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.
2. MULTIVARIATE ESTIMABLE FUNCTIONS

Consider the multivariate linear model

\[ Y = XB + U \]

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

Let \( \mu_i \) stand for the mean vector of \( Y \) in the population \( H_i, i = 1, \ldots, k \). Then we have

\[ \mu = XB \]

where

\[ \mu = (\mu_1, \ldots, \mu_k)^t. \]

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

An estimable function \( \psi^* \) (see Cuadras (1974), definition 2.4.1.) is an element of the dual space \( F^* \) of \( F \) given by a linear combination of the \( \mu_i^* \)'s, i.e.,

\[ \psi^* = d_1\mu_1^* + \ldots + d_k\mu_k^* = D^t\mu^*. \]

If \( E\beta \) stands for the real vector space spanned by \( \beta_1, \ldots, \beta_m \) and \( E\beta^* \) stands for its dual space then a parametric function is just an element of \( E\beta^* \) (see Cuadras (1974), 3.2.2.).

We recall that a parametric function is said to be estimable if it has a linear unbiased estimate. Otherwise a parametric function, say \( \psi^* = P_1\beta_1^* + \ldots + P_m\beta_m^* = p^TB^* \) 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 \( \psi = p_1\beta_1 + \ldots + p_m\beta_m = p^tB \) is estimable if and only if

\[ p^t(X^tX)^{-1}X^tX = p^t \]
where \((X'X)^{-1}\) is a generalized inverse of \(X'X\).

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 \(\Delta\) the diagonal matrix with entries \(n_1, ..., n_k\).

Let \(\psi^*\) be an estimable function and \(\hat{\mu}^*\) the usual estimation of the mean \((i = 1, ..., k)\). If the Gauss-Markov estimation of \(\psi^*\) is given by \(\hat{\psi}^* = p^t \hat{\beta}^*\), we say that \(\psi^* = \hat{D}^t \hat{\mu}^*\) is the "intrinsic expression" of \(\psi^*\), where \(\hat{D} = \Delta X(X'\Delta X)^{-1}p^t\), (see Cuadras (1974), 3.3.4.).

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

\[
V = \begin{bmatrix}
y_{111} & \cdots & y_{p11} \\
\vdots & \ddots & \vdots \\
y_{1n_1} & \cdots & y_{p1n_1} \\
\vdots & \ddots & \vdots \\
y_{1k1} & \cdots & y_{pk1} \\
\vdots & \ddots & \vdots \\
y_{1nk1} & \cdots & y_{pknk}
\end{bmatrix}
\]

An unbiased estimation of the covariance matrix \(\Sigma\) is

\[
\hat{\Sigma} = \frac{R_p^2}{n - r} = \frac{(V - X\Delta \hat{B}^*)^t(V - X\Delta \hat{B}^*)}{n - r}
\]

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 \(\psi\) 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.), \((n - r)\hat{\Sigma}\) has a Wishart distribution with \(n - r\) degrees of freedom, \((n - r)\hat{\Sigma} \sim W(n - r, \Sigma)\) and \(\hat{\psi}^*, \hat{\Sigma}\) are independent.

### 3. DISTANCE BETWEEN ESTIMABLE FUNCTIONS

For any given estimable functions \(\psi_1^*, \psi_2^*\), we consider the following distance (squared) introduced by C.M. Cuadras (1974)

\[
D_p^2(\psi_1^*, \psi_2^*) = (\psi_1 - \psi_2)^t \Sigma^{-1} (\psi_1 - \psi_2)
\]
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 $\psi_1^*$ and $\psi_2^*$ induced by $\sum$ on the dual space $F^*$.

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 $Z_i^* = a_i Y_i^* + \ldots + a_p Y_p^*(i = 1, \ldots, p)$, and consider an estimable function $\psi^* = p^t B^*$, where $B^* = MY^*$ and $M$ is an $m \times p$-matrix. Then it is easily seen that $A$ transforms estimable functions into estimable functions. Moreover, given two estimable functions $\psi_1^*, \psi_2^*$ if $\psi_i(i = 1, 2)$ is the component vector of $\psi_i^*$ with respect to $Y_1^*, ..., Y_p^*$, we have that,

$$D_{Z}^2(\psi_1^*, \psi_2^*) = (A^t \psi_1 - A^t \psi_2)^t (A^t \sum A)^{-1} (A^t \psi_1 - A^t \psi_2) = (\psi_1 - \psi_2)^t A A^{-1} \sum^{-1} A^t (\psi_1 - \psi_2) = D_Y^2(\psi_1^*, \psi_2^*).$$

**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

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

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 $\pi_1(\psi_1^*)$, resp. $\pi_2(\psi_2^*)$, 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 $(\psi_1^i, \psi_2^i)$ $(i = 1, 2)$ denote the component vector of $\psi_i^*$, where $\psi_i^1$ consists of the coordinates of $\psi_i^*$ with respect to $Y_1^*, ..., Y_p^*$ and $\psi_i^2$ of the corresponding
ones for $Z_1^*, ..., Z_q^*$. As the covariance matrix of $(M_1, M_2)$ is

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

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

$$D_{p+q}^2 (\psi_1^*, \psi_2^*) = ((\psi_1^1, \psi_1^2) - (\psi_2^1, \psi_2^2))^t \Sigma_i^{-1} ((\psi_1^1, \psi_1^2) - (\psi_2^1, \psi_2^2))$$

$$= \left((\sigma_1^1 - \sigma_2^1)^t, (\sigma_1^2 - \sigma_2^2)^t\right) \begin{pmatrix}
\Sigma_1^{-1} & 0 \\
0 & \Sigma_2^{-1}
\end{pmatrix} \left(\sigma_1^1 - \sigma_2^1, (\sigma_1^2 - \sigma_2^2)^t\right)$$

$$= (\sigma_1^1 - \sigma_2^1)^t \Sigma_1^{-1} (\sigma_1^1 - \sigma_2^1) + (\sigma_1^2 - \sigma_2^2)^t \Sigma_2^{-1} (\sigma_1^2 - \sigma_2^2)$$

$$= D_p^2 (\psi_1^*, \psi_2^*) + D_q^2 (\psi_1^*, \psi_2^*).$$

**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

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

**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 $Z$ the covariance matrix of $(Y, Z)$, i.e.,

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

where $\Sigma_Y$ and $\Sigma_Z$ are the covariance matrices of $Y$ and $Z$ respectively.

The distance (5) between two estimable functions $\psi_1, \psi_2$ 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 $\Sigma$ is positive definite, there is a unique vector $X$ in $F$ such that $\Omega = \Sigma_X X$. Therefore

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

on the other hand,

$$D_{p+q}^2 (\psi_1^*, \psi_2^*) = \Omega^t \Sigma_Y^{-1} \Omega$$
where $\tilde{\Omega}$, which lies in the dual space $E^*$ of $E$, is of the form $(\Omega, M)^t$ with $M$ arbitrary.

Note that the matrix $\Sigma^{-1}$ has the form
\[
\begin{pmatrix}
A & B \\
B^t & C
\end{pmatrix}
\]

and satisfies
\[
\Sigma_Y A + DB^t = D^t B + \Sigma_Z C = AD + BD^t = B^t D + C \Sigma_Z = I
\]
\[
\Sigma_Y B + DC = D^t A + \Sigma_Z B^t = AD + C \Sigma_Z = B^t \Sigma_Y + CD^t = 0
\]

where as usual, $I$ is the identity matrix and $0$ is the null matrix. Consider now the following decomposition
\[
\begin{pmatrix}
\Omega \\
M
\end{pmatrix} = \begin{pmatrix}
\Sigma_Y \times X \\
M
\end{pmatrix} = \begin{pmatrix}
\Sigma_Y & X \\
D^t & X
\end{pmatrix} + \begin{pmatrix}
0 \\
M - D^t X
\end{pmatrix}.
\]

If $\Omega_i$ stands for the $i$th term of the right hand side of the preceding expression ($i = 1, 2$), we have
\[
D^2_{p+q} (\psi_1^*, \psi_2^*) = (\Omega_i, M^t) \Sigma^{-1} \begin{pmatrix}
\Omega \\
M
\end{pmatrix} = (\Omega_1 + \Omega_2) \Sigma^{-1} (\Omega_1 + \Omega_2)
\]
\[
= \Omega_1 \Sigma^{-1} \Omega_1 + \Omega_2 \Sigma^{-1} \Omega_2 + 2 \Omega_1 \Sigma^{-1} \Omega_2.
\]

But,
\[
\Omega_1 \Sigma^{-1} \Omega_1 = (X^t \Sigma_Y, X^t D) \Sigma^{-1} \begin{pmatrix}
\Sigma_Y & X \\
D^t & X
\end{pmatrix}
\]
\[
= (X^t \Sigma_Y, X^t D) \begin{pmatrix}
(A \Sigma_Y + BD^t)X \\
(B^t \Sigma_Y + CD^t)X
\end{pmatrix} = (X^t \Sigma_Y, X^t D) \begin{pmatrix}
X \\
0
\end{pmatrix} = X^t \Sigma_Y X.
\]

If we put $R = M - D^t X$ then
\[
\Omega_2 \Sigma^{-1} \Omega_2 = (0, R)^t \Sigma^{-1} \begin{pmatrix}
0 \\
R
\end{pmatrix} = (0, R^t) \begin{pmatrix}
B & R \\
C & R
\end{pmatrix} = R^t C R.
\]
and as $\Sigma^{-1}$ is positive definite, $R^tCR \geq 0$.

Finally,

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

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

And this concludes the proof, since

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

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

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

Then we have

**Proposition 4**

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

**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 $\mu_1^*, \ldots, \mu_k^*$ 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 $N_p (\mu, \Sigma)$ where $\Sigma$ is fixed and $\mu \in \mathbb{R}^p$ is the parameter, the fundamental metric tensor is just $\Sigma^{-1}$.

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

$$g_{ij} (\psi) = (\mu_i^*)^t \Sigma^{-1} \mu_j^*$$
and where $\mu^*_i$ is the component vector of $\mu^*_i$ 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 $N_p(\psi^*_1, \Sigma)$, $N_p(\psi^*_2, \Sigma)$ 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 $\psi^*_i$ 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 $p \times r$-matrix such that

$$\mu^*_j = a_{1j} Y^*_1 + \ldots + a_{pj} Y^*_r , \quad j = 1, \ldots, r$$

Then we can write

$$D^2_p(\psi^*_1, \psi^*_2) = (\psi_1 - \psi_2)^t \Sigma^{-1} (\psi_1 - \psi_2) = (A \psi'_1 - A \psi'_2)^t \Sigma^{-1} (A \psi'_1 - A \psi'_2)$$

$$= (\psi'_1 - \psi'_2)^t G (\psi'_1 - \psi'_2) = R^2(1, 2). \quad \#$$

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

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

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

where $\hat{\psi}^*_i(i = 1, 2)$ is the Gauss-Markov estimation of $\psi^*_i$ and $\hat{\Sigma}$ is the unbiased estimation of $\Sigma$ 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 $\psi^*_1, \psi^*_2$, the Gauss-Markov estimation $\psi^*$ 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 $\psi$. Then, under the hypothesis $H_0 : \psi^*_1 = \psi^*_2$,

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

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

$$D^{-1} \frac{n - r - p + 1}{p(n - r)} \hat{D}^2_p(\psi^*_1, \psi^*_2)$$

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

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

5. REFERENCES


