QÜESTIIÓ, Vol 13, nº 1,2,3 pp. 3-11, 1989

ON A DISTANCE BETWEEN ESTIMABLE FUNCTIONS

C. ARENAS

Universitat de Barcelona

In this paper we study the main properties of a distance introduced by C.M. Cuadras (1974). This distance is a generalization of the well-known Mahalanobis distance between populations to a distance between parametric estimable functions inside the multivariate analysis of variance model. Reduction of dimension properties, invariant properties under linear automorphisms, estimation of the distance, distribution under normality as well as the interpretation as a geodesic distance are studied and commented.

Keywords: Mahalanobis distance. Multivariate parametric estimable functions. MANOVA. Geodesic distance.

1. INTRODUCTION

Cuadras (1974) has generalized to the case of multivariate estimable functions the distance introduced by Mahalanobis. This allows to apply the canonical analysis to the representation of estimable functions. This distance together with techniques of dimension reduction (canonical analysis, principal coordinate analysis) makes clear the interpretation of principal effects in multivariate analysis of variance designs. For applications in Medicine see Ballús and others (1980); in Agriculture see Cuadras, Oller (1982) and other applications in Cuadras (1981).

In this paper we study the main properties of this distance for multivariate estimable functions.

–C. Arenas - Dept. d'Estadística - Facultat de Biologia - Universitat de Barcelona.

–Article rebut el novembre de 1988.

2. MULTIVARIATE ESTIMABLE FUNCTIONS

Consider the multivariate linear model

$$Y = XB + U \tag{1}$$

where $Y = (y^{(1)}, ..., y^{(p)})$ is a kxp-matrix of data with $y^{(j)}$ representing the observations of the variable j. We will suppose that these observations come from k different populations $H_1, ..., H_k$ and that every p-dimensional observation is assimilated to a random vector $Y = (Y_1, ..., Y_p)$ with covariance matrix $\sum$ of rank p, assumed to be the same for the k populations. X is a kxm-design matrix of rank r. $B = (\beta_1, ...\beta_m)^t$ is an mxp-parametric matrix, i.e., each β_i is a p-dimensional parametric vector. Finally the kxp matrix U is the error matrix and it is assumed that $E(U) = 0$.

Let μ_i stand for the mean vector of Y in the population H_i, $i = 1, ..., k$. Then we have

$$\mu = XB \tag{2}$$

where

$$\mu = (\mu_1, ..., \mu_k)^t.$$

In the following F will be the real vector space generated by $Y_1, ..., Y_p$, and μ_i^* stands for the mean function on random vectors of F in the population $H_i (i = 1, ..., k)$.

An estimable function ψ^* (see Cuadras (1974), definition 2.4.1.) is an element of the dual space F^* of F given by a linear combination of the μ_i^*'s, i.e., $\psi^* = d_1\mu_1^* + ... + d_k\mu_k^* = D^t\mu^*$.

If $E\beta$ stands for the real vector space spanned by $\beta_1, ..., \beta_m$ and $E\beta^*$ stands for its dual space then a parametric function is just an element of $E\beta^*$ (see Cuadras (1974), 3.2.).

We recall that a parametric function is said to be estimable if it has a linear unbiased estimate. Otherwise a parametric function, say, $\psi^* = P_1\beta_1^* + ... + P_m\beta_m^* = p^tB^*$ is estimable if the vector p^t belongs to the space spanned by the rows of X. In Cuadras (1974), theorem 3.2.1., it is proved that a parametric function is estimable if and only if it is an estimable function. The relation between the expression $\psi^* = p^tB^*$ and $\psi^* = D^t\mu^*$ is given by $P = X^tD$ (see Cuadras (1974), theorem 3.2.2.). If the matrix X (the reduced design matrix) has full rank every parametric function is estimable. Otherwise a parametric function $\psi = p_1\beta_1 + ... + p_m\beta_m = p^tB$ is estimable if and only if

$$p^t(X^tX)^-X^tX = p^t \tag{3}$$

where $(X^tX)^-$ is a generalized inverse of X^tX.

We consider a sample of size n_1, (resp. $n_2, ..., n_k$) in the population H_1, (resp. $H_2, ..., H_k$) and denote by Δ the diagonal matrix with entries $n_1, ..., n_k$.

Let ψ^* be an estimable function and $\hat{\mu}_i^*$ the usual estimation of the mean ($i = 1, ..., k$). If the Gauss-Markov estimation of ψ^* is given by $\hat{\psi}^* = p^t\hat{\beta}^*$, we say that $\psi^* = \hat{D}^t\hat{\mu}^*$ is the "intrinsic expression" of ψ^*, where $\hat{D} = \Delta X(X^t\Delta X)^- p^t$, (see Cuadras (1974), 3.3.4.).

Suppose that Y has a multivariate normal distribution and consider the random matrix

$$V = \begin{bmatrix} y_{111} & \cdots & y_{p11} \\ \vdots & & \vdots \\ y_{11n_1} & \cdots & y_{p1n_1} \\ \vdots & & \vdots \\ y_{1k1} & \cdots & y_{pk1} \\ \vdots & & \vdots \\ y_{1kn_1} & \cdots & y_{pkn_k} \end{bmatrix}$$

An unbiased estimation of the covariance matrix $\sum$ is

$$\hat{\sum} = \frac{R_o^2}{n-r} = \frac{(V - X\Delta\hat{B}^*)^t(V - X\Delta\hat{B}^*)}{n-r} \tag{4}$$

with $n = n_1 + ... + n_k$ and $\hat{B}^*$ the least square estimation of B^*. In Cuadras (1974), p. 23, it is shown that the Gauss-Markov estimation $\hat{\psi}^*$ of an estimable function ψ with intrinsic expression $\psi^* = d_1\mu_1^* + ... + d_k\mu_k^*$, follows a multivariate normal distribution, $\hat{\psi}^* \sim N_p(\psi^*, \bar{D})$ where $\bar{D} = \sum_{i=1}^{k} d_i^2/n_i$. Moreover, (theorema 4.6.1.), $(n-r)\hat{\sum}$ has a Wishart distribution with $n-r$ degrees of freedom, $(n-r)\hat{\sum} \sim W(n-r, \sum)$ and $\hat{\psi}^*, \hat{\sum}$ are independent.

3. DISTANCE BETWEEN ESTIMABLE FUNCTIONS

For any given estimable functions ψ_1^*, ψ_2^*, we consider the following distance (squared) introduced by C.M. Cuadras (1974)

$$D_p^2(\psi_1^*, \psi_2^*) = (\psi_1 - \psi_2)^t\,\Sigma^{-1}(\psi_1 - \psi_2) \quad , \tag{5}$$

where $\psi_i (i = 1, 2)$ is the vector of components of $\psi_i^* (i = 1, 2)$ in the dual basis $Y_1^*, ..., Y_p^*$.

Note that (5) is just the square of the distance between ψ_1^* and ψ_2^* induced by $\sum$ on the dual espace F^*.

As $\sum$ has full rank, the expression in (5) is strictly positive whenever $\psi_1^* \neq \psi_2^*$ (and zero otherwise).

PROPOSITION 1

The distance (5) is invariant under linear automorphisms of the variables $Y_1^*, ..., Y_p^*$.

Proof

Let $Z^* = A^t Y^*$ be a linear automorphism, where $Z_i^* = a_{i1} Y_1^* + ... + a_{ip} Y_p^* (i = 1, ..., p)$, and consider an estimable function $\psi^* = p^t B^*$, where $B^* = MY^*$ and M is an mxp-matrix. Then it is easily seen that A transforms estimable functions into estimable functions. Moreover, given two estimable functions ψ_1^*, ψ_2^* if $\psi_i (i = 1, 2)$ is the component vector of ψ_i^* with respect to $Y_1^*, ..., Y_p^*$, we have that, $D_Z^2(\psi_1^*, \psi_2^*) = (A^t \psi_1 - A^t \psi_2)^t (A^t \sum A)^{-1} (A^t \psi_1 - A^t \psi_2) = (\psi_1 - \psi_2)^t A\, A^{-1} \sum^{-1} (A^t)^{-1} A^t (\psi_1 - \psi_2) = D_Y^2 (\psi_1^*, \psi_2^*)$.

PROPOSITION 2

Let $M_1 = X_1 B_1 + U_1$ and $M_2 = X_2 B_2 + U_2$ be two multivariate linear models with $M_1 = (Y_1, ..., Y_p)$ and $M_2 = (Z_1, ..., Z_q)$ noncorrelated. If $\psi^* = p_1^t B_1 + p_2^t B_2$ is an estimable function with respect to the multivariate linear model

$$(6) \qquad \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} + \begin{pmatrix} U_1 \\ U_2 \end{pmatrix},$$

then, $D_{p+q}^2 (\psi_1^*, \psi_2^*) = D_p^2 (\pi_1 (\psi_1^*), \pi_1(\psi_2^*)) + D_q^2 (\pi_2(\psi_1^*), \pi_2(\psi_2^*))$ where we have set $\pi_i(\psi^*) = p_i^t B_i^*$ for $i = 1, 2$.

Proof

Using (3) it is easy to see that $\pi_1 (\psi_i^*)$, resp. $\pi_2 (\psi_i^*)$, is an estimable function with respect to $M_1 = X_1 B_1 + U_1$, resp. $M_2 = X_2 B_2 + U_2$.

Let (ψ_i^1, ψ_i^2) $(i = 1, 2)$ denote the component vector of ψ_i^*, where ψ_i^1 consists of the coordinates of ψ_i^* with respect to $Y_1^*, ..., Y_p^*$ and ψ_i^2 of the corresponding ones for $Z_1^*, ..., Z_q^*$. As the covariance matrix of (M_1, M_2) is

$$\sum = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}$$

where $\sum_i$ is the covariance matrix of M_i $(i = 1, 2)$ we have,

$$D^2_{p+q}(\psi_1^*, \psi_2^*) = ((\psi_1^1, \psi_1^2) - (\psi_2^1, \psi_2^2))^t \sum^{-1} ((\psi_1^1, \psi_1^2) - (\psi_2^1, \psi_2^2))$$
$$= \left((\psi_1^1 - \psi_2^1)^t, (\psi_1^2 - \psi_2^2)^t\right) \begin{pmatrix} \Sigma_1^{-1} & 0 \\ 0 & \Sigma_2^{-1} \end{pmatrix} ((\psi_1^1 - \psi_2^1), (\psi_1^2 - \psi_2^2))$$
$$= (\psi_1^1 - \psi_2^1)^t \sum_1^{-1} (\psi_1^1 - \psi_2^1) + (\psi_1^2 - \psi_2^2)^t \sum_2^{-1} (\psi_1^2 - \psi_2^2)$$
$$= D_p^2(\psi_1^*, \psi_2^*) + D_q^2(\psi_1^*, \psi_2^*).$$

PROPOSITION 3

Let $Y = XB + U$ be a multivariate linear model. If we add new variables Z, such that $Y = (Y_1, ..., Y_p)$ and $Z = (Z_1, ..., Z_q)$ are not necessarily noncorrelated, then

$$D_p^2(\psi_1^*, \psi_2^*) \leq D^2_{p+q}(\psi_1^*, \psi_2^*) \ .$$

Proof

Let E be the real vector space spanned by $Y_1, ..., Y_p, Z_1, ..., Z_q$ and let F be the subspace generated by $Y_1, ..., Y_p$, and $\sum$ the covariance matrix of (Y, Z), i.e.,

$$\sum = \begin{pmatrix} \Sigma_Y & D \\ D^t & \Sigma_Z \end{pmatrix} \quad ,$$

where $\sum_Y$ and $\sum_Z$ are the covariance matrices of Y and Z respectively.

The distance (5) between two estimable functions ψ_1, ψ_2 is given by

$$D_p^2(\psi_1^*, \psi_2^*) = (\psi_1 - \psi_2)^t \ \Sigma_Y^{-1}(\psi_1 - \psi_2)$$

But $\Omega = \psi_1 - \psi_2$ is an element of the dual space F^* of F and as $\sum_Y$ is positive definite, there is a unique vector X in F such that $\Omega = \sum_Y X$. Therefore

$$D_p^2(\psi_1^*, \psi_2^*) = \Omega^t \Sigma_Y^{-1} \ \Omega = X^t \Sigma_Y^t \ \Sigma_Y^{-1} \Sigma_Y X = X^t \Sigma_Y X$$

on the other hand,

$$D^2_{p+q}(\psi_1^*, \psi_2^*) = \bar{\Omega}^t \Sigma^{-1} \ \bar{\Omega}$$

where $\bar{\Omega}$, which lies in the dual space E^* of E, is of the form $(\Omega, M)^t$ with M arbitrary.

Note that the matrix $\sum^{-1}$ has the form

$$\begin{pmatrix} A & B \\ B^t & C \end{pmatrix}$$

and satisfies

$$\Sigma_y A + DB^t = D^t B + \Sigma_Z C = A\Sigma_Y + BD^t = B^t D + C\Sigma_Z = I$$
$$\Sigma_Y B + DC = D^t A + \Sigma_Z B^t = AD + B\Sigma_Z = B^t\Sigma_Y + CD^t = 0$$

where as usual, I is the identity matrix and 0 is the null matrix. Consider now the following decomposition

$$\begin{pmatrix} \Omega \\ M \end{pmatrix} = \begin{pmatrix} \Sigma_Y X \\ M \end{pmatrix} = \begin{pmatrix} \Sigma_Y & X \\ D^t & X \end{pmatrix} + \begin{pmatrix} 0 \\ M - D^t X \end{pmatrix} .$$

If Ω_i stands for the ith term of the right hand side of the preceding expression $(i = 1, 2)$, we have

$$D^2_{p+q}(\psi_1^*, \psi_2^*) = (\Omega^t, M^t)\Sigma^{-1} \begin{pmatrix} \Omega \\ M \end{pmatrix} = (\Omega_1^t + \Omega_2^t)\Sigma^{-1}\ (\Omega_1 + \Omega_2)$$
$$= \Omega_1^t\Sigma^{-1}\Omega_1 + \Omega_2^t\Sigma^{-1}\ \Omega_2 + 2\Omega_1^t\Sigma^{-1}\Omega_2.$$

But,

$$\Omega_1^t\ \Sigma^{-1}\Omega_1 = (X^t\Sigma_Y, X^t D)\ \Sigma^{-1} \begin{pmatrix} \Sigma_Y & X \\ D^t & X \end{pmatrix}$$

$$= (X^t\Sigma_Y,\ X^t D) \begin{pmatrix} (A\Sigma_Y + BD^t)X \\ (B^t\Sigma_Y + CD^t)X \end{pmatrix} = (X^t\Sigma_Y,\ X^t D) \begin{pmatrix} X \\ 0 \end{pmatrix} = X^t\Sigma_Y X.$$

If we put $R = M - D^t X$ then

$$\Omega_2^t\ \Sigma^{-1}\ \Omega_2 = (0, R^t)\Sigma^{-1} \begin{pmatrix} 0 \\ R \end{pmatrix} = (0, R^t) \begin{pmatrix} B & R \\ C & R \end{pmatrix} = R^t C R$$

and as $\sum^{-1}$ is positive definite, $R^t C R \geq 0$.

Finally,

$$\Omega_1^t \, \Sigma^{-1} \Omega_2 = (X^t \Sigma_Y, \, X^t D) \, \Sigma^{-1} \begin{pmatrix} 0 \\ R \end{pmatrix}$$

$$= (X^t \Sigma_Y, \, X^t D) \begin{pmatrix} B & R \\ C & R \end{pmatrix} = X^t \left(\Sigma_Y B + DC \right) R = 0$$

And this concludes the proof, since

$$D^2_{p+q} \left(\psi_1^*, \, \psi_2^* \right) = X^t \Sigma_Y X + R^t \, C \, R \geq X^t \Sigma_Y X = D^2_p (\psi_1, \, \psi_2). \; \#$$

Consider the parametric family of p-multivariate normal distributions $p(\, \cdot \mid \mu, \, \sum)$ with $\sum$ fixed and $\mu \in \mathbf{R}^p$ being the parameter and identify this family with F^* by means of

$$Z^* \to p(\, \cdot \mid \mu, \, \Sigma) : Z^*(W) = E(W) = \omega_1 E(Y_1) + \ldots + \omega_p E(Y_p) = \mu^t W$$

Then we have

PROPOSITION 4

Distance (5) is Rao's distance between the distributions $N_p \, (\psi_1^*, \sum)$ and $N_p \, (\psi_2^*, \, \sum)$.

Proof

With the preceding notations, let M be the subspace of F^* generated by $\mu_1^*, \ldots, \, \mu_k^*$, i.e., the space of estimable functions. Supposes that $\mu_1^*, \ldots, \mu_r^*$ is a basis of M

Recall that in order to find Rao's distance between two probability distributions, we take (Cuadras, 1988) Fisher's information matrix as the fundamental metric tensor in the manifold generated by the parameters. If we considerer the parametric family of multivariate normal distributions $N_p(\mu, \sum)$ where $\sum$ is fixed and $\mu \in \mathbf{R}^p$ is the parameter, the fundamental metric tensor is just $\sum^{-1}$.

Now consider the parametric family of multinormal distributions $N_p(\psi, \sum)$ with $\sum$ fixed and the parameter ψ varying in M. As M is a submanifold of $\mathbf{R}^p$ and $\mathbf{R}^p$ is isomorphic to F^* the fundamental metric tensor restricted to M is given by the rxr-matric $G = (g_{ij})$, with

$$g_{ij}(\psi) = (\mu_i^*)^t \, \Sigma^{-1} \; \mu_j^* \quad ,$$

and where μ_i^* is the component vector of μ_i^* in the basis of F^*.

As G does not depend on the parameter, the Christoffel symbols vanish, so the geodesics are the straight lines (properly parametrized).

Then Rao's distance between the distributions $N_p(\psi_1^*, \sum)$, $N_p(\psi_2^*, \sum)$ is given by $R^2(1,2) = (\psi_2' - \psi_1')^t\, G(\psi_2 - \psi_1)$.

If $\psi_i (i = 1, 2)$ is the component vector of ψ_i^* in the basis of F^*, then

$$\psi_i = A\psi_i' \quad \text{and} \quad G = A^t\Sigma^{-1}A$$

where A is a pxr-matrix such that

$$\mu_j^* = a_{1j} \quad Y_1^* + \ldots + a_{pj} \quad Y_p^*\ , \ j = 1, \ldots, r$$

Then we can write

$$D_p^2(\psi_1^*, \psi_2^*) = (\psi_1 - \psi_2)^t\Sigma^{-1}(\psi_1 - \psi_2) = (A\psi_1' - A\psi_2')^t\ \Sigma^{-1}(A\psi_1' - A\psi_2')$$

$$= (\psi_1' - \psi_2')^t\ G\ (\psi_1' - \psi_2') = R^2(1,2)\,. \qquad \#$$

For properties of Rao's distance see Cuadras (1988).

Finally, as an estimation of distance (5) we may take

$$\hat{D}_p^2(\psi_1^*, \psi_2^*) = \left(\hat{\psi}_1^* - \hat{\psi}_2^*\right)^t\ \hat{\Sigma}^{-1}\left(\hat{\psi}_1^* - \hat{\psi}_2^*\right)$$

where $\hat{\psi}_i^* (i = 1, 2)$ is the Gauss-Markov estimation of ψ_i^* and $\hat{\sum}$ is the unbiased estimation of $\sum$ given in (4). Suppose that $(Y_1, \ldots, Y_p)$ has multivariate normal distribution and take samples of size n_i in each population H_i. Given two estimable functions ψ_1^*, ψ_2^*, the Gauss-Markov estimation $\hat{\psi}^*$ of $\psi^* = \psi_1^* - \psi_2^*$ follows the distribution $N_p(\psi_1^* - \psi_2^*, D)$, where $D = \sum_{i=1}^{k} d_i^2/n_i$ and where $d_1, \ldots, d_k$ are the coefficients of the intrinsic expression of ψ. Then, under the hypothesis $H_o : \psi_1^* = \psi_2^*$,

$$D^{-1}\left(\hat{\psi}_1^* - \hat{\psi}_2^*\right)^t\ \hat{\Sigma}^{-1}\left(\hat{\psi}_1^* - \hat{\psi}_2^*\right)$$

has a Hotteling distribution with parameters p and $n - r$, and

$$D^{-1}\,\frac{n - r - p + 1}{p(n - r)} \quad \hat{D}_p^2(\psi_1^*, \psi_2^*)$$

has a Fisher-Snedecor distribution with p and $n - r - p - 1$ degrees of freedom.

4. ACKNOWLEDGEMENT

The autor wishes to thank Prof. C.M. Cuadras for his helpful suggestions and constant encouragement in the preparation of this paper.

5. REFERENCES

[1] **Anderson, T.W.** (1958). "An introduction to Multivariate Statistical Analysis". Wiley and Sons, New York.

[2] **Ballús, C.; Cuadras, C.M.; Malgá, A.; Sánchez Turet, M.; Vallvé, C.** (1980). "Estudio de dos ansiolíticos (Diazepan y Clobazam) mediante una prueba de conducción de automóviles". Rev. Dept. Psiquiatría Facult. Med. Barcelona 7,2. 107-122.

[3] **Cuadras, C.M.** (1974). "Análisis discriminante de funciones paramétricas estimables". Trabajos de Estadística y de Investigación Operativa, vol. 25(3), 3-31.

[4] **Cuadras, C.M.** (1981). "Métodos de Análisis Multivariante".Ed. Eunibar, 642 pp.

[5] **Cuadras, C.M.** (1988). "Distancias Estadísticas". Estadística Española, Vol. 30, n^o 119, 295-378.

[6] **Cuadras, C.M.; Oller, J.M.** (1982). "Representación canónica en MANOVA: Aplicación a una clase de diseño anidado". Qüestiio, 6(3), 221-229.

[7] **Cuadras, C.M.; Rios, M.** (1986). "Distancia entre modelos lineales normales". Qüestiio, 10(2), 83-92.

[8] **Cuadras, C.M.; Sánchez, P.** (1988). "Métodos de regresión y análisis de la varianza". Pub. Dept. Estadística, Univ. Barcelona.

[9] **Mahalanobis,P.C.** (1936). "On the generalized distance in Statistics". Proc. Nat. Inst. Sci. India, 12, 49-55.

[10] **Rao, C.R.** (1945). "Information and the accuracy attainable in the estimation of statistical parameter". Bull. Calcutta Math. Soc. 17, 81-91

[11] **Seber, G.A.F** (1977). "Linear regression analysis". John Wiley and Sons, New York.

[12] **Seber, G.A.F.** (1984). "Multivariate observations". John Wiley and Sons, New York.