Una propuesta para la comprensión del Análisis en Componentes Principales

Giovanni Sanabria Brenes[1]
Félix Núñez Vanegas[2]

Resumen

Se propone el análisis de una tabla de diez individuos por tres variables a través de situaciones problemas guías que conducen a un análisis en componentes principales (ACP), tomando como base didáctica la Teoría de Situaciones de Brousseau. La resolución de las situaciones problema junto con las etapas de institucionalización, permiten la comprensión de la técnica ACP. Se pretende que esta propuesta sirva de ayuda y referencia a aquellos investigadores que tengan que realizar estudios similares y no cuentan con los conocimientos necesarios para tales efectos.

Palabras clave: Análisis en Componentes Principales (ACP), didáctica de la matemática, centro de gravedad, nube de puntos, inercia, plano principal, círculo de correlaciones.

1. Introducción

El auge que en los últimos años han tenido los ordenadores, ha provocado que las técnicas del análisis multivariado de datos sean, hoy por hoy, herramientas útiles en las investigaciones, tanto cuantitativas como cualitativas. Se aplican en disciplinas como la Sociología, Biología, Farmacia entre otras.

Dicho análisis podría revelar relaciones entre los datos que el investigador no había tomado en cuenta, lo cual implicaría proponer nuevas hipótesis y de pronto desechar otras.

No obstante, por su naturaleza, para algunos investigadores las técnicas del Análisis Multivariado de Datos no son fáciles de entender y más difícil todavía es su correcta aplicación.

Debido a lo anterior, se presenta ante la Vicerrectoría de Investigación y Extensión del Instituto Tecnológico de Costa Rica el proyecto "ESTUDIO DE MÉTODOS DE ANÁLISIS MULTIVARIADO DE DATOS" que pretende abordar el problema de cómo comprender y aplicar las principales técnicas del Análisis Multivariado de Datos, entre ellas, el ACP y el AFC. Dicho proyecto fue aprobado en agosto de 2007 y finaliza en junio de 2009. El proyecto tiene por objetivo: Desarrollar una metodología para

[1] Instituto Tecnológico de Costa Rica – Universidad de Costa Rica, gsanabria@itcr.ac.cr
[2] Instituto Tecnológico de Costa Rica – Universidad de Costa Rica, fnunez@itcr.ac.cr

abordar las principales técnicas del Análisis de Datos desde una perspectiva didáctica tales como: Análisis en Componentes Principales (ACP), Análisis Factorial de Correspondencias (AFC), Análisis Factorial de Correspondencias Múltiples (ACM), Análisis Factorial Discriminante (AFD) y Clasificación Jerárquica Ascendente (CJA).

El presente trabajo brinda una propuesta para la comprensión de una de las técnicas más utilizada en análisis de datos: El Análisis en Componentes Principales. Este análisis es la base para otras técnicas como el AFC y AFD. Se pretende, desde un punto de vista didáctico, analizar tres variables asociadas a diez individuos y que tal análisis sirva de ayuda y referencia a aquellos investigadores que tengan que llevar a cabo estudios similares y no cuentan con los conocimientos necesarios para tales efectos.

2. Marco de Referencia

- **Teoría didáctica**. Toda intención didáctica debe estar apoyada en concepciones sobre la enseñanza y el aprendizaje. En nuestro caso, nuestra intención toma la Teoría de Situaciones, de Guy Brousseau [2] y el trabajo realizado por Carmen Batanero para proponer una didáctica para el análisis multivariado de datos. El lector que lo desea puede consultar Núñez & Sanabria [5] donde se justifica y desarrolla esta didáctica. Las ideas expuestas en Núñez & Sanabria [5] están plasmadas en el presente trabajo y viene a ser un aporte desde el polo del estudiante, aunque desde luego es también un esfuerzo para el profesor interesado en una metodología para el establecimiento de tales conocimientos, y por ello no se puede abandonar el polo epistemológico. Por otro lado, la propuesta se base en las pautas metodológicas expuestas en Sanabria & Núñez [6].

- **El Análisis en Componentes Principales**. El Análisis en Componentes Principales fue propuesto por Karl Pearson a inicios del siglo XX, esta técnica es base de la mayoría de los métodos de análisis de datos y su importancia surgió con el desarrollo de los computadores. El lector interesado en abordar el desarrollo teórico de estos métodos puede consultar Trejos [7] y Trejos [8].

- **Público meta**. El presente material va dirigido al sector docente y estudiantil de las universidades y a investigadores principiantes. Estos deben tener

conocimientos básicos de estadística descriptiva (análisis univariado) y de algebra lineal.

3. La tabla de datos

Este análisis se utiliza en tablas de individuos × variables cuantitativas y consiste en hallar un número menor de variables nuevas que conserven la mayor cantidad de información de los datos.

Supongamos que se tomaron diez individuos y se determinó su edad, su estatura y peso[3]:

# de individuo	Edad (años)	Peso (kg)	Estatura (cm)
1	20	55	155
2	18	52	160
3	24	60	162
4	23	53	170
5	22	50	157
6	40	48	140
7	45	60	157
8	26	49	180
9	49	65	175
10	38	67	186

4. Centro de gravedad, nube de puntos e inercia

Cada individuo se puede representar como un vector de tres entradas. Por ejemplo el individuo seis es (40,48,140) pues tiene 40 años, un peso de 48 kilos y una estatura de 140 cm.

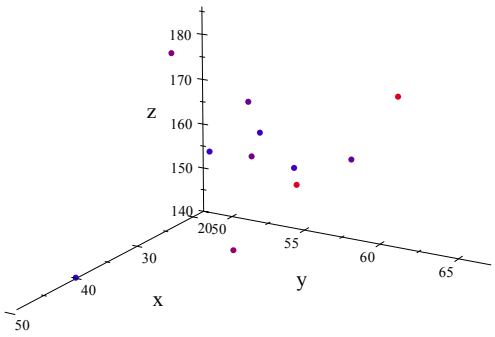
Situaciones problema #1: Para estos diez individuos:

1. ¿Cuál es el individuo promedio ? (Concepto de centro de gravedad)
2. Represente gráficamente a los individuos (Concepto de nube de puntos)
3. ¿Qué utilidad tiene una variable con varianza cero?
4. Si una variable tiene varianza cercana a cero, ¿Qué se puede decir sobre la cantidad de información que aporta?

[3] Estos datos fueron inventados y no proviene de la encuesta. Su fin es didáctico.

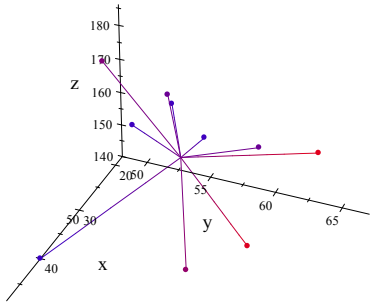
5. En un análisis univariado, la cantidad de información de una variable se puede cuantificar por medio de su varianza. ¿Cómo podemos medir la cantidad de información en la nube de puntos? (Concepto de inercia)

En estos datos el centro de gravedad o el individuo típico está dado por el vector (30.5,55.9,164.2). La nube de puntos es:



Cada punto círculo es un individuo y el punto cuadrado representa el centro de gravedad.

La inercia de la nube de puntos es el promedio de la suma de las distancias al cuadrado de cada individuo al centro de gravedad que es igual a la suma de las varianzas de las variables:



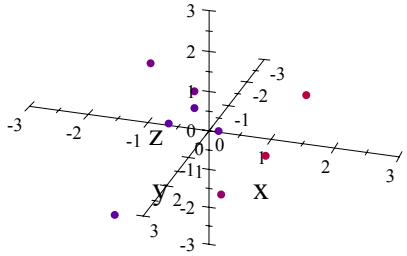
La inercia de la nube de puntos es 115.65+40.89+167.16 = 323. 7

5. Nube de puntos centrada y estandarizada

Situaciones problema #2: Al trabajar con la nube de puntos, se presenta el problema de que las variables que describen a los individuos pueden ser muy diferentes, tiene diferentes unidades de medida, lo cuál amplifica la inercia, ¿cómo se puede evitar esto?

Para proceder con el análisis se debe trabajar con una nube de puntos centrada y estandarizada por variables, esto es, a cada valor de la variable se le debe restar la media y dividir entre su desviación estándar. Esto permite eliminar el efecto de las unidades de medida dándole importancia al poder discriminante de las variables:

individuo	Variable 1	Variable 2	Variable 3
1	-0.976374617	-0.140745317	-0.711576581
2	-1.162350735	-0.609896375	-0.324850178
3	-0.604422382	0.641173112	-0.170159617
4	-0.697410441	-0.453512689	0.448602627
5	-0.7903985	-0.922663746	-0.55688602
6	0.883386558	-1.23543112	-1.871755788
7	1.348.326.852	0.641173112	-0.55688602
8	-0.418446264	-1.079047432	1.222055432
9	1.720279087	1.423091541	0.835329029
10	0.697410441	1735858912	1.686127115



Dado que las variables centradas y estandarizadas tienen media cero y varianza 1 entonces el centro de gravedad de la nube centrada y estandarizada es cero y la inercia es : I=1+1+1=3

6. El plano principal y las componentes principales

Nuestra nube de puntos está en un espacio de tres dimensiones, se busca hallar un espacio de menor dimensión que contenga la mayor cantidad de inercia. Así, se busca el plano (espacio de dos dimensiones) en el cual las proyecciones de los individuos tengan inercia máxima. Este plano es llamado el plano principal.

La matriz de correlaciones es una matriz simétrica que contiene en las entradas (i,j) la correlación entre la variables i y j ya centradas y estandarizadas Note que esta matriz tiene unos en la diagonal. Un resultado algebraico importante es el siguiente:

Dados los valores propios de la matriz de correlaciones, el plano principal está generado por los dos vectores propios asociados a los dos mayores valores propios.

Estos vectores propios son llamados los ejes principales del plano, la inercia de las proyecciones de los individuos sobre cada eje corresponde al valor propio al que esté

asociado y la inercia de las proyecciones de los individuos sobre el plano corresponde a la suma de las inercias asociadas a los ejes.

Situaciones problema #3:

1. Determine la matriz de correlaciones de nuestra tabla de datos.

2. Para lectores con conocimiento en Algebra Lineal: Determine los vectores y valores propios de la matriz de correlaciones.

La matriz de correlaciones para nuestros datos es

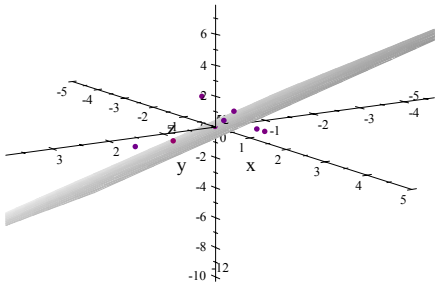
$$\begin{pmatrix} 1 & 0.538774 & 9.99714\times10^{-2} \\ 0.538774 & 1 & 0.525188 \\ 9.99714\times10^{-2} & 0.525188 & 1 \end{pmatrix}$$

Los valores y vectores propios de esta matriz son

$$\text{Valor propio}: \quad 1.80403 \qquad 0.900062 \qquad 0.295911$$

$$\text{Vector propio asociado}: \quad v_1 = \begin{pmatrix} 0.521412 \\ 0.683273 \\ 0.511143 \end{pmatrix} \quad v_2 = \begin{pmatrix} 0.697682 \\ 3.51684\times10^{-3} \\ -0.716399 \end{pmatrix} \quad v_3 = \begin{pmatrix} 0.491294 \\ -0.730155 \\ 0.474873 \end{pmatrix}$$

Por lo tanto, el plano principal es el plano generado por los vectores v_1 y v_2. Para poder representar el plano principal en el espacio[4], es necesario conocer algún punto que contega el plano, por ejemplo el centro de gravedad o origen:



Es indiferente cual punto contenga el plano, pues las proyecciones de los individuos en el plano serán las mismas. En realidad, para un estudio no interesa representar el plano principal en el espacio. Basta representarlo en dos dimensiones con las proyecciones de

[4] El plano generado por los vectores v_1 y v_2 que pasa por el punto P tiene ecuación:
$((x, y, z) - P)\,(v_1 \times v_2) = 0$

los individuos. Así en el plano principal, el individuo es visto como un par de ordenado, por lo que se tiene dos nuevas variables llamadas componentes principales:

	Nube de puntos	Plano principal
Individuos:	vectores de 3 entradas	vectores de 2 entradas
Variables	Son 3: edad,peso,estatura	Son 2: C1,C2

En general existen tres nuevas variables C1,C2,C3 cada una asociada a un valor propio que indica la parte de la inercia que representa

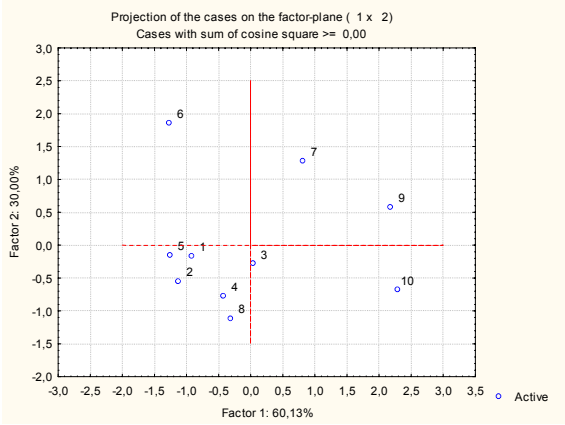
Ejes	**Componente principal**	**Inercia del eje**	**% de inercia (Calidad de la representación)**
Eje 1	C1	I_1=1. 80403	1. 80403)/I = 60.1%
Eje 2	C2	I_2=0.900062	0.900062/I = 30%
Eje 3	C3	I_3=0.295911	0.295911/I = 9.9%
Total:		I=3	100%

Estas variables, a diferencia de las anteriores, tienen correlación cero y la suma de sus varianzas es 3, que es la inercia de la nube centrada y estandarizada. El plano principal está formado por los ejes 1 y 2, éste tiene inercia: I_1+I_2 = 1.80403 + 0.900062 = 2. 70409 que equivale al 90.1% de la inercia. Otros planos son:

Plano	Inercia del plano	% de inercia del plano
1 y 3	I_1+I_3=1. 80403+0.295911= 2.09994	70%
2 y 3	I_2+I_3=0.900062+0.295911= 1. 19597	40.9%

Así, el plano principal es el que contiene la mayor inercia, y las variables C□ y C□, son las que mejor describen a las originales (son combinación lineal de estas) y no tiene información redundante. Para hallar el valor de estas nuevas variables para cada individuo se multiplica la tabla de datos (centrada y estandarizada) por el vector propio asociado. Con estos datos, podemos representar a los individuos en el plano principal:

Individuo	C1	C2
1	-0.919254	-0.163098
2	-1. 12783	-0.550589
3	3. 41214×10^{-2}	-0.282268
4	-0.421416	-0.768001
5	-1. 2591	-0.147747
6	-1. 27148	1. 85268
7	0.812532	1. 27305
8	-0.313844	-1. 11111
9	2. 17847	0.575645
10	2. 2878	-0.678559



7. Interpretación de resultados en el plano principal

Algunas interpretaciones de los resultados que se obtienen a partir del plano principal y la tabla de datos original son:

1. Los individuos 1,5,2 son muy parecidos.

2. La primer componente separa los delgados-jóvenes de los viejos-más pesados.

3. La segunda componente divide los grupos de edad según la relación de su peso y estatura. De un lado están los individuos cuarentones (6,7 y 9), para los cuales el peso es directamente proporcional a la estatura, y del otro lado están los individuos de veinte y resto, un grupo de edad para el cual, el peso y la estatura no se correlacionan (ver en especial los individuos 4 y 8).

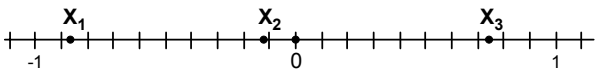
Situaciones problema #4:

1. Realice la interpretación de resultados para los otros planos.
2. De las interpretaciones obtenidas, ¿cuales tienen mayor peso?

8. El círculo de correlaciones

Situaciones problema #5:

1. ¿Comó se mide la relación de una variable original con una componente principal?
2. De las tres variables originales, ¿cuáles tienen mayor relación con la primer componente?
3. Suponga que se tiene tres variables X_1, X_2, X_3, representadas con un punto en el segmento [-1,1] que indica su correlación con una componente principal:

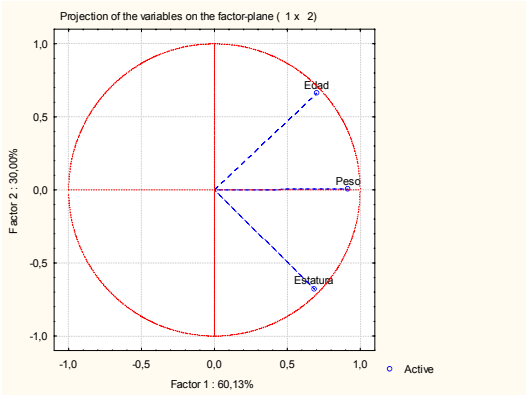


   a. ¿Qué relación tiene cada variable con la componente principal?
   b. ¿Qué relación tienen las variable X_1 y X_3? Explique
   c. ¿Cuáles variables están bien representadas por la componente principal?
4. A partir del punto anterior ¿Cómo se pueden medir las relaciones entre variables en el plano conformado por dos componentes principales?

Otra representación gráfica del ACP es el círculo de correlaciones. Para el caso de dos componentes principales, cada variable se representa como un punto en $\Re^2$ donde: la coordenada X es el coeficiente de correlación de la variable y la primer componente, y la coordenada Y es el coeficiente de correlación de la variable y la segunda componente. En nuestro caso se tiene que:

Variable	Correlación con C1	Correlación con C2	Coordenadas en $\mathbb{R}^2$
Edad	0.700331	0.661903	(0.700331,0.661903)
Peso	0.917732	$3.\ 33691\times10^{-3}$	$(0.917732,3.\ 33691\times10^{-3})$
Estatura	0.686537	-0.679659	(0.686537,-0.679659)

Dado que una correlación varía de -1 a 1 y las componentes no se correlacionan entonces se traza un círculo centrado en cero de radio 1. Cuanto más cerca esté la representación de la circunferencia, mejor representada esta la variable en ambas o alguna de las componentes. El círculo de correlaciones principal para nuestros datos es



Interpretación de resultados en el círculo de correlaciones

9. Algunas interpretaciones que se pueden realizar son:

1. Relaciones entre el plano principal y el círculo de correlaciones.

    a. Note que las tres variables están bien representadas por estas componentes, especialmente por la primera que mantiene una correlación positiva con todas.

    b. Si vemos el plano principal, los individuos que tiene alta la primer componente son el 9 y 10.

c. Combinando los dos puntos anteriores, se concluye que los individuos 9 y 10 están por encima del promedio en edad, peso y estatura. Contrario a los individuos 2, 5 y 1.

d. Así, si se transpone el círculo de correlaciones sobre el plano principal, entonces los individuos cercanos a una variable, indican que están por encima del promedio con respecto a esa variable.

2. Relaciones entre variables.

a. Considere en el círculo de correlaciones el ángulo θ formado entre las variables A y B (cercanas a la circunferencia) con centro en el origen.

b. Si θ es cercano a 90°, esto quiere decir que si una variable está muy correlacionada con una componente entonces la otra estará muy correlacionada con la otro componente, pero las componentes no se correlacionan, entonces las variables no se correlacionan (coeficiente de correlación se acerca 0).

c. Si θ es pequeño . En este caso ambas variables están correlacionadas con las componentes de la misma manera. Es decir, por ejemplo si A tiene una correlación positiva con la componente 1, B también. Por lo tanto las variables tienen una correlación positiva. (coeficiente de correlación se acerca 1).

d. Si θ es grande cercano a 180° . En este caso ambas variables están correlacionadas con las componentes de manera inversa. Es decir, por ejemplo si A tiene una correlación positiva con la componente 1, B tendrá una correlación negativa. Por lo tanto las variables tiene una correlación negativa. (coeficiente de correlación se acerca -1).

e. Lo anterior justifica que el coeficiente de correlación entre las variables A y B está muy asociado al cosθ.

f. En nuestro ejemplo, las variables están correlacionadas positivamente.

10. ¿En qué consiste el análisis en componentes principales?

Situaciones problema #6: Ejecute varias veces el software "Comprendiendo el ACP" adjunto en el CD (Este software requiere de la instalación previa del JAVA 3D, también adjunto en CD).

Situaciones problema #7: Suponga que se tiene una tabla de datos de 200 individuos por 60 variables y se le desea aplicar un ACP.

1. ¿Cuál es la dimensión de la nube de puntos? ¿Se puede visualizar?
2. ¿Cuánto es la inercia de la nube centrada y estandarizada?
3. Suponga que se determinan los valores propios de la matriz de correlaciones. ¿Cuántos valores propios son? ¿Cuánto es su suma?
4. ¿Cuántas variables nuevas se obtienen con el ACP? ¿Cómo medir el porcentaje de información que tendrá cada nueva variable?
5. Suponga que se quiere hallar el espacio de dimensión 3 que mejor represente la información. ¿Cuántas y cuáles componentes escogería para representar a las variables? ¿Cómo medir la calidad de esta representación?

En general, el ACP se aplica a tablas con varios individuos y variables, así se tiene una tabla de n individuos y m variables cuantitativas, se tiene que:

1. El ACP obtiene m nuevas variables que contengan toda la información de las anteriores y que no contengan información redundante (tiene correlación cero). Estás son llamadas componentes principales y son ordenadas de mayor a menor de acuerdo a la cantidad de información que conserven (dada por los valores propios de la matriz de correlaciones): $C_1, C_2, ..., C_m$.

2. La idea del ACP es elegir de esas nuevas variables aquellas que mejor me representen la información y me ayuden a describirla.

3. ¿Cómo elegir esas variables? Un resultado importante es que el espacio que mejor representa la información (espacio óptimo) de dimensión k está contenido en el espacio óptimo de dimensión k+1 cuya inercia es la del espacio k más la inercia de la componente agregada. Así:

Dimensión del espacio	espacio óptimo
1	1° componente principal: C_1

2	Plano principal: C_1 y C_2
3	Espacio de dim 3 principal: C_1,C_2 y C_3
⋮	⋮

4. ¿Cómo describir la información? A través de planos y círculos de correlaciones. Los software del mercado (SPSS, PIMAD, entre otros) realizan el ACP a partir de la tabla de datos, brindan los valores propios y permiten visualizar los diferentes planos y círculos de correlaciones. Así, por ejemplo, se puede escoger el espacio de dim3 principal y analizar los planos 1-2,1-3 y 2-3 junto con sus respectivos círculos de correlaciones. La interpretación de los resultados del ACP le corresponde al usuario.

El Pimad 3.0 permite realizar el ACP a partir de la tabla de datos. Para ello, siga los siguientes pasos:

1. Abrir el programa Análisis de Componentes Principales de Pimad 3.0
2. Cargar los datos (primer botón de izquierda a derecha):
    a. Un archivo para la tabla de datos extensión *.TXT, este archivo debe encabezarse con 2 números enteros, el primero indica en número de Individuos y el segundo para el numero de variables, y luego los datos.
    b. Un archivo de etiquetas para los individuos extensión *.ETI, que debe ser encabezado por un número entero que indica la cantidad de individuos. Luego en cada fila se coloca la etiqueta del individuo correspondiente.
    c. Un archivo de etiquetas para las variables extensión *.ETV, que debe ser encabezado por un número entero que indica la cantidad de variables. Luego en cada fila se coloca la etiqueta de la variable correspondiente.
3. Generar un plano (segundo botón de izquierda a derecha).
4. Generar un círculo de correlaciones (tercer botón de izquierda a derecha).
5. Adicionalmente, se pueden generar otros elementos, como la matriz de correlaciones, utilizando secuencialmente el menú ACP-Paso-a-Paso.

11. Conclusiones del trabajo

La teoría de Situaciones nos permitió abordar el problema de la enseñanza -- aprendizaje del ACP. A través de 6 grupos de situaciones problema, se desarrollaron los elementos principales de esta técnica, todas contextualizadas en el análisis de cierta tabla de datos. Se pretende que este trabajo lleve a una mayor comprensión de las

técnicas del Análisis Multivariado de datos presentadas, para que los académicos puedan utilizarlos en su quehacer profesional.

12. Bibliografia

1. Antibí, A. 2000. Didáctica de las Matemáticas: Métodos de Resolución de problemas. Serie Cabecar, Costa Rica.
2. Brousseau, Guy. 1986. Fundamentos y Métodos de la Didáctica de las Matemáticas, traducción de "Fondements et méthodes de la didactiques des mathématiques". Revista Recherches en Didactique des Mathématiques, Vol 7, n 2, pp.33-111 .
3. Chevallard, Yves.1991. La Transposición Didáctica. Del saber sabio al saber enseñado. Aique grupo Editor S.A., Argentina.
4. Lam, Longhow. 1999. An Introduction to S-PLUS for Windows. Amsterdam, Holanda.
5. Núñez, F, Sanabria, G. 2009. Didáctica del Análisis Multivariado de Datos. Capítulo 2 del documento "ESTUDIO DE MÉTODOS DE ANÁLISIS MULTIVARIADO DE DATOS", proyecto de investigación del ITCR.
6. Sanabria, G. Núñez, F. 2009. Propuesta metodológica para la enseñanza de las principales Técnicas del Análisis Multivariado de Datos. Capítulo 3 del documento "ESTUDIO DE MÉTODOS DE ANÁLISIS MULTIVARIADO DE DATOS", proyecto de investigación del ITCR.
7. Trejos, Javier 2004. Notas del curso de Análisis de Datos Multivariados, correspondiente al programa de Maestría en Matemática con énfasis en Matemática Educativa. Universidad de Costa Rica.
8. Trejos, J. 1998. Introducción al análisis de datos. Programa de investigación en modelos y análisis de datos, Escuela de matemática, UCR.
9. Trejos, Javier 2000. Manual del Usuario. PIMAD. Versión 3.0, Costa Rica.
10. Vergnaud, G. 1990. "La théorie des champs conceptuels", Recherches en Didactique des Mathémathiques Vol. 10 (23): 133-170.
11. Visauta, B. 1999 Análisis estadístico con SPSS para Windows. McGraw-Hill. España.