

# Typicality in Discriminant Analysis with Mixed Variables

C.M. Cuadras and J. Fortiana  
Department of Statistics, University of Barcelona.

**Summary:** The problem of allocating one observation to one of two given populations is fundamental in multivariate analysis. In some applications (distinctiveness studies) it is also interesting to determine whether one observation is related to both populations or to an entirely different population. The classic test to detect atypical observations presupposes  $p$ -variate normality. This contribution suggests a generalization of this test by using the distance-based approach in discrimination, which can deal with mixed variables.

**Keywords:** Classification. Distinctiveness studies. Outliers with mixed data. Goodness of fit.

## 1. Typicality in Linear Discrimination

Suppose that  $\mathbf{X}$  is a  $p$ -dimensional random vector,  $\Pi_i$ ,  $i = 1, 2$ , are two known populations and  $\Pi_3$  is a third population. Typicality is described as a test to determine whether an observation  $\mathbf{x}$  of  $\mathbf{X}$  is typical of a mixture of  $\Pi_1$  and  $\Pi_2$  or belongs to  $\Pi_3$ . When  $\Pi_i \sim N_p(\mu_i, \Sigma)$ ,  $i = 1, 2, 3$ , the appropriate hypotheses to be tested are

$$\begin{aligned} H_0 &: \mathbf{x} \in N_p(\alpha\mu_1 + (1 - \alpha)\mu_2, \Sigma), 0 \leq \alpha \leq 1, \\ H_1 &: \mathbf{x} \in N_p(\mu_3, \Sigma). \end{aligned}$$

For testing the hypotheses  $H'_0(\alpha = 1)$ ,  $H''_0(\alpha = 0)$ ,  $H_0(0 \leq \alpha \leq 1)$ , Rao(1973) proposes the statistics

$$\begin{aligned} U_i(\mathbf{x}) &= ((\mathbf{x} - \mu_i)' \Sigma^{-1} (\mu_2 - \mu_1))^2 / (\mu_2 - \mu_1)' \Sigma^{-1} (\mu_2 - \mu_1) \sim \chi_1^2 \quad (i = 1, 2), \\ W(\mathbf{x}) &= (\mathbf{x} - \mu_1)' \Sigma^{-1} (\mathbf{x} - \mu_1) - U_1(\mathbf{x}) \sim \chi_{p-1}^2. \end{aligned}$$

These tests work with  $p$ -variate normal data when  $\Sigma$  is known or estimated from large samples. If  $W(\mathbf{x})$  is significant, then  $\mathbf{x}$  is atypical and cannot belong to the predefined populations, otherwise  $U_1(\mathbf{x})$  and  $U_2(\mathbf{x})$  can be used to decide whether  $\mathbf{x}$  comes from  $\Pi_1$  or  $\Pi_2$ . See also Bar-Hen and Daudin (1997). The aim of this contribution is to present the typicality test for general data using distances and proximity functions.

## 2. The Proximity Function in Discrimination

Suppose that  $\mathbf{X}$  has a probability density  $f(\mathbf{x})$  and support  $S$ . Let  $\delta(\mathbf{x}_1, \mathbf{x}_2)$  be a distance function between observations of  $\mathbf{X}$ . The geometric variability of  $\mathbf{X}$  and the proximity function of an observation  $\mathbf{x}_0$  of  $\mathbf{X}$  to the population  $\Pi_1$  represented by  $\mathbf{X}$ , are respectively defined by

$$V_\delta(\mathbf{X}) = \frac{1}{2} \int_{S \times S} \delta^2(\mathbf{x}_1, \mathbf{x}_2) f(\mathbf{x}_1) f(\mathbf{x}_2) d\mathbf{x}_1 d\mathbf{x}_2, \quad (1)$$

$$D_1^2(\mathbf{x}_0) = \int_S \delta^2(\mathbf{x}_0, \mathbf{x}) f(\mathbf{x}) d\mathbf{x} - V_\delta(\mathbf{X}). \quad (2)$$

$V_\delta(\mathbf{X})$  is a general measure of variability and  $D_1^2(\mathbf{x}_0)$  can be used in discriminant analysis. It can be proved that if  $(S, \delta)$  can be represented in a Hilbert space  $L$ , i.e., there exists  $\phi : S \rightarrow L$  such that  $\delta^2(\mathbf{x}_1, \mathbf{x}_2) = \|\phi(\mathbf{x}_1) - \phi(\mathbf{x}_2)\|^2$ , then

$$\begin{aligned} V_\delta(\mathbf{X}) &= E(\|\phi(\mathbf{X})\|^2) - \|E(\phi(\mathbf{X}))\|^2, \\ D_1^2(\mathbf{x}_0) &= \|\phi(\mathbf{x}_0) - E(\phi(\mathbf{X}))\|^2. \end{aligned}$$

Thus  $D_1^2(\mathbf{x}_0)$  is the squared distance from  $\mathbf{x}_0$  to the  $\delta$ -mean  $E(\phi(\mathbf{X}))$ . If  $\Pi_2$  is a second population represented by  $\mathbf{Y}$ , with density  $g(\mathbf{y})$  and same support  $S$ , the distance-based (DB) rule allocates  $\mathbf{x}_0$  to the nearest population

$$\mathbf{x}_0 \text{ comes from } \Pi_i \text{ if } D_i^2(\mathbf{x}_0) = \min\{D_1^2(\mathbf{x}_0), D_2^2(\mathbf{x}_0)\}.$$

The squared distance between two populations from distances between observations, can be defined by

$$\Delta^2(\Pi_1, \Pi_2) = \int_{S \times S} \delta^2(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) g(\mathbf{y}) d\mathbf{x} d\mathbf{y} - V_\delta(\mathbf{X}) - V_\delta(\mathbf{Y}). \quad (3)$$

When  $\delta$  is the Mahalanobis distance we have

$$\begin{aligned} D_i^2(\mathbf{x}) &= (\mathbf{x} - \mu_i)' \Sigma^{-1} (\mathbf{x} - \mu_i), \\ \Delta^2(\Pi_1, \Pi_2) &= (\mu_2 - \mu_1)' \Sigma^{-1} (\mu_2 - \mu_1), \end{aligned}$$

and the DB rule is equivalent to performing a Linear Discriminant Analysis. Other classic discriminant rules can be obtained as particular cases of DB by choosing suitable distances. The real advantage arises in nominal or mixed data, when it is difficult to obtain a probabilistic model, while a proximity function, which can be estimated from a sample without knowing the density, is more accessible. This approach was introduced, studied and applied by Cuadras and Fortiana(1995), Cuadras *et al.*(1997a, 1997b).

### 3. Typicality using Proximity Functions

Suppose that we have multivariate mixed data and a suitable distance  $\delta$ . The DB versions for  $P_1(\mathbf{x}) = (\mu_2 - \mu_1)' \Sigma^{-1}(\mathbf{x} - \mu_1)$  and  $P_2(\mathbf{x}) = (\mu_2 - \mu_1)' \Sigma^{-1}(\mathbf{x} - \mu_2)$ , when  $\delta$  is a general distance, are

$$\begin{aligned} P_1(\mathbf{x}) &= [\Delta^2(\Pi_1, \Pi_2) + D_1^2(\mathbf{x}) - D_2^2(\mathbf{x})]/2, \\ P_2(\mathbf{x}) &= [\Delta^2(\Pi_1, \Pi_2) + D_2^2(\mathbf{x}) - D_1^2(\mathbf{x})]/2. \end{aligned}$$

The test that  $\mathbf{x}$  comes from the convex linear combination  $\alpha\Pi_1 + (1 - \alpha)\Pi_2$ , i.e., from a population with  $\delta$ -mean  $\alpha E(\phi(\mathbf{X})) + (1 - \alpha)E(\phi(\mathbf{Y}))$ ,  $0 \leq \alpha \leq 1$ , and the other two related tests, may be performed by using the statistics

$$\begin{aligned} H_0'(\alpha = 1) : \quad U_1(\mathbf{x}) &= (P_1(\mathbf{x}))^2 / \Delta^2(\Pi_1, \Pi_2), \\ H_0''(\alpha = 0) : \quad U_2(\mathbf{x}) &= (P_2(\mathbf{x}))^2 / \Delta^2(\Pi_1, \Pi_2), \\ H_0(0 \leq \alpha \leq 1) : \quad W(\mathbf{x}) &= D_1^2(\mathbf{x}) - U_1(\mathbf{x}) = D_2^2(\mathbf{x}) - U_2(\mathbf{x}). \end{aligned}$$

$W(\mathbf{x})$  significant means that  $\mathbf{x}$  comes from a different population  $\Pi_3$ .

The sampling distributions of  $U_1, U_2$  and  $W$  can be difficult to find for mixed data and may be obtained by resampling methods. But we prefer to follow a procedure proposed by Cuadras and Fortiana(1994).

### 4. Distribution of $W$

Suppose that the sample sizes are  $N_1, N_2$ . Then we have  $N = N_1 + N_2$  observations of  $U_1, U_2$  and  $W$ , respectively. Let  $x_1 \leq \dots \leq x_N$  be an ordered sample with mean  $\bar{x}$  and sample variance  $s$ , of observed values of  $W$ , say. Let  $F_N$  be the empirical cumulative distribution function and suppose that  $F$  is the theoretical distribution of  $W$ , which has mean  $\mu$  and variance  $\sigma^2$ . To measure the agreement between  $F_N$  and  $F$ , we propose the maximal Hoeffding correlation between the sample and  $W$ . This correlation  $r_N^+$  is the maximum correlation which we can obtain considering all bivariate distributions with marginals  $F_N$  and  $F$ . The value of  $r_N^+$  is invariant from linear transformations of  $W$ , that is, it does not depend of the particular values of  $\mu$  and  $\sigma$ .

Assume, for example, that  $F$  is the exponential distribution. Then we can suppose that  $\mu = \sigma = 1$  and we obtain

$$r_N^+ = \frac{1}{sN} \left[ \sum_{i=1}^{N-1} x_i \left( \log(N - i)^{N-i} - \log(N + 1 - i)^{N+1-i} + \log(N) \right) + x_N \log(N) \right].$$

$r_N^+$  close to 1 indicates that the sample comes from an exponential distribution.

## 5. One mixed data example

Cuadras, Fortiana and Oliva(1997a) used a cancer data set, consisting of 11 mixed measurements (7 continuous, 2 binary and 2 categorical) to illustrate a distance-based discrimination rule, based on (2). This rule is applied to allocate individuals in two groups of tumours: benign ( $N_1 = 78$ ) and malignant ( $N_2 = 59$ ). The data comes from Krzanowski(1980) and the distance is computed from Gower(1971) similarity coefficient for mixed variables.

In order to ascertain whether or not one observation is atypical we consider  $W(\mathbf{x})$ , expressed as

$$W(\mathbf{x}) = D_1^2 - \frac{1}{4}(\Delta^2 + D_1^2 - D_2^2)^2/\Delta^2 = D_2^2 - \frac{1}{4}(\Delta^2 + D_2^2 - D_1^2)^2/\Delta^2.$$

where  $D_1^2 = D_1^2(\mathbf{x})$ ,  $D_2^2 = D_2^2(\mathbf{x})$  are the proximity functions, see (2), and  $\Delta^2$  is a distance between the two populations, see (3). These quantities can be easily estimated from the sample.  $W(\mathbf{x})$  significant is interpreted as  $\mathbf{x}$  comes from another different population (the individual may not have this tumour).

The distribution of  $W$  for this cancer data seems to be exponential and taking the sample of  $N = N_1 + N_2 = 137$  individuals, we obtain  $r_N^+ = 0.9285$ . Individual 11 of the malignant group gives one extreme  $W$  value and may be atypical, so is removed. The computations for the  $N - 1 = 136$  remaining individuals give  $r_{N-1}^+ = 0.9732$ . The fit to the exponential distribution is quite good. Assuming this distribution, individual 11 is clearly atypical.

## References

- Bar-Hen, A. and Daudin, J.-J. (1997). A test of a special case of typicality in linear discriminant analysis. *Biometrics*, **53**, 39-48.
- Cuadras, C. M. and J. Fortiana (1994). Ascertaining the underlying distribution of a data set. In R. Gutiérrez and M. J. Valderrama (Eds.), *Selected Topics on Stochastic Modelling*, pp. 223–230. World Scientific, Singapore.
- Cuadras, C.M and Fortiana, J. (1995) A continuous metric scaling solution for a random variable. *J. of Multivariate Analysis*, **52**, 1-14, 1995.
- Cuadras, C. M., Fortiana, J., Oliva, F. (1997a) The proximity of an individual to a population with applications in discriminant analysis. *J. of Classification*, **14**, 117-136.
- Cuadras, C. M., Atkinson, R. A., Fortiana, J. (1997b) Probability densities from distances and discriminant analysis. *Statistics & Probability Letters*, **33**, 405-411.
- Gower, J. C. (1971) A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857-874.
- Krzanowski, W.J. (1980) Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, **36**, 493-499.
- Rao, C.R. (1973) *Linear Statistical Inference and Its Applications*. New York: Wiley.