Open Journal of Statistics, 2015, 5, 665-679
Published Online December 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.57068

Statistical Classification Using the Maximum Function

T. Pham-Gia[1], Nguyen D. Nhat[2], Nguyen V. Phong[3]

[1]Université de Moncton, Moncton, Canada
[2]University of Economics and Law, Hochiminh City, Vietnam
[3]University of Finance and Marketing, Hochiminh City, VietNam
Email: thu.pham-gia@umoncton.ca, nhatnd12@gmail.com, nv.phongbmt@ufm.edu.vn

Received 8 October 2015; accepted 14 December 2015; published 17 December 2015

Abstract

The maximum of k numerical functions defined on R^p, $p \geq 1$, by $f_{\max}(x) = \max\{f_1(x), \cdots, f_k(x)\}$, $\forall x \in R^p$ is used here in Statistical classification. Previously, it has been used in Statistical Discrimination [1] and in Clustering [2]. We present first some theoretical results on this function, and then its application in classification using a computer program we have developed. This approach leads to clear decisions, even in cases where the extension to several classes of Fisher's linear discriminant function fails to be effective.

Keywords

Maximum, Discriminant Function, Pattern Classification, Normal Distribution, Bayes Error, L_1-Norm, Linear, Quadratic, Space Curves

1. Introduction

In our two previous articles [1] and [2], it is shown that the maximum function can be used to introduce new approaches in Discrimination Analysis and in Clustering. The present article, which completes the series on the uses of that function, applies the same concept to develop a new approach in classification that can be shown to be versatile and quite efficient.

Classification is a topic encountered in several disciplines of applied science, such as Pattern Recognition (Duda, Hart and Stork [3]), Applied Statistics (Johnson and Wichern [4]), Image Processing (Gonzalez, Woods and Eddins [5]). Although the terminologies can differ, the approaches are basically the same. In R^p, we are in the presence of training data sets to build discriminant functions that will enable us to do some classification of a future data set into one of the C classes considered. Several approaches are proposed in the literature. The Baye-

How to cite this paper: Pham-Gia, T., Nhat, N.D. and Phong, N.V. (2015) Statistical Classification Using the Maximum Function. *Open Journal of Statistics*, **5**, 665-679. http://dx.doi.org/10.4236/ojs.2015.57068

sian Decision Theory approach starts with the determination of normal (or non-normal) distributions f_i governing these data sets, and also prior probabilities q_i (with sum $\sum_{i=1}^{C} q_i = 1$) assigned to these distributions. More general considerations include the cost $c_{i|j}$ of misclassifications, but since in applications we rarely know the values of these costs they are frequently ignored. We will call this approach the common Bayesian one. Here, the comparison of the related posterior probabilities of these classes, also called "class conditional distribution functions", is equivalent to compare the values of $g_i = q_i f_i$, and a new data point $\boldsymbol{x}_0$ will be classified into the distribution g_{i_0} with highest value of $g_i(\boldsymbol{x}_0)$, *i.e.* $g_{i_0}(\boldsymbol{x}_0) = \max_j \{g_j(\boldsymbol{x}_0)\}$.

On the other hand, Fisher's solution to the classification problem is based on a different approach and remains an interesting and important method. Although the case of two classes is quite clear for the application of Fisher's linear discriminant function, the argument and especially the computations, become much harder when we are in the presence of more than two classes.

At present, the multinormal model occupies, and rightly so, a position of choice in discriminant analysis, and various approaches using this model have led to the same results. We have, in R^p, $p \geq 1$

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \quad \boldsymbol{x} \in R^p \tag{1}$$

1) For discrimination and classification into one of the two classes, we have the two equations:

$$g_i(\boldsymbol{x}) = q_i f_i(\boldsymbol{x}) = \frac{q_i}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)}, \quad i = 1, 2,$$

and their ratio $g_1(\boldsymbol{x})/g_2(\boldsymbol{x}) = q_1 f_1(\boldsymbol{x})/q_2 f_2(\boldsymbol{x})$, supposing the cost of misclassification can be ignored.

2) In general, using the logarithm of $g(\boldsymbol{x})$ we have:

$$\phi_i(\boldsymbol{x}) = \ln g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{1}{2}(p\ln 2\pi + \ln|\boldsymbol{\Sigma}_i|) + \ln q_i. \tag{2}$$

Expanding the quadratic form, we obtain:

$$\phi_i(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{A}_i\boldsymbol{x}' + \boldsymbol{B}_i'\boldsymbol{x} + \boldsymbol{C}_i,$$

where

$$\boldsymbol{A}_i = -\boldsymbol{\Sigma}_i^{-1}/2, \boldsymbol{B}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i, \text{ and } \boldsymbol{C}_i = -\frac{1}{2}(\boldsymbol{\mu}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i' + p\ln 2\pi + \ln|\boldsymbol{\Sigma}_i|) + \ln q_i. \tag{3}$$

This function $\phi_i(\boldsymbol{x})$ is called the quadratic discriminant function of class π_i, by which we will assign a new observation to class π_{i_0} when $\phi_{i_0}(\boldsymbol{x}_0)$ has the highest value among all $\phi_i(\boldsymbol{x}_0)$. Ignoring $\boldsymbol{A}_i$, $\theta_i(\boldsymbol{x}) = \boldsymbol{B}_i'\boldsymbol{x} + \boldsymbol{C}_i$ is called the linear discriminant function of class *i*. We will essentially use this result in our approach.

An equivalent approach considers the ratio of two of these functions

$$\Delta_{1,2}(\boldsymbol{x}) = \phi_1(\boldsymbol{x})/\phi_2(\boldsymbol{x}) \tag{4}$$

and leads to the decision of classifying a new observation as in class π_1 if this ratio is larger than 1.

The presentation of our article is as follows: in Section 2, we recall the classical discriminant function in the two-class case when training samples are used. Section 3 formalizes the notion of classification and recalls several important results presented in our two previous publications, which are useful to the present one. Section 4 presents the intersections of two normal surfaces and their projections on Oxy. Section 5 deals with classification into one of *C* classes, with $C > 2$. Fisher's approach for multilinear classification is briefly presented there, together with some advantages of our approach. In Section 6, we present an example in classification with $C = 4$, solved with our software Hammax. The minimum function is studied in Section 7 while Section 8 presents the non-parametric approach, as well as the non-normal case, proving the versatility of Hammax.

2. Classification Rules Using Training Samples

Working with samples, since the values of $\boldsymbol{\Sigma}_i, i = 1, 2,$ are unknown, we use plug-in values of the means and

variances and obtain the following results:

1) $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ using $\boldsymbol{S}^*$ and $\bar{\boldsymbol{x}}_i, i = 1, 2$

The decision rule is then:

For a new vector $\boldsymbol{x}_0$, allocate it to class π_1 if

$$\left(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2\right)' \left(\boldsymbol{S}^*\right)^{-1} \boldsymbol{x}_0 - \frac{1}{2}\left(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2\right)' \left(\boldsymbol{S}^*\right)^{-1} \left(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2\right) \geq \ln\left(\frac{1-q}{q}\right), \tag{5}$$

and to class π_2, otherwise. Here $\boldsymbol{S}^*$ is the estimate of the common variance matrix $\boldsymbol{\Sigma}$, and can be obtained by pooling $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$.

We can see that the discriminant function $\boldsymbol{ld}(\boldsymbol{x})$ is linear in $\boldsymbol{x}$, since

$$\boldsymbol{ld}(\boldsymbol{x}) = \left(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2\right)' \left(\boldsymbol{S}^*\right)^{-1} \boldsymbol{x} - \boldsymbol{A}, \tag{6}$$

where $\boldsymbol{A} = -\frac{1}{2}\left(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2\right)' \left(\boldsymbol{S}^*\right)^{-1} \left(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2\right)$ and $\boldsymbol{x}_0 \in \pi_1$ if $\boldsymbol{ld}(\boldsymbol{x}_0) \geq 0$.

2) Different covariance matrices: $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

For a new vector $\boldsymbol{x}_0$, we consider the quadratic discriminate function $\boldsymbol{qd}(\boldsymbol{x})$, and allocate it to π_1 if

$$\boldsymbol{qd}(\boldsymbol{x}_0) = \left[\bar{\boldsymbol{x}}_1' \boldsymbol{S}_1^{-1} - \bar{\boldsymbol{x}}_2' \boldsymbol{S}_2^{-1}\right] \boldsymbol{x}_0 - \frac{1}{2} \boldsymbol{x}_0' \left[\boldsymbol{S}_1^{-1} - \boldsymbol{S}_2^{-1}\right] \boldsymbol{x}_0 - k \geq \ln\left(\frac{1-q}{q}\right), \tag{7}$$

and to π_2, otherwise, where

$$k = \frac{\ln\left(|\boldsymbol{S}_1| / |\boldsymbol{S}_2|\right) + \left(\bar{\boldsymbol{x}}_1' \boldsymbol{S}_1^{-1} \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2' \boldsymbol{S}_2^{-1} \bar{\boldsymbol{x}}_2\right)}{2}.$$

3. Classification Functions

3.1. Decision Surfaces and Decision Regions

Let a population consist of *C* disjoint classes. We now present our approach and prove that for the two class case it coincides with the method in the previous section.

Definition 1. A decision surface $D(\boldsymbol{x})$ is a surface defined in R^{p+1} that separates points assigned to a specific class, from those assigned to other classes.

Definition 2. Let $\{g_i = q_i f_i\}_{i=1}^C$ be a finite family of densities $\{f_i\}_{i=1}^C$, with prior weights $\{q_i\}_{i=1}^C$, with $g_{\max}(\boldsymbol{x}) = \max\{g_1(\boldsymbol{x}), \cdots, g_C(\boldsymbol{x})\}$, $\boldsymbol{x} \in R^p$.

A max-classification function $\varphi_{\{g_i\}}$ is a mapping from a domain $\Omega \subset R^p$ into the discrete family $\{1, 2, \cdots, C\}$, defined as follows:

For a value $\boldsymbol{x}_0 \in \Omega$, $\varphi_{\{g_i\}}(\boldsymbol{x}_0) = i_0$, s.t. $g_{i_0}(\boldsymbol{x}_0) = g_{\max}(\boldsymbol{x}_0)$.

3.2. Properties of $g_{\max}(\cdot)$

There are several other properties associated with the max function and we invite the reader to look at these two articles [2] and [1], to find:

1) Clustering of densities using the width of successive clusters. L^1-distance between 2 densities is well-known but does not apply when there are more than 2 densities. Let us consider *k* densities $f_i(\boldsymbol{x}), i = 1, \cdots, k,$ with $k \geq 3$ and let

$$f_{\max}(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots, f_k(\boldsymbol{x})\}, \forall \boldsymbol{x} \in R^p.$$

A L^1-distance between all densities taken at the same time, cannot really be defined, and the closest to it is a weighted sum of pairwise L^1-distances. However, using $f_{\max}$, we can devise a measure which could be considered as generalized L^1-distance between these *k* functions, since it is consistent with other considerations related to distances in general. This measure is

$$\|f_1, f_2, \cdots, f_k\|_1 \equiv \int_{R^p} f_{\max}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - 1$$

and is slightly different than the case $k = 2$. We have the double inequality

$$\max_{i<j}\|f_i - f_j\|_1 \le 2\|f_1, f_2, \cdots, f_k\|_1 \le \sum_i\sum_j \|f_i - f_j\|_1 .$$

2) Considering now $g_{\max} = \max\{g_i\}$, we study the basic properties of $g_{\max}$, and its role as a classifier. Several original results related to L^1-distances, overlapping coefficients and Bayes errors, are established, for two and more densities. This error can be shown to be $1 - \int g_{\max}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ and several applications were presented.

From [6] and [7], we have the double inequality

$$\frac{1}{2}\max_{i<j}\left\{\|g_i - g_j\|_1\right\} + \min_i\{q_i\} \le \int_{R^p} g_{\max}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \le \frac{1}{k}\left\{\sum_i\sum_j \|g_i - g_j\|_1 + 1\right\},$$

with Bayes error given by

$$Pe_{1,2,\cdots,k}^{[(q)]} = 1 - \int_{R^p} g_{\max}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\ (q) = \left(\frac{1}{C}, \cdots, \frac{1}{C}\right)$$

since $\int_{R^p} g_{\max}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ still represents the unconditional probability of correct classification.

4. Discrimination between 2 Classes

For simplicity and for graphing purpose we will consider only the bivariate case $p = 2$ in the rest of the article. However, all arguments can be applied to the case $p > 2$, and the basic answer on the classification of a new data point is still provided.

4.1. Determining the Function $g_{\max}$

Our approach is to determine the function $g_{\max}$ and use it with the max-classification function $\phi_{\{g_i\}}$. This is achieved by finding the regions of definition of $g_{\max}$ in R^2, *i.e.* by determining their boundaries as projections onto the horizontal plane of intersections between transformed normal surfaces $\{x, g_i(x)\} \in R^3$, and the value of $\{x, g_{\max}\}$ there.

1) For the two-class case we show that this approach is equivalent to the common Bayesian approach recalled earlier in Section 2. First, from Equation (6), equation $\boldsymbol{ld}(\boldsymbol{x}) = 0$ determines precisely the linear boundary of the two adjacent regions where $g_{\max} = qf_1$ and $g_{\max} = (1-q)f_2$ respectively, and hence the two approaches are equivalent in this case. Second, from Equation (7), $\boldsymbol{qd}(\boldsymbol{x}) = 0$ also determines the quadratic boundary (ies) of the region separating $g_{\max} = qf_1$ from $g_{\max} = (1-q)f_2$ since the two surfaces g_1 and g_2 intersect each other along curves which have quadratic projections (Straight lines, Ellipses, Parabolas or Hyperbolas) on the horizontal plane. But whereas the common Bayesian approach only retains only the linear, or quadratic, boundary for decision purpose, $g_{\max}$ retains a partial surface on each side of the boundary and atop of the horizontal plane. This fact makes the max-classification function much more useful.

When the dimension of *p* exceeds 2 we have these projections as hyperquadrics, which are harder to visualize and represent graphically.

2) For the C classes case, $C > 2$: In general, when there are *C* classes the intersections between each of the $\binom{C}{2}$ couples of normal surfaces $\{f_i\}$ are space curves in R^3, and their projections into the horizontal plane determine definition regions of $f_{\max}$.

These regions are given below. Once they are determined they are clearly marked down as assigned to class *i*, or to class *j*, and the family of all these regions will give the classification regions for all observations. Naturally, definition regions for $g_{\max}$ are deformations of those of $f_{\max}$, but have to be computed separately since there is no rules to go from one set of regions to the other. They are identical only in the case $q_1 = q_2 = \cdots = q_C = 1/C$.

4.2. Intersections of Normal Surfaces

In the non-normal case, the intersection space curve(s), and its projection, can be quite complex (see example 6).

Below are some examples for the normal case.

Two normal surfaces, representing f_1 and f_2, always intersect each other along a curve, or two curves in R^3, which, when projected in the (*x*, *y*)-plane, give(s) a quadratic curve, whose equation can be obtained by solving $f_1 = f_2$, where

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)}, i = 1, 2.$$

In R^2, taking the logarithm, we have:

$$\ln f_i(x, y) = -\frac{1}{2(1-\rho_i^2)}\left[\left(\frac{x-\mu_{x_i}}{\sigma_{x_i}}\right)^2 - 2\rho_i\left(\frac{x-\mu_{x_i}}{\sigma_{x_i}}\right)\left(\frac{y-\mu_{y_i}}{\sigma_{y_i}}\right) + \left(\frac{y-\mu_{y_i}}{\sigma_{y_i}}\right)^2\right] + \ln\left(\frac{1}{2\pi\sigma_{x_i}\sigma_{y_i}\sqrt{1-\rho_i^2}}\right), i = 1, 2.$$

Equaling the two expressions we obtain equations of the projections (in the horizontal plane) of the intersections curves in R^3. There are several cases for these intersections, depending on the values of the mean vectors and the covariance matrices. We do not give them here, to avoid confusion, but they are sketched in the appendix and are available upon request. Instead, we give clear examples and graphs of the different cases.

1) A straight line (when covariance matrices are equal), or a pair of straight lines, parallel or intersecting each other.

2) A parabola: This happens when $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and $\sigma_{x_1} \gg \sigma_{x_2}$ and $\sigma_{y_1} \gg \sigma_{y_2}$.

Example 1.

Let $\boldsymbol{\mu}_1 = \begin{pmatrix} 5.00 \\ 1.43 \end{pmatrix}$, $\boldsymbol{\mu}_2 = \begin{pmatrix} 4.58 \\ 2.97 \end{pmatrix}$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 2 & 2 \\ 2 & 5 \end{pmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}$ and $\rho_1 = 0.6325$, $\rho_2 = 0.5$.

Figure 1 shows the graph of $\max\{f_1, f_2\}$ in 3D where the intersection of these two normal surfaces is a parabola.

$f_{\max}$'s boundary: The equation of this parabola is

$$-0.5x^2 + 2.85x + 1.3066y - 4.4 = 0$$

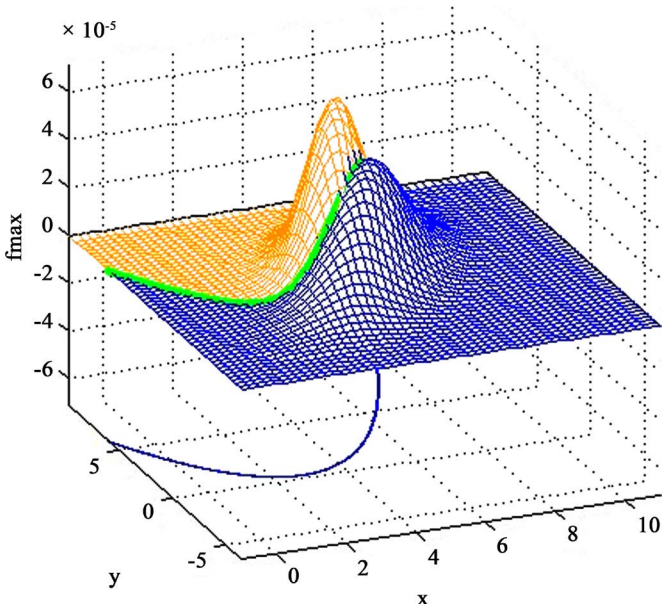

Figure 1. $f_{\max}$ 3D-view.

3) An ellipse: When $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, and $\rho_1 = \rho_2$.

Example 2.

Let $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \begin{pmatrix} 4 \\ 4 \end{pmatrix}$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.3525 & 0.2714 \\ 0.2714 & 0.3790 \end{pmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 2.2031 & 1.6962 \\ 1.6962 & 2.3687 \end{pmatrix}$, $\rho_1 = \rho_2 = 0.7425$, ($\sigma_{x_2} = 2.5\sigma_{x_1}$ and $\sigma_{y_2} = 2.5\sigma_{y_1}$).

Figure 2 shows the graph of $\max\{f_1, f_2\}$ in 3D, where the intersection of these two normal surfaces is an ellipse.

The equation of this ellipse is

$$5.3114x^2 - 7.6069y^2 + 4.94xy - 12.0634x - 9.0923y + 38.6461 = 0$$

4) A hyperbola: This happens when $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and $\sigma_{x_1} \neq \sigma_{x_2}$, $\sigma_{y_1} \neq \sigma_{y_2}$.

Example 3. Let $\boldsymbol{\mu}_1 = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1.80 & 0.27 \\ 0.27 & 1.00 \end{pmatrix}$, $\rho_1 = 0.16$, $\boldsymbol{\mu}_2 = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 3.00 & 1.45 \\ 1.45 & 1.00 \end{pmatrix}$, $\rho_2 = 0.84$.

Figure 3 shows the graph of $\max\{f_1, f_2\}$ in 3D, where the intersection of these two normal surfaces is a

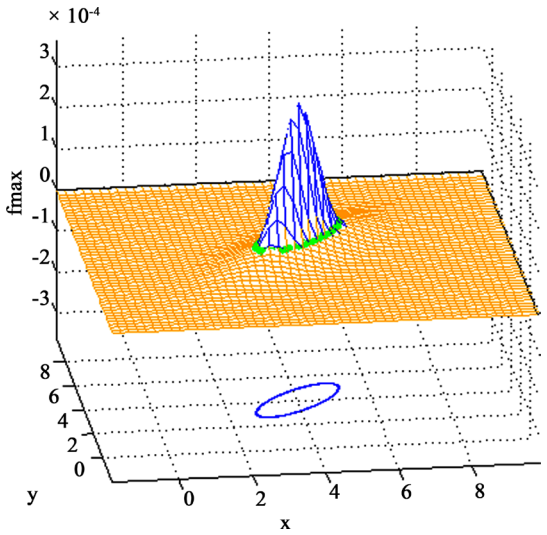

Figure 2. $f_{\max}$'s 3D view.

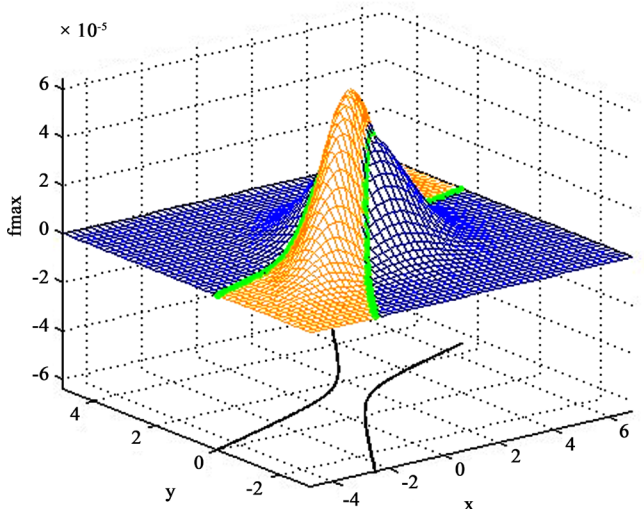

Figure 3. $f_{\max}$'s 3D view.

hyperbola.

The equation of this hyperbola is

$$-0.562x^2 + 3.049y^2 - 2.3702xy - 2.4957x + 1.8133y + 1.3936 = 0.$$

5. Classification into One of C Classes ($C \geq 3$)

The $g_{\max}$ function is quite simple when the three class covariance matrices are equal, as can be seen from **Figure 4(a)**. Then the discriminant functions are all straight lines intersecting at a common point. These lines are projections of normal surface intersection curves.

In the case these matrices are unequal they can intersect according to a complicated pattern, as shown in **Figure 4(b)**.

5.1. Our Approach

For normal surfaces of different means and covariance matrices, in the common Bayesian approach we can use (6) or (7), or equivalently, classify a new value $\boldsymbol{x}_0$ into the class j_0 such that $\phi_{j_0}(\boldsymbol{x}_0) = \max_j \{\phi_j(\boldsymbol{x}_0)\}$. In the common Bayesian approach, we have the choice between:

1) One against all, using the $(C-1)$ discriminant functions (6), with the dichotomous decision each time: $\boldsymbol{x}_0$ is in group j or not in group *j*.

2) Two at a time, using $C(C-1)/2$ expressions (6) or (7) with regions delimited by straight lines or quadratic curves, each expression classifies new data as in C_i or C_j.

As pointed out by Fukunaga ([8], p. 171) these methods can lead to regions not clearly assignable to any group.

In our approach, we use the second method and compile all results so that R^2 is now divided into disjoint sub-regions, each having a surface atop of it, which constitute the graph of $g_{\max}$ in R^3. Then, for a new observation $\boldsymbol{x}_0$, to classify it we just use the max-classification function $\varphi_{\{g_i\}}$ given in Definition 2.

5.2. Fisher's Approach

It is the method suggested first [9], in the statistical literature for discrimination and then for classification. It is still a very useful method. The main idea is to find, and use, a space of lesser dimensions in which the data is projected, with their projections exhibiting more discrimination, and being easier to handle.

1) Case of 2 classes. It can be summarized as follows:

$C = 2,\ p = 2,\ r = p - 1 = 1$: Projection into a direction which gives the best discrimination: Decomposition of total variation

$$\boldsymbol{S}_T = \boldsymbol{S}_W + \boldsymbol{S}_B,$$

with $\boldsymbol{S}_W = \boldsymbol{S}_1 + \boldsymbol{S}_2$ and $\boldsymbol{S}_i = \sum_{\boldsymbol{x} \in D_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)'$, $\boldsymbol{m}_i = \frac{1}{n_i} \sum_{\boldsymbol{x}_k \in D_i} \boldsymbol{x}_k$, $i = 1, 2$. We then search for a direction $\boldsymbol{w}$ such that

$$J(\boldsymbol{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

is maximum, where $\tilde{m}_i$ and $\tilde{s}_i^2$ are projected values into that direction. We have $\boldsymbol{w} = \boldsymbol{S}_W^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$ with $\boldsymbol{S}_B = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)'$.

Fisher's method in this case reduces to the common Baysian method if we suppose the populations normal. Implicitly it already supposes the variances equal. But Fisher's method allows the consideration that variables can be can enter individually, so as to measure their relative influence, as in analysis of variance and regression.

2) Fisher's multilinear method (extension of the above approach due to CR Rao): *C* classes, of dimension *p* and $r = C - 1 < p$.

Projection into space of $\dim r$: Decomposition of total variation in original space:

$\boldsymbol{S}_T = \sum_{\boldsymbol{x}} (\boldsymbol{x} - \boldsymbol{m})(\boldsymbol{x} - \boldsymbol{m})'$, $\boldsymbol{m} = \frac{1}{n} \sum_{i=1}^{C} n_i \boldsymbol{m}_i$, $\boldsymbol{S}_T = \boldsymbol{S}_W + \boldsymbol{S}_B$, with

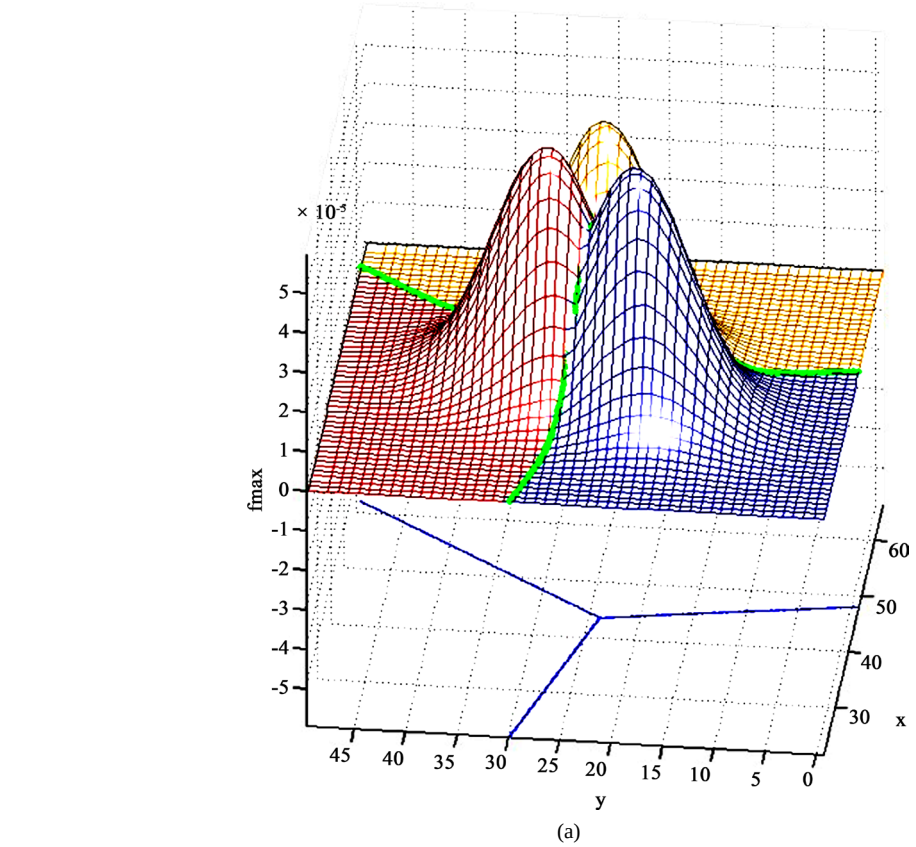

Figure 4. (a). $g_{\max}$ 3D-view in the case of three equal covariance matrices; (b) $g_{\max}$ 3D-view in the case of three unequal covariance matrices.

$$\boldsymbol{S}_W = \sum_{i=1}^{C} \sum_{\boldsymbol{x} \in D_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)',$$

$$\boldsymbol{S}_B = \sum_{i=1}^{C} n_i (\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})'.$$

The projection from a p-dim space to a $(C-1)$-dim space is done with a matrix $\boldsymbol{W}$ and we have $\boldsymbol{y} = \boldsymbol{W}'\boldsymbol{x}$. Using $\boldsymbol{y}$, let the projected quantities be $\tilde{\boldsymbol{S}}_W = \boldsymbol{W}'\boldsymbol{S}_W\boldsymbol{W}$, $\tilde{\boldsymbol{S}}_B = \boldsymbol{W}'\boldsymbol{S}_B\boldsymbol{W}$. We want to find the matrix $\boldsymbol{W}$ so that the ratio $J(\boldsymbol{W}) = |\tilde{\boldsymbol{S}}_B| / |\tilde{\boldsymbol{S}}_W|$ is maximum. Solving $|\boldsymbol{S}_B - \lambda_i \boldsymbol{S}_W| = 0$ to obtain λ_i and then solving $(\boldsymbol{S}_B - \lambda_i \boldsymbol{S}_W)\boldsymbol{w}_i = 0$ to have eigenvectors $\boldsymbol{w}_i$, we obtain the matrix $\boldsymbol{W}$, which often is not unique. Within the $(C-1)$-dim space a probability distribution can be found for the projected data, which will provide cut-off values to classify a new observation into one of the *C* classes.

We can see that Fisher's multilinear method can be quite complicated.

5.3. Advantages of Our Approach

Our computer-based approach offers the following advantages:

1) It uses concepts at the base: Max of $\{g_i\}_{i=1}^{C}$, and is self-explanatory in simple cases. It avoids several matrix transformations and projections of Fisher's method, which could, or could not be done.

2) The determination of the maximum function is essentially machine-oriented, and can often save the analyst from performing complex matrix or analytic manipulations. This point is of particular interest when this analysis concerns vectors of high dimensions. To classify a new observation $\boldsymbol{x}_0$ into the appropriate group, say j_0, it suffices now to find the index j_0 so that $g_{j_0}(\boldsymbol{x}_0) = g_{\max}(\boldsymbol{x}_0)$. This operation can always be done since *C* is finite.

3) Complex cases arise when there are a large number of classes, or a large number of variables (high value for p). But as long as the normal surfaces can be determined the software Hammax can be used. In the case where *p* is much larger than the sample sizes, we have to find the most significant dimensions and use them only, before applying the software.

4) It offers a visual tool very useful to the analyst when $p = 2$. The full use of the function $g_{\max}$ in R^3 necessitates the drawing of its graph, which could be a complex operation in the past, but not now. In general, the determination of the intersections between densities (or between g_i) in R^{p+1}, and their projections into R^p, gives more insights into the problem: in classical statistical discriminant analysis, we only deal with these projections, and do not consider the curves in R^p, of which they are projections (Equation (6) and Equation (7)). Hence, for any other family of densities which has the same intersections in R^p as those already considered, we would have the same classification rule. For $p \geq 3$ integration of $g_{\max}$ is carried out using an appropriate approach (see [1]) and classification of a new data point can again be made.

5) Regions not clearly assignable to any group, are removed with the use of $g_{\max}$, as already mentioned.

6) For the non-normal case, $g_{\max}$ can still provide a simple practical approach to classification, as can be seen in Example 6, where $g_{\max}$ does allow us to derive classification rules. [10] can be consulted for this case.

7) It permits the computation of the Bayes error, which can be used as a criterion in ordering different classification approaches. Naturally, the error computed by our software from data is an estimation of the theoretical, but unknown, Bayes error obtained from population distributions.

6. Output of Software Hammax in the Case of 4 Classes

The integrated computer software developed by our group is able to handle most of the computations, simulations and graphic features presented in this article. This software extends and generalizes some existing routines, for example the Matlab function Bayes Gauss ([5], p. 493), which is based on the same decision principles.

Below are some of its outputs, first in the case of classification into a four-class population.

Example 4. Numerical and graphical results determining $g_{\max}$ in the case of four classes in two dimensions f_1, f_2, f_3, f_4, *i.e.* $\boldsymbol{X}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1,2,3,4$, with

$$\bar{X}_1 = \begin{pmatrix} 40 \\ 20 \end{pmatrix}, \bar{X}_2 = \begin{pmatrix} 48 \\ 24 \end{pmatrix}, \bar{X}_3 = \begin{pmatrix} 43 \\ 32 \end{pmatrix}, \bar{X}_4 = \begin{pmatrix} 38 \\ 28 \end{pmatrix}$$

$$S_1 = \begin{pmatrix} 35 & 18 \\ 18 & 20 \end{pmatrix}, S_2 = \begin{pmatrix} 28 & -20 \\ -20 & 25 \end{pmatrix},$$

$$S_3 = \begin{pmatrix} 15 & 25 \\ 25 & 65 \end{pmatrix}, S_4 = \begin{pmatrix} 5 & -10 \\ -10 & 70 \end{pmatrix}$$

$g_{\max} = \max\{q_1 f_1, q_2 f_2, q_3 f_3, q_4 f_4\}$, where $q_1 = 0.25, q_2 = 0.20, q_3 = 0.40, q_4 = 0.15$. **Figure 5** gives the 3D graph of $g_{\max}$ in Oxyz (with projections of the intersection curves onto Oxy):

To obtain **Figure 6** we use all intersection curves given in **Figure 7** below.

In this example we have all hyperbolas as boundaries in the horizontal plane. Their intersections will serve to determine the regions of definition of $g_{\max}$. **Figure 8** below shows us these regions.

Classification: For the new observation, for example (25, 35), we can see that it is classified in C_4.

Note: In the above graph, for computation purpose we only consider $g_{\max}$ within a window $[a,b]\times[c,d]$ in Oxy, with $a = 18.06$, $b = 61.94$, $c = 3.32$, $d = 56.68$. We can show that outside this window the values of the integrals of $q_1 f_1, \cdots, q_4 f_4$ are negligible and using these results we can compute $\int_{R^2} g_{\max}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0.7699$.

7. Risk and the Minimum Function

1) When risk, as the penalty in misclassification, is considered in decision making we aim at the min risk rather than the max risk. In the literature, to simplify the process, we usually take the average risk, also called Bayes

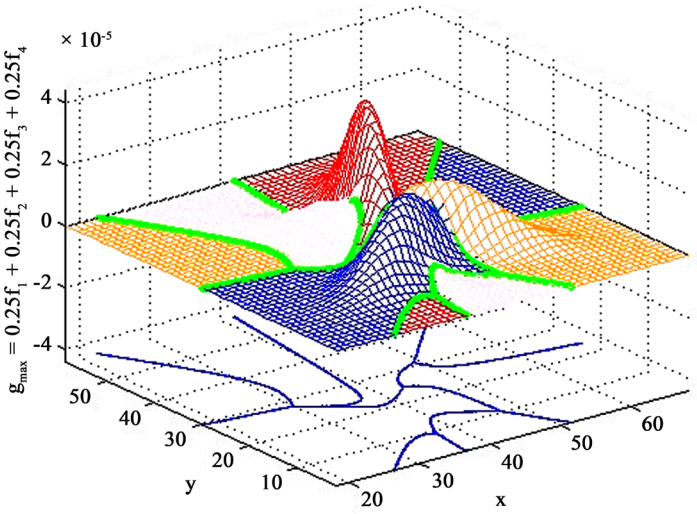

Figure 5. 3D Graph of $g_{\max}$.

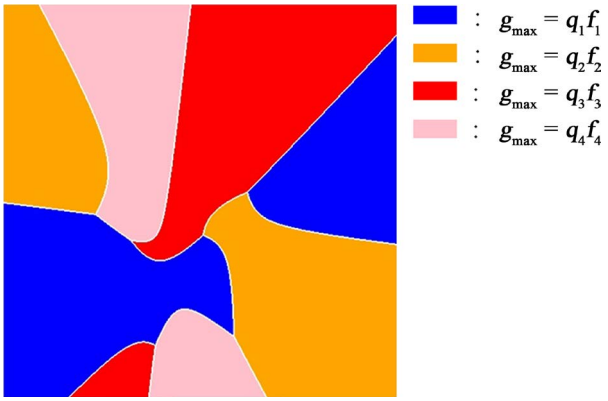

Figure 6. Regions in Oxy of definition of $g_{\max}$.

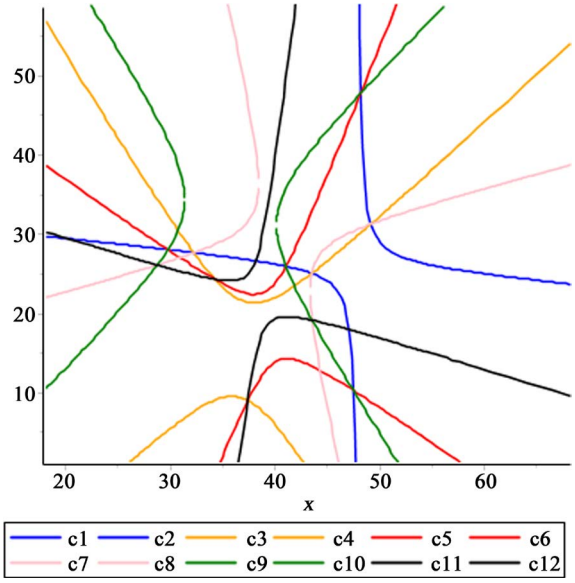

Figure 7. Projections of intersection curves of g_i surfaces onto Oxy.

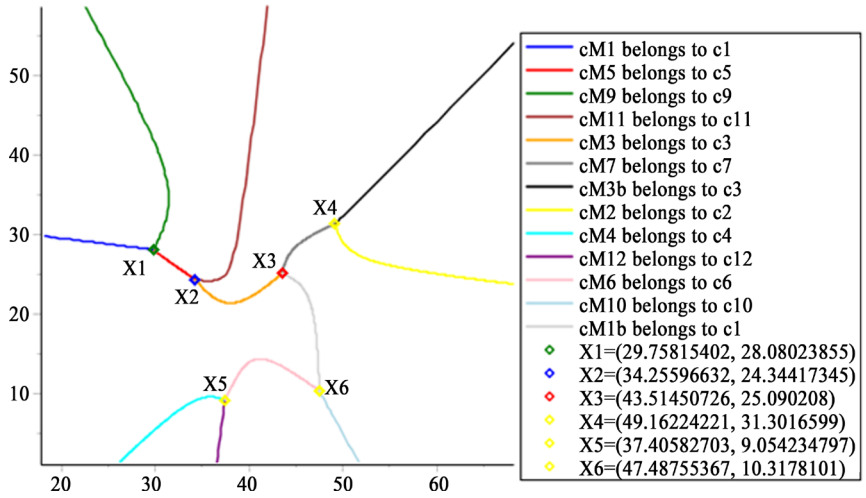

Figure 8. Points used to determine definition regions of $g_{\max}$.

risk, or the min of all max values of all different risks, according to the minimax principle.

We suppose here that risk R_i has f_i as its normal probability distribution, function of 2 variables *x* and *y*, and various competing risks $\{R_k\}_{k=1}^{C}$ are present.

A minimum-classification function $\psi_{\{g_i\}}(x,y)$ is defined similarly to $\phi_{\{g_i\}}(x,y)$.

Definition 4. A min-classification function $\psi_{\{g_i\}}$ is a mapping from a domain $\Omega \subset R^p$ into the discrete family $\{1,2,\cdots,C\}$, defined as follows:

For a value $\boldsymbol{x}_0 \in \Omega$, $\psi_{\{g_i\}}(\boldsymbol{x}_0) = i_0$, s.t. $g_{i_0}(\boldsymbol{x}_0) = g_{\min}(\boldsymbol{x}_0)$.

2) A relation between the max and min functions can be established by using the inclusion-exclusion principle:

$$f_{\max} = \sum_{i=1}^{k} f_i - \sum_{i<j} \min\left(f_i, f_j\right) + \sum_{i<j<l} \min\left(f_i, f_j, f_l\right) + \cdots + (-1)^{k-1} \min\left(f_1, \cdots, f_k\right).$$

Integrating this relation we have a relation between $\int f_{\max}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ and various integrals on the minimums of

subgroups of $\{f_i\}_{i=1}^{C}$. For classification purpose we classify a new set of data (x_0, y_0) as belonging to the class having the lowest risk at that point. The function $g_{\min}$ represents the minimum risk, but is, however, the invisible part of the graphs since it lies below all surfaces g_i.

Example 5. The four normal distributions are the same as in **Example 1** but represent the densities of the risks associated with the problem. Using the same prior probabilities the function $g_{\min}$ is given by **Figure 9(a)** while its definition regions are given by **Figure 9(b)**.

For $g_{\min}$, similarly to $g_{\max}$, we can approximately compute $\int g_{\min}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0.007546$.

Remarks: a) For the two-population case this integral is also the overlap coefficient and can be used for inferences on the similarity, or difference between the two populations.

b) The boundaries between regions defining $g_{\max}$ are in general linear or quadratic curves coming from the intersections of normal surfaces. Boundary between regions defining $g_{\min}$ can be simpler since they might not come from these intersections.

8. Applications and Other Considerations

8.1. The Software Hammax

This software has been developed by our research group and is part of a more elaborate software to deal with discrimination, classification and cluster analysis, as well as with other applications related to the multinormal distribution. This software is in further development to be interactive and more user-friendly, and has its own

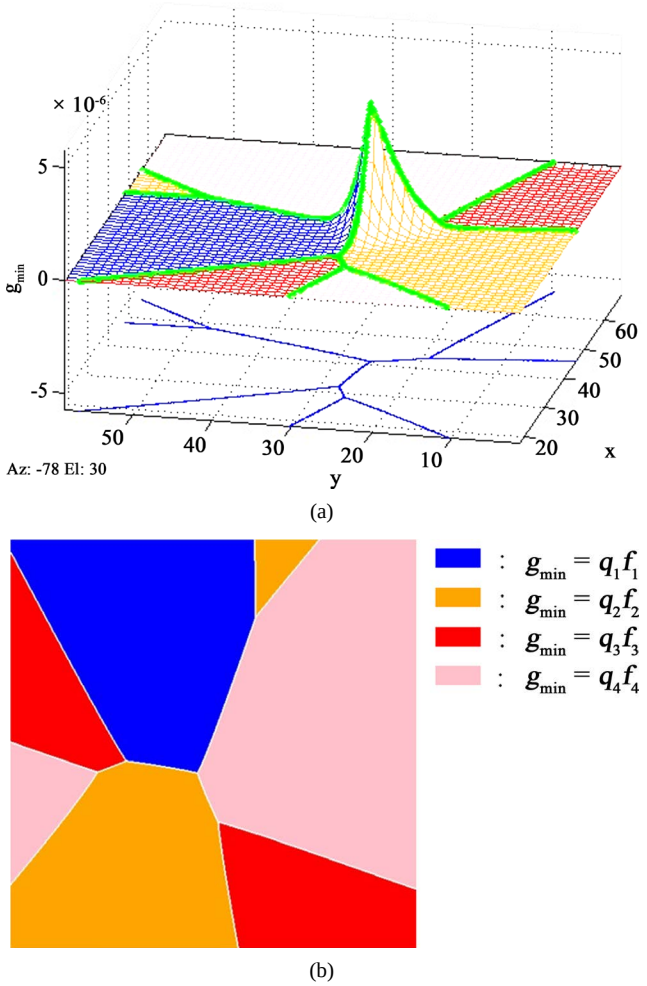

Figure 9. (a). 3D-graph of $g_{\min}$ in Oxyz; (b). Regions defining $g_{\min}$.

copyright. It will also have more connections with social sciences applications.

8.2. The Non-Parametric Density Estimation Approach

A more general approach would directly use data available in each group to estimate the density of its distribution. The $g_{\max}$ function approach to classification would then follow, exactly as for the normal case. But, unless we approximate the density obtained by parametric methods, different regions of definition of $g_{\max}$ can only be obtained empirically, to be used in the classification of a new data x_0. Densities of all classes are now estimated by the classical kernel density estimation method for two variables, x_1 and x_2. Using the kernel

$$K(t)=\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-t^2}{2}\right),$$

they are estimated by

$$f_i(x_1,x_2)=\frac{1}{nh_1h_2}\sum_{j=0}^{n-1}\left\{K\left(\frac{x_1-\hat{x}_{1j}}{h_1}\right)K\left(\frac{x_2-\hat{x}_{2j}}{h_2}\right)\right\},\ i=1,\cdots,C,$$

where $(\hat{x}_{1j},\hat{x}_{2j})$ is the j-th observation. Optimal values for h_1 and h_2 has been discussed by various authors ([11]). We refer to [1] where a numerical example was redone, using density estimation. Also, we have the associated function $g^*_{\max}$. The Bayes error

$$Pe_{1,2}^{[(1/3)]}=1-\int_{R^2}g^*_{\max}(\mathbf{x})\,\mathrm{d}\mathbf{x}=0.125,$$

computed by simulation, gives the same value as for the parametric normal case.

8.3. Non-Normal Model

As stated earlier an approach based on the maximum function is valid for non-normal populations. We construct here an example for such a case.

Let us consider the case where the population density $f_1(x_1,x_2)$ in $[0,1]^2$, given by

$$f_1(x_1,x_2)=\prod_{i=1}^{2}h_i(x_i;\alpha_i,\beta_i),$$

where $h_i(x_i;\alpha_i,\beta_i)$, $i=1,2$, are independent standard beta densities of the first kind, *i.e.*

$$h_i(x_i;\alpha_i,\beta_i)=x_i^{\alpha_i-1}(1-x_i)^{\beta_i-1}/Beta(\alpha_i,\beta_i),$$

with $\alpha_i,\beta_i>0$, and $0\le x_i\le 1$.

Similarly, we have:

$$f_2(x_1,x_2)=\prod_{i=1}^{2}k_i(x_i;\gamma_i,\delta_i),$$

where $k_i(x_i;\gamma_i,\delta_i)$ are also independent beta densities.

Example 6. For $\alpha_1=3$, $\beta_1=6$, $\alpha_2=4$, $\beta_2=7$, $\gamma_1=4$, $\delta_1=5$ and $\gamma_2=6$, $\delta_2=5$ and $q=0.35$, the $g_{\max}$ function is defined in $[0,1]^2$ by

$$g_{\max}(x_1,x_2)=\max\{0.35f_1(x_1,x_2),0.65f_2(x_1,x_2)\}$$

We can see that the last two functions intersect each other along a curve in R^3, the projection of which in R^2 is the discriminant curve giving the boundary between the two classification regions, as given by **Figure 10**, with an equation which is neither linear nor quadratic, since its expression is

$$f(x_1)=\frac{7A(1-2x_1+x_1^2)}{7A(1-2x_1+x_1^2)+13Bx_1^3},\ 0\le x_1\le 1,$$

where $A=beta(4,5)beta(6,5)$ and $B=beta(3,6)beta(3,7)$. **Figure 10** illustrates this case.

This curve will serve in the classification of a new observation in either of the two groups.

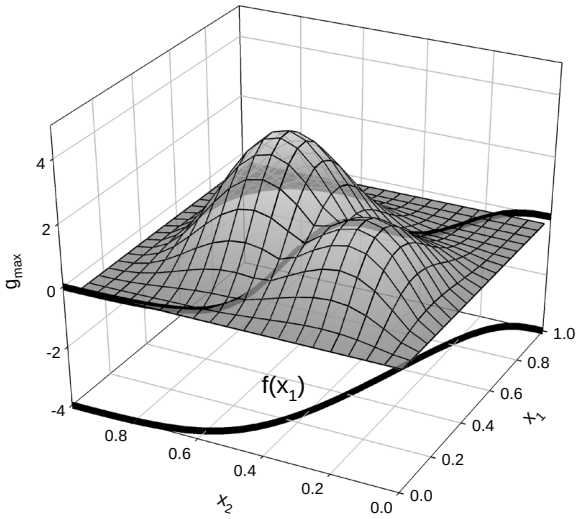

Figure 10. Two bivariate beta densities, their intersection and its projection.

Any data above the curve, e.g. (0.2, 0.6), is classified as in Class 1. Otherwise, e.g. (0.2, 0.2), it is in Class 2. Numerical integration gives

$$Pe_{1,2}^{(0.35)} = 1 - \int_{R^2} g_{\max}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0.1622.$$

9. Conclusion

The maximum function, as presented above, gives another tool to be used in Statistical Classification and Analysis, incorporating discriminant analysis and the computation of Bayes error. In the two-dimensional case, it also provides graphs for space curves and surfaces that are very informative. Furthermore, in higher dimensional spaces, it can be very convenient since it is machine oriented, and can free the analyst from complex analytic computations related to the discriminant function. The minimum function is also interested, has many applications of its own, and will be presented in a separate article.

References

[1] Pham-Gia, T., Turkkan, N. and Vovan, T. (2008) Statistical Discrimination Analysis Using the Maximum Function, Communic. in Stat., *Computation and Simulation*, **37**, 320-336. http://dx.doi.org/10.1080/03610910701790475

[2] Vovan, T. and Pham-Gia, T. (2010) Clustering Probability Densities. *Journal of Applied Statistics*, **37**, 1891-1910.

[3] Duda, R.O., Hart, P.E. and Stork, D.G. (2001) Pattern Classification. John Wiley and Sons, New York.

[4] Johnson and Wichern (1998) Applied Multivariate Statistical Analysis. 4th Edition, Prentice-Hall, New York. http://dx.doi.org/10.2307/2533879

[5] Gonzalez, R.C., Woods, R.E. and Eddins, S.L. (2004) Digital Image Processing with Matllab. Prentice-Hall, New York.

[6] Glick, N. (1972) Sample-Based Classification Procedures Derived from Density Estimators. *Journal of the American Statistical Association*, **67**, 116-122. http://dx.doi.org/10.1080/01621459.1972.10481213

[7] Glick, N. (1973) Separation and Probability of Correct Classification among Two or More Distributions. *Annals of the Institute of Statistical Mathematics*, **25**, 373-382. http://dx.doi.org/10.1007/BF02479383

[8] Fukunaga (1990) Introduction to Statistical Pattern Recognition. 2nd Edition, Academic Press, New York.

[9] Fisher, R.A. (1936) The Statistical Utilization of Multiple Measurements. *Annals of Eugenic*, **7**, 376-386.

[10] Flury, B. and Riedwyl, H. (1988) Multivariate Statistics. Chapman and Hall, New York. http://dx.doi.org/10.1007/978-94-009-1217-5

[11] Martinez, W.L. and Martinez, A.R. (2002) Computational Statistics Handbook with Matlab. Chapman & Hall/CRC, Boca Raton.

Appendix

In R^2, taking the logarithm, we have the equations for a family of quadratic curves:

$$a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_1x + 2a_2y + a = 0,$$

where,

$$a_{11} = \frac{1}{2\sigma_{x_2}^2\left(1-\rho_2^2\right)} - \frac{1}{2\sigma_{x_1}^2\left(1-\rho_1^2\right)},$$

$$a_{22} = \frac{1}{2\sigma_{y_2}^2\left(1-\rho_2^2\right)} - \frac{1}{2\sigma_{y_1}^2\left(1-\rho_1^2\right)},$$

$$a_{12} = \frac{\rho_2}{\sigma_{x_2}\sigma_{y_2}\left(1-\rho_2^2\right)} - \frac{\rho_1}{\sigma_{x_1}\sigma_{y_1}\left(1-\rho_1^2\right)},$$

$$a_1 = \frac{1}{4\left(1-\rho_2^2\right)}\left(-\frac{\mu_{x_2}}{\sigma_{x_2}^2} + \frac{\rho_2\mu_{y_2}}{\sigma_{x_2}\sigma_{y_2}}\right) - \frac{1}{4\left(1-\rho_1^2\right)}\left(-\frac{\mu_{x_1}}{\sigma_{x_1}^2} + \frac{\rho_1\mu_{y_1}}{\sigma_{x_1}\sigma_{y_1}}\right),$$

$$a_2 = \frac{1}{4\left(1-\rho_2^2\right)}\left(-\frac{\mu_{y_2}}{\sigma_{y_2}^2} + \frac{\rho_2\mu_{x_2}}{\sigma_{x_2}\sigma_{y_2}}\right) - \frac{1}{4\left(1-\rho_1^2\right)}\left(-\frac{\mu_{y_1}}{\sigma_{y_1}^2} + \frac{\rho_1\mu_{x_1}}{\sigma_{x_1}\sigma_{y_1}}\right),$$

$$a = \frac{1}{4\left(1-\rho_2^2\right)}\left(\frac{\mu_{x_2}^2}{\sigma_{x_2}^2} + \frac{\mu_{y_2}^2}{\sigma_{y_2}^2} - 2\rho_2\frac{\mu_{x_2}\mu_{y_2}}{\sigma_{x_2}\sigma_{y_2}}\right) - \frac{1}{4\left(1-\rho_1^2\right)}\left(\frac{\mu_{x_1}^2}{\sigma_{x_1}^2} + \frac{\mu_{y_1}^2}{\sigma_{y_1}^2} - 2\rho_1\frac{\mu_{x_1}\mu_{y_1}}{\sigma_{x_1}\sigma_{y_1}}\right)$$
$$- \ln\left(\frac{\sigma_{x_2}\sigma_{y_2}\sqrt{1-\rho_2^2}}{\sigma_{x_1}\sigma_{y_1}\sqrt{1-\rho_1^2}}\right).$$

The forms of these quadratic curves depend on the values of the above coefficients, which, in turn, depend on the parameters of the two normal surfaces.