PESQUIMAT, Revista de la Fac.CC.MM. de la
UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Vol.III,$N°$1,pag.29-39,LIMA-PERÚ. Julio 2000

DISTANCIA ENTRE POBLACIONES NORMALES QUE TIENEN ESTRUCTURA COMÚN DE COVARIANZAS

Doris Gomez Ticeran
e-mail: yakov@terra.com.pe

Universidad Nacional Mayor de San Marcos
Facultad de Ciencias Matemáticas

ABSTRACT Se presenta el resultado teórico Krazanowski(1996) y una aplicación de una manera de establecer Distancias entre Poblaciones Normales Multivariantes cuyas matrices de covarianzas satisfacen el "Modelo de Componentes Principales Comunes(Flury, 1988)". Según el Modelo de Componentes Principales Comunes (Flury,1988), los autovalores de las matrices de covarianzas de las poblaciones concurrentes son iguales pero los autovectores son arbitrarios.

1. INTRODUCCIÓN

En muchas situaciones donde se realizan análisis estadísticos, una manera de describir distancias entre poblaciones normales multivariantes es la Distancia de Mahalanobis (1936). Se sabe que la Distancia de Mahalanobis es usada para establecer distancias entre grupos donde las estructuras de los vectores de medias de las poblaciones concurrentes son diferentes, pero las estructuras de las matrices de covarianzas deben de ser iguales.

Sin embargo, con mucha frecuencia, no siempre es sostenible la suposición de que las matrices de covarianzas de las poblaciones concurrentes son iguales, puesto que es poco probable que dos o más poblaciones tengan estructuras de covarianzas iguales. Este es un problema con el que se enfrentan investigadores de las otras áreas y por facilidad y/o desconocimiento de propuestas alternativas abordan el problema usando la Distancia de Mahalanobis. En éste tipo de situaciones, calcular la Distancia de Mahalanobis como parte del análisis estadístico es irreal y conduce a conclusiones equivocadas, por lo tanto se hace necesaria la búsqueda de otra medida para establecer distancias entre grupos.

Una alternativa es la Distancia de Rao (Krzanowski, 1996), que emula la Distancias de Mahalanobis en poblaciones normales cuyas estructuras de covarianzas se ajustan al Modelo de Componentes Principales Comunes(Flury,

1988). Sin embargo, aún no existe ningún desarrollo teórico para establecer distancias entre poblaciones normales multivariantes cuyos vectores de medias y matrices de covarianzas son completamente arbitrarios.

En el contexto descrito, se presenta el desarrollo teórico y una aplicación al problema de establecer distancias entre poblaciones que tienen estructuras de covarianzas que siguen el Modelo de Componentes Principales Comunes-CPC (Flury, 1988).

2. PLANTEAMIENTO

- Se tiene una familia de distribuciones para el vector aleatorio $\vec{x}$ donde: $f(\vec{x}\ /\Theta)$: función de densidad de probabilidad de la familia, $\Theta = (\vartheta_1, \ldots, \vartheta_r)'$: vector de parámetros en el espacio paramétrico Θ, que satisface las condiciones de regularidad estadística; $g(\Theta)$: matriz de información (definida positiva), cuyos elementos tienen la forma:

$$g_{ij}(\Theta) = E\left\{\frac{\partial}{\partial\vartheta_i} lnf(x/\Theta)\frac{\partial}{\partial\vartheta_j} lnf(x/\Theta)\right\}, \quad j = 1, \ldots, r.$$

- La Distancia de Rao entre las poblaciones identificadas por los parámetros $s(\Theta_i\Theta_j)$ es :

$$d_S^2 = \sum_{i.j=1}^{r} g_{ij}(\Theta)\, d\vartheta_i\, d\vartheta_j.$$

- En el presente trabajo, para una población, la familia paramétrica considerada es la Familia de Distribuciones Normales Multivariantes con parámetros:

$$\Theta' = (\vec{\mu}\ ', [vech\Sigma]'), \quad \text{donde:}$$

$\vec{\mu}\ ' = (\mu_1, \ldots, \mu_p)$ es el vector de medias; $\Sigma = (\sigma_{ij})$ es la matriz de dispersiones de orden $p \times p$; $vech(\Sigma) = (\sigma_{11}, \sigma_{12}, \ldots, \sigma_{1p}, \sigma_{22}, \ldots, , \sigma_{2p}, \sigma_{33}, \ldots, \sigma_{3p}, \ldots, \sigma_{(p-1)(p-1)}, \sigma_{(p-1)p}, \sigma_{pp})'$ es el vector colum- na con los $p(p+1)/2$ elementos de la matriz Σ.

- Siguiendo el mismo planteamiento, dos miembros de esa familia Normal Multivariante, con vectores de medias y matrices de covarianzas diferentes, se representan mediante sus correspondientes parámetros, es decir: $(\vec{\mu}_1, \Sigma_1)$ y $(\vec{\mu}_2, \Sigma_2)$
- Cabe observar, que un aspecto muy importante previo a la obtención de la Distancia de Rao, es saber si estamos frente a muestras que provienen de poblaciones con iguales o diferentes estructuras de covarianzas. En ese sentido, hay necesidad de hacer algunas pruebas de hipótesis que permitan saber si nos enfrentamos a poblaciones cuyas

matrices de covarianzas son diferentes o iguales, y según el caso escoger la Distancia adecuada. Así:

- Ha : Será la hipótesis más general donde se supone que no existe ninguna restricción para los elementos de las matrices de covarianzas.
- Ho: Es la situación más restrictiva, en la que se postula que las matrices de covarianzas son iguales. Es decir, Ho:$\Sigma_1 = \Sigma_2$.
- Hc: Es el caso intermedio, especificado por el Modelo de Componentes Principales Comunes (Flury, 1988) corresponde a la hipótesis Hc:$\Sigma_i = \Gamma\,\Lambda_i\Gamma'$ para todo i, donde Γ es una matriz ortogonal y las matrices son diagonales.

- Cabe observar que tomando $\Lambda_i = \Lambda$ para todo i, se genera la hipótesis Ho; mientras que al considerar para cada Σ un Λ arbitrario se reproduce la hipótesis Ha. Así, $Ho \subset Hc \subset Ha$.
- En el contexto descrito, la Distancia de Rao será desarrollada y aplicada en situaciones donde las poblaciones siguen el modelo de Componentes Principales Comunes, es decir cuando la hiptesis Hc es verdadera.

3. MÉTODO

Supongamos que tenemos g-poblaciones normales multivariantes π_1, π_2, $\ldots$, π_g y que las distribuciones son normales multivariantes, es decir: $\vec{X} \sim N((\vec{\mu}_i, \Sigma_i)$ en $\pi_i = 1, 2, \ldots, g$.

En el Modelo de Componentes Principales Comunes las matrices de covarianzas de las poblaciones tienen la siguiente estructura: $\Sigma_i = \Gamma\,\Lambda_i\,\Gamma'$ para todo i. Observamos que estamos frente a la estructura de covarianzas de un nuevo vector que es el transformado del vector original. Es decir: $\vec{Z} = \Gamma' X \sim N((\vec{\nu}_i, \Lambda_i)$ en π_i con $(\vec{\nu}_i = \Gamma'\ \vec{\mu}_i)$. Este aspecto es considerado y analizado en su verdadera dimensión para dar solución al problema planteado.

La observación anterior es suficiente para reemplazar y usar el vector $\vec{Z}$ en lugar del vector X. Así, en lugar de encontrar la distancia de Rao en la familia de distribuciones $f(x/\Theta)$ se puede encontrar dicha distancia en la familia de distribuciones $f(\vec{x}\ /\Theta)$. Es decir, lo que se debe de encontrar es la Distancia de Rao para la familia de distribuciones normales multivariantes , $N(\vec{\nu}, \Lambda)$ donde:

$$\vec{\nu} = (\nu_1, \ldots, \nu_p) \quad \text{y} \quad \Lambda = diag\,(\delta_1^2, \ldots, \delta_p^2).$$

Así, si tomamos Z en lugar de X, en una única población, la función de densidad toma la siguiente forma:

$$f(\vec{z}\ /\Theta) = \frac{1}{(2\pi)^{p/2}\partial_1,\ldots\ldots,\partial_p} \quad e^{-\frac{1}{2}\sum_{i=1}^{p}\left(\frac{z_i-\nu_i}{\partial_i}\right)^2}$$

Para formar la matriz de información hagamos lo siguiente:

- para los elementos ν_i tenemos:

$$g_{jj}(\Theta) = E\left(\frac{d}{d\,\nu_j} lnf(x/\Theta)\frac{d}{d\,\nu_j} lnf(x/\Theta)\right) = \frac{1}{\partial_j^2}$$

- para los ∂_j tenemos:

$$g_{kk}(\Theta) = E\left(\frac{d}{d\,\partial_j} lnf(x/\Theta)\frac{d}{d\,\partial_j} lnf(x/\Theta)\right) = -\frac{2}{\partial_j^2}$$

- y para los elementos fuera de la diagonal:

$$\begin{aligned} g_{jk}(\Theta) &= E\left(\frac{d}{d\,\nu_j} lnf(x/\Theta)\frac{d}{d\,\partial_k} lnf(x/\Theta)\right) \\ &= E\left(\frac{1}{\partial_j^2}(z_j-\nu_j)\right)\left(-\frac{1}{\partial_j}-\frac{1}{\partial_j^3}(z_k-\nu_k)\right) \\ &= 0 \end{aligned}$$

A partir de estos resultados la matriz de información tiene la siguiente estructura:

$$\mathrm{g}(\Theta) = \begin{pmatrix} \frac{1}{\partial_1^2} & 0 & \ldots & 0 & 0 & \ldots & \ldots & 0 \\ 0 & \frac{1}{\partial_2^2} & \ldots & 0 & 0 & \ldots & \ldots & 0 \\ 0 & 0 & \ldots & \frac{1}{\partial_p^2} & 0 & \ldots & \ldots & 0 \\ 0 & 0 & \ldots & 0 & \frac{2}{\partial^2} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & \frac{2}{\partial_p^2} \end{pmatrix}$$

Luego, la Distancia de Rao para el caso planteado toma la siguiente forma:

$$\begin{aligned} d\,S^2 &= \sum_{i,j=1}^{r} g_{ij}(\Theta) d\,\theta_i\, d\,\theta_j \\ & \frac{1}{\partial_1^2}(d\nu_1)^2 + \ldots + \frac{1}{\partial_p^2}(d\,\nu_p)^2 + \frac{2}{\partial_1^2}(d\,\partial_1)^2 + \ldots + \frac{2}{\partial_p^2}(d\,\partial_p)^2 \end{aligned}$$

El resultado obtenido en el paso anterior, llevado a dos poblaciones normales multivariantes con representación $\vec{x_1} \sim N(\vec{\mu_1}, \Sigma_1)$ y $(\vec{\mu_2}, \Sigma_2)$, cuyos vectores de parámetros son:

$$\begin{aligned} \Theta_1' &= (\nu_{11}, \ldots, \nu_{1p}, \partial_{11}^2, \ldots, \partial_{1p}^2) \\ \Theta_2' &= (\nu_{21}, \ldots, \nu_{2p}, \partial_{21}^2, \ldots, \partial_{2p}^2) \end{aligned}$$

es la Distancia de Rao entre esas poblaciones con coordenadas Θ_1 y Θ_1', definida por:

$$dS^2 = \sum_{j=1}^{p} \frac{(d\,\nu_j)^2 + 2(d\,\partial_j)^2}{\partial_j^2}.$$

Desarrollando y usando la fórmula (8.1) de la página 80 de Iversen (1992), se obtiene:

$$d\{(\nu_{11}, \partial_{11}), (\nu_{21}, \partial_{21})\} = d_1 = \sqrt{2} cosh^{-1} \left[\frac{(\nu_{ll} - \nu_{21})^2 + 2\partial_{11}^2 + 2\partial_{21}^2}{4\partial_{11}\partial_{21}} \right]$$

Para el caso general p, la distancia requerida es:

$$dS_j^2 = \frac{[(d\,\nu_j)^2 + 2(d\,\partial_j)^2]}{\partial_j^2}$$

para todo j.

Finalmente, la Distancia de Rao entre las poblaciones 1 y 2, p-variantes es:

$$d\{(\nu_1, \Delta_1), (\nu_2, \Delta_2)\} = \sqrt{d_1^2 + d_2^2 + \ldots + d_p^2}$$

donde:

$$d_j = \sqrt{2} cosh^{-1} \left[\frac{(\nu_{1j} - \nu_{2j})^2 + 2\partial_{1j}^2 + 2\partial_{2j}^2}{4\partial_{1j}\partial_{2j}} \right].$$

Se sigue el mismo esquema para encontrar la distancia entre éstas y las otras poblaciones.

4. APLICACIÓN

La aplicación de la presente metodología se ha realizado en el campo de Biología [2]. Así, se dispone de muestras aleatorias de tamaños 173, 141, 88 y 76 de 4 poblaciones de roedores de las siguientes especies:

Población 1: Roedores machos de la especie *Microtus Californicus*;

Población 2: Roedores hembras de la especie *Microtus Californicus*;

Población 3: Roedores machos de la especie *Microtus Ochrogaster*;

Población 4: Roedores hembras de la especie *Microtus Ochrogaster*.

Se midieron las siguientes caractersticas:

X_1: Longitud del cráneo de los roedores,

X_2: Ancho del cráneo de los roedores,

X_3: Altura del cráneo de los roedores.

1. Al realizar el análisis descriptivo de los datos originales se encontró que existía mucha variabilidad entre las variables, probablemente debido a que el factor edad de los roedores no había sido controlado, por lo que se consideró conveniente trabajar con los logaritmos de las variables anteriores. En el siguiente cuadro se presentan las matrices de covarianzas y los vectores de medias de los 4 grupos.

Cuadro 1
Vectores de Medias y Matrices de Covarianzas de las Tres Variables en cada uno de los grupos

Especie	*Machos*			*Hembras*		
M. Californicus Medias	328.80	271.9	237.1	326.3	269.7	235.1
Matriz de Covarianzas	112.01	106.64	52.97	86.08	81.66	40.24
	106.64	108.13	54.73	81.66	85.54	42.08
	52.97	54.75	33.86	40.24	42.08	26.66
M. Ochrogaster Medias	324.70	266.6	231.3	323.2	265.7	231.7
Matriz de Covarianzas	65.4	60.23	24.69	88.66	79.11	41.32
	60.23	62.27	23.47	79.11	80.57	38.81
	24.69	23.47	16.33	41.32	38.81	23.81

2. La hipótesis nula que se plantea es que no existe diferencias entre las matrices de covarianzas:

$$\mathbf{H_0} = \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4.$$

- Según el test de razón de verosimilitud [7], se usa la estadística: $-2ln\lambda = n\,ln|\hat{\Sigma}| - \sum_{i=1}^{g} n_i\,ln|\hat{\Sigma}_i|$, para tomar la decisión de aceptar o rechazar la hipótesis planteada. Para n suficientemente grande $-2\,ln\,\lambda \sim ji - cuadrado\left(\frac{1}{2}p(p-1)(g-1)\right)$.

- La hipótesis planteada fue rechazada puesto que el valor muestral de $-2ln\lambda = 54.47$ fue altamente significativo al 5 % y $\frac{1}{2}p(p-1)(g-1) = \frac{1}{2}(3)(2)(3) = 9$ grados de libertad.

- Puesto que la hipótesis de igualdad de matrices de covarianzas ha sido rechazada, la Distancia de Mahalabobis no es una métrica para cuantificar diferencias entre los grupos. Por consiguiente, se hace necesario utilizar otra métrica para cuantificar las diferencias entre los grupos.

En el siguiente paso se analiza la posibilidad de que los datos se ajusten al Modelo de Componentes Principales Comunes (Fleury, 1988).

3. Se postula la hipótesis de que se cumple Modelo de Componentes Principales Comunes, es decir:

$$\mathbf{H_c} : \Sigma = \Gamma\Delta_i\Gamma' \quad \text{para todo} \quad i = 1, \ldots, 4.$$

Aplicando la metodología de Componentes Principales Comunes se encuentra que la estadística de la razón de verosimilitud para docimar la hipótesis planteada es:

$$X^2_{cpc} = \sum_{l}^{g} n_i \left[log\left(\frac{det(\text{diagonal}F_1)}{det(F_1)}\right)\right],$$

donde:

$$\mathbf{F_i} = (\hat{\Delta}_i) \quad \text{y} \quad \hat{\Delta}_\mathbf{i} = \hat{\Gamma}' S_i \hat{\Gamma}.$$

El valor de la estadística muestral fue $X^2_{cpc} = 15.6516$ mientras que el valor teórico para $\frac{1}{2}p(p-1)(g-1) = \frac{1}{2}(3)(2)(3) = 9$ grados de libertad y con 5 % de nivel de significación fue 16.92. Finalmente, se acepta la hipótesis de que las matrices de covarianzas de las poblaciones concurrentes siguen el modelo de Componentes Principales Comunes.

Puesto que se acepta la hipótesis de Componentes Principales Comunes, la Distancia de Rao será la métrica usada para cuantificar las diferencias entre los grupos.

4. Obtención de la Distancia de Rao

Previo al cálculo de la Distancia de Rao entre dos poblaciones, se presentan las siguientes estimaciones de los parámetros.

- Vectores de medias muestrales transformados:

$$\hat{\nu}_l = \hat{\Gamma}'\overline{X}_1 = \hat{\Gamma}' \begin{bmatrix} 328.8 \\ 271.9 \\ 237.1 \end{bmatrix} = \begin{bmatrix} 479.5232 \\ 60.4424 \\ -67.838 \end{bmatrix} = \begin{bmatrix} \hat{\nu}_{11} \\ \hat{\nu}_{12} \\ \hat{\nu}_{13} \end{bmatrix}$$

$$\hat{\nu}_2 = \begin{bmatrix} 475.7243 \\ 59.8022 \\ -67.3435 \end{bmatrix} = \begin{bmatrix} \hat{\nu}_{21} \\ \hat{\nu}_{22} \\ \hat{\nu}_{23} \end{bmatrix}$$

$$\hat{\nu}_3 = \hat{\Gamma}'\overline{X}_3 = \hat{\Gamma}' \begin{bmatrix} 324.7 \\ 266.6 \\ 231.3 \end{bmatrix} = \begin{bmatrix} 471.3413 \\ 57.0848 \\ -67.2972 \end{bmatrix}$$

$$\hat{\nu}_4 = \begin{bmatrix} 469.868 \\ 58.1096 \\ -67.3164 \end{bmatrix} = \begin{bmatrix} \hat{\nu}_{41} \\ \hat{\nu}_{42} \\ \hat{\nu}_{43} \end{bmatrix}$$

Cuadro 2.

Varianzas de cada una de las variables transformadas en cada uno de los cuatro grupos

Varianzas Grupos	Variable (1)	Variable (j = 2)	Variable (j = 3)
$\hat{\partial}^2_{1j}$	244.28	7.21	2.52
$\hat{\partial}^2_{2j}$	188.40	6.46	3.43
$\hat{\partial}^2_{3j}$	133.58	6.01	4.43
$\hat{\partial}^2_{4j}$	183.88	3.91	5.42

- Cálculos relacionados con la distancia entre las poblaciones 1 y 2:

$$\frac{(\hat{\nu}_{11} - \hat{\nu}^2_{21} + 2\hat{\partial}^2_{11} + 2\hat{\partial}_{21})^2}{4\hat{\partial}_{11}\hat{\partial}_{21}} =$$

$$= \frac{(479.52 - 475.72)^2 + 2(244.28) + 2(188.4)}{4\sqrt{244.28(188.4)}} = 1.0253$$

$$d_1 = \sqrt{2}cosh^{-1}(1.0253) = 0.3172$$

$$\frac{\hat{\nu}_{12} - \hat{\nu}_{22})^2 + 2\hat{\partial}^2_{12} + 2\hat{\partial}^2_{22}}{4\hat{\partial}_{12}\hat{\partial}_{22}} = 1.0165$$

$$d_2 = \sqrt{2}cosh^{-1}(1.0165) = 0.2565$$

$$\frac{\hat{\nu}_{13} - \hat{\nu}_{23})^2 + 2\hat{\partial}^2_{13} + 2\hat{\partial}^2_{23}}{4\hat{\partial}_{13}\hat{\partial}_{23}} = 1.032$$

$$d_3 = \sqrt{2}cosh_{-1}(1.032) = 0.3567$$

- Finalmente, la Distancia de Rao entre los grupos 1 y 2 alcanza el valor de:

$$d[(\nu_1, \Delta_1)(\nu_2, \Delta_2)] = \sqrt{d_1^2 + d_2^2 + d_3^2} = 0.539$$

- Usando la información muestral que corresponde a los respectivos grupos muestrales y emulando la metodología aplicada para calcular la distancia entre las poblaciones 1 y 2 se genera la siguiente matriz de Distancias de Rao.

Cuadro 3.

Distancias de Rao entre los Grupos de Roedores

Grupo	*M.Californicus* Macho	*M.Californicus* Hembra	*M.ochrogaster* Macho	*M.ochrogaster* Hembra
M.Californicus Macho	0			
M.Californicus Hembra	0.5390	0		
M.Ochrogaster Macho	1.5454	1.1590	0	
M.Ochrogaster Hembra	1.4048	0.9810	0.6246	0

5. CONCLUSIONES

Las muestras provienen de poblaciones con diferentes estructuras de covarianzas, puesto que se rechazó la hipótesis de homogeneidad de matrices de covarianzas.

Se aceptó la hipótesis de que las muestras provienen de poblaciones de roedores con estructuras de covarianzas que siguen el Modelo de Componentes Principales Comunes.

La Distancia de Mahalanobis(Anderson, 1978) que es muy utilizada para calcular distancias entre poblaciones, no pudo ser usada en la presente investigación, puesto que el supuesto en el que se basa -estructuras de covarianzas

iguales- no se cumplió. Por tal razón fue necesario la búsqueda de una distancia diferente.

Observamos que las poblaciones más cercanas son: *M.Californicus* - Hembras con *M.Californicus* Machos; *M.ochrogaster* Hembras con *M.ochrogaster* Machos.
y las poblaciones más alejadas son:
M.ochrogaster Machos con *M.Californicus* Machos.

Se podrían continuar el presente trabajo, tratando de generalizar el trabajo de Gómez(1998), que plantea un método de discriminación en dos poblaciones heteroscedásticas, para 4 poblaciones que siguen el modelo de Componentes Principales Comunes.

BIBLIOGRAFÍA

[1] T.W. Anderson, *Introduction to the Multivariate Analisis.* John and Willey.(1984).

[2] J. Airoldi and B. Flury,*Common Principal Component Analysis to cranial morphometry of Microtus Californicus and M.ochrogaster.* Journal of Zoology. London, in Press.(1988).

[3] B. Flury, *Common Principal Components and Related Models.* New York. John and Willey.(1988)

[4] D. Gómez, *Discriminación Cuadrática basada en Estructura de Covarianzas.* Pesquimat $N°2$. Revista de Investigación de la Facultad de Ciencias Matemáticas. UNMSM.(1998).

[5] B. Iversen, *Hiperbolic Geometry.* Cambridge: University Press.(1992).

[6] W. Krazanowski, *Rao's Distance Between Normal Populations that have Common Principal Components.* Biometrics $N°51$,(1999), p. 1467-71.

[7] W. Krazanowski, *Principles of Multivariate Analysis: A Users Perspective.* Oxford: Clarendon Press.(1988).