PIMAD: UN SISTEMA PARA EFECTUAR ANÁLISIS EN COMPONENTES PRINCIPALES

Oldemar Rodríguez Rojas*

En este artículo se presenta una de las técnicas más importantes de la estadística matemática de orientación francesa, el Análisis en Componentes Principales (ACP). Además se ofrece el sistema computacional PIMAD desarrollado por el autor el cual permite efectuar ACP.

INTRODUCCIÓN

El Análisis en Componentes Principales (ACP) es la técnica básica de un conjunto amplio de métodos de la estadística matemática francesa conocidos como *Análisis de Datos,* entre los que se incluyen: el Análisis Canónico (AC), Clasificación Automática (CA), Análisis Factorial Múltiple (AFM) y otros métodos.

En este artículo se hace una presentación geométrica del ACP, en contraposición a la presentación clásica en la que el ACP es presentado como una consecuencia del álgebra lineal [Pages76].

En la última sección se describe mediante un ejemplo el funcionamiento del sistema PIMAD desarrollado por el autor para efectuar análisis de datos mediante el ACP. Este sistema fue desarrollado en Borland C++ para el ambiente Windows, una descripción detallada se puede encontrar en [Rodri94].

LA NATURALEZA DE LOS DATOS, CONCEPTOS FUNDAMENTALES

Se hace una distinción entre dos conjuntos de datos: *individuos* y *caracteres (variables)* referidas a esos individuos. El término "individuo" se refiere a los "objetos" en estudio: empleados, clientes, animales, planetas, etc. El conjunto de individuos es por lo general una muestra de una población entera.

Los caracteres o variables son observaciones *cualitativas* o *cuantitativas* del individuo, por ejemplo, para un empleado, los caracteres pueden ser el sexo, la profesión, el salario y la edad. Los caracteres cuantitativos son numéricos, por ejemplo, el salario y la edad. Por su parte, los caracteres cualitativos no son numéricos, por ejemplo, su sexo o profesión. Los caracteres cualitativos se representan por medio de *modalidades;* por ejemplo, para el sexo sus modalidades son: masculino y femenino.

Los datos son representados en una tabla de individuos y caracteres denominada *Tabla individuos×caracteres* [Diday82], la cual se presenta en la figura 1. Los individuos se numeran de 1 a n, mientras que los caracteres se denotan en general por $X^1, X^2, \ldots, X^p$. Así x_i^j es el valor tomado por el individuo i en el caracter X^j

Caracteres

	X^1	X^2	...	X^j	...	X^p
1	x_1^1	x_1^2	...	x_1^j	...	x_1^p
2	x_2^1	x_2^2	...	x_2^j	...	x_2^p
...	...	...	...	...	...	...
i	x_i^1	x_i^2	...	x_i^j	...	x_i^p
...	...	...	...	...	...	...
n	x_n^1	x_n^2	...	x_n^j	...	x_n^p

Figura 1. Tabla individuosXcaracteres

* M.Sc. en computación. Profesor en la Escuela de Matemática e investigador del Programa de Investigación en Modelos y Análisis de Datos de la Universidad de Costa Rica (PIMAD).

En la figura 2 se presenta una tabla de individuos×caracteres con n individuos y 4 caracteres, a saber, la edad, el sueldo, el peso y la estatura. Así, se tiene que X^1 = edad, X^2 = sueldo, X^3 = peso, X^4 = estatura. Por lo tanto, x_i^2 es el sueldo del individuo *i-ésimo.*

Caracteres

Individuos	Edad	Sueldo	Peso	Estatura
1	x_1^1	x_1^2	x_1^3	x_1^4
2	x_2^1	x_2^2	x_2^3	x_2^4
...	...	...	...	...
i	x_i^1	x_i^2	x_i^3	x_i^4
...	...	...	...	...
n	x_n^1	x_n^2	x_n^3	x_n^4

Figura 2. Ejemplo de una tabla individuos×caracteres

Para el caso en que X y Y son variables (caracteres) cuantitativos, presentamos algunas definiciones de la estadística clásica que serán útiles posteriormente; además, en posteriores secciones se generalizan estas definiciones.

Sea $p_i \in \mathbb{R}$, el conjunto de los números reales, para $i = 1, 2, \ldots, n$, tal que $\sum_{i=1}^{n} p_i = 1$, (los p_i se denominan los pesos), entonces la *media aritmética* [Pages76] se define como:

$$\bar{X} = \sum_{i=1}^{n} p_i x_i .$$

Por su parte, la *varianza* S de X se define como [Pages76]:

$$S^2 = var(X) = \sum_{i=1}^{n} p_i (x_i - \bar{X})^2 ,$$

se de denota además: $\sigma_x = \sqrt{var(X)}$.

La *covarianza* entre dos caracteres X y Y se define como [Pages76]:

$$S_{XY} = cov(X, Y) = \sum_{i=1}^{n} p_i (x_i - \bar{X})(y_i - \bar{Y}).$$

El *coeficiente de correlación* entre dos caracteres X y Y se define como [Pages76]:

$$r = cor(X, Y) = \frac{cov(X, Y)}{\sigma_x \sigma_y}.$$

Presentación del Análisis en Componentes Principales (ACP)

Suponga que tenemos dos caracteres X^1 y X^2; la representación gráfica de éstos es fácil, y se puede hacer en un plano, pues, dado un individuo e_i este tiene coordenadas x_i^1 y x_i^2. Así, representando el conjunto de individuos en el plano podemos analizar intensidad de la relación entre X^1 y X^2 por medio de un simple estudio visual, como se muestra en la figura 3.

Si se tienen tres caracteres todavía es posible hacer un análisis geométrico, sin embargo, cuando se tienen p caracteres con

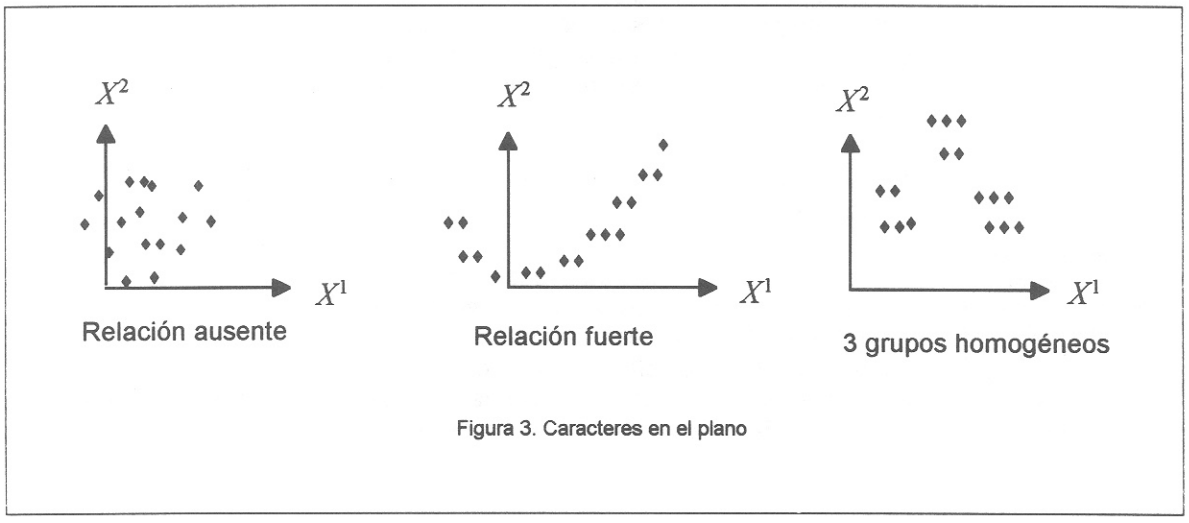

Figura 3. Caracteres en el plano

$p \geq 4$, como es lo usual, el análisis es muy complejo, al menos en forma visual.

Por ejemplo, supongamos que se tienen 24 individuos los cuales son descritos por 11 caracteres, entonces, se dice que los 24 individuos forman una "nube" en $\mathbb{R}^{11}$. La idea central es graficar éstos 24 individuos en un plano; esta representación por supuesto deformará en un cierto grado la representación original, sin embargo, se busca una distorsión mínima.

Geométricamente, el gráfico se obtendrá proyectando los individuos (puntos de $\mathbb{R}^p$) $e_1, e_2, \ldots, e_n$ en un plano como se observa en la figura 4.

Si f_i es la proyección de e_i en el plano, entonces se tiene que $d(f_i, f_j) \leq d(e_i, e_j)$, pero se busca que esta disminución sea mínima.

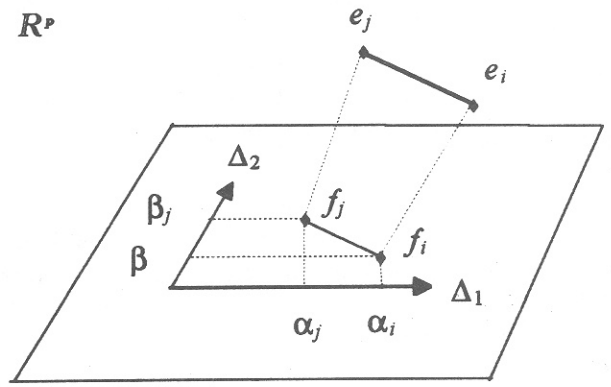

Figura 4. Proyecciones en el plan principal

Para determinar en qué plano se proyectan los e_i, se deben "escoger" dos rectas Δ_1 y Δ_2. Si éstas rectas son perpendiculares, entonces se tiene:

$$d^2(f_i, f_j) = d^2(\alpha_i, \alpha_j) + d^2(\beta_i, \beta_j),$$

donde α_i y β_i son las proyecciones de los f_i en Δ_1 y Δ_2 respectivamente.

El método consiste en buscar Δ_1 para el cual el promedio de las distancias al cuadrado $d^2(\alpha_i, \alpha_j)$ sea máximo. Ahora, como Δ_1 y Δ_2 son perpendiculares, entonces el promedio de las distancias $d^2(\beta_i, \beta_j)$ también será máximo.

Continuando de este modo, se pueden encontrar $\Delta_3, \Delta_4, \ldots, \Delta_p$ perpendiculares dos a dos. Los Δ_i se llaman *ejes principales de la nube*.

La proyección de $e_i = (x_i^1, x_i^2, \ldots, x_i^p)$ sobre los ejes principales permiten obtener un vector de coordenadas $(c_i^1, c_i^2, \ldots, c_i^p)$ que es el vector de coordenadas de e_i en los nuevos ejes. Así, se puede construir un nuevo conjunto de caracteres $C^1, C^2, \ldots, C^p$ denominados *componentes principales*, en donde $C^k = (c_1^k, c_2^k, \ldots, c_n^k) \in \mathbb{R}^n$. Es decir, se forma una nueva tabla de "individuos-caracteres" en donde las filas son los individuos e_i con $i = 1, 2, \ldots, n$ y las columnas son los nuevos caracteres C^j para $j = 1, 2, \ldots, p$.

El esquema denominado *Análisis en Componentes Principales* (ACP), consiste básicamente en la reducción del número de caracteres, con el fin de lograr interpretaciones geométricas de esos caracteres. Esta reducción no será posible si los p caracteres iniciales no son independientes y los coeficientes de correlación son no nulos.

Geometría de los caracteres y de los individuos.

Se define el *centro de gravedad* [Bouro80] de una nube de individuos como:

$$g = (\overline{X^1}, \overline{X^2}, \ldots, \overline{X^p}),$$

es decir, un vector de $\mathbb{R}^p$, cuya i-ésima entrada es la media aritmética del caracter X^i.

Las varianzas y las covarianzas son agrupadas en una tabla (matriz) V llamada *matriz de varianza covarianzas* [Bouro80] de los p caracteres, cuya entrada (i,j) es la covarianza entre el caracter X^i y el caracter X^j es decir:

$$V = \begin{pmatrix} S_1^2 & S_{12} & \cdots & S_{1p} \\ & S_2^2 & \cdots & S_{2p} \\ & & \ddots & \vdots \\ & & & S_p^2 \end{pmatrix}.$$

De la misma forma los coeficientes de correlación se agrupan en una matriz R, llamada *matriz de correlación* [Bouro80] cuya entrada (i,j) es la correlación $r_{ij} = r(X^i, X^j)$ entre el caracter X^i y el caracter X^j, o sea:

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ & 1 & \cdots & r_{2p} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix}.$$

Claramente tanto R como V son matrices simétricas, por esta razón se acostumbra no escribir la parte inferior de la matriz.

Se define también la matriz $D_{1/S}$ como la siguiente matriz diagonal [Bouro80]:

$$D_{1/S} = \begin{pmatrix} 1/S_1 & 0 & \cdots & 0 \\ 0 & 1/S_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/S_p \end{pmatrix}.$$

Se deduce fácilmente la siguiente relación matricial:

$$R = D_{1/S} V D_{1/S}.$$

Generalmente, para facilitar los cálculos se hace un cambio de variable, y se estudian los caracteres centrados, es decir, $X^j - \overline{X^j}$, de modo tal que la nube de individuos tiene como origen el centro de gravedad g.

Sea X la tabla de datos de n líneas y p columnas de los datos centrados, y sea D la siguiente matriz, denominada *matriz de pesos* [Bouro80]:

$$D = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n \end{pmatrix},$$

entonces se tiene la relación matricial $V = X^t D X$.

Como se ha visto, cada individuo e_i se considera como un punto o un vector de $\mathbb{R}^p$, es decir:

$$e_i = (x_i^1, x_i^2, \ldots, x_i^p),$$

en donde x_i^k es el valor tomado por el individuo i en el caracter k.

Es fundamental para efectuar los análisis estadísticos definir una métrica sobre el espacio de individuos. Usando el teorema de Pitágoras, y suponiendo que los ejes generados por los X^j son perpendiculares, la distancia se escribe como:

$$d^2(e_1, e_2) = a_1(x_1^1 - x_2^1)^2 + \cdots + a_p(x_1^p - x_2^p)$$

Sin embargo, si los ejes no son perpendiculares, sino que forman un ángulo θ, la fórmula en general se expresa como sigue:

$$d^2(e_1, e_2) = \sum_{k=1}^{p} \sum_{j=1}^{p} m_{kj}(x_1^k - x_2^k)(x_1^j - x_2^j).$$

Se denota por M la matriz cuya entrada (k,j) es m_{kj} entonces:

$$d^2(e_1, e_2) = (e_1 - e_2)^t M (e_1 - e_2),$$

la matriz M se conoce como *matriz de métricas* [Bouro80].

Obsérvese que si se usa la métrica euclídea, es decir, si todos los ejes son perpendiculares, entonces M es la matriz identidad $I_{p\times p}$.

Además, se puede probar que M es una matriz simétrica y definida positiva. De donde, la métrica M induce un producto escalar sobre el espacio de individuos, definido por:

$$\langle e_i, e_j \rangle_M = e_i^t M e_j.$$

Se denota por $\|e_i\|_M^2 = \langle e_i, e_i \rangle_M$ y se denomina la *longitud del vector* (individuo) con la M-norma.

La métrica más utilizada en el ACP es la siguiente:

$$M = D_{1/S^2} = \begin{pmatrix} 1/S_1^2 & 0 & \cdots & 0 \\ 0 & 1/S_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/S_p^2 \end{pmatrix}.$$

Esta métrica tiene la ventaja que la distancia entre los individuos no dependerá de la unidad de medida que se utilice.

Se puede probar que para toda matriz simétrica definida positiva M existe al menos

una matriz T tal que $M = T^t T$. Usando esta relación puede probar que:

$$\langle e_i, e_j \rangle_M = \langle Te_i; Te_j \rangle_I$$

Donde $\langle x, y \rangle_I$ es el producto interno inducido por la matriz identidad, es decir, es el producto punto usual de $\mathbb{R}^p$. Por esta razón es conveniente efectuar el cambio de variable $Y = XT^t$ para poder utilizar la métrica inducida por la matriz identidad $I_{p\times p}$.

Un paso fundamental en el ACP es cómo se calculan las coordenadas de un individuo e_i en el nuevo sistema de ejes? Para responder a esto, supongamos que los caracteres iniciales forman un conjunto ortonormal.

Dado un eje Δ la proyección de todos los individuos sobre este eje generan un nuevo caracter $C = (c_1, c_2, \ldots, c_n)$ donde c_i es la longitud algebraica de la proyección e_i sobre Δ.

Sea *a* un vector unitario, es decir, de M-norma 1, que genera Δ, entonces c_1 es la longitud de la proyección de e_1 sobre el eje Δ generado por *a*, de donde usando resultados del álgebra lineal se tiene:

$$c_1 = \left\| \frac{\langle e_1; a \rangle_M}{\|a\|_M} a \right\|_M = \langle e_1; a \rangle_M = e_1^t M a,$$

y usando el hecho de que la matriz M es simétrica se tiene:

$$c_1 = (Ma)^t e_1.$$

Si denotamos por $u = Ma$, entonces:

$$c_1 = u^t e_1 = \sum_{j=1}^{p} u_j x_1^j = \sum_{j=1}^{p} x_1^j u_j,$$

así, el caracter C con la nuevas n coordenadas $c_1, c_2, \ldots, c_n$ se puede obtener de la fórmula:

$$C = Xu,$$

donde X es la tabla de datos inicial.

Para medir la dispersión global de una nube de puntos se utiliza una generalización de la varianza, llamada *inercia*, que se define como el promedio de los cuadrados de las distancias de los n puntos (individuos) al centro de gravedad g, es decir:

$$\mathcal{I} = \sum_{i=1}^{n} p_i \|e_i\|_M^2 = \sum_{i=1}^{n} p_i e_i^t M e_i.$$

El ***plano principal*** de la nube de puntos se interpreta como aquel que permita inercia máxima en el conjunto de las n proyecciones de los e_i sobre este plano. Se puede probar que buscar un plano con inercia máxima en las proyecciones de los n puntos es equivalente a buscar un plano para el cual el promedio de las distancias al cuadrado de los puntos de la nube al plano sea mínima, es decir el plano con inercia mínima.

El otro espacio involucrado en el ACP es el "espacio de caracteres", un caracter es considerado como un vector de $\mathbb{R}^n$, donde la entrada i es la "respuesta" del individuo e_i al caracter. Para medir la proximidad entre los caracteres es importante proveer a $\mathbb{R}^n$ de una métrica, usualmente se escoge la *métrica de pesos* dada por la matriz:

$$D = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n \end{pmatrix},$$

D es una matriz simétrica definida positiva. Así, el producto escalar entre el caracter X^j y X^k está dado por:

$$\langle X^j; X^k \rangle_D = (X^j)^t D X^k = \sum_{i=1}^{n} p_i X_i^j X_i^k,$$

que no es otra cosa que la covarianza entre los caracteres X^j y X^k si se asume que estos fueron previamente centrados. Se deduce también que $\|X^j\|_D^2 = S_j^2$, o sea que la longitud de un caracter es el tipo de desviación de éste.

Del álgebra lineal se sabe que el ángulo θ entre dos vectores está dado por la relación:

$$\cos(\theta) = \frac{\langle X^j; X^k \rangle_D}{\|X^j\|_D \|X^k\|_D} = \frac{S_{jk}}{S_j S_k} = r_{jk},$$

es decir, el coseno del ángulo entre dos caracteres es el coeficiente de correlación entre estos, se deduce entonces la propiedad conocida del coeficiente de correlación $-1 \leq r_{ik} \leq 1$.

En el espacio de los individuos interesa la distancia entre los puntos (individuos), mientras que en el espacio de los caracteres interesa el ángulo entre estos.

Si $X^1, X^2, \ldots, X^p$ son p caracteres medidos a los n individuos entonces se pueden generar nuevos caracteres mediante combinaciones lineales de éstos $C = u_1X^1 + u_2X^2 + \cdots + u_pX^p$. Como ya se estudió previamente, esto permite escoger nuevos ejes en los cuales podrán ser representados los individuos.

Del álgebra lineal se sabe también que el conjunto de todas éstas combinaciones lineales forman un subespacio vectorial del espacio de caracteres, que denominaremos por W. La dimensión de W es q, con $q \leq p$.

Búsqueda de los factores, los componentes y los ejes principales

El eje principal Δ_1, como ya se dijo, debe maximizar el promedio de los cuadrados de las distancias de las proyecciones de la nube de puntos. Esto es equivalente a maximizar la inercia de las proyecciones dada por $\sum_{i=1}^{n} p_i c_i^2$, donde los c_i son las medidas algebraicas de las proyecciones de los individuos e_i en Δ_1. Además, tal como se dijo en la sección anterior Δ_1 se debe escoger de modo tal que pase por el centro de gravedad de la nube y de modo que los c_i sean los más grandes posibles, es decir, que se debe escoger $C = (c_1, c_2, \ldots, c_n)$ como una combinación lineal de los X^j, $j = 1, 2, \ldots, p$ con varianza máxima.

Tómese $M = I_{p\times p}$, es decir, supóngase que la tabla de datos fue centrada con el cambio de variable $Y = XT^t$, donde $M = T^tT$

Si C es el componente principal que se describió en la sección anterior, entonces $C = Yv$, pues, sabemos que $C = Xu$, con $u = T^tv$.

Sea V_Y la matriz de varianza asociada a la tabla Y, entonces se puede probar que:

$$V_Y = TVT^t,$$

donde V es la matriz de varianzas de X.

La componente principal C debe maximizar la varianza v^tV_yv, además el vector v debe tener norma 1 para generar el primer eje principal. Ahora, el que v^tV_yv sea máximo es equivalente a que el cociente $\frac{1}{v^tv}(v^tV_yv)$ sea máximo. Esto es máximo cuando las derivadas de las p componentes son cero. De las fórmulas de derivación se deduce que la derivada de este cociente es nulo si:

$$2(v^tv)V_yv - 2(v^tV_yv)v = 0.$$

o sea si:

$$V_yv = (v^tV_yv)v = \lambda v.$$

es decir, el cociente es cero si v es el vector propio asociado a la matriz V_y, con valor propio asociado λ.

Como la matriz de varianza es simétrica y definida semi-positiva, se sabe del álgebra lineal que tiene p vectores propios dos a dos ortogonales, y los valores propios son positivos o cero.

Lo anterior se puede resumir como sigue: si $M = I$ los ejes principales $v_1, v_2, \ldots, v_p$, son los vectores propios de la matriz de varianzas, con valores propios asociados $\lambda_1, \lambda_2, \ldots, \lambda_p$. Además, se puede probar [Pages76] que la matriz de varianzas de los ejes principales está dada por:

$$V_c = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}.$$

El ACP reemplaza los p caracteres iniciales por los caracteres no correlacionados de varianza máxima encontrados anteriormente.

Para encontrar los factores y los componentes principales directamente de X considere lo siguiente: $V_yv = \lambda v = TVT^tv$, y

multiplicando por T^t se tiene que $T^tTVT^tv = \lambda T^tv$, es decir, $MVu = \lambda u$ con $u = T^tv$, pero el eje a es tal que $u = Ma$ de donde $MVMa = \lambda Ma$, así suponiendo que si M es invertible se tiene que:

$$VMa = \lambda a.$$

Se puede concluir que los ejes principales son también los vectores propios de VM con valor propio asociado λ.

Sin embargo, se puede probar que los valores y vectores propios de VM son los mismos que los de la matriz de correlaciones R, por lo que en la práctica es a partir de esta matriz que se obtienen los ejes principales.

Se había probado que las componentes principales satisfacen la relación $C = Xu$, además como $MVu = \lambda u$ y $V = X^tDX$ se tiene que $MX^tDXu = \lambda u$, luego multiplicando por X a ambos lados de la igualdad se tiene que $XMX^tDXu = \lambda Xu$, lo cual implica que:

$$XMX^tDC = \lambda C.$$

de donde la componente principal C es el vector propio asociado a la matriz XMX^tD, con valor propio asociado λ.

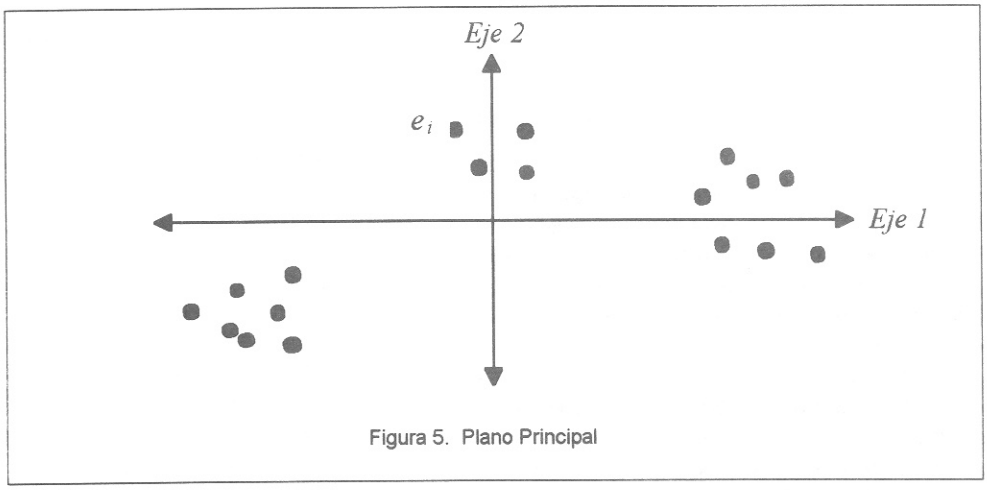

Figura 5. Plano Principal

Se puede probar [Rodri94] que la inercia total satisface $\mathscr{I}$ =traza(MV), se puede probar además que la traza de MV y de V_Y está dada por la suma $\lambda_1 + \lambda_2 + \cdots + \lambda_p$, es decir, la suma de los valores propios mide la inercia total de la nube [Pages76]. Así, el cociente $\lambda_k / \mathscr{I}$ mide la *inercia explicada* por el k-ésimo eje. Por ejemplo, el cociente:

$$\frac{\lambda_1 + \lambda_2}{\mathscr{I}}$$

mide la inercia explicada por los dos primeros ejes principales.

Interpretación de Resultados

Una vez calculados los ejes principales (vectores propios de la matriz de correlación) se calcula la inercia explicada por cada uno de ellos, luego, se grafican los planos principales generados por estos ejes y se grafican las proyecciones de los individuos en estos planos. Estos planos constituyen verdaderas "radiografías" del fenómeno que se está analizando, y deben por lo tanto ser interpretados por expertos. En el caso del estudio de la opinión pública, por ejemplo, los resultados son interpretados por sociólogos, politólogos o economistas.

Por ejemplo, en la figura 5 se presenta un plano principal generado por los dos primeros vectores propios de la matriz R, y se observan tres agrupaciones de individuos, que podrían tener ciertas interpretaciones, así también los ejes principales son interpretados por los expertos.

Otra herramienta estadística importante es el *círculo de correlación* [Pages76], para graficarlo se calculan las correlaciones entre los caracteres originales X^j y los componentes principales C^k. Así, por ejemplo, para graficar el círculo de correlación asociado a los dos primeros componentes principales se deben calcular $r(C^1, X^j)$ y $r(C^2, X^j)$ para $j = 1, 2, \ldots, p$, y se grafican los pares ordenados $(r(C^1, X^j), r(C^2, X^j))$. Este círculo de correlación se muestra en la figura 6 y

corresponde por "dualidad" al plano principal de la figura 5 generado por los dos primeros vectores propios de R.

Además, se puede probar que $r^2(C^1,X^j)+r^2(C^2,X^j)\leq 1$, debido a esto todos los puntos se ubican dentro del círculo de radio 1.

Se puede probar que si se usa la métrica D_{1/S^2} entonces $r(C^k,X^j)$ es igual a la j-ésima entrada del k-ésimo vector propio v_k multiplicado por $\sqrt{\lambda_k}$ [Pages76].

En el círculo de correlación se interpreta a "grosso modo" que dos caracteres cuyos puntos son muy próximos tienen el mismo significado o están muy correlacionados.

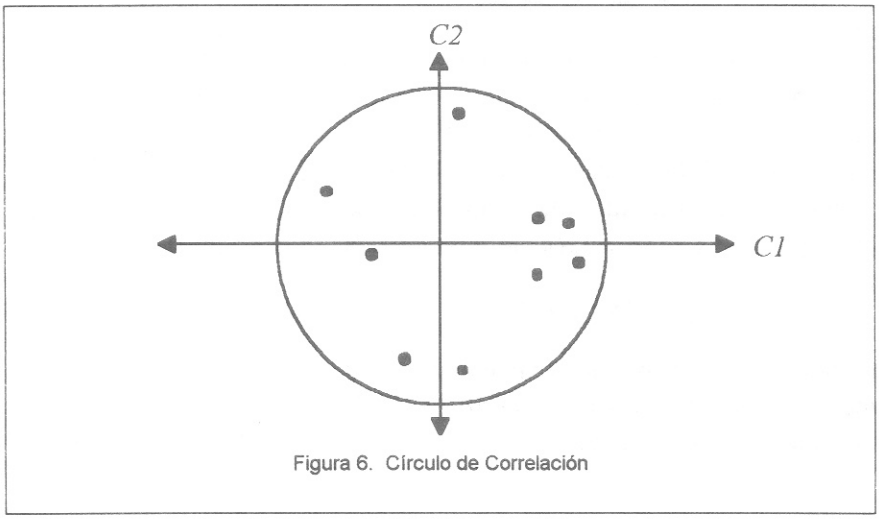

Figura 6. Círculo de Correlación

Un Ejemplo Mediante el Sistema PIMAD

En esta sección presentaremos un ejemplo de análisis estadísticos mediante el Análisis en Componentes (ACP). Este ejemplo tiene un doble propósito, por una parte ilustrar el método y por otra ilustrar la utilización del sistema PIMAD.

El sistema de cómputo PIMAD se encuentra actualmente en desarrollo en el Programa de Investigaciones en Modelos y Análisis de Datos de la U.C.R., el núcleo central de este es el ACP y fue desarrollado por el autor. Actualmente están en proceso de diseño y desarrollo otra serie de módulos que permiten efectuar análisis estadísticos mediante otras técnicas de orientación francesa, como son: el Análisis Factorial Múltiple (AFM) mediante el método "STATIS", las tablas cruzadas, métodos para el "pegue" o análisis simultáneo de encuestas, entre otros.

El sistema PIMAD ha sido y es desarrollado como un sistema Orientado a Objetos de principio a fin, es decir tanto el análisis, el diseño como la implementación del sistema se han hecho con métodos Orientados a Objetos. La implementación se desarrolla en lenguaje C++ en la versión 3.1 de Borland para ambiente Windows.

Si el lector desea mayores detalles sobre el desarrollo del sistema PIMAD puede consultar [Rodri94], aquí se pueden ver con detalle los algoritmos numéricos que el autor implementó, el análisis y diseño del sistema, así como el código C++ de este.

En [Gonva93] se presenta un ejemplo completo de aplicación del ACP a estudio de la calidad del agua de algunos ríos en Costa Rica, una reseña de este puede consultarse también en [Rodri94].

En este artículo presentaremos un ejemplo que pretende analizar la producción agrícola de Costa Rica de hace 100 años, y mediante estas variables estudiar el grado de desarrollo de los diferentes cantones de nuestro país en esa época. Los datos de este ejemplo fueron recopilados por el Centro de Investigaciones Históricas de la Universidad de Costa Rica y tomados por el autor del artículo [Casti91].

Debido al doble propósito de este ejemplo presentaremos las salidas del sistema mediante la captura de la pantalla completa, no obstante el sistema puede imprimir los gráficos directamente. Una guía completa de como usar PIMAD puede encontrarse en la ayuda del sistema o en [Rodri94].

En la tabla de la figura 7 se presentan los datos del ejemplo, aquí los individuos del estudio son cantones de Costa Rica y los caracteres o variables estadísticas son las cantidades de producción en los diferentes productos agrícolas de la época.

Producción Agropecuaria por Cantones en Costa Rica
AÑO 1892

(Recopilados por el Centro de Investigaciones Históricas de la UCR)

	Variables Agropecuarias (Caracteres)						
Cantón (Individuos)	**Café (kilos)**	**Dulce (kilos)**	**Maíz (litros)**	**Frijoles (litros)**	**Arroz (litros)**	**Ganado vacuno**	**Ganado caballar**
Santa Cruz	0	18032	462892	6221	4292	15823	2592
Santo Domingo	1509260	0	178895	1101	0	6927	1427
Cañas	0	0	35823	7280	0	6927	1023
Puriscal	11638	603750	732597	199482	18462	4932	2712
Limón	0	8510	0	0	0	9099	268
Barva	542432	183908	202626	6325	0	3428	527
La Unión	749570	0	216423	23645	0	2428	981
Santa Barbara	147246	102488	282321	692	0	0	503
San Rafael	207414	0	462027	9896	0	7228	496
Aserrí	32246	3727	523309	61328	9325	7028	3125
Nicoya	0	10258	382745	19693	21890	13927	2628
Tarrazú	5112	17572	182400	7275	0	2809	1282
Palmares	1057908	134550	823692	73328	0	2987	589
Atenas	47242	143888	182425	109647	135620	4720	1482
Puntarenas	0	37720	634745	25425	165850	9667	1721
Esparta	0	42458	255421	47329	445725	0	0
San Mateo	8694	33166	252328	18029	232925	6002	1909
Mora	8786	114632	289342	112428	109432	3245	1720
Paraiso	229310	266708	2382621	945321	19897	21120	3892
Grecia	505862	1870042	1006548	62625	12325	11321	2728
Cartago	392334	64032	2147625	108725	0	25007	5027
Escazú	705088	951602	2328615	308625	0	4461	1480
San Ramón	254242	870688	2682491	262325	78693	10125	3212
Alajuela	1221110	806462	2123452	298897	228625	21328	5128
Heredia	1697078	165370	1492325	245682	0	14892	3427
San José	4223904	10350	749818	57325	0	21782	4203
Naranjo	1854766	171396	1125489	83308	7887	5927	1726
Desamparados	1317670	611110	375896	162324	1896	7627	3020
Liberia	0	78062	369227	25647	65728	77892	143929

Figura 7. Tabla de Datos X.

El logotipo de PIMAD se presenta en la figura 8. Se puede ver claramente que el sistema tiene una interfaz que sigue el estandar de Windows, tiene además una "barra de iconos" (tool bar) que permite obtener resultados rápidamente.

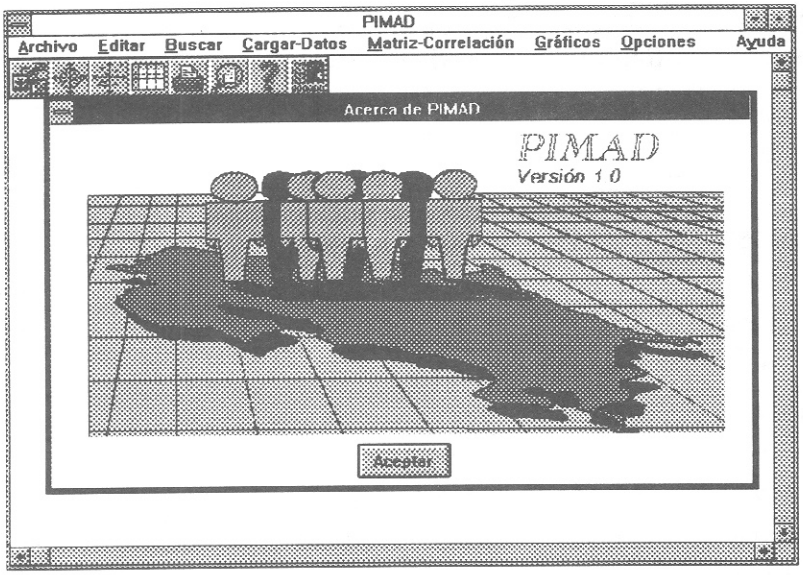

Figura 8. Logotipo de PIMAD

El sistema PIMAD recibe los datos mediante archivos tipo ASCII, la tabla de datos se introduce mediante un archivo extensión DAT, en la primera fila de este se deben indicar cuántos individuos y cuántas variables conforman el estudio. Las "etiquetas" de los individuos (la hilera de caracteres que lo identifica) y de las variables se introducen en dos archivos extensión ETI y ETV respectivamente, en la primera fila de este archivo se debe indicar cuántas etiquetas contiene.

PIMAD tiene incorporado un editor de textos que permite editar la tabla de datos y los archivos de etiquetas en caso de que el número de datos no sea muy grande, no obstante, los datos pueden provenir de cualquier otro programa que produzca archivos de tipo ASCII, las tres primeras opciones del menú principal se dedican a esto.

La opción **Cargar-Datos** permite colocar los datos en memoria principal para que el sistema pueda efectuar los cálculos necesarios y también permite cargar los archivos de etiquetas. Esto se puede efectuar en forma rápida mediante el primer icono de la barra de iconos.

Una vez cargados los datos se pueden calcular la matriz de correlaciones, los valores y vectores propios mediante la opción **Matriz-Correlación**.

Mediante La opción **Gráficos** se pueden calcular las principales correlaciones entre las variables originales y los componentes principales para luego graficar el círculo de correlaciones. También se pueden calcular las componentes principales para generar luego el plano principal.

Por ejemplo, los datos de tabla de la figura 7 se encuentran en archivo denominado EJEMPLO.DAT, las etiquetas de los individuos en el archivo denominado EJEMPLO.ETI y las etiquetas de las variables en el archivo EJEMPLO.ETV. Una vez cargados estos archivos con las opción **Cargar** se deben efectuar los cálculos necesarios como son la matriz de correlaciones, los vectores y valores propios de esta matriz, las principales correlaciones y los componentes principales para luego generar los gráficos.

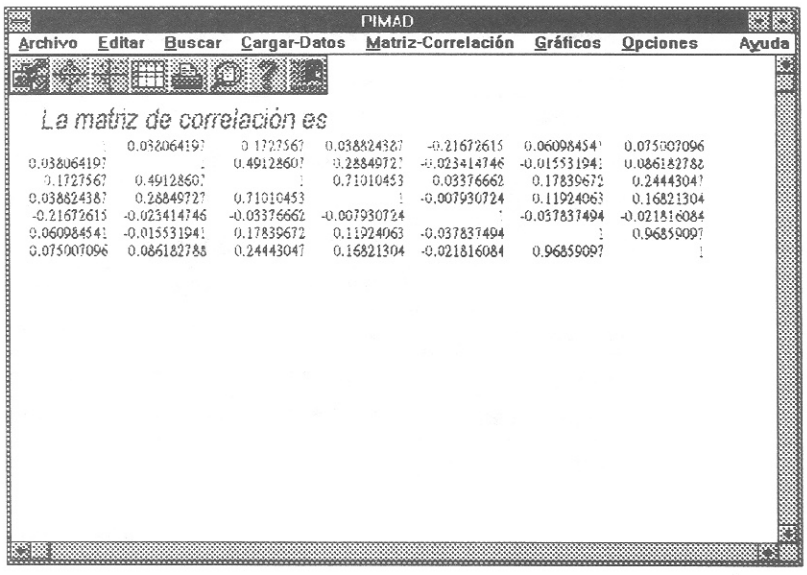

Figura 9. Matriz de correlaciones

Todos estos cálculos son desplegados por pantalla, así por ejemplo en la figura 9 se muestra la matriz de correlaciones.

Una vez efectuados los cálculos necesarios mediante al opción **Gráficos** se pueden generar el círculo de correlaciones y el plano principal. En la figura 10 se presenta el círculo de correlaciones correspondiente a los datos de la tabla de la figura 7. Aquí se identifican claramente tres grupos de variables o caracteres, el primero formado por las variables *frijoles, maíz* y *dulce,* el segundo formado por las variables *café* y *arroz,* y un tercer grupo por las variables *cabezas de ganado caballar* y *cabezas de ganado vacuno.* Los grupos indican cuáles variables estaban altamente correlacionadas hace un siglo, a mayor cercanía mayor correlación entre estas.

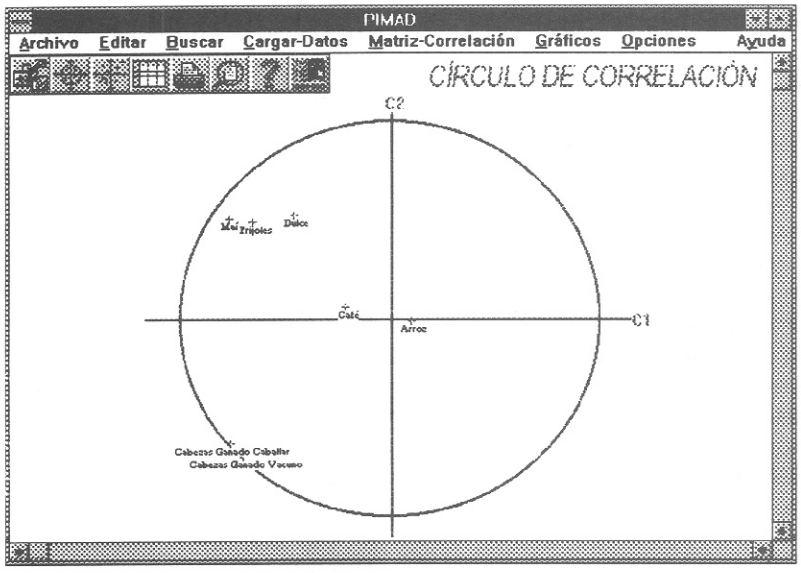

Figura 10. Círculo de correlaciones

En la figura 11 se presenta el plano principal asociado a los datos de la tabla 7, en este se representan gráficamente los

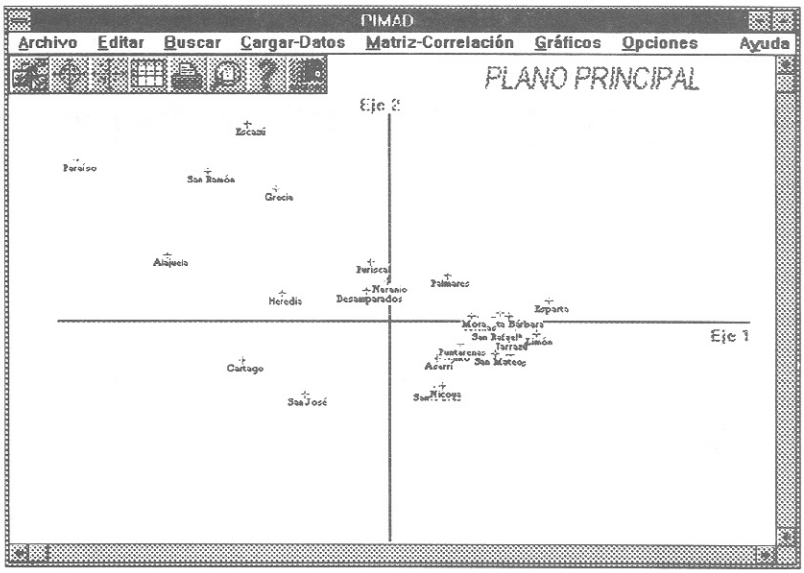

Figura 11. Plano Principal

individuos del estudio, en este ejemplo, los cantones de Costa Rica en 1892. Aquí se observa un grupo de variables en el segundo cuadrante cercano al origen del sistema conformado por los cantones de *Grecia, Heredia, Puriscal, Naranjo, Desamparados* y *Palmares* que eran los cantones de gran productividad de café en aquellos años, la ubicación de este grupo de individuos coincide con la ubicación de la variable *café* en el círculo de correlaciones de la figura 10, en este sentido se dice que el círculo de correlaciones de la figura 10 corresponde por **dualidad** al plano principal de la figura 11. Algebraicamente se puede demostrar que uno el dual algebraico del otro. Se puede observar otro grupo de individuos (cantones) en el tercer cuadrante formado por los cantones centrales de *San José* y *Cartago* que ya en ese entonces comenzaban a ser cantones con características más urbanas.

Se ubica un tercer grupo de cantones en el cuarto cuadrante y parte del primero, formado entre otros por *Esparta, Puntarenas, San Mateo* y *Limón,* es decir los cantones más cercanos a las costas del país. No obstante en el gráfico de la figura 11 no se pueden observar con claridad los cantones de este grupo pues existe sobreposición de etiquetas. Para resolver este problema PIMAD ofrece una lupa en la barra de iconos que permite efectuar "Zoom" en los gráficos, en la figura 12 se presenta un acercamiento

al tercer grupo de cantones que permite distinguir con claridad que cantones lo conforman.

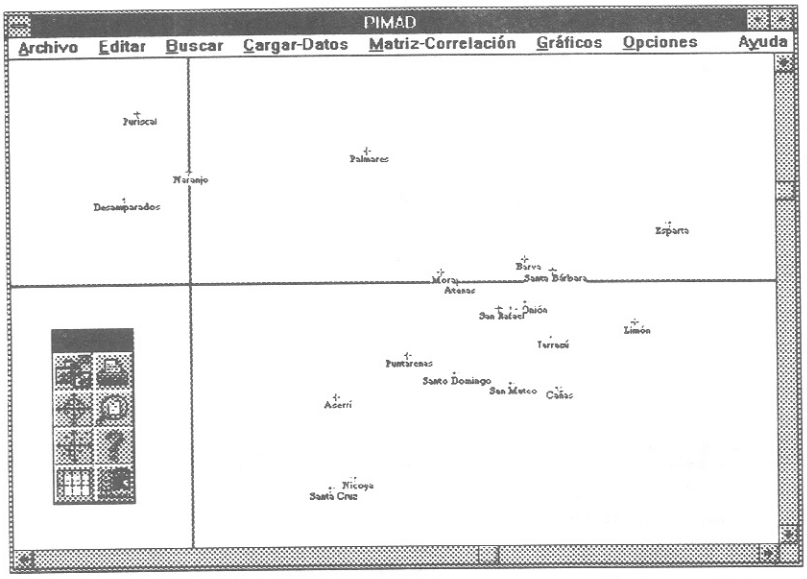

Figura 12. Plano Principal con "Zoom"

PIMAD ofrece otra serie de opciones que permitir escoger que variables o que individuos se desean graficar, también permite escoger en que ejes se desea graficar y con que dirección, para mayor detalle consulte [Rodri94] o la ayuda del sistema.

Actualmente el sistema de cómputo PIMAD se encuentra en crecimiento, y como ya lo habíamos mencionado se está desarrollando un módulo para que el "pegue" de tablas de datos, por ejemplo, como se analizarían los datos si tuviésemos otra tabla de producción agrícola correspondiente a 1994, o si tuviésemos una secuencia de diez tablas de datos correspondientes a cada diez años. Es decir como estudiar la evolución de los datos, para interpretar como está evolucionando la situación en estudio. En una posterior publicación presentaremos una descripción teórica de estos métodos así como el software asociado.

BIBLIOGRAFÍA

[Borla93] Borland International, *Object-Windows® for C++, User's Guide, Borland International Inc., USA, 1993.*

[Casti91] Castillo W. Clasificación automática, *Ciencias Matemáticas U.C.R.* vol 1, No.1, Diciembre 1991.

[Bouro80] Bouroche J. y Saporta G. *Que sais-je? L' analyse des données,* Presses Universitaires de France, París,1980.

[Cavar92] Cavarero A. y González C. Unification of conceptual schemata in a fuzzy object-oriented approch, *Indo-French workshop on object-oriented systems*, Goa, India, november 1992.

[Coad91] Coad P. y Yourdon E. *Object-Oriented analysis,* Yourdon Press,Texas USA, 1991.

[Diday82] Diday E. y otros. *Éléments d' anlyse de données,* Bordas, París 1982.

[Duart91] Duarte M. *Analyse de données symboliques, pyramides d'héritage,* Tesis para obtener el grado de Doctor en Matemáticas, Universidad París IX, París, 1991.

[Gonza90] González G. Vectores y valores propios asociados a un par de matrices, *Ciencias Matemáticas U.C.R.* vol 1, No.1, pp 14-20. Diciembre 1990.

[Gonva93] González J. y Morales V. Análisis multivariado de la calidad del agua, proyecto hidroeléctrico Ventanas Garita, *VI Congreso Latinoaméricano de Biomatemáticas*, UNED, San José 1993.

[Mccor92] McCord J. *Borland C++ tools,* Sams Publishing, Indiana, 1992.

[Pages76] Pages J. *Introduction à L' analyse des données,* SMASH, París 1976.

[Rodri94] Rodríguez O. *Desarrollo Orientado a Objetos: Una Aplicación al Análisis de Datos.* Tesis para optar a grado de master en Computación, I.T.C.R. 1994.