QÜESTIIÓ, Vol 15, 3. pp. 279-298, 1991

REPRESENTACIÓN CANÓNICA SIMULTÁNEA DE POBLACIONES Y VARIABLES: UNA APLICACIÓN AL ESTUDIO DE ENFERMEDADES RENALES

M. RÍOS, A. VILLARROYA, E. VEGAS y J.M. OLLER

Universitat de Barcelona

En este trabajo se describe una metodología que nos permite representar simultáneamente poblaciones estadísticas y variables aleatorias, así como transformaciones admisibles de las mismas. Para ello se representan las variables aleatorias, o sus transformadas, en el espacio tangente a la variedad de las poblaciones estadísticas, a continuación se proyectan en dicha variedad y, finalmente, se aplica un análisis canónico de poblaciones para su representación en un espacio de dimensión reducida. Los resultados son aplicados a la representación de grupos de enfermos renales y a las variables observadas sobre ellos (concentraciones de inmunoglobulinas).

Simultaneous canonical representation of populations and variables: an application to the study of kidney illness.

Keywords: Análisis canónico de poblaciones, representación simultánea de poblaciones y variables.

Clasificación AMS (1980): 62-07, 62H25.

-Departament d'Estadística. Universitat de Barcelona.

-Article rebut el juliol de 1991.

1. INTRODUCCIÓN

Como es sabido, el Análisis Canónico de Poblaciones es un método que nos permite definir una distancia entre poblaciones estadísticas, realizar una representación gráfica de las mismas en espacios euclídeos de dimensión reducida (generalmente en un plano) y hallar e interpretar las variables o combinaciones lineales de variables que más influyen en las diferencias entre las poblaciones. Véase, entre otros, Pearson (1901) que fue el primero en estudiar de manera rigurosa el problema de reducción de la dimensión en espacios euclídeos, Mahalanobis (1936) quien introduce la bien conocida distancia que lleva su nombre y que es la generalización del coeficiente racial de Pearson (Pearson, 1926) y Rao (1948) que realiza una clasificación de castas y tribus indias relacionándolas con sus características antropológicas y sociales, basándose en la distancia de Mahalanobis.

Sin embargo, en muchos casos resulta de interés representar simultáneamente las poblaciones y las variables sobre ellas observadas. Esta metodología está desarrollada en el Análisis Factorial de Correspondencias, Benzecri (1976), y en las técnicas descritas en Rios *et al.* (1991), entre otros, para poblaciones asociadas a distribuciones multinomiales. Otros métodos que realizan esta representación simultánea son los basados en el método Biplot, Gabriel (1971, 1981a, 1981b).

Siguiendo las ideas desarrolladas en Rios *et al.* (1991), en este trabajo realizamos una representación simultánea de variables y poblaciones asociadas a distribuciones normales multivariantes, con igual matriz de covarianzas, utilizando técnicas de geometría diferencial.

Por último, esta metodología es utilizada en la representación de grupos de enfermos renales y ciertas variables observadas sobre ellos.

2. REPRESENTACIÓN DE LAS POBLACIONES EN UN ESPACIO DE DIMENSIÓN REDUCIDA

Sean las poblaciones P_i $(i = 1, \ldots, k)$ sobre las que observamos el vector aleatorio $X = (X_1, \ldots, X_n)'$ que sigue una distribución normal multivariante de vector de medias $\mu_i = (\mu_i^1, \ldots, \mu_i^n)'$ y matriz de covarianzas $\Sigma = (\sigma_{ij})$ común para todas las poblaciones. En la variedad constituida por las funciones de densidad $p(x \mid \mu_i, \Sigma)$ las poblaciones quedan definidas por las coordenadas $\mu_i = (\mu_i^1, \ldots, \mu_i^n)'$. La distancia al cuadrado entre dos poblaciones cualesquiera

viene dada, utilizando como tensor métrico la matriz de información de Fisher, Mahalanobis (1936) y Burbea (1986), por:

$$d^2(P_i, P_j) = d_{ij}^2 = (\mu_i - \mu_j)'\Sigma^{-1}(\mu_i - \mu_j) \tag{1}$$

Las coordenadas centradas de las poblaciones en un espacio de dimensión q vienen dadas por las filas de la matriz:

$$Y = (M - \mathbf{1}\bar{\mu}')U \tag{2}$$

siendo M una matriz cuyos vectores fila son las coordenadas de las poblaciones P_i, es decir, $M = (\mu_i^j) \quad i = 1, \ldots, k \quad j = 1, \ldots, n$; $\bar{\mu} = (\bar{\mu}_1, \ldots, \bar{\mu}_n)'$ con $\bar{\mu}_j = \frac{1}{k}\sum_{h=1}^{k} \mu_{hj}$ y $\mathbf{1} = (1, \ldots, 1)'$. U es una matriz cuyas columnas son los q vectores propios u_i obtenidos de la expresión:

$$Ru_i = \lambda_i \Sigma u_i \tag{3}$$

siendo $\lambda_1 \geq \cdots \geq \lambda_q$ los q mayores valores propios de R respecto de Σ, con la condición:

$$u_i'\Sigma u_j = \delta_{ij} \tag{4}$$

con $R = (M - \mathbf{1}\bar{\mu}')'(M - \mathbf{1}\bar{\mu}')$ y δ_{ij} las deltas de Kronecker.

La demostración, así como otros detalles de estos resultados pueden verse entre otros en Seber (1984) y tambien Ríos y Oller (1986).

El porcentaje de la dispersión explicado por los q vectores es:

$$P = 100\,\frac{\lambda_1 + \cdots + \lambda_q}{\lambda_1 + \cdots + \lambda_n} \tag{5}$$

siendo $\lambda_i \quad (i = 1, \ldots, n)$ los n valores propios de (3) con la condición (4).

3. REPRESENTACIÓN DE VARIABLES ALEATORIAS EN LA VARIEDAD

Dado un modelo estadístico paramétrico dominado, con las usuales condiciones de regularidad, es bien conocido que puede ser representado por la variedad $\mathcal{M}$ constituida por las correspondientes funciones de densidad $p(x, \theta)$ respecto de una medida de referencia ν, ver por ejemplo Amari (1985), Burbea(1986) y Oller (1989) entre otros. Vamos ahora a estudiar el problema de la representación de variables aleatorias mediante curvas en la variedad.

Sea $F = \{X : \mathcal{E}_p(|X|^2) < \infty \quad \forall p \in \mathcal{M}\}$. Definamos F_p como el espacio vectorial formado por las variables aleatorias con momentos de segundo orden finitos en p, es decir:

$$F_p = \{X : \mathcal{E}_p(|X|^2) < \infty\} \tag{6}$$

donde $\mathcal{E}_p$ es la esperanza matemática calculada utilizando la medida probabilística correspondiente al punto $p = p(\cdot;\theta)$, $dP(x) = p(x,\theta)\ d\nu(x)$, por lo tanto $F = \bigcap_{p\in\mathcal{M}} F_p$.

Sea, a su vez, H_p el subespacio vectorial de F_p constituido por el conjunto de variables aleatorias centradas con momentos de segundo orden finitos:

$$H_p = \{X \in F_p : \mathcal{E}_p(X) = 0\} \tag{7}$$

Resulta evidente que $F_p = H_p \oplus < 1 >$ (donde $< 1 >$ simboliza el subespacio vectorial generado por la variable aleatoria constante igual a 1) ya que toda variable aleatoria puede descomponerse de manera única en la forma:

$$X = (X - \mathcal{E}_p(X)) + \mathcal{E}_p(X) \tag{8}$$

donde $(X - \mathcal{E}_p(X)) \in H_p$ y $\mathcal{E}_p(X) \in < 1 >$.

Además, H_p con el producto escalar definido a través de la covarianza, (H_p, cov_p) constituye un espacio de Hilbert.

Consideremos ahora la representación del espacio vectorial tangente en un punto de la variedad $\mathcal{M}$, véase entre otros Amari (1985) y Oller (1989), definida por

$$E_p =< Z_1, \ldots, Z_n > \qquad \text{donde} \qquad Z_i = \frac{\partial \log p(x,\theta)}{\partial \theta_i} \tag{9}$$

Bajo las habituales condiciones de regularidad se cumple que

$$\mathcal{E}_p(Z_i) = 0 \qquad i = 1, \ldots, n \tag{10}$$

Por tanto, $E_p \subset H_p$. Adicionalmente, la restricción de la covarianza a E_p, permite definir un producto escalar en el espacio tangente, tal que

$$g_{ij} =< Z_i, Z_j >= \mathcal{E}_p(Z_i \cdot Z_j) \qquad i,j = 1, \ldots, n \tag{11}$$

siendo $G = (g_{ij})$ la matriz de información de Fisher.

Por todo lo anteriormente expuesto podemos representar de una manera natural a cualquier variable aleatoria, Y, de F en E_p mediante la aplicación i

(inclusión natural), su centrado y su posterior proyección ortogonal:

$$\begin{array}{ccccccc} & i & & \xi_p & & \pi_p & \\ F & \longrightarrow & F_p & \longrightarrow & H_p & \longrightarrow & E_p \\ Y & \rightarrow & Y & \rightarrow & Y-\mathcal{E}_p & \rightarrow & \pi(Y-\mathcal{E}_p) \end{array}$$

es decir, como una combinación lineal de $Z_1, \ldots, Z_n$:

$$Y \longrightarrow (\pi_p \circ \xi_p \circ \mathtt{i})(Y) = \alpha_1 Z_1 + \cdots + \alpha_n Z_n$$

donde los coeficientes $\alpha = (\alpha_1, \ldots, \alpha_n)'$ se obtiene resolviendo el sistema:

$$G\alpha = V_Y \quad \text{donde} \quad V_Y = \begin{pmatrix} cov_p(Y, Z_1) \\ \vdots \\ cov_p(Y, Z_n) \end{pmatrix} \tag{12}$$

Finalmente:

$$\alpha = G^{-1} V_Y \tag{13}$$

Esta representación de las variables aleatorias de F en el espacio tangente satisface debido al caracter lineal de ξ_p, π_p y $\mathtt{i}$.

$$(\pi_p \circ \xi_p \circ \mathtt{i})(\alpha X + \beta Y) = \alpha\,(\pi_p \circ \xi_p \circ \mathtt{i})(X) + \beta\,(\pi_p \circ \xi_p \circ \mathtt{i})(Y)$$

donde $X, Y \in F$ y $\alpha, \beta \in \Re$,

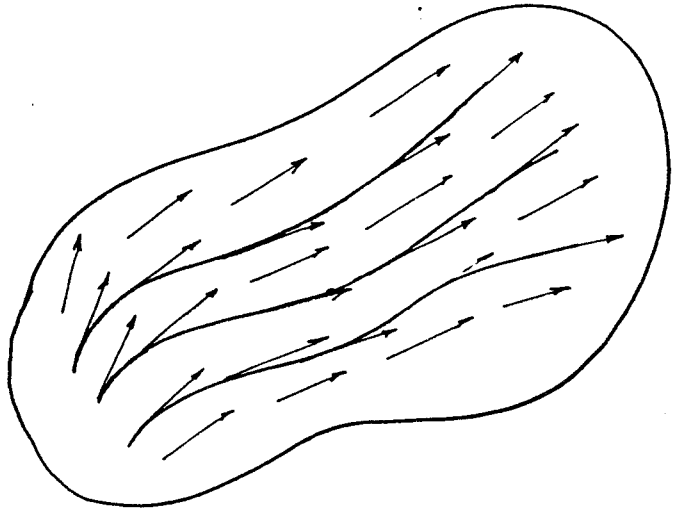

Figura 1. Curvas integrales de un campo vectorial.

con

$$cov(X_\beta^2, Z_\gamma) = cov\left(X_\beta^2, \sum_{j=1}^{n} \sigma^{\gamma j}(X_j - \mu^j)\right)$$

$$= \sum_{j=1}^{n} \sigma^{\gamma j}\left\{\mathcal{E}\left[(X_\beta - \mu^\beta)^2(X_j - \mu^j)\right] + 2\mu^\beta \mathcal{E}\left[(X_\beta - \mu^\beta)(X_j - \mu^j)\right]\right\}$$

y puesto que los momentos de tercer orden son nulos

$$cov\,(X_\beta^2, Z_\gamma) = \sum_{j=1}^{n} \sigma^{\gamma j} 2\mu^\beta \sigma_{\beta j} = 2\mu^\beta \delta_{\beta\gamma} \tag{22}$$

donde $\delta_{\beta\gamma}$ simboliza las deltas de Kronecker y $\beta, \gamma = 1, \ldots, n$.

Tomemos como ejemplo el caso $\beta = 1$, es decir, vamos a representar la variable X_1^2. En este caso:

$$V_{X_1^2} = 2\mu^1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Según (14) obtenemos el siguiente sistema de ecuaciones diferenciales:

$$\frac{d\mu^1(t)}{dt} = 2\mu^1 \Sigma \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = 2\mu^1 \begin{pmatrix} \sigma_{11} \\ \sigma_{21} \\ \vdots \\ \sigma_{n1} \end{pmatrix} \tag{23}$$

Que expresado en forma de componentes:

$$\dot{\mu}_i^1(t) = 2\mu^1 \sigma_{i1} \qquad i = 1, \ldots, n \tag{24}$$

de la integración de la primera componente $(\mu_1^1(t))$ se deduce $\mu^1 = Ke^{2\sigma_{11}t}$ y finalmente

$$\left.\begin{array}{l} \dot{\mu}_2^1(t) = 2\sigma_{21} K e^{2\sigma_{11}t} \\ \vdots \\ \dot{\mu}_n^1(t) = 2\sigma_{n1} K e^{2\sigma_{11}t} \end{array}\right\} \Longrightarrow \mu_\alpha^1(t) = K\frac{\sigma_{\alpha 1}}{\sigma_{11}} e^{2\sigma_{11}t} + C_\alpha \tag{25}$$

$$\alpha = 2, \ldots, n$$

Reparametrizando: $s = e^{2\sigma_{11}t}$ obtenemos que la proyección de X_1^2 en la variedad viene dada por la siguiente curva expresada en forma paramétrica:

$$X_1^2 \longrightarrow (Ks, K\frac{\sigma_{21}}{\sigma_{11}}s + C_2, \ldots, K\frac{\sigma_{n1}}{\sigma_{11}}s + C_n) \tag{26}$$

Es decir, X_1^2 se representa en la variedad $\mathcal{M}$ por rectas. Por ejemplo, si queremos que la recta se origine en el punto de la variedad $\mathcal{M}$ correspondiente al centro de masas, las constantes deben valer:

$$\begin{aligned} K &= \bar{\mu}_1 \\ C_\alpha &= \bar{\mu}_\alpha - \bar{\mu}_1 \tfrac{\sigma_{21}}{\sigma_{11}} \end{aligned} \qquad \alpha = 2, \ldots, n$$

Ahora, sólo nos queda proyectar estas rectas en el espacio de dimensión reducida. Las curvas asociadas a la variable X_β^2 en el espacio de dimensión reducida vendrán dadas por las filas q_β de la matriz Q:

$$Q(s) = (PU)s \tag{27}$$

donde la fila p_β $(\beta = 1, \ldots, n)$ vendrá dada por:

$$p_\beta = (K_\beta \frac{\sigma_{11}}{\sigma_{\beta 1}}s + C_{\beta 1}, \ldots, K_\beta\, s, \ldots, K_\beta \frac{\sigma_{n1}}{\sigma_{\beta 1}}s + C_{\beta n})$$

Se comprueba fácilmente que las pendientes de las rectas correspondientes a X_β^2 coinciden con las obtenidas anteriormente para X_β, aunque los sentidos de ambas no tienen porqué coincidir.

4. CORRELACIÓN ENTRE LAS VARIABLES ALEATORIAS Y RECTAS EN LA VARIEDAD

Con el objecto de interpretar mejor las influencias de las variables $(X_1, \ldots, X_n)$ en las diferencias y analogías de las poblaciones estudiadas, conviene calcular el coeficiente de correlación entre las variables aleatorias observadas $X = (X_1, \ldots, X_n)'$ y las variables aleatorias asociadas a las direcciones de las rectas de la variedad, en especial las direcciones de los ejes.

Si una determinada dirección esta asociada a la variable aleatoria V con

$$V = v_1 Z_1 + \ldots + v_n Z_n$$

teniendo en cuenta (15), podemos expresar la variable aleatoria V como:

$$v'\Sigma^{-1}(X-\mu)$$

donde $v = (v_1, \ldots, v_n)'$. por lo tanto la correlación entre X_i y V vendrá dada por

$$\rho_i = \frac{v_i}{\sqrt{\sigma_{ii}}\,||v||}$$

siendo $\sigma_{ii} = var(X_i)$ y $||v||^2 = v'\Sigma^{-1}v$

En el caso de que la variable aleatoria V_i sea la que corresponde al eje de coordenadas $e_i = (0, \ldots, 1, \ldots, 0)'$ de la nueva representación, tendremos que por (2) las componentes de V_i respecto de la base $Z_1, \ldots, Z_n$ vienen dadas por el vector

$$v_i = \Sigma U e_i$$

Por lo tanto los coeficientes de correlación entre la variable X_i y V_j será

$$\rho_{ij} = \frac{\Sigma U e_i}{\sqrt{\sigma_{ii}}}$$

matricialmente las correlaciones entre las variables $X_1, \ldots, X_n$ y las $V_1, \ldots, V_n$ vienen dadas por los elementos de la matriz Δ

$$\Delta = D^{\frac{1}{2}}\Sigma U$$

siendo

$$D^{\frac{1}{2}} = diag(\frac{1}{\sqrt{\sigma_{11}}}, \ldots, \frac{1}{\sqrt{\sigma_{nn}}})$$

5. APLICACIÓN MÉDICA

A continuación aplicamos la metodología anteriormente desarrollada, al estudio de 61 niños, 45 atendidos en un servicio de Nefrología Infantil por presentar enfermedades relacionadas con el riñón o las vías urinarias y el resto correspondiente a un grupo control.

A cada paciente se le midieron las variables IgA sérica, IgG, IgE e IgA secretora, IgA secretora/clerance, IgA secretora/gr. de creatinina. Las tres primeras medidas en mg/dl, la IgA secretora en mg/l, la IgA secretora/clerance en mg/clerance y la IgA secretora/gr de creatinina en mg/gr. de creatinina y que

vienen representadas por el vector aleatorio $X = (X_1, X_2, X_3, X_4, X_5, X_6)$, que suponemos se distribuye en cada población según una normal multivariante de vector de medias μ_i y matriz de covarianzas común $\sum_i$ y donde X_1 representa la concentración de IgA sérica; X_2 la de IgG ; X_3 la de IgE, X_4 la de IgA secretora, X_5 la de IgA secretora/clerance y X_6 la de IgA secretora/gr. de cretinina.

Los nombres de las poblaciones estudiadas así como el tamaño de las muestras (n_i) obtenidas de ellas, vienen dados en la tabla 1.

Tabla 1

Población	Enfermedad	N^{o} de individuos
(1)	Control	16
(2)	Hematurias	6
(3)	Inf. agudas del tracto urinario	9
(4)	Inf. crónicas del tracto urinario	7
(5)	Enfermos con patología renal	7
(6)	Sindrome Nefrótico	6
(7)	Otras enfermedades	10

Los vectores medios μ_i $(i = 1, \ldots, 7)$ son substituidos por sus estimadores $\bar{x}_i = (\bar{x}_{i1}, \bar{x}_{i2}, \bar{x}_{i3}, \bar{x}_{i4}, \bar{x}_{i5}, \bar{x}_{i6})$, siendo

$$\bar{x}_{ij} = \frac{1}{n_i} \sum_{h=1}^{n_i} x_{ijh} \tag{28}$$

y donde x_{ijh} es la h observación de la variable j en la población i.

Estas estimaciones de los vectores medios vienen dadas en la tabla 2.

Tabla 2

Población	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$	$\bar{x}_5$	$\bar{x}_6$
(1)	118.8	873.1	145.7	0.1516	1.946	0.2610
(2)	152.8	1212.0	222.0	0.7150	6.617	1.283
(3)	167.6	1198.0	214.8	3.812	36.32	5.069
(4)	213.1	1015.0	206.9	2.293	29.71	4.427
(5)	165.3	1267.0	216.3	4.839	75.19	6.671
(6)	289.8	1422.0	249.3	4.603	31.35	8.979
(7)	120.4	832.7	152.0	0.3910	5.045	1.145

Sean ahora Σ_i las matrices de covarianzas poblacionales de la variable X en la población P_i $(i = 1, \ldots, 7)$ y sean S_i las estimaciones máximo verosímiles de las Σ_i, cuyos elementos (j, h) vienen dados por:

$$s^i_{jh} = \frac{1}{n_i} \sum_{r=1}^{n_i} (x_{ijr} - \bar{x}_{ij})(x_{ihr} - \bar{x}_{ih}) \qquad j, h = 1, \ldots, 6 \tag{29}$$

Si todas las poblaciones tienen la misma matriz de covarianzas una estimación insesgada de Σ, la matriz de covarianzas común a todas las poblaciones, es:

$$\hat{S} = \frac{1}{61 - 7} \sum_{i=1}^{7} n_i S_i \tag{30}$$

y que viene dada en la tabla 3.

Tabla 3

$$\hat{S} = \begin{pmatrix} 5391. & 6682. & -240.8 & 7.130 & -119.8 & -8.743 \\ & 64760. & 2812. & -54.87 & 276.1 & -155.8 \\ & & 4881. & -32.63 & -213.9 & -50.02 \\ & & & 1.112 & 8.006 & 1.055 \\ & & & & 103.7 & 7.726 \\ & & & & & 3.246 \end{pmatrix}$$

Para que tenga sentido la aplicación del ACP al menos dos poblaciones deben ser diferentes. Para comprobarlo vamos a realizar el contraste de hipótesis:

$$\begin{aligned} H_0 &: \mu_i = \mu_j \\ H_1 &: \exists i, j \in \{1, \ldots, 7\} / \quad \mu_i \neq \mu_j \end{aligned}$$

Este contraste puede realizarse a través de la razón Λ de Wilks:

$$\Lambda = \frac{|W|}{|W + B|} \tag{31}$$

siendo W la matriz de dispersión dentro de grupos y B la matriz de dispersión entre grupos que vienen dadas en la tabla 4.

Tabla 4

$$B = \begin{pmatrix} 1644 \times 10^2 & 4886 \times 10^2 & 98110 & 4276. & 33630 & 8037. \\ & 2484 \times 10^3 & 4404 \times 10^2 & 19930 & 1969 \times 10^2 & 31200 \\ & & 85470 & 3563. & 34930 & 5756. \\ & & & 217.7 & 2437 & 327.0 \\ & & & & 33300 & 3390 \\ & & & & & 528.8 \end{pmatrix}$$

$$W = \begin{pmatrix} 2911 \times 10^2 & 3608 \times 10^2 & -13000 & 385.0 & -6469 & -472.1 \\ & 3497 \times 10^2 & 1519 \times 10^2 & -2963 & 14910 & -8413 \\ & & 2636 \times 10^2 & -1762 & -11550 & -2701 \\ & & & 60.03 & 432.3 & 56.98 \\ & & & & 5600 & 417.2 \\ & & & & & 175.3 \end{pmatrix}$$

El valor de Λ en nuestro caso es 0.0103 equivalente a un valor de 11.0921 de una distribución F de Fisher-Snedecor con 36 y 217 grados de libertad, Rao (1951). Con lo que se concluye que las poblaciones son diferentes y por tanto, tiene sentido realizar el Análisis Canónico de Poblaciones.

A continuación calculamos los valores propios de R respecto de S. En la práctica la matriz R es substituida por la matriz:

$$A = \bar{X}'\bar{X} \tag{32}$$

donde

$$\bar{X} = (\bar{x}_{ij} - \bar{x}_{.j}) \qquad i = 1, \ldots, 7; \quad j = 1, \ldots, 6$$

siendo

$$\bar{x}_{.j} = \frac{1}{7} \sum_{i=1}^{7} \bar{x}_{ij}$$

Los valores propios de A respecto a S y los porcentajes de dispersión acumulados vienen dados en la tabla 5.

Tabla 5

Valores propios	Porcentaje acumulado
57.14	57.40
38.46	96.03
2.87	98.92
0.84	99.77
0.20	99.97
0.03	100

Observamos que con los dos primeros ejes se representa el 96.03 % de la dispersión, por lo que son suficientes para representar las poblaciones. Las componentes de las dos primeras variables canónicas, es decir, los vectores aleatorios propios (U_i) asociados a los dos mayores valores propios se muestran en la tabla 6.

Tabla 6

U_1	U_2
0.4121×10^{-2}	-0.7593×10^{-2}
0.4557×10^{-3}	0.4009×10^{-2}
0.8040×10^{-2}	0.3824×10^{-2}
0.2803×10^{-1}	1.477
0.7229×10^{-1}	-0.1820
0.2760	0.3099

Las correlaciones entre las variables observadas, $X_1, X_2, X_3, X_4, X_5, X_6$, y las dos primeras variables canónicas vienen dados en la tabla 7.

Tabla 7

	U_1	U_2
X_1	0.1696	0.1983
X_2	0.2164	0.1574
X_3	0.1338	0.1004
X_4	0.6099	0.1068
X_5	0.7626	-0.3401
X_6	0.5410	0.2265

De las correlaciones obtenidas se aprecia que las variables relacionadas con IgA secretora X_4, X_5, X_6 son las que más influyen en la discriminación.

Las coordenadas de las poblaciones en el espacio de dimensión 2, es decir, las coordenadas canónicas son:

Tabla 8

Población	Coordenadas	
(1)	2.275	3.106
(2)	3.819	4.797
(3)	7.094	4.940
(4)	6.438	2.592
(5)	10.41	0.1765
(6)	8.720	8.328
(7)	2.789	3.019

La matriz de interdistancias entre las poblaciones viene dada en la tabla 9.

Tabla 9

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1)							
(2)	2.5145						
(3)	5.2560	3.5147					
(4)	4.3825	3.8611	3.2646				
(5)	8.6545	8.0767	5.8886	4.8741			
(6)	8.3144	6.1687	4.0706	6.2229	8.3596		
(7)	0.6620	2.4169	4.8928	3.8145	8.1684	7.9679	

Los círculos de centros las coordenadas canónicas y radios $R_i = \frac{R_\epsilon}{\sqrt{n_i}} \quad i = 1, \ldots, 7$ definen regiones confidenciales al $100(1-\epsilon)\%$ (Cuadras, 1991), siendo:

$$R_\epsilon^2 = F_\epsilon \frac{(61-7)\,6}{(61-7-6+1)} \tag{33}$$

y donde F_ϵ es el valor tal que $P(F > F_\epsilon) = \epsilon$ siendo F una variable aleatoria que se distribuye según una F de Fisher-Snedecor con 4 y (61-7-6+1) grados de libertad.

Los radios de los círculos de confianza con coeficiente de confianza 0.95 se dan en la tabla 10.

Tabla 10

Población	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Radio	0.9729	1.589	1.297	1.471	1.471	1.589	1.231

La representación de las variables $X_\alpha\,(\alpha = 1, \ldots, 6)$ en el espacio de dimensión reducida, viene dada por las filas de la matriz

$$Y(t) = (\hat{S}U)t$$

que resulta de substituir en la ecuación (22) la matriz Σ por su estimación máximo verosimil y U la matriz cuyos vectores columna son U_1 ,U_2.

En nuestro caso la matriz $Y(t)$ fue

Tabla 11

Variables	Componentes	
(1)	12.4517	14.5586
(2)	55.0766	40.0639
(3)	9.3494	7.0007
(4)	0.6431	0.1134
(5)	7.7656	-3.4557
(6)	0.9748	0.4085

Por lo tanto las pendientes de las rectas que representan a las variables vienen dadas en la tabla 12.

Tabla 12

Variables	X_1	X_2	X_3	X_4	X_5	X_6
Pendientes	1.1992	0.7274	0.7488	0.1763	-0.4450	0.4191

La representación gráfica simultánea de las variables y las poblaciones viene dada en la fig. 2.

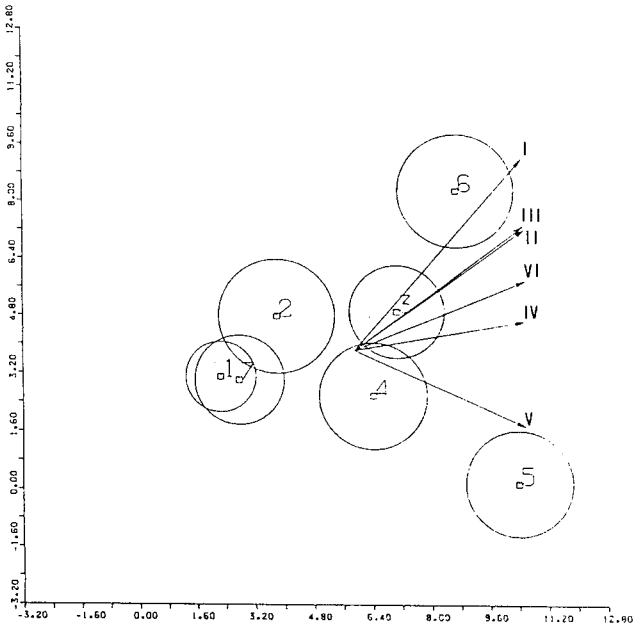

Figura 2. Representación mediante las dos primeras variables canónicas de las poblaciones de enfermos renales $(1, \ldots, 7)$ y de las variables estudiadas $(I, \ldots, VI)$.

De esta representación podemos hacer las siguientes interpretaciones

1. Existe cuatro grupos de poblaciones perfectamente diferenciados, uno constituido por la población control, la constituida por pacientes que presentaban hematurias sin otra complicación añadida y los enfermos asignados a otras enfermedades no renales (poblaciones 1, 2 y 7). Otro grupo al que pertenecen dos poblaciones de enfermos, una con infecciones renales agudas y otra con infecciones crónicas, sin patología renal añadida (poblaciones 3 y 4). Un tercer grupo formado por los pacientes con infecciones crónicas y con patología renal añadida (población 5) y finalmente otro grupo de enfermos con Síndrome Nefrótico (población 6)

2. Las diferencias entre las poblaciones (1, 2, 7), las poblaciones (3, 4) y la población (5) son debidas a la variable X_5. Mientras que las diferencias entre población (6) y las poblaciones (1, 2, 7) y (3, 4) son debidas a las demás variables.

3. Las diferencias entre la población (5) y la población (6) son debidas a todas las variables estudiadas.

Por lo dicho anteriormente se deduce que la variable X_5 es la que más influye en las diferencias entre los grupos, hecho éste esperado si tenemos en cuenta la correlación entre la variable X_5 y el eje de abscisas, considerado como variable aleatoria, es la mayor.

6. DISCUSIÓN Y CONCLUSIONES

Los métodos que permiten realizar una representación simultánea de individuos estadísticos y variables sobre ellas observados es el método de Análisis de Correspondencias y el método Biplot, de éste último cabe señalar que la representación de los individuos obtenida por el método JK'-Biplot coincide con la obtenida por el análisis de componentes principales.

El método que aquí desarrollamos es un método más general pues es utilizable, con más o menos dificultad técnica, siempre que tengamos individuos estadísticos representados en una variedad con una geometría definida.

En este artículo lo hemos aplicado al caso de tener poblaciones y variables aleatorias definidas sobre ellas que se distribuyen según una normal multivariante y con matriz de covarianzas común y como distancia definida entre ellas la de Mahalanobis.

Los resultados son aplicados a la representación de poblaciones constituida por enfermos asistidos en un servicio de nefrología pediátrica y un grupo control que nos permiten, no solo visualizar las diferencias entre las poblaciones estudiadas, sino también, las variables que influyen en estas diferencias.

AGRADECIMIENTOS

A la Dra. Beatriz Ramos por habernos autorizado a utilizar los datos que aparecen en la aplicación médica de este trabajo.

7. BIBLIOGRAFÍA

[1] **Amari S.** (1985). "Differential Geometrical Methods in Statistics". Springer-Verlag, New York.

[2] **Arenas C., Cuadras C.M. y Fortiana J.** (1991). "Multicua" *Pub. Dep. Estadística,* **n. 4**. Barcelona.

[3] **Benzecri, J.P.** (1976). *L'analyse des données.* Tomo II. L'analyse des correspondances. Dunod. París.

[4] **Burbea, J.** (1986). "Informative geometry of probability spaces". *Exposition Math.,* **4**, 347–378.

[5] **Cuadras, C.M.** (1974). "Análisis discriminante de funciones paramétricas estimables". *Trab. Estad. Inv. Oper.,* **25** (3), pp 3–31.

[6] **Cuadras, C.M.** (1991). *Métodos de análisis multivariante.* (2^a Edición) P.P.U. Barcelona, 1991.

[7] **Gabriel, K.R.** (1971). "The biplot graphic display of matrices with application to principal component analysis". *Biometrika,* **58**, pp 453–464.

[8] **Gabriel, K.R.** (1981a). "Biplot. Encyclopedia of Statistical Sciencies". S. Kotz, N.L. Johnson, Ed. Wiley, New York.

[9] **Gabriel, K.R.** (1981b). "Biplot Display of Multivariate Matrices for Inspeccion of Data and Diagnosis". Ed. Wiley, London 1981.

[10] **Hicks, N.J.** (1965). "Notes on differential geometry". Van Nostran, Princenton.

[11] **Mahalanobis, P.C.** (1936). "On the generalized distance in statisties". *Proc. Natl. Inst. Sei. India,* **2** (1), 49–55.

[12] **Oller, J.M.** (1989). "Some geometrical aspects on data analysis and statistics". *Statistical Data Analysis and Inference.* J. Dodge Ed. 41–58. North-Holland, Amsterdam.

[13] **Pearson, K.** (1901). "On lines and planes of closest fit to systems of points in space". *Phil Mag.,* Ser. 6, **2** (11), 559–572.

[14] **Pearson, K.** (1926). "On the coefficient of racial likeness". *Biometrika,* **18**, 105–117.

[15] **Rao, C.R.** (1948). "The utilization of Multiple Measurements in Problems of Biological Classification". *J. Roy. Stat. Soc.,* **B10** (2), 159–203.

[16] **Rao, C.R.** (1951). "An asymptotic expansion of the distribution of Wilks's criterion". *Bull. Inst. Inter. Statist.,* **XXXIII** (2), pp 1771–180.

[17] **Ríos, M. y Oller, J.M.** (1986). "Análisis Canónico de Poblaciones". *Publicaciones de Bioestadística,* **N°21** Barcelona.

[18] **Ríos, M., Villarroya, A., Azuara, A. y Oller, J.M.** (1991). "Representación de Caracteres y poblaciones en el modelo multinominal". Proceeding del XIX Congreso Nacional de Estadística, Investigación Operativa e Informática, pág. 145–146. Segovia, 1991.

[19] **Seber G.A.F.** (1984). *Multivariate Observations.* Wiley, New York.