{Y} = \mathbf{S}$ es la matriz de varianzas y covarianzas. Si se descompone en valores singulares la matriz $\mathbf{Y}$ se obtiene el ACP, donde los vectores singulares por la derecha son los coeficientes de las componentes y los valores singulares son sus desviaciones estándar.

Cuando se trabaja con variables estandarizadas, para eliminar así el efecto que podría causar el que las variables tengan unidades de medida muy diferentes, o el que alguna de ellas tenga varianza mayor que el resto, el ACP recibe el nombre de Análisis de Componentes Principales Normado (ACPN). Partiendo del AG el ACPN se obtiene si se considera la SVD de la matriz,, $\tilde{\mathbf{Y}} = \left[\frac{y_{ij}}{s_j} \right]$, donde s_j es la desviación estándar de la *j*-ésima variable, pues $\tilde{\mathbf{Y}}^t\tilde{\mathbf{Y}} = \mathbf{R}$, la matriz de correlaciones.

Aplicación al ACS

Sea $\mathbf{F}$ la matriz de frecuencias relativas obtenidas de la tabla de contingencia. Defina los vectores $f = \mathbf{F}1$ y $c = \mathbf{F}^t 1$, los elementos de *f* son las frecuencias marginales relativas de las filas de $\mathbf{F}$ y los elementos de *c* son las frecuencias marginales relativas de las columnas. Además $\mathbf{D}_f$ y $\mathbf{D}_c$ y son matrices diagonales con los elementos de *f* y *c* en la diagonal principal, respectivamente.

El ACS se obtiene al considerar la SVD generalizada de $(\mathbf{F} - fc^t)$, donde $\mathbf{N}^t\mathbf{D}_f^{-1}\mathbf{N} = \mathbf{I}$ y $\mathbf{M}^t\mathbf{D}_c^{-1}\mathbf{M} = \mathbf{I}$, o bien la SVD de la matriz de residuales $\mathbf{A} = \mathbf{D}_f^{-1/2}(\mathbf{F} - fc^t)\mathbf{D}c^{-1/2}$. Las coordenadas de las filas de la tabla de contingencia son las filas de $\mathbf{C}_f = \mathbf{D}_f^{-1/2}\mathbf{U}_1\mathbf{D}$ y las coordenadas de las columnas son $\mathbf{C}_c = \mathbf{D}_c^{-1/2}\mathbf{V}_1\mathbf{D}$.

Ejemplo

La tabla de contingencia del Cuadro 1 contiene la clasificación de los 300 municipios más representativos de México, determinados por el volumen de captación y aplicación de recursos, de acuerdo al porcentaje de sus egresos destinados a obra pública durante el año 2001 y a su región geográfica.

Los datos se analizaron con el paquete estadístico S-PLUS para obtener las coordenadas de las categorías de las dos variables, vea la representación gráfica de la tabla de contingencia en la Figura (1).

Cuadro 1. Tabla de contingencia con clasificación de los 300 municipios más representativos de México, determinados por el volumen de captación y aplicación de recursos, de acuerdo al porcentaje de sus egresos destinados a obra pública durante el año 2001 y a su región geográfica.

Región	**Bajo** (0-24.40) %	**Medio** (24.41-48.80) %	**Alto** (48.81-73.20) %	**Total**
Norte	54	21	5	80
Centro	78	56	10	144
Sur	38	28	10	76
Total	170	105	25	300

Fuente: Elaboración propia de acuerdo a las Finanzas públicas estatales y municipales de México, (2001). INEGI.

Los municipios de la región norte se asocian más con un bajo porcentaje de egresos destinado a obra pública, mientras que los municipios del centro se asocian con porcentajes medios, los municipios de la región sur se asocian a porcentajes altos y medios. La aproximación de la matriz a la matriz de residuales es de 100 %.

Conclusiones

De lo anterior se desprende que el ACP, el ACPN y el ACS pueden ser obtenidos partiendo del AG, aplicando la SVD a la matriz adecuada. El AG es una poderosa herramienta que facilita considerablemente el estudio de los métodos multivariados aquí presentados. Obtener más métodos multivariados a través del AG puede ser considerado para futuros estudios.

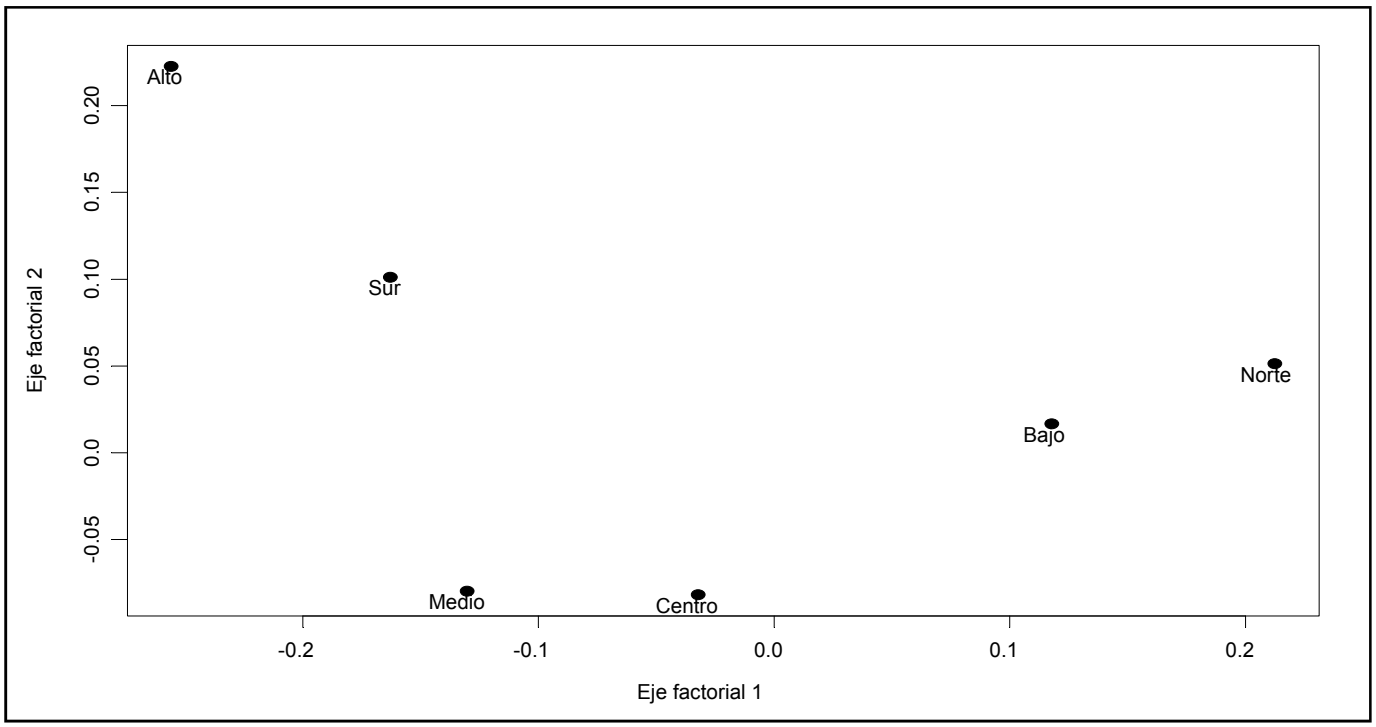

Figura 1. Tabla de contingencia de 300 municipios más representativos de México, determinados por el volumen de captación y aplicación de recursos, de acuerdo al porcentaje de sus egresos destinados a obra pública durante el año 2001 y a su región geográfica.

Literatura Citada

Greenacre, M. J. y Blasius. 1994. Correspondence analysis in the social science. Academic Press. 370 p.

Johnson, R. M. 1963. On a theorem stated by Eckart and Young. Psychometrika 28:259-263.

Lebart, L., Morineau, A. yWarwick, K. 1984. Multivariate descrptive statistical analysis. John Wiley and Sons. 231 p.