Revista de Matemática: Teoría y Aplicaciones 1999 **6**(1) : 35–50
cimpa – ucr – ccss issn: 1409-2433

ANÁLISIS DISCRIMINANTE DESCRIPTIVO: TEORÍA, ALGORITMO Y SOFTWARE

William Castillo Elizondo*– Oldemar Rodríguez Rojas†

Recibido: 23 Octubre 1998

Resumen

El propósito de este artículo es presentar una implementación computacional del Análisis Discriminante Descriptivo. Para ello hemos desarrollado los aspectos teóricos que nos permiten formular el algoritmo que ha sido programado en C++. Finalmente se ilustra el funcionamiento del software y el método, mediante un ejemplo.

Palabras clave: Análisis discriminante, algoritmo, variable discriminante, grupo a priori, baricentro.

Abstract

The objective of this article is to present a computational implementation of the Descriptive Discriminant Analysis. We have developed some theoretical aspects in order to formulate an algorithm that was implemented using C++. Finally, the method and the software are illustrated by an example.

Keywords: Discrimination, algorithm, discriminant variable, a priori group, baricenter.

AMS Subject Classification: 62-07,62H30,68N99.

1 Introdución

Un problema de Análisis Discriminante (también decimos: un problema de discriminación) se presenta cuando es necesario 'explicar' una variable cualitativa con base en un cierto número de variables cuantitativas llamadas variables explicativas o predictores. Pero es

* CIMPA, Escuela de Matemática, Universidad de Costa Rica, 2060 San José, Costa Rica; Tel.: +(506) 2075574, Fax: +(506) 207 4397; E-Mail: wcastill@cariari.ucr.ac.cr

† CEREMADE, Université de Paris IX-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 París Cedex 16, Francia. E-Mail: orodrigu@ceremade.dauphine.fr

bien conocido que existe una amplia gama de métodos para hacer discriminación en presencia de predictores cualitativos [1]. Uno de estos métodos, importante por sus aplicaciones, es el método 'Disqual' de G. Saporta usado para el cálculo de puntajes (*credit scoring*) [6]. El Análisis Discriminante se puede entender como un conjunto de métodos y procedimientos estadístico-matemáticos orientados a la búsqueda de dos objetivos, que pueden ser complementarios:

1. Determinar si las variables observadas permiten distinguir (discriminar) los r grupos a priori. Este objetivo es de carácter descriptivo y se relaciona con el Análisis en Componentes Principales. Es natural entonces que se dé la mayor importancia a la construcción de representaciones bidimensionales de los individuos, de las variables y de los grupos a priori.

2. Construir reglas de clasificación -reglas decisionales- para asignar cada nuevo objeto a uno de los grupos a priori. Este objetivo es de carácter decisional y su nexo es con los métodos probabilísticos. Esencial a este énfasis es la construcción de reglas de decisión y los procedimientos para su evaluación

Este artículo trata el Análisis Discriminante solo en relación con el primer objetivo. Es decir, desde un punto de vista descriptivo, por eso le llamaremos Análisis Discriminante Descriptivo (ADD). En la sección siguiente se presentarán las definiciones y resultados concernientes al ADD. En la sección 3 se halla un algoritmo a partir del cual se desarrolló la implementación computacional tal como se explica en la sección 4. Por último en la sección 5, se usa un ejemplo para ilustrar la aplicación del software y algunos aspectos metodológicos del ADD.

2 Propiedades y resultados del ADD

Se consideran p variables continuas (variables explicativas) $x^1, \ldots, x^p$ observadas en una muestra E de n individuos. Cada individuo $i \in E$ se identifica con su vector de mediciones en $\mathbb{R}^p$, $x_i = (x_{i1}, \ldots, x_{ip})$ y cada variable x^j con su vector de valores asumidos $x^j = (x_{1j}, x_{2j}, \ldots, x_{nj})^t$. La variable cualitativa Y (a explicar) determina una partición de E, $P = \{E_1, \ldots, E_r\}$; cada clase E_i se llama categoría o grupo a priori.

2.1 Definiciones básicas

Se supondrá que las variables $\left\{x^1, \ldots, x^p\right\} \subset \mathbb{R}^n$ son centradas.

$V = X^t D X$ es la matriz de correlaciones de las p variables $x^1, \ldots, x^p$, donde $D = \operatorname{diag}(p_i)$ es la matriz diagonal de pesos de los individuos y $X = \left[x^1 \cdots x^p\right]$.

$W = \sum_{l=1}^{r} \sum_{i \in E_l} p_i \left(x_i - g_l\right)^t \left(x_i - g_l\right)$ es la matriz de inercia intraclases, donde $g_l = \sum_{i \in E_l} \frac{p_i}{q_l} x_i$ con $q_l = \sum_{i \in E_l} p_i$ el peso de la clase E_l.

$B = \sum_{l=1}^{r} q_l g_l^t g_l$ es la matriz de inercia interclases.

$\text{inter}(x^j) = \sum_{l=1}^{r} q_l g_{jl}^2$ es la varianza interclases de la variable $x^j = (x_{1j}, \ldots, x_{nj})^t$ con respecto a la partición $P = \{E_1, \ldots, E_r\}$, donde g_{jl} es la entrada j de g_l. Por su parte la varianza intraclases de x^j es: $\text{intra}(x^j) = \sum_{l=1}^{r} \sum_{i \in E_l} p_i \left(x_i - g_{jl}\right)^2$.

Son bien conocidas las siguientes relaciones: ([2])

Teorema 1:

1. $B = C_g^t D_q C_g$, donde C_g es la matriz $r \times p$ cuyas filas son $g_1, \ldots, g_r$ y $D_q = \text{diag}(q_l)$.
2. $0 = \sum_{l=1}^{r} q_l g_l$. Es decir, C_g es de rango $\leq r - 1$.
3. $\text{rango}(C_g) = \text{rango}(B) \leq r - 1$ y si $r = 2$, entonces $\text{rango}(B) = 1$.
4. $V = B + W$.
5. $\text{var}(z) = \alpha^t V \alpha = \alpha^t W \alpha + \alpha^t B \alpha$ para todo $z = X\alpha \in \mathbb{R}^n$.
6. $\text{intra}(z) = \alpha^t W \alpha$ e $\text{inter}(z) = \alpha^t B \alpha$.

2.2 Cálculo de las funciones discriminantes: un ACP particular

El objetivo principal del ADD consiste en determinar m variables $z^1, \ldots, z^m$ llamadas funciones discriminantes, que poseen (y son determinadas por) las propiedades siguientes (ver [2]):

1. Cada $z^j \in \mathbb{R}^p$ es una combinación lineal de las p variables. Esto es,

$$z^j = \sum_{s=1}^{p} \alpha_{js} x^s = X\alpha_j,$$

donde $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jp})^t$; $j = 1, \ldots, m$.

2. Las variables z^j son D ortonormadas. Es decir, no correlacionadas y de varianza 1.

3. Los valores de cada variable z^j en los individuos de un mismo grupo, deben ser lo más próximos posible. Es decir, se debe minimizar $\text{intra}(z^j)$ (la varianza intraclases).

4. Los valores de cada variable z^j en los individuos pertenecientes a clases distintas, deben ser lo más diferentes posible. Esto es, se debe maximizar $\text{inter}(z^j)$ (la varianza interclases).

De acuerdo con la ecuación $1 = \text{Var}(z) = \alpha^t V \alpha = \alpha^t B \alpha + \alpha^t W \alpha$ se ve que las propiedades 3. y 4. de las funciones discriminantes son equivalentes:

$$\max \left\{ \alpha^t B \alpha \,\middle|\, \alpha^t V \alpha = 1 \right\} \Leftrightarrow \min \left\{ \alpha^t W \alpha \,\middle|\, \alpha^t V \alpha = 1 \right\}.$$

Diremos que (α_1, λ_1) es la primera solución del problema de máximo (salvo por el signo de α_1) si $\lambda_1 = \alpha_1^t B \alpha_1 = \max\{\alpha^t B\alpha \,|\alpha^t V\alpha = 1\}$. La primera variable discriminante es $z_1 = X\alpha_1$. Las otras variables discriminantes se obtienen resolviendo el problema de máximo secuencialmente, con restricción de $V-$ ortogonalidad sobre α. Esto es, (α_k, λ_k) es la $k-$ésima solución si

$$\lambda_k = \alpha_k^t B \alpha_k \max\left\{\alpha^t B\alpha \,\middle|\, \alpha^t V\alpha = 1,\ \alpha^t V\alpha_s = 0,\ s = 1, \ldots, k-1\right\}.$$

donde los $\alpha_1, \ldots, \alpha_{k-1}$ fueron previamente calculados. Así, λ_k es la inercia interclases de la $k-$ésima variable discriminante. Nótese que $\lambda_k \in [0, 1]$.

Se va a probar que las variables z^j se obtienen a partir de un ACP y que las representaciones bidimensionales de los individuos, de las clases y de las variables son consecuencia de este resultado. Estas representaciones ayudan a verificar si las variables discriminan las clases a priori y si es posible describirlas en términos de las variables originales.

Teorema 2: Sea X de rango p –es decir, V es invertible–. Si $\beta_1, \ldots, \beta_t$ son los vectores propios $V^{-1}-$ ortonormados del ACP del triplete $\left(C_g, V^{-1}, D_q\right)$, entonces las variables discriminantes son $z^j = XV^{-1}\beta_j, j = 1, \ldots, t$.

Prueba: Sea $\beta = V\alpha$ entonces $\|\alpha\|_V = \|\beta\|_{V^{-1}}$ e $\text{inter}(z) = \alpha^t B\alpha = \beta^t V^{-1} B V^{-1}\beta$.

Es claro entonces que

$$\max\left\{\alpha^t B\alpha \,|\|\alpha\|_V = 1, \alpha \in \mathbb{R}^p\right\} = \max\left\{\beta^t V^{-1} B V^{-1}\beta \ \mid\ \|\beta\|_{V^{-1}} = 1,\ \beta \in \mathbb{R}^p\right\}.$$

Por otra parte, $\beta^t V^{-1} B V^{-1}\beta = \beta^t V^{-1}\left(C_g^t D_q C_g\right) V^{-1}\beta$ es la inercia de la nube de baricentros de las clases (es decir, las filas de la matriz C_g), V^{-1} proyectada sobre la recta determinada por β.

De lo anterior es claro que los vectores propios $V^{-1}-$ ortonormados del ACP del triplete $\left(C_g, V^{-1}, D_q\right)$ proveen la solución al problema de máximo. Sean $\beta_1, \ldots, \beta_t$ dichos vectores propios asociados a los valores propios $\lambda_1 \geq \cdots \geq \lambda_t > 0$.

Definiendo $\alpha_j = V^{-1}\beta_j$ resulta que las variables $z^j = X\alpha_j = XV^{-1}\beta_j$, $j = 1, \ldots, t$ son $D-$ ortonormadas y satisfacen las otras propiedades que caracterizan las variables discriminantes[1].

Cada valor propio λ_l se llama *poder discriminante* y el vector propio correspondiente β_l, *eje discriminante*. Los ejes discriminantes son entonces los ejes de máxima inercia de la nube de baricentros. En este sentido decimos que son los ejes que más discriminan los grupos a priori.

El siguiente teorema tiene un interés práctico en la implementación computacional del ADD. Se sabe que el proceso de diagonalización de una matriz representa un esfuerzo computacional significativo. Para lograr más eficiencia en ese proceso se debe procurar diagonalizar siempre una matriz simétrica del menor tamaño posible, aún cuando posteriormente se deban hacer ciertas transformaciones. El teorema 3 nos garantiza que esto siempre es posible en ADD [4].

[1]Mediante un procedimiento similar al empleado en la prueba del teorema 2, se pueden deducir las funciones discriminantes resolviendo el problema de máximo $\max\left\{\frac{u^t B u}{u^t W u} \mid u \neq 0\right\}$, el cual es equivalente a: $\max\left\{v^t B v \,\middle|\, v^t W v = 1\right\}$.

Teorema 3: Sea $C = C_g^t D_q^{\frac{1}{2}}$, $p \times r$.

1. Se tiene que $B = CC^t$
2. Sean $w_1, \ldots, w_t$ tales que $C^t V^{-1} C w_j = \lambda_j w_j$ con $\lambda_j \neq 0$ y $w_j^t I_r w_s = \delta_{js}$. Entonces existen $\beta_1, \ldots, \beta_t$ tales que $BV^{-1}\beta_j = \lambda_j \beta_j$ y $\beta_j^t V^{-1} \beta_s = \delta_{js}$.
3. Recíprocamente, sean $\beta_1, \ldots, \beta_t$ tales que $BV^{-1}\beta_j = \lambda_j \beta_j$ con $\lambda_j \neq 0$ y $\beta_j^t V^{-1} \beta_s = \delta_{js}$. Entonces existen $w_1, \ldots, w_t$ tales que $C^t V^{-1} C w_j = \lambda_j w_j$ con $w_j^t I_r w_s = \delta_{js}$.
4. Existe una matriz S tal que $SBV^{-1}S^{-1} = HBH$ donde $H = U\Delta^{-\frac{1}{2}}U^t$ con $U = (u_1, \ldots, u_p)$, $\Delta = \text{diag}(\mu_i)$ y $u_1, \ldots, u_p$ son los vectores propios de V asociados a $\mu_1 \geq \cdots \geq \mu_p > 0$. Los vectores propios de BV^{-1}, V^{-1} ortonormados son: $\beta_j = H^{-1} w_j$ donde los w_j son los vectores propios de HBH, I_p ortonormados.

Prueba: 2. Definiendo $\beta_j = \frac{Cw_j}{\sqrt{\lambda_j}}$ se obtiene el resultado. 3. Definiendo $w_j = \frac{1}{\sqrt{\lambda_j}} C^t V^{-1} \beta_j$ se obtiene el resultado. 4. Se sabe que $V^{-1} = \left(U\Delta^{-\frac{1}{2}}U^t\right)^2 = H^2$ por lo tanto $BV^{-1} \sim HBV^{-1}H^{-1} = HBH$. De aquí sigue que $BV^{-1}\beta_j = \lambda_j \beta_j$ con $\beta_j = H^{-1} w_j$. Además,

$$\beta_j V^{-1} \beta_s = w_j^t H^{-1} V^{-1} H w_s = w_j^t I_p w_s = \delta_{js}.$$

2.3 Representaciones en ADD

A partir de los resultados obtenidos con el teorema 2, se pueden construir las siguientes representaciones bidimensionales para el caso de más de dos grupos a priori. Es decir, $r > 2$.

2.3.1 Representación de los grupos a priori

Hemos visto que el ADD se puede interpretar como la búsqueda de los ejes (en $\mathbb{R}^p$) más discriminantes de los grupos a priori, en el sentido de maximización de la inercia interclases (teorema 2). Estos ejes son los vectores propios del ACP de $\left(C_g, V^{-1}, D_q\right)$, lo que nos permite al mismo tiempo calcular las funciones discriminantes. Para obtener las representaciones bidimensionales de los baricentros de los grupos, se proyectan éstos, $V^{-1}-$ ortogonalmente sobre los planos principales del ACP.

De lo anterior sigue que la coordenada del baricentro g_l del grupo E_l, sobre el eje $j-$ ésimo es: $\text{coord}_j(g_l) = g_l V^{-1} \beta_j$.

2.3.2 Representación de las variables como son definidas por los grupos a priori

Las columnas de la matriz C_g representan las variables tal como son determinadas por los grupos a priori ya que cada columna y^j de C_g es el vector $(g_{j1}, \ldots, g_{jr})$, donde g_{jl} es el promedio de la variable x^j en el grupo E_l. Se les llamará variables promedio.

Por las fórmulas de dualidad del ACP se sabe que las columnas de la matriz $(\beta_1 \dots \beta_t) D_{\sqrt{\lambda}}$ son las coordenadas de las columnas y^j de C_g. Es decir, $\text{coord}_s\left(y^j\right) = \sqrt{\lambda_s}\beta_{js}$ para $j = 1, \dots, p$ y $s = 1, \dots, t$.

Superponiendo el gráfico de los grupos a priori y el de las variables promedio, es posible analizar la influencia de las variables en la determinación de los grupos a priori.

2.3.3 Representación de los individuos

Los individuos se proyectan en suplementario sobre los ejes discriminantes. Es decir, sobre los ejes principales del ACP de $\left(C_g, V^{-1}, D_q\right)$. Sea x_i el $i-$ésimo individuo, su coordenada sobre el $j-$ ésimo eje es $\text{coord}_j\left(x_i\right) = x_i V^{-1}\beta_j$.

Es claro que el vector de coordenadas de los individuos sobre el $j-$ ésimo eje es la función discriminante z^j.

2.3.4 Representación de las variables

Las variables se representan en el sistema $D-$ ortonormado determinado por las variables discriminantes. La coordenada de la variable x^j (columna $j-$ ésima de X) sobre el eje $s-$ ésimo es: $\text{coord}_s\left(x^j\right) = x^j D z^s = \beta_{js}$. En efecto, como $z^s = XV^{-1}\beta_s$ entonces el vector de coordenadas de las variables en la dirección de la variable discriminante z^s es,

$$X^t D z^s = X^t D X V^{-1}\beta_s = \beta_s.$$

Si las variables son estandarizadas (varianza igual a 1) entonces $\text{coord}_s\left(x^j\right) = corr\left(x^j, z^s\right)$ y las variables se pueden representar como en ACP normado, en un círculo de correlaciones.

2.3.5 El caso de dos grupos a priori

Cuando solo hay dos grupos a priori, las representaciones se simplifican ya que el rango de B (y por tanto el de BV^{-1}) vale 1. En este caso $g_2 - g_1$ es un vector propio de BV^{-1} asociado al único valor propio $q_1 q_2 \left\|g_2 - g_1\right\|_{V^{-1}}$. Se puede leer una prueba de este resultado por ejemplo en [7]. En consecuencia, las representaciones tanto de los individuos y baricentros como de las variables, se hacen sobre una recta.

2.3.6 Indices de calidad

La calidad de la discriminación en un subespacio principal de dimensión q es el porcentaje de inercia explicada:

$$100\frac{\sum_{j=1}^{q}\lambda_j}{\text{Inercia Total}} = 100\frac{\sum_{j=1}^{q}\lambda_j}{\sum_{j=1}^{t}\lambda_j}$$

donde t es el número de valores propios positivos del ACP de $\left(C_g, V^{-1}, D_q\right)$.

La calidad de la representación de los baricentros y de los individuos en cada eje discriminante se mide por medio de los cosenos cuadrados de la misma forma como se hace en ACP. Las fórmulas para el cálculo de estos cosenos son:

- Baricentros:

$$cos^2(g_l, s) = \frac{\|\mathrm{Pr}_s(g_l)\|^2_{V^{-1}}}{\|g_l\|^2_{V^{-1}}} = \frac{\left\|\left(g_l V^{-1}\beta_s\right)\beta_s\right\|^2_{V^{-1}}}{g_l V^{-1} g_l^t} = \frac{\left(g_l V^{-1}\beta_s\right)^2}{g_l V^{-1} g_l^t}.$$

- Individuos:

$$cos^2(i, s) = \frac{\|\mathrm{Pr}_s(x_i)\|^2_{V^{-1}}}{\|x_i\|^2_{V^{-1}}} = \frac{\left\|\left(x_i V^{-1}\beta_s\right)\beta_s\right\|^2_{V^{-1}}}{x_i V^{-1} x_i^t} = \frac{\left(x_i V^{-1}\beta_s\right)^2}{x_i V^{-1} x_i^t}.$$

3 Algoritmo para la implementación computacional del ADD

En esta implementación se usaron las mismas notaciones introducidas en el texto precedente y se asume que todos los individuos tienen peso igual: $p_i = \frac{1}{n}, \quad i = 1, 2, \ldots, n$. El número de grupos a priori debe ser mayor que 2, es decir $r > 2$.

Para el cálculo de los cosenos de individuos y baricentros, se utilizaron las identidades:

$$x_i V^{-1} x_i^t = \sum_{h=1}^{t} \left[\mathrm{coord}_h(x_i)\right]^2 \quad \text{y} \quad g_l V^{-1} g_l^t = \sum_{h=1}^{t} \left[\mathrm{coord}_h(g_l)\right]^2.$$

1. Operación de centraje y reducción:
   - 1.1 Para $j = 1, \ldots, p$ se calcula $\overline{x}^j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ y $\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 - \left(\overline{x}^j\right)^2$.
   - 1.2 Para $i = 1, \ldots, n$ y $j = 1, \ldots, p$ se sustituye la entrada x_{ij} de X por $x_{ij} - \overline{x}^j$ para el centraje, y por $\frac{x_{ij} - \overline{x}^j}{\sigma_j}$ para la reducción.
2. Baricentros de los grupos: para $l = 1, \ldots, r$; $j = 1, \ldots, p$ calcular $g_{jl} = \frac{1}{n_l}\sum_{i \in E_l} x_{ij}$ donde $n_l = |E_l|$.
3. Calcular $V = \frac{1}{n}X^t X = (v_{ij})_{p\times p}$, donde $v_{ij} = \frac{1}{n}\sum_{s=1}^{n} x_{si}x_{sj}$ para i, $j = 1, \ldots, p$.
4. Calcular V^{-1} :
   - 4.1 Para V calcular sus valores propios $\mu_1 \geq \cdots \geq \mu_p > 0$ y los vectores propios correspondientes $u_1, \ldots, u_p$, I_p- ortonormados.
   - 4.2 Sea $V^{-1} = (t_{ij})$; para i, $j = 1, \ldots, p$; calcular $t_{ij} = \sum_{s=1}^{p} \frac{1}{\mu_s} u_{is}u_{js}$, con $u_s^t = (u_{1s}, \ldots, u_{ps})$.
5. Sea C la matriz definida en el teorema 3. Si $r \leq p$ calcular $A = C^t V^{-1} C$, sus valores propios no nulos, los vectores propios I_r- ortonormados, y los β_j correspondientes (ver la prueba del teorema 3, parte 2.):
   - 5.1 Para $l, k = 1, \ldots, r$ calcular

$$\left(C^t V^{-1} C\right)_{lk} = \frac{\sqrt{n_l n_k}}{n} \sum_{b=1}^{p} g_{bk} \left[\sum_{f=1}^{p} g_{fl}\ t_{fb}\right]$$

5.2 Calcular los valores propios $\lambda_1 \geq \cdots \geq \lambda_t > 0$ de A y los vectores propios correspondientes $w_1, \ldots, w_t$, I_r- ortonormados.

5.3 Calcular $\beta_j = \frac{Cw_j}{\sqrt{\lambda_j}}$: Es decir, para $j = 1, \ldots, t$ y $s = 1, \ldots, p$ calcular:

$$\beta_{sj} = \frac{1}{\sqrt{n\lambda_j}} \sum_{k=1}^{r} \sqrt{n_k} g_{sk} w_{kj}.$$

6. Si $r > p$ calcular $A = HBH$ y sus valores propios no nulos, sus vectores propios I_p- ortonormados y los β_h correspondientes (ver teorema 3, parte 4.):

6.1 Cálculo de B y H: para $i, j = 1, \ldots, p$ calcular

6.1.1 $B_{ij} = \frac{1}{n} \sum_{s=1}^{r} n_s g_{is} g_{js}$

6.1.2 $H_{ij} = \sum_{k=1}^{p} \frac{1}{\sqrt{\mu_k}} u_{ik} u_{kj}$

6.2 Cálculo de HB y HBH :

6.2.1 para $i, j = 1, \ldots, p$ calcular $(HB)_{ij} = \sum_{s=1}^{p} H_{is} B_{sj}$

6.2.2 para $i, j = 1, \ldots, p$ calcular $(HBH)_{ij} = \sum_{s=1}^{p} (HB)_{is} H_{sj}$

6.3 Calcular los valores propios $\lambda_1 \geq \cdots \geq \lambda_t > 0$ de A y los vectores propios correspondientes $w_1, \ldots, w_t$, I_p- ortonormados.

6.4 Cálculo de $\beta_j = H^{-1} w_j$: para $j = 1, \ldots, t$, $s = 1, \ldots, p$ calcular $\beta_{sj} = \sum_{k=1}^{p} w_{kj} \left[\sum_{h=1}^{p} \sqrt{\mu_h} u_{sh} u_{hk} \right]$.

7. Cálculo de coordenadas:

7.1 Baricentros: para $s = 1, \ldots, t$; $l = 1, \ldots, r$; calcular $\text{coord}_s(g_l) = \sum_{j=1}^{p} \beta_{js} \sum_{k=1}^{p} g_{kl} t_{kj}$.

7.2 Individuos: para $s = 1, \ldots, t$; $i = 1, \ldots, n$; calcular $\text{coord}_s(x_i) = \sum_{j=1}^{p} \beta_{js} \sum_{k=1}^{p} x_{ik} t_{kj}$.

7.3 Variables: para $s = 1, \ldots, t$; $j = 1, \ldots, p$; calcular $\text{coord}_{z^s}\left(x^j\right) = \beta_{js}$.

8. Cálculo de los cosenos cuadrados:

8.1 Individuos: para $s = 1, \ldots, t$; $i = 1, \ldots, n$; calcular $\cos^2(i, s) = \frac{\left[\text{coord}_s(x_i)\right]^2}{\sum_{h=1}^{t} \left[\text{coord}_h(x_i)\right]^2}$.

8.2 Baricentros: para $s = 1, \ldots, t$; $l = 1, \ldots, r$; calcular $\cos^2(l, s) = \frac{\left[\text{coord}_s(g_l)\right]^2}{\sum_{h=1}^{t} \left[\text{coord}_h(g_l)\right]^2}$.

4 Acerca de la implementación [5]

El algoritmo anterior fue implementado en lenguaje C++ utilizando la Programación Orientada a Objetos. Se programaron las clases `Matriz`, `TablaDatos`, `MatrizV`, `MatrizA`, `Plano-Principal`, `CirculoCorrelacion`, `CosenosIndividuos` y `CosenosVariables`. La clase `Matriz` se encarga de aspectos comunes a todas las clases, como son manipular la matriz en memoria, guardarla y recuperarla de disco, además tiene métodos para calcular estadísticas básicas como la media, la correlacion, la desviación estándar entre otras. La clase `TablaDatos` hereda de la clase `Matriz` y se encarga de manipular la tabla datos y la variable cualitativa. Tiene métodos (además de los heredados) para centrar y reducir la tabla, para calcular y almacenar los baricentros de los grupos (paso 2 del algoritmo). La clase `MatrizV` hereda de la clase `Matriz` y se encarga de los pasos 3 y 4 del algoritmo, es decir, tiene métodos para calcular V , V^{-1} y los valores y vectores propios de V. La clase `MatrizA` hereda de clase `MatrizV` y se encarga de los pasos 5 y 6 del algoritmo por lo que tiene métodos para calcular tanto la matriz A como los β_j. Los métodos para calcular los valores y vectores propios los hereda de la clase `MatrizA`. La clase `PlanoPrincipal` hereda de la clase `Matriz` y tiene métodos para calcular las coordenadas de los individuos y de los baricentros (pasos 7.1 y 7.2 del algoritmo), así como para generar el plano principal en formato LaTeX y en formato Windows. Similarmente, la clase `CirculoCorrelacion` hereda de la clase `Matriz` y tiene métodos para calcular las coordenadas de las variables (paso 7.3 del algoritmo) y para generar el círculo de correlación en formato LaTeX y en formato Windows. Finalmente, las clases `CosenosIndividuos` y `CosenosVariables` se encargan de los pasos 8.1 y 8.2 del algoritmo. Es decir, calculan los cosenos cuadrados de los individuos y los baricentros.

5 Ejemplo ilustrativo

Esta sección tiene el propósito de ilustrar el uso del software y de mostrar los productos que se pueden obtener mediante su aplicación.

5.1 Sobre los datos de contaminación de ríos

Vamos a considerar 13 variables relacionadas con la contaminación de aguas (entre paréntesis se coloca el nombre abreviado de la variable): Fosfato (FOS), Nitratos (NITA), Calidad del agua (CAL), Sólidos totales (STOT), Acidez (PH), Manganeso (MN), Zinc (ZN), Sólidos Sedimentales (SS), Alcalinidad (ALCA), Cloro (CL), Caudal (CAU), Demanda Bioquímica de Oxígeno (DBO) y Porcentaje de Saturación de Oxígeno (PORS). Estas variables fueron medidas en los ríos que forman el embalse La Garita, en 9 puntos de muestreo que corresponden a la Presa, los ríos Alajuela, Ciruelas, Virilla y Quebrada Soto; tres en el Embalse (Orilla, Centro y Salida) y uno en el Desfogue-Garita. Se hace una medición por estación: Verano (V), Verano-Invierno (VI), Invierno (I) e Invierno-Verano (IV). El nombre de un punto de muestreo cualquiera se forma con las primeras letras del nombre del sitio seguido por el nombre abreviado de la estación. Por ejemplo,

PV es la Presa en Verano, VVI es el río Virilla en Verano Invierno y EOV es el Embalse Orilla en Verano.

La tabla de datos contiene el promedio de cada variable para cada caso "sitio-estación" dando como resultante una matriz de datos de 36 filas (sitio-estación) y 13 columnas (ver [3]).

Los grupos a priori que se quieren discriminar son los definidos por las 4 estaciones: V, VI, I e IV. La composición de estos grupos se presenta en la tabla siguiente:

G1	PV – DV – EOV – ESV – ECV – AV – CV – QV – VV
G2	PVI – DVI – EOVI – ESVI – ECVI – AVI – CVI – QVI – VVI
G3	PI – DI – EOI – ESI – ECI – AI – CI – QI – VI
G4	PIV – DIV – EOIV – ESIV – ECIV – AIV – CIV – QIV - VIV

5.2 Los archivos de datos y etiquetas

Para hacer los cálculos el sistema necesita cuatro archivos tipo texto, los cuales se describen a continuación:

- El archivo de datos: debe tener extensión TXT. El primer registro (fila) debe contener dos números separados por blancos que indican respectivamente el número de individuos y variables. Los demás registros contienen los datos propiamente dichos. En el caso de los datos de contaminación de ríos el archivo tiene 37 registros y 13 columnas. Los primeros 4 registros del archivo son los siguientes:

```
36 13
1.44 0.78 58.25 126.75 7.29 0.21 0.02 0.66 120.50 3.92 17.85 25.00 53.25
2.70 1.13 62.50 195.00 7.50 0.20 0.02 1.15 104.50 4.10 28.20 90.70 77.00
3.89 0.35 79.25 176.50 7.41 0.37 0.28 0.62 80.25 3.35 53.27 3.77 89.75
```

- El archivo de etiquetas de los individuos: debe tener extensión ETI (de ETiqueta de los Individuos). Consta de una sola columna que contiene en la primera posición un número que indica cuantos individuos hay. En las restantes posiciones vienen las etiquetas de los individuos. En el caso de los datos de contaminación de ríos el archivo tiene 37 registros. Los primeros 4 registros del archivo son los siguientes:

```
36
PV
PVI
PI
```

- El archivo de etiquetas de las variables: debe tener extensión ETV (de ETiqueta de las Variables). Consta de una sola columna que contiene en la primera posición un número que indica cuantas variables hay. En las restantes posiciones vienen las etiquetas de las variables (las cuales no pueden tener espacios en blanco). En el caso de los datos de contaminación de ríos el archivo tiene 14 registros. Los primeros 4 registros del archivo son los siguientes:

```
13
```

```
NITA
FOS
CAL
```

- El archivo de la variable cualitativa: debe tener extensión CUA (de variable CUAlitativa) y consta de una sola columna. En la primera posición contiene un número que indica cuantas clases hay. En las restantes posiciones se escribe un número entero que indica el número de clase a la que pertenece el individuo en cuestión. En el caso de los datos de contaminación de ríos el archivo tiene 37 registros. Los primeros 5 registros del archivo son los siguientes:

```
4
1
2
3
4
```

5.3 El submenú `Discriminante-Paso-a-Paso`

El módulo de Análisis Discriminante se integró al Programa Interactivo para Métodos de Análisis de Datos en su versión 3.0 (PIMAD 3.0). Se dispone de tres variantes para ejecutar los cálculos:

- El uso de los botones para conseguir los gráficos en pantalla, los cosenos de los individuos y de los baricentros.
- El submenú `Discriminante-Directo` con el cual se despliegan los gráficos en pantalla.
- El submenú `Discriminante-Paso-a-Paso` permite ejecutar el algoritmo de la sección 3 paso a paso a través de 16 opciones.

A continuación se describe cada opción del submenú `Discriminante-Paso-a-Paso`. Todos los archivos de 'salida' son tipo ASCII con extensión `TXT`.

1. Opción: `Centraje y Reducción de la Tabla de Datos.` Con esta opción se crea un archivo de nombre **tabla.txt** que contiene en columna las variables explicativas centradas y reducidas (es decir, estandarizadas).
2. Opción: `Calcula los Baricentros`. Calcula los baricentros de cada grupo a priori.
3. Opción: `Calcular la matriz V`. Crea un archivo de nombre **mat_V.txt** que contiene la matriz de correlaciones de las variables explicativas (V).
4. Opción: `Calcular los Vectores y Valores Propios de V`. Crea dos archivos de nombres **vectorp.txt** y **valorp.txt** que contienen respectivamente los vectores propios I_{13} ortonormados de V y los valores propios correspondientes.

5. Opción: `Calcular la Matriz V^(-1)`. Crea un archivo de nombre **vinv.txt** que contiene la matriz inversa de V.

6. Opción: `Calcular la Matriz A`. Crea un archivo de nombre **mat_A.txt** que contiene la matriz A (ver pasos 5. y 6. del algoritmo). En nuestro ejemplo, este archivo es:

$$A = \begin{pmatrix} 0.547 & -0.223 & -0.260 & -0.064 \\ -0.223 & 0.654 & -0.223 & -0.207 \\ -0.260 & -0.224 & 0.650 & -0.166 \\ -0.064 & -0.207 & -0.166 & 0.436 \end{pmatrix}$$

7. Opción: `Calcular los Vectores y Valores Propios de A`. Crea dos archivos de nombres **vec_p_A.txt** y **val_p_A.txt** que contienen respectivamente los vectores propios de A, I_5 ortonormados y los valores propios correspondientes (que son al mismo tiempo los valores propios positivos de BV^{-1}). Para nuestro ejemplo, estos valores propios son:

λ_1	λ_2	λ_3
0.9283	0.8617	0.4203

El archivo de los correspondientes vectores propios de A, tiene la apariencia siguiente:

w_1	w_2	w_3
0.425	0.499	-0.566
0.309	-0.808	-0.038
-0.843	-0.004	-0.196
0.109	0.313	0.800

8. Opción: `Calcular la matriz Beta`. Con esta opción se calcula la matriz de vectores propios de BV^{-1}, V^{-1} ortonormados y la guarda en un archivo de nombre **beta.txt**. Estos vectores propios son:

β_1	-0.39	0.60	-0.89	0.14	0.25	-0.29	-0.18	0.24	0.32	0.19	-0.33	0.45	-0.50
β_2	-0.19	-0.73	0.10	-0.53	0.06	-0.31	0.13	-0.34	0.28	-0.26	0.04	-0.69	-0.17
β_3	-0.27	-0.18	0.25	-0.21	-0.05	-0.30	0.81	-0.00	-0.41	-0.17	0.41	-0.14	-0.26

Aquí se presentan los β_j en las filas, sin embargo debe notarse que en el archivo **beta.txt** los vectores $\beta_1, \ldots, \beta_4$ se almacenan en las columnas.

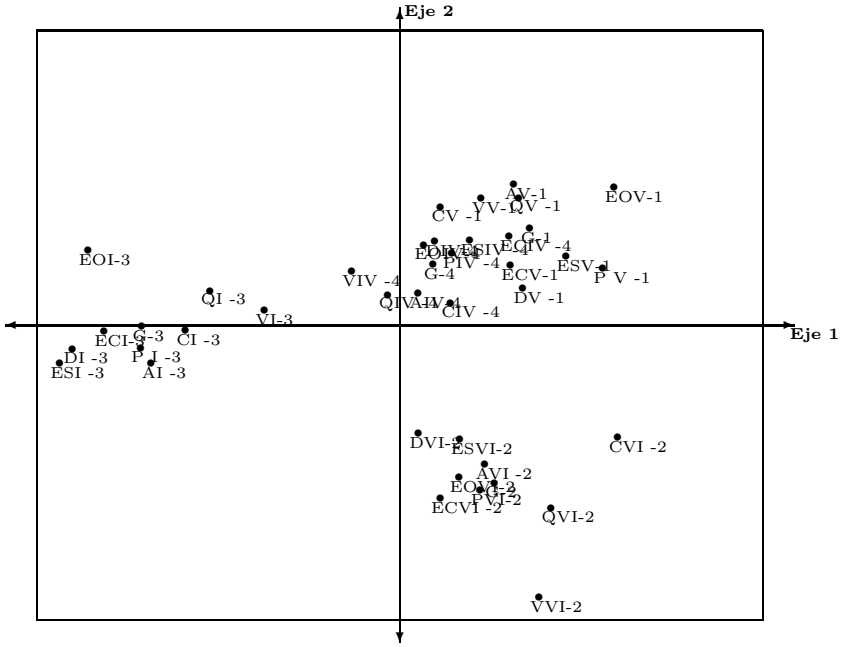
9. Opción: `Calcular Componentes Principales y las Coordenadas de Baricentros`.
Crea un archivo de nombre **compone.txt** que contiene las coordenadas de los individuos en los ejes determinados por los vectores propios de BV^{-1}. Los primeros 5

registros de este archivo son:

	Eje 1	Eje 2	Eje 3
PV	1.326	0.589	-0.751
PVI	0.524	-1.689	-0.757
PI	-1.695	-0.229	0.806
PIV	0.340	0.737	3.656
AV	0.745	1.444	-2.288

10. Opción: `Graficar el Plano Principal`. Despliega una ventana donde el usuario puede escoger dos ejes para representar simultáneamente los individuos y los baricentros de los grupos a priori. Opcionalmente el usuario puede escoger los individuos que desea representar y el tamaño del gráfico.

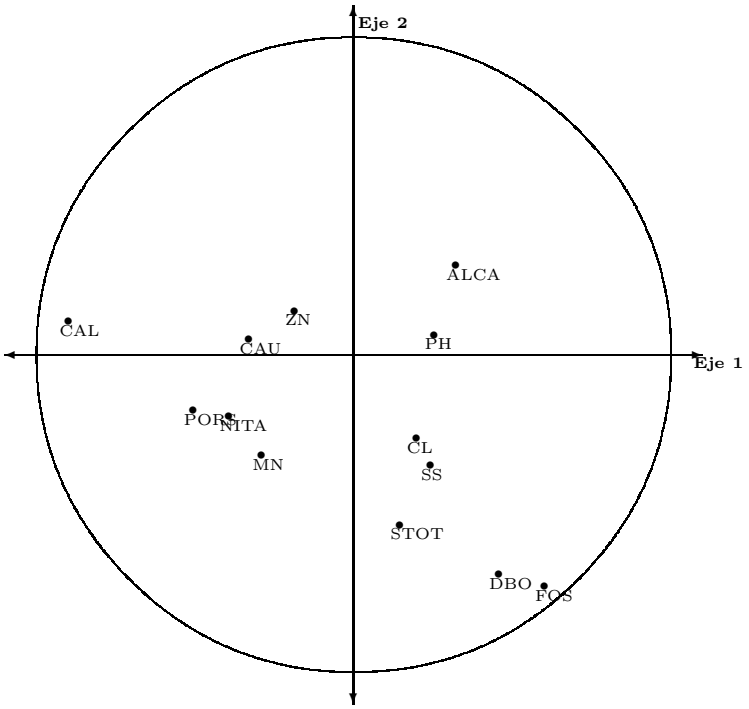
11. Opción: `Generar Archivo LaTeX del Plano Principal`. Crea un archivo tipo LaTeX con las coordenadas de los individuos en el plano principal escogido (plano discriminante). El número al final de las etiquetas indica el grupo a priori al que pertenece el individuo. Por ejemplo, QIV-4 en el extremo inferior del eje vertical indica que el individuo QIV pertenece a la clase a priori número 4. Para nuestro ejemplo el plano principal es:



12. Opción: `Calcular Principales Correlaciones`. Crea un archivo de nombre **princorr.txt** que contiene las coordenadas de las variables explicativas centradas y reducidas, en los ejes determinados por las variables discriminantes.

13. Opción: `Graficar el Círculo de Correlación`. Despliega una ventana donde el usuario puede escoger dos ejes, las variables que desea representar y el tamaño del gráfico. Las coordenadas son tomadas del archivo **princorr.txt**.

14. Opción: `Generar Archivo LaTeX del Círculo de Correlación`. Crea un archivo tipo LaTeX con las coordenadas de las variables, para nuestro ejemplo el círculo de correlación es:



15. Opción: `Calcula Cosenos Cuadrados de los Individuos`. Crea un archivo de nombre **cos_ind.txt** que contiene los cuadrados de los cosenos del ángulo que forma cada individuo con su proyección V^{-1} - ortogonal, sobre los ejes discriminantes. Los cinco primeros registros de **cos_ind.txt** son:

	$\cos^2(i,1)$	$\cos^2(i,2)$	$\cos^2(i,3)$
PV	0.659	0.130	0.211
PVI	0.074	0.771	0.155
PI	0.804	0.015	0.182
PIV	0.008	0.039	0.953
AV	0.070	0.265	0.665

16. Opción: `Calcula Cosenos Cuadrados de los Baricentros`. Crea un archivo de nombre **cos_bar.txt** que contiene los cuadrados de los cosenos del ángulo que forma cada baricentro con su proyección V^{-1} ortogonal, sobre los ejes discriminantes. Para nuestro ejemplo estos cosenos valen:

	$\cos^2(g_l, 1)$	$\cos^2(g_l, 2)$	$\cos^2(g_l, 3)$
G-1	0.241	0.332	0.427
G-2	0.128	0.870	0.002
G-3	0.949	0.000	0.051
G-4	0.016	0.131	0.853

References

[1] Celeux, G.; Nakache, J.P. (1994) *Analyse Discriminante sur Variables Qualitatives.* Polytechnica, Paris.

[2] Diday, E.; Lemaire, J.; Pouget, J.; Testu, F. (1982) *Eléments d'Analyse de Données.* Dunod, Paris.

[3] González, J.; Morales, V. (1993) "Análisis multivariado de la calidad del agua: proyecto hidroeléctrico Ventanas-Garita", *VI Congreso Internacional de Biomatemáticas*, J. Oviedo et al. (Eds.), Universidad de Costa Rica: 227–236.

[4] Lebart, L.; Morineau, A.; Piron, M. (1994) *Statistique Exploratoire Multidimensionnelle.* Dunod, Paris.

[5] Rodríguez, O. (1997) *Introducción a la Programación C++ para Ambiente Windows.* Editorial Tecnológica de Costa Rica, Cartago, Costa Rica.

[6] Saporta, G. (1994) "Los métodos y las aplicaciones del credit-scoring", VII y VIII Simposios, W. Castillo & J. Trejos (Eds.), Editorial de la Universidad de Costa Rica, San José: 103–109.

[7] Saporta, G. (1994) "Análisis discriminante", *Métodos Matemáticos Aplicados a las Ciencias*, VII y VIII Simposios, W. Castillo & J. Trejos (Eds.), Editorial de la Universidad de Costa Rica, San José: 75–102.

[8] Saporta, G. (1980) *L'Analyse des Données.* Que sais-je?, Presses Universitaires de France, Paris.