PESQUIMAT, Revista de la F.C.M. de la
Universidad Nacional Mayor de San Marcos
Vol. I - N°2 Pgs. 111-116 Lima – Perú Dic. 1998

DISCRIMINACIÓN CUADRÁTICA MEDIANTE MATRICES DE COVARIANZAS

Doris Gómez Ticerán
Universidad Nacional Mayor de San Marcos

RESUMEN

Maximizando y minimizando la función $\frac{a\,S_1\,a}{d\,S_2 a}$, donde $S_1\,y\,S_2$ son las matrices de covarianzas muestrales de dos poblaciones p-variantes, con vectores de medias iguales o diferentes, se consiguen dos combinaciones lineales de las componentes del vector p-variante. Si los dos grupos tienen estructuras de dispersiones diferentes, esas combinaciones son usadas para la clasificación de indivíduos que proceden de una de dos poblaciones normales multivariantes. Este método comparado con el método de clasificación usando el cociente de verosimilitudes (Mardia, 1976), para poblaciones normales simuladas por Monte Carlo con p=2, p=3 y p=4, resultan equivalentes.

INTRODUCCIÓN

Uno de los tópicos de la Estadística Multivariante se presenta con el nombre de Análisis Discriminante o Clasificación, cuyo objetivo principal es ubicar un individuo, objeto o grupo de objetos en una de las poblaciones concurrentes. También los criterios de discriminación frecuentemente son utilizados para la reducción de la dimensión del problema estadístico.

La clasificación consiste en la identificación del grupo al cual pertenece el nuevo individuo llevando en cuenta sus características observadas. Cuando esas características son mediciones numéricas, la designación a los grupos recibe el nombre de Discriminación y las combinaciones de esas mediciones recibe el nombre de Función Discriminante. Más específicamente hablando, en la discriminación se trata de describir de manera gráfica o algebraica con el uso de funciones llamadas discriminantes, los aspectos que sirven para diferenciar o discriminar a los individuos u objetos de varias poblaciones. Cuando esas funciones discriminantes son lineales lo que se tiene son combinaciones de las variables

originales que escogidas convenientemente proporcionan información importante respecto a las poblaciones (Anderson, 1984; Dillon et al, 1984; Seber, 1984 entre otros autores.).

Por otro lado, esas combinaciones simplifican las estructuras de las matrices de covarianzas facilitando la interpretación de los datos.

En problemas prácticos de discriminación y clasificación, los métodos mas usados son el de Fisher (Fisher, 1936); el de razón de verosimilitud (Mardia, 1976) y el de Bayes (Anderson, 1984). Los criterios anteriores proporcionan funciones discriminantes lineales para el caso homoscedástico y funciones discriminantes cuadráticas para el caso heteroscedásticos, en las coordenadas del indivíduo a ser clasificado. Para reducir la dimensión solamente el caso lineal es útil; y cuando la dimensión de las poblaciones es mayor que dos la visualización gráfica de las separatrices de las regiones de clasificación así como de los puntos muestrales de las poblaciones, solo puede realizarse mediante proyecciones en dos o tres dimensiones.

El método que se presenta en el presente artículo no posee esas limitaciones, pues para cualquier dimensión $p \geq 2$, siempre es posible encontrar dos combinaciones lineales de las componentes del individuo que queremos clasificar. Estas combinaciones pueden ser usadas en representaciones gráficas y reducción de dimensión, toda vez que las dispersiones entre las dos poblaciones concurrentes son diferentes.

En el presente contexto, el objetivo del presente trabajo es proponer un criterio de clasificación basado en las matrices de covarianzas de las poblaciones concurrentes.

MÉTODO

Sean Π_1 y Π_2 dos poblaciones normales p-variantes con vectores de medias y matrices de covarianzas μ_g y $\sum_g$, $g = 1, 2$, respectivamente.

Si las matrices de covarianzas de las dos poblaciones son diferentes, entonces, las discrepancias entre los dos grupos pueden ser analizadas a través de las combinaciones lineales de las componentes de $X = (X_1, X_2, ..., X_n)'$ que más la exacerba, esto es, de las combinaciones $\alpha = (d_1, d_2, ..., d_p)$ que maximizan o minimizan el cociente de varianzas:

$$\frac{a' S_1 a}{a' S_1 a} \qquad (1)$$

Estos son los autovectores asociados al mayor y menor autovector de la matriz $\sum_1^{-1} \sum_2$ y son denominados Componentes Principales Generalizados (Flury, 1983).

Para proponer el presente método de discriminación se usan las combinaciones que dan el mayor y el menor cociente de varianzas en (1), para formar transformaciones lineales de R^p en R^2, en cada una de las poblaciones, que expresen las mayores discrepancias entre las poblaciones concurrentes. En el nuevo espacio R^2 pueden ser utilizados el Método de Bayes o el de Razón de Verosimilitud para efectuar análisis discriminante.

Para tal fin, se toman muestras aleatorias de tamaño n_g de las poblaciones $\Pi_g\,(g=1,2)$ y se obtiene los estimadores máximo verosímiles $\hat{\mu}_g = \bar{X}^g$, y $\hat{\Sigma}_g = S_g$ (Anderson, 1984). (2)

Luego, se calculan los autovalores y autovectores de la matriz $S_1^{-1} S_2$ normalizados por $\alpha' S_1 \alpha = 1$, donde los α' son los autovectores asociados a los autovalores λ_i, de la matriz $S_1^{-1} S_2$. Se escogen los autovalores asociados a los autovalores máximos y mínimos: α_{max} y α_{min}

Estos autovectores, α_{max} y α_{min} son tales que $\alpha'_{max} S\, y\, \alpha_{min} = 0$; $g=1,2$ y corresponden al mayor y menor autovalores respectivamente.

Usando el resultado de Flury, se definen combinaciones lineales en el espacio de dimensión 2, en cada una de las dos poblaciones. Con esas acciones se genera una regla de clasificación cuadrática muestral.

A continuación se presenta el resumen de la metodología poblacionalmente.

Sea $A = \left(\alpha'_{max}, \alpha'_{min}\right)$ de orden 2xp, la matriz de la transformación de R^p en R^2; $Y^{(g)} = AX^{(g)}$, $g = 1, 2$

Entonces, si

$\overline{X^g}$ tiene distribución normal p-variante, con vector de medias: $\overline{\Psi_g} = A\overline{\mu_g}$, y matriz de covarianzas: $\phi_g = A\sum_g A$ (en cada uno de los grupos) donde

$$\begin{aligned} \Phi_1 &= I_{2x2} \\ \Phi_2 &= \text{Diagonal}\left(\lambda_{max}, \lambda_{min}\right) \end{aligned} \tag{3}$$

con λ_{max} y $\lambda_{\min}$ mayor y menor autovalores de $\sum_1^{-1} \sum_2$ respectivamente.

Lo que indica la ecuación (3) es que se consigue una transformación con matriz de covarianza circular en Π_1 y elíptica en Π_2 con mayor varianza a lo largo del eje X.

Luego se trabaja en la obtención de la regla de clasificación y se propone la siguiente.

Dada una observación p-variante $\vec{x}$, sabiendo que pertenece a Π_1 a Π_2, con $y = Ax$, la regla de clasificación de Bayes que se propone es:

Clasificar $\overline{X}$ en Π_1 si :

$$\begin{aligned} -\frac{1}{2} y'\left(\Phi_1^{-1} - \Phi_2^{-1}\right)y + y'\left(\phi_1^{-1}\psi_1 - \phi_2^{-1}\psi_2\right) \geq \ln\left(\frac{q_2\, C(1/2)}{q_1\, C(2/1)}\right) - \ln\frac{\phi_2}{\phi_1} \\ + \frac{1}{2}\left(\psi_2'\,\phi_1^{-1}\,\psi_2 \quad - \psi_1'\,\phi_1^{-1}\,\psi_1\right) \end{aligned} \tag{4}$$

Caso contrario, clasificar x en Π_2.

---

donde: q_1 y q_2 son las probabilidades a priori de Π_1 y Π_2; $C(i/j)$ es el costo de clasificar equivocadamente un individuo de Π_j en Π_i y Ln es el logaritmo natural.

Observamos que dentro del espacio transformado por la matriz **A,** la regla (4) propuesta es admisible (Anderson, 1984), por tanto, minimiza el costo esperado de mala clasificación, es decir, es la mejor regla de clasificación.

Cabe señalar que para proponer (4) se supone que los parámetros son conocidos, pero en la práctica se reemplazan los parámetros por los estimadores correspondientes.

SIMULACIONES

Se hicieron muchas simulaciones para p=2,3,4 con la finalidad de comparar los resultados de la regla (4) con el método implementado en los paquetes estadísticos como por ejemplo el SAS.

Se simularon muestras de poblaciones normales esféricas versus esféricas, esféricas versus elípticas, elípticas versus elípticas, con varios ángulos de diferencias entre los respectivos ejes principales. Todas las poblaciones fueron simuladas con los mismos vectores de medias justamente para valuar el efecto de la clasificación basada en las estructuras de las matrices de covarianzas.

Los casos considerados fueron simulados con costos de mala clasificación iguales y probabilidades a priori iguales.

Las muestras simuladas todas fueron de tamaño 100 para cada uno de los dos grupos.

La siguiente Tabla nos muestra algunos resultados.

TABLA

Población 1			Población 2			Método Propuesto		SAS	
σ_1	σ_2	θ^0	σ_1	σ_2	θ^0	m_1	m_2	m_1	m_2
1	1	0	1	1	0	35	37	42	40
1	1	0	1	1	45°	31	56	42	43
4	4	0	2	4	0	53	19	53	19
4	4	0	2	4	45°	42	20	42	20
3	0.5	0	1.75	1.75	45°	16	12	46	12
3	0.5	0	0.5	3	90°	8	10	8	11
8	0.05	0	4.25	4.25	45°	1	0	1	0
8	0.05	0	0.05	8	90°	0	0	0	0

donde

σ_j : Desviación estándar de la coordenada i

θ : Angulo entre el primer eje principal y el de las dos primeras coordenadas (variables generadas antes de la transformación **A**).

m_j : Indivíduos mal clasificados de Π_g $(g = 1,2)$.

CONCLUSIÓN

El método presentado es un nuevo método de discriminación cuadrática para poblaciones normales. Se observa que dentro del espacio transformado por la matriz A, la regla obtenida es admisible y minimiza el costo esperado de mala clasificación.

Este método de discriminación y clasificación tendrá excelente aplicación en situaciones donde las estructuras de covarianzas de las poblaciones involucradas son diferentes. Aplicaciones interesantes se encuentran en el campo de la medicina y es la continuación del presente trabajo.

BIBLIOGRAFÍA

1. ANDERSON, T.W. (1984) An Introduction to multivariate Statistical Analysis. John Wiley & Sons, New York.
2. DILLON *et. al* (1984) Multivariate Analysis :methods and applications. John Wiley & Sons.
3. FISHER, R.A. (1936) The use of multiple measurement in taxonomic problems. In: Atchley,W.R. (1975) Multivariate Statistical Methods: Amoung-Groups Covariation. Dowden Huchinson, Stroudsburg, Pennsylvania.
4. FLURY, B. (1983) Some relations between the comparison of covariance matrices and principal component analysis. Computational Statistics & Data Analysis. 1: 97-109.
5. GÓMEZ, D. (1988) Discriminaçâo de duas polulaçôes multivariadas com base nas matrizes de dispersôes. Tese de Mestrado. IMECC, UNICAMP- Sao Paulo- Brasil.
6. MARDIA, K.V.; KENT, J.T. AND BIBBY,J.M. (1976) Multivariate Analysis. Academic Press, London.
7. SEBER, G. A. (1984) Multivariate Observations. Wiley Series in Probability and Mathematical Statistics. Recibido en octubre de 1998