Estadística avanzada en medicina: el análisis de componentes principales

Pablo F. Argibay

En general, cualquiera que haya intentado registrar y posteriormente analizar en forma más o menos avanzada sus datos, se encuentra con un problema (A Grané 2003):

- Cómo describir la información contenida en un conjunto de datos mediante algún conjunto menor de variables que sean manipulables y signifiquen algo para el investigador.

La idea subyacente en el método que analizaremos es que:

- Si una variable es función de otras, contiene información redundante.

Por lo tanto, si muchas de las variables observadas están fuertemente correlacionadas, será posible sustituirlas por menos variables sin gran pérdida de la información.

El problema práctico sería el típico de una hoja de Excel en la que tenemos registrados gran cantidad de pacientes y gran cantidad de variables de cada paciente: cómo hacer que los datos "hablen" y nos muestren algún patrón que los relacione y a su vez elimine los datos de alguna manera redundantes. Dentro de los métodos de análisis de más de una variable, el análisis de componentes principales es una forma rápida y sencilla de encontrar un patrón de relación entre los datos y a su vez de reducir gran cantidad de variables. Sus fundamentos modernos se encuentran en el álgebra lineal y su gran difusión en los últimos años se debe al poder de cómputo actual.

CONCEPTOS

La idea central del análisis de componentes principales (PCA) es reducir la dimensionalidad de un conjunto de datos consistente en un número elevado de variables interrelacionadas. Se trata de mantener de la mejor manera posible la variación contenida en los datos. Esto se logra transformando el conjunto original en un nuevo conjunto de datos, los componentes principales, que son no correlacionados y que están ordenados de tal manera que los primeros pocos retengan la mayor parte de la variación presente en todas las variables originales.

HISTORIA (Jollife, 2010)

Se dice que Beltrami y Jordan de forma independiente llegaron a la descomposición de valores singulares (SVD), de una manera que subyace en el PCA. La SVD ha sido usada inicialmente por Fisher y Mackenzie en el contexto de un análisis bivariado en estudios de agricultura. Sin embargo, se acepta en general que la técnica de PCA tal como la conocemos ahora se debe a Pearson y Hotelling. En particular, Pearson comenta (medio siglo antes del desarrollo masivo de las computadoras) que estos métodos pueden ser fácilmente aplicables a problemas numéricos, a pesar de su preocupación de que se limitara el método por las limitaciones computacionales a 4 ó 5 variables.

A su vez Hotelling parte de que debería existir un conjunto fundamental de variables independientes, que determinaran los valores del conjunto original y mayor de *p* variables. A estas variables independientes, Hotelling las denomina "factores" en la literatura de investigación en psicología, pero también introduce el término "componentes" para evitar confusiones con los factores en matemática. Hotelling elige sus "componentes" de manera de maximizar sus sucesivas contribuciones al total de las varianzas de los valores originales. Por otra parte denomina a los componentes derivados de esta manera como "componentes principales". De algún modo la derivación de Hotelling de los componentes principales es similar a los métodos que utilizan multiplicadores de Lagrange,[1] finalizando el problema de encontrar autovalores y autovectores. Sin embargo, el método de Hotelling difiere en tres aspectos:

1. Trabaja con correlaciones y no con matrices de covarianza.
2. Trabaja con las variables originales expresadas como funciones lineales de los componentes más que como componentes expresados en términos de las variables originales.
3. No utiliza notación matricial.

1. En los problemas de optimización (problemas en los cuales se desea elegir el mejor entre un conjunto de elementos), el método de los multiplicadores de Lagrange, llamados así en honor de Joseph Louis Lagrange, es un procedimiento para encontrar los máximos y mínimos de funciones de varias variables sujetas a restricciones. Este método reduce el problema restringido con n variables a uno sin restricciones de n + k variables, donde k es igual al número de restricciones y cuyas ecuaciones pueden ser resueltas más fácilmente. Estas nuevas variables escalares desconocidas, una para cada restricción, son llamadas multiplicadores de Lagrange. El método dice que buscar los extremos condicionados de una función con k restricciones, es equivalente a buscar los extremos sin restricciones de una nueva función construida como una combinación lineal de la función y las restricciones, donde los coeficientes de las restricciones son los multiplicadores.

Laboratorio de aprendizaje biológico y artificial, ICBME-IU Hospital Italiano de Buenos Aires
Correspondencia: pablo.argibay@hospitalitaliano.org.ar

Durante varios años, luego de Hotelling, no hubo grandes avances en PCA. Sin embargo, en los últimos años se han sucedido una explosiva cantidad de desarrollos teóricos y aplicaciones en relación con PCA. Tal vez esto refleje el crecimiento general de los métodos estadísticos y sus aplicaciones y básicamente el crecimiento del "poder de cómputo". Esto es básico ya que, como lo predijo Pearson, no es factible efectuar PCA a "mano" salvo que *p* sea menor a cuatro.

ALGUNAS MEDIDAS ESTADÍSTICAS SUBYACENTES EN PCA (Smith, 2002)

La covarianza

El desvío estándar (ds) y la varianza (v) son conocidas medidas de dispersión de los datos en estadística descriptiva. Por otra parte son medidas unidimensionales. Sin embargo, en gran parte de los problemas científicos en general y en particular en biología y medicina, las variables son multidimensionales (analizan varias variables). El ejemplo más simple sería el análisis conjunto del índice de masa corporal (variable *x*) con la presión arterial (variable *y*) de un grupo de individuos. El objetivo, más allá de analizar cómo varían en sí mismos *x* o *y*, sería analizar el efecto de *x* sobre *y*, o viceversa. Un estudio interesante consiste en analizar cómo varían las dimensiones de una variable con respecto a la media en relación con lo mismo en la otra variable. La covarianza es la medida útil para este tipo de análisis estadístico. La covarianza siempre se mide entre dos dimensiones y si uno calcula la covarianza de una dimensión consigo misma obtiene la varianza.

Ejemplo 1:

Si uno tiene tres dimensiones correspondiente a tres variables (*x, y* y *z*), a través de la covarianza uno puede analizar la covarianza entre *x* e *y*; o *x* y *z;* o *y* y *z*. Por otra parte el análisis combinado de esas covarianzas también nos da las varianzas de *x*, *y* y *z*.

La fórmula de la covarianza es similar a la de la varianza:

$$\mathrm{cov}(X, Y) = \frac{\sum_i^n = (Xi - X)(Yi - Y)}{(n - 1)}$$

En general, la covarianza es una medida de cómo una variable varía con respecto a la otra, en términos de la propia variación de sus valores con respecto a la media: esto podrá dar un valor positivo, negativo o cero; en este último caso ambas variables no varían una con respecto a la otra, es decir, son independientes. ¡Claro!, es fácil visualizar esto en un gráfico (*x*,*y*) elemental; sin embargo, cuando nos manejamos con más de tres variables, la visualización gráfica es imposible y la covarianza cobra un valor importante.

La matriz de covarianza

Un análisis interesante es entre múltiples variables. Para un conjunto *n*-dimensional de datos, uno puede calcular un número de covarianzas determinado por la fórmula:

$$\frac{n!}{(n - 2)! * 2}$$

Una forma útil de obtener todos los posibles valores de covarianza entre las diferentes dimensiones es calcularlos todos juntos y agruparlos en una matriz.[2]

La matriz de covarianza de un conjunto de n dimensiones es:

$$C^{n \times n} = (ci, j, ci, j = cov(Dim\ i, Dim\ j))$$

Donde $C^{n \times n}$ es una matriz con n filas y n columnas, y Dim_x es la dimensión x^{th}.

Ejemplo 2: Matriz de covarianza de 3 filas × tres columnas.

$$C = \begin{matrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{matrix}$$

ÁLGEBRA DE MATRICES (Antón, 2010)

Autovectores (*Eigenvectors*)

Se pueden multiplicar dos matrices entre ellas, siempre, que tengan tamaños compatibles.[3] Los autovectores son un caso especial. En el ejemplo 3 se observan las multiplicaciones entre una matriz y un vector.[4]

Ejemplo 3: ejemplo de un no autovector y un autovector (abajo).

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 11 \\ 5 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

En este ejemplo, en la operación de arriba, se observa que el vector resultante no es un múltiplo entero del vector original. En la operación de abajo, el vector resultante es exactamente 4 veces el vector de origen. Lo importante es que la matriz actúa como transformación de un vector bidi-

2. Una matriz es una tabla cuadrada o rectangular de datos (llamados elementos o entradas de la matriz) ordenados en filas y columnas, donde una fila es cada una de las líneas horizontales de la matriz y una columna es cada una de las líneas verticales. A una matriz con m filas y n columnas se la denomina matriz *m*-por-*n* (escrito *m*×*n*), y a *m* y *n* dimensiones de la matriz. Las dimensiones de una matriz siempre se dan con el número de filas primero y el número de columnas después.
3. El producto de dos matrices se puede definir solo si el número de columnas de la matriz izquierda es el mismo que el número de filas de la matriz derecha. Si *A* es una matriz *m*×*n* y *B* es una matriz *n*×*p*, entonces su producto matricial *AB* es la matriz *m*×*p* (*m* filas, *p* columnas) dada por: $(AB)[i,j] = A[i,1]\,B[1,j] + A[i,2]\,B[2,j] + \ldots + A[i,n]\,B[n,j]$ para cada par i y j.
4. A los fines prácticos de este trabajo, un vector es una matriz de n filas y una sola columna (vector columna).

mensional con origen en 0.0 y finalización en 3.2 (Fig. 1). Por otra parte el vector $\binom{12}{8}$ tiene diferente longitud pero la misma orientación (Fig. 2).

Figura 1. Gráfico de un vector bidimensional con origen en 0.0 y finalización en 3.2.

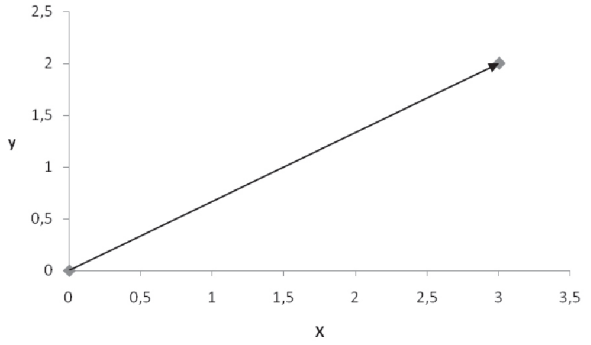
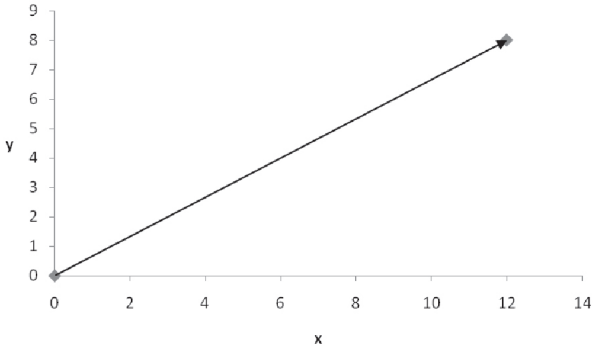


Figura 2. Gráfico de un vector bidimensional con coordenadas (x=12 e y=8), con diferente longitud e igual orientación que el vector representado en la figura 1.



Una propiedad de los autovectores es que, si uno hace una escala del vector, solamente cambia su longitud pero no su dirección.
Ejemplo 4:

$$2\times \begin{pmatrix}3\\2\end{pmatrix}=\begin{pmatrix}6\\4\end{pmatrix}$$

$$\begin{pmatrix}2&3\\2&1\end{pmatrix}\times\begin{pmatrix}6\\4\end{pmatrix}=\begin{pmatrix}24\\16\end{pmatrix}=4\times\begin{pmatrix}6\\4\end{pmatrix}$$

Una propiedad de los autovectores de una matriz es que estos son perpendiculares (ortogonales[5]) entre sí. Lo importante es que uno puede expresar sus datos en términos de autovectores, una propiedad utilizable en PCA. En general, para grandes matrices, la única forma razonablemente rápida y segura de encontrar sus autovectores es utilizar un *software* apropiado. La detección de autovectores es fácil de implementar en MatLab.

Autovalores (*Eigenvalues*)
Es fácil entender el concepto de autovalor asociado a un autovector observando el ejemplo 3 en su parte de abajo:

$$\begin{pmatrix}2&3\\2&1\end{pmatrix}\times\begin{pmatrix}3\\2\end{pmatrix}=\begin{pmatrix}12\\8\end{pmatrix}=4\times\begin{pmatrix}3\\2\end{pmatrix}$$

En este caso "4" es el autovalor del autovector $\begin{pmatrix}3\\2\end{pmatrix}$.

EL ANÁLISIS DE COMPONENTES PRINCIPALES

PCA es un método de identificación de patrones en un conjunto de datos. Por otra parte, expresa los datos de manera tal que resalta las similaridades y diferencias en los datos. La gran utilidad de PCA reside en el análisis de datos de elevada dimensionalidad, donde la graficación de las variables es dificultosa si no imposible. En relación con el presente trabajo práctico, la principal ventaja de PCA es la posibilidad de compresión de los datos, reduciendo el número de dimensiones sin gran pérdida de información. Surge claramente la ventaja en la compresión de imágenes.

El método

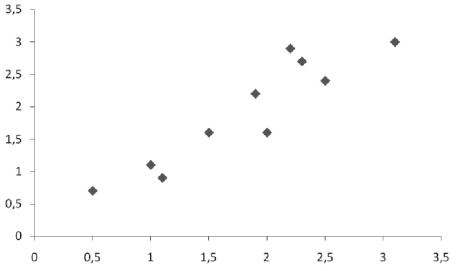
Supongamos un conjunto de datos bidimensional (Figs. 3A y 3B)

Figura 3. A. Conjunto bidimensional de datos. Ejemplo tomado de L. Smith (2002). **B.** Gráfico de puntos de los datos de la figura. **C.** Representación de los datos de la figura 3A, con las medias sustraídas de cada valor.

A.

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

B.



Sustracción de la media

El primer paso es sustraer las medias (Xm=1.81, Ym=1.91) de cada uno de los datos de cada dimensión (Fig. 3C).

Matriz de covarianza

$$\mathrm{cov}=\begin{pmatrix}0.616555556 & 0.615444444\\0.615444444 & 0.616555556\end{pmatrix}$$

5. Dos vectores son ortogonales si su producto escalar es cero. $\vec{u}.\vec{v}=0 \qquad u_1 . v_1 + u_2 . v_2 = 0$

Figura 3. C. Representación de los datos de la figura 3A, con las medias sustraídas de cada valor.

C.	
x	y
0.69	0.49
-1.31	-1.21
.39	0.99
.09	0.29
1.29	1.09
.49	0.79
.19	-0.31
-0.81	-0.81
-0.31	-0.31
-0.71	-1.01

Dado que los elementos no diagonales son positivos, deberíamos suponer que *x* e *y* aumentan juntos.

- Cálculo de autovectores y autovalores de la matriz de covarianza:

$$\text{autovalores} = \begin{pmatrix} 0.4990839\mathbf{89} \\ 1.2840277\mathbf{1} \end{pmatrix}$$

$$\text{autovectores} = \begin{pmatrix} -0.735178\mathbf{656} & -0.677873\mathbf{399} \\ 0.677873\mathbf{399} & -0.735178\mathbf{656} \end{pmatrix}$$

Se debe notar que estos autovalores son de magnitud "1". La figura 4 muestra el patrón de distribución de los datos ajustados y la inclusión de los autovectores. Como se puede observar, uno de los autovalores cruza por la mitad de los datos. Este vector relaciona los datos. Es decir, a través de los autovectores obtenemos dos líneas perpendiculares que caracterizan a los datos.

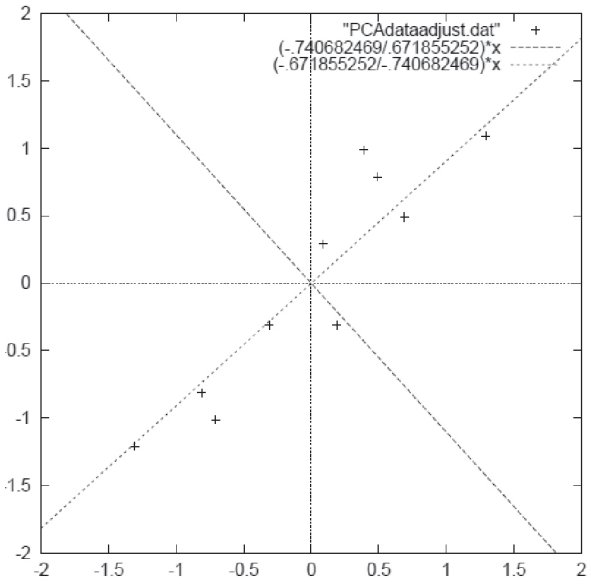
- Seleccionando componentes y formando un vector de rasgos (*feature vector)*:

Acá aparece la noción de "compresión de datos" y reducción de dimensionalidad. Los autovalores previos tienen valores bastante diferentes. El autovector con el mayor autovalor es el "componente principal" del conjunto de datos. En muchos casos se descartan los autovalores de menor significancia y en este caso se está reduciendo la dimensionalidad del problema. El nuevo tamaño será el de los *p* autovectores seleccionados.

La formación de una matriz de vectores (vector de rasgos) se formará con los autovectores seleccionados en las columnas: *Vector de rasgos* = $(eig1\ eig2\ eig3\ ...eign)$

Dado que en nuestro ejemplo explicativo solo existen dos autovectores, tenemos dos elecciones:

Figura 4. Diagrama de puntos donde se observan los datos normalizados con la sustracción de medias y los autovectores de la matriz de covarianza.



- Matriz con ambos autovectores:

$$\begin{pmatrix} -0.677873\mathbf{399} & -0.735178\mathbf{656} \\ -0.735178\mathbf{656} & 0.677873\mathbf{399} \end{pmatrix}$$

- Matriz de una columna con el mayor vector:

$$\begin{pmatrix} -0.677873\mathbf{399} \\ -0.735178\mathbf{656} \end{pmatrix}$$

Derivando el nuevo conjunto de datos

En esta etapa, una vez seleccionados los componentes y formado el vector de rasgos, tomamos la transposición[6] del vector y la multiplicamos con el conjunto original de datos transpuestos: *Datos finales = Vector de rasgos* (fila) × *Datos ajustados* (fila).

Los datos finales mostrarán los ítems en las columnas y las dimensiones en las filas. El significado de todo esto es que tenemos los datos originales en términos de los vectores elegidos. En el caso de una nueva matriz con ambos vectores tendremos la figura 5.

Como podemos ver, los autovectores ahora son los ejes.

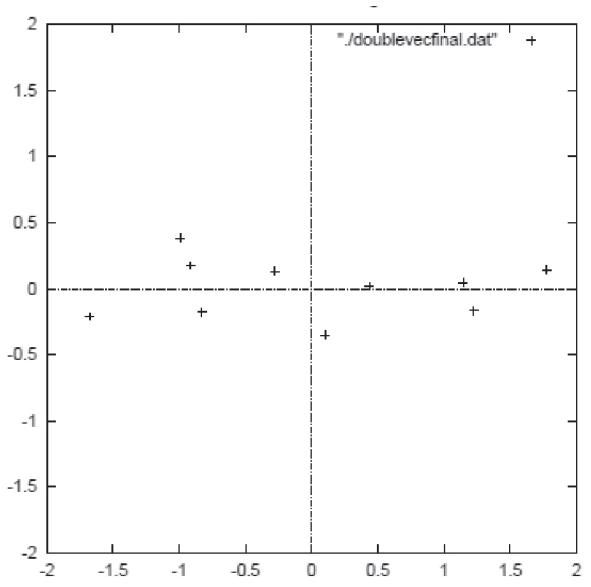
En el caso de la otra transformación, usando un solo vector (el autovector mayor), solo tenemos una dimensión y por lo tanto hemos reducido la dimensionalidad de los datos (Fig. 6).

Como se puede observar y es esperable, esta tabla de datos es igual a la primera columna de la tabla anterior.

Hasta ahora lo que hemos hecho conceptualmente es transformar nuestros datos de tal manera que expresen los

6. La transposición de un vector columna da lugar a un vector fila.

Figura 5. Tabla de datos y gráfico de puntos, aplicando PCA con el uso ambos autovectores.

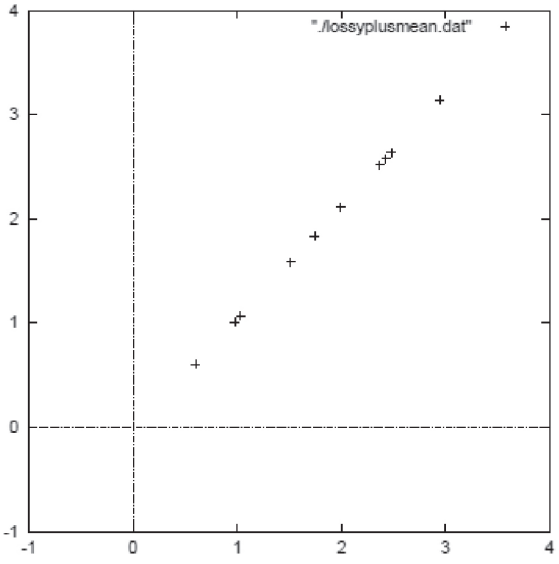


x	y
-0.827970186	-0.175115307
1.77758033	0.142857227
-0.992197494	0.384374989
-0.274210416	0.130417207
-1.67580142	-0.209498461
-0.912949103	0.175282444
0.991094375	-0.349824698
1.14457216	0.0464172582
0.438046137	0.0177646297
1.22382056	-0.162675287

Figura 6. Tabla de datos transformados usando el autovector más significativo.

x
-0.827970186
1.77758033
-0.992197494
-0.274210416
-1.67580142
-0.912949103
0.0991094375
1.14457216
0.438046137
1.22382056

Figura 7. Reconstrucción de datos que fueron derivados de la reducción de la dimensionalidad utilizando solo un autovector.



patrones entre ellos. Estos patrones son las líneas que mejor definen las relaciones entre los datos. En realidad, lo que hemos hecho es una clasificación de nuestros puntos de representación como una combinación de las contribuciones de cada una de las líneas.

Regresando a los datos originales

Teniendo en cuenta que PCA se utiliza habitualmente para compresión de datos, volver a los datos originales es del mayor interés. En primer lugar, la única manera de tener los datos originales consiste en que se pueden utilizar todos los autovectores. En el caso de la selección de los más significativos y por lo tanto de reducción de dimensionalidad, los datos recuperados pierden algo de información. Basándonos en la fórmula *Datos finales = Vector de rasgos (fila) × Datos ajustados* (*fila*), podemos volver atrás de la siguiente manera:

Datos ajustados (*fila*) = *Vector de rasgos* (*fila*)$^{-1}$ × *Datos Finales* o = *vector de rasgos* (*fila*) ^{T} × *Datos finales*.

Y finalmente: *Datos originales* = (*Vector de rasgos* (*fila*) ^{T} × *Datos finales*) + *Media original.*

En la figura 7 se puede observar cómo, a partir de la reducción de vectores, se pueden recuperar los datos. Compárese con la figura 4.

En conclusión, el análisis de componentes principales es un procedimiento matemático que usa una transformación ortogonal para convertir una muestra o conjunto de variables, potencialmente correlacionadas en un conjunto de valores de variables pobremente correlacionadas y denominadas componentes principales. La ventaja es que, en general, el número de componentes principales es menor que el número de variables originales. En síntesis, se trata de "aislar" las variables que no correlacionan y potencialmente no se influencian.

BIBLIOGRAFÍA

- Anton H. Elementary linear algebra: applications version. 10th ed. Hoboken, NJ: Wiley; 2010.
- Grané A. Análisis de componentes principales [Internet]. Madrid: Universidad Carlos III de Madrid. Departamento de Estadística; [s.d.] [citado: 05/09/2010]. Disponible en: http://halweb.uc3m.es/esp/Personal/personas/agrane/ficheros_docencia/MULTIVARIANT/slides_comp_reducido.pdf
- Jolliffe T. Principal component analysis. 2nd ed. New York: Springer; 2002. (Springer Series in Statistics).
- Smith L. A tutorial of principal components analysis [Internet]. [Citado: 05/09/2010]. Disponible en: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf