{file logi00.tex appeared *Biometrical Letters* 37(2000), No. 1, 21–40 }

# LOGISTIC DISCRIMINANT FUNCTION – HOW MUCH ADVANTAGES OVER THE LINEAR ONE ?

Anna Bartkowiak, Adam Szustalewicz and Joanna Zdziarek

Institute of Computer Science, University of Wrocław, Przesmyckiego 20, Wrocław 51-151 Poland e-mail: {aba,asz,zdziarek}@ii.uni.wroc.pl

Running title: `Logistic discriminant function -- applicability`

**Abstract**

The logistic discriminant function receives much attention, especially in epidemiological research. It is thought as being more general than the ordinary linear Fisherian discriminant function. However, one has to pay for the assumed generality by applying more complicated iterative computations.

Our aim is to issue a warning, that the generality attained when using the logistic formula is very restricted.

We show in more detail, for what kind of data the logistic discriminant function is better, and when it is equivalent to the linear Fisherian one. We show also, for what kind of data both the logistic and the linear Fisherian discriminant function are worse than the quadratic discriminant function.

**Key words:** Two-group discriminant function, probability a posteriori, exponential family of distributions, shape of covariance matrices, long–tailed distributions.

# 1 Introduction

The logistic discriminant function receives much attention in statistical data analysis. It is especially popular in medicine when dealing with epidemiological data. For a review of the topic, its methodology, applications and further references see, e.g. [16, 11, 13, 14]. However it still seems, that the role of the logistic discriminant function – as opposed to the linear and quadratic ones – is perhaps overestimated.

The aim of our paper is to show wherefrom the logistic discriminant function has descended. We want to state clearly what can and what cannot be expected when applying that function.

It happens, especially in epidemiology, when comparing a disease group with a control group of (healthy) subjects, that the disease group exhibits different variances and covariances than the control. In such case the fundamental assumption underlying the derivation of the logistic discriminant function (on the equality of the covariance matrices ) is not satisfied, and the logistic function may be inappropriate for the analyzed data.

Let $\mathbf{x} = (x_1, \ldots, x_p)$ denote an observed vector of data values. In *Section 2* we introduce briefly the Fisherian linear discriminant function for two groups of data. In *Section 3* we consider generally the Bayesian rule and the concept of the probability á *posteriori* for two groups of data and we arrive at the logistic discriminant function with argument $a(\mathbf{x})$;

this argument being generally a non-linear function of the observed variables $X_1, \ldots, X_p$. In *Section 4* we consider probability distributions which are specific members of the exponential family of distributions. For such distributions the argument of the logistic discriminant function (derived in Section 3) reduces to a linear function of the observed variables: $a(\mathbf{x}) = b_0 + b_1 x_1 + \ldots b_p x_p$. *Section 5* considers the practice of using the linear, logistic and quadratic discriminant functions. We generate 2-dimensional data following specific distributions: Banana shape, Higleyman shape and Lithuania sausage shape, also independent negative binomial and gamma distributions and compare in their context the role of the logistic discriminant function with the two other ones. *Section 6* gives an overall summary of the results and some indications how to proceed in practice.

# 2 The linear Fisherian discriminant function as a classification rule

Already in 1936 R.A. Fisher (quoted after [16]) has considered the problem : Suppose that we have to do with given measurements of two groups of data, coming from two populations with expected vectors $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ ($\boldsymbol{\mu_k} = (\mu_{k1}, \ldots, \mu_{kp}), k = 1, 2$) and a common covariance matrix $\mathbf{\Sigma}$. Suppose for the moment, that the expected values $\boldsymbol{\mu_1}, \boldsymbol{\mu_2}$ and the covariance matrix $\mathbf{\Sigma}$ are known. Let $\mathbf{x} = (x_1, x_2, \ldots x_p)$ denote a vector of observations in $p$ variables coming from one of the populations.

Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ denote a column vector of coefficients. To discriminate between the two populations, Fisher proposed to use the linear combination $\mathbf{x}\boldsymbol{\beta}$ that maximizes the ratio

$$(\boldsymbol{\mu_1}\boldsymbol{\beta} - \boldsymbol{\mu_2}\boldsymbol{\beta})^2 / \boldsymbol{\beta}^T \mathbf{\Sigma} \boldsymbol{\beta}. \tag{1}$$

After some algebra we arrive at the result, that the vector $\boldsymbol{\beta}$ maximizing the ratio above is given as

$$\boldsymbol{\beta}^T = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})\mathbf{\Sigma}^{-1}. \tag{2}$$

The linear function $z = \mathbf{x}\boldsymbol{\beta}$ is called the *linear discriminant function*. The elements of the vector $\boldsymbol{\beta}$ are called *coefficients* of the linear discriminant function. The maximal ratio attained after substituting into (1) the expression for $\boldsymbol{\beta}$ evaluated from formula (2) is then equal to $D^2 = \boldsymbol{\beta}^T \mathbf{\Sigma}^{-1} \boldsymbol{\beta}$, which is the *Mahalanobis distance* between the two populations.

Let us point out that this procedure does not evaluate the intercept term $\beta_0$, which is calculated by using the condition, that the discriminant plane $z = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ should pass through the arithmetic mean of the expected values $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ of the considered two groups of data.

In the case, when only sample values $\bar{\mathbf{x}}_{\mathbf{1}.}$, $\bar{\mathbf{x}}_{\mathbf{2}.}$ and $\mathbf{S}$ instead of the population values $\boldsymbol{\mu_1}, \boldsymbol{\mu_2}$ and $\mathbf{\Sigma}$ are known, the ratio (1) and the formula (2) for the coefficients of the linear discriminant function can be expressed in terms of the sample means $\bar{\mathbf{x}}_{\mathbf{1}.}$, $\bar{\mathbf{x}}_{\mathbf{2}.}$ and the sample within covariance matrix $\mathbf{S}$. The vector $\mathbf{b}$ designating the empirical discriminant function takes then the form:

$$\mathbf{b}^T = (\bar{\mathbf{x}}_{1.} - \bar{\mathbf{x}}_{2.})\mathbf{S}^{-1}.$$

Let $\bar{\bar{\mathbf{x}}}$ denote the arithmetic mean of the sample means $\bar{\mathbf{x}}_{1.}$ and $\bar{\mathbf{x}}_{2.}$:

$$\bar{\bar{\mathbf{x}}} = (\bar{\mathbf{x}}_{1.} + \bar{\mathbf{x}}_{2.})/2.$$

The classification rule is then formulated as follows. For the given vector $\mathbf{x}$ compute:

$$z = (\mathbf{x} - \bar{\bar{\mathbf{x}}})\,\mathbf{b}. \tag{3}$$

If $z > 0$, assign $\mathbf{x}$ to group (population) 1, otherwise assign $\mathbf{x}$ to group (population) 2.

As pointed out in the original 1936 paper by Fisher and confirmed later a.o. by Kendall 1961 and Cramer 1967 (see [16] for exact references) the same results concerned the form of the discriminating vector $\mathbf{b}$ and the goodness of discrimination may be obtained in a regression layout, taking as $Y$ the group indicator variable, and $\mathcal{X} = (X_1, \ldots, X_p)$, the observed vector of features, as explanatory variables. For more details see, e.g. [11].

Let us underline here, that the above results, coming from Fisher, are obtained on a purely algebraical ground, without any special assumptions on the probability distributions of the variables; all we need is the assumption that the covariances in both groups (populations) are equal.

# 3 Probabilities a posteriori obtained from the Bayesian rule

## 3.1 The general Bayesian rule

Shortly after publication of Fisher's work another statistician, B.L. Welch, has shown in 1939 the link of Fisher's discriminant rule with the Bayesian approach (see [16] for exact references).

Suppose that we have to do with two groups, called also classes, denoted as $C_1$ and $C_2$ characterized by multivariate vectors of features (traits) denoted as $\mathbf{X} = (X_1, X_2, \ldots, X_p)$. Suppose for the moment that we know the conditional distribution functions $P(\mathbf{X} \mid C_k)$ and the *á priori* probabilities $P(C_k)$ for $k = 1, 2$ (Welch has assumed, that these are multivariate normal distributions with the same covariance matrix).

Now let us assume that we have observed the feature vector $\mathbf{X}$ for one individual $\mathcal{I}$ and we denote the gathered values by $\mathbf{x} = (x_1, \ldots, x_p)$. On the basis of them we want to assign the individual $\mathcal{I}$ to one of the two classes $C_1$ or $C_2$.

Then the probability *á posteriori* (alias: the *posterior* probability), that the observation $\mathbf{x}$ belongs to class $C_1$ is given by the Bayesian rule:

$$P(C_1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_1)P(C_1)}{p(\mathbf{x} \mid C_1)P(C_1) + p(\mathbf{x} \mid C_2)P(C_2)} \,. \tag{4}$$

Dividing the numerator and the denominator of the right side of the above formula by $p(\mathbf{x} \mid C_1)P(C_1)$ we may write alternatively (4) as:

$$P(C_1 \mid \mathbf{x}) = \frac{1}{1 + exp\{-a(\mathbf{x})\}} \,, \tag{5}$$

with $a(\mathbf{x})$ equal to

$$a(\mathbf{x}) = ln\, \frac{p(\mathbf{x} \mid C_1)P(C_1)}{p(\mathbf{x} \mid C_2)P(C_2)} \,. \tag{6}$$

Notice, that formula (5) was obtained under quite general assumptions without stating exactly the form of the involved probability density functions — thus that formula holds for **every** probability density function $P(\mathbf{x} \mid C_k), k = 1, 2$. The formula shows, that the posterior distribution $P(C_1 \mid \mathbf{x})$ depends from the observational vector $\mathbf{x}$ only through the score $a(\mathbf{x})$ as given above.

We may consider the probability $P = P(C_1 \mid \mathbf{x})$ as function of the evaluated score $a = a(\mathbf{x})$. Then we obtain the *logistic function* $f(a) = 1/(1 + e^{-a})$. The shape of this function is shown in Figure 1.

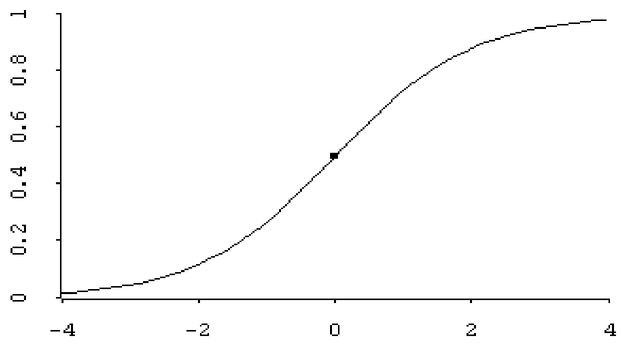



*Figure 1. The logistic function f(a) evaluated for the argument* $a \in [-4.0,\ 4.0]$

One can see that the function $f(a)$ has a sigmoidal shape. The function yields probabilities $P \in [0, 1]$ depending on the values of its argument $a$:

- if $a = 0$ then $f(a) = 0.5$;
- if $a > 0$ then $f(a)$ is greater than 0.5 and approaches $P = 1.0$ with $a$ tending to $+\infty$;
- if $a < 0$ then $f(a)$ is smaller than 0.5 and approaches $P = 0.0$ with $a$ tending to $-\infty$.

## 3.2 Definition of the 'ODDS' and the 'LOGITS'

Having defined the *á posteriori* probability $P_1 = P(C_1 \mid \mathbf{x})$ – we may easily obtain $P_2 = P(C_2 \mid \mathbf{x})$, the posterior probability of the complementary event (which is that $\mathbf{x}$ belongs to the complementary class $C_2$)

$$P(C_2 \mid \mathbf{x}) = 1 - P(C_1 \mid \mathbf{x}) = 1 - \frac{1}{1 + exp\{-a(\mathbf{x})\}} = \frac{exp\{-a(\mathbf{x})\}}{1 + exp\{-a(\mathbf{x})\}}\,. \tag{7}$$

The score $a(\mathbf{x})$ appearing in (5) and (7) has an interesting property: it is directly the (natural) logarithm of the *Odds* favoring the event $C_1$:

$$Odds(C_1 \mid \mathbf{x}) = \frac{P(C_1 \mid \mathbf{x})}{P(C_2 \mid \mathbf{x})} = \frac{1}{exp\{-a(\mathbf{x})\}} = exp\{a(\mathbf{x})\},$$

thus

$$ln(Odds(C_1|\mathbf{x})) = a(\mathbf{x})\,.$$

A positive value of $a(\mathbf{x})$ favors the class $C_1$, a negative – the class $C_2$.

Considering problems in bioassay, C.I. Berkson introduced in 1944 (for exact references see ([16])) the *logit* transformation as

$$logit(p) = ln(p/(1-p)).$$

Substituting for $p$ the posterior probability $P_1 = P(C_1|\mathbf{x})$ we obtain as the logit for $P_1$:

$$logit(P_1) = a(\mathbf{x}).$$

Looking at the above definitions of the logistic discriminant function defined by formula (5) and of the odds we see that they depend on the observed vector $\mathbf{x}$ only through the score $a = a(\mathbf{x})$.

An important question is: for which conditional distributions $p(\mathbf{x}|C_k)$ $(k = 1, 2)$ the score $a = a(\mathbf{x})$ reduces to a linear function of the observed vector $\mathbf{x}$, i.e. to the form $a(\mathbf{x}) = b_0 + \mathbf{x}\mathbf{b}$?

The problem will be considered in the next chapter.

# 4 Probabilities a posteriori for distribution functions being a specific member of the exponential family of distributions

Suppose that we have $K$ groups of data indexed as $k = 1, \ldots, K$. Usually the probability distribution function $p(\mathbf{x}|C_k)$ describing the data observed in the $k$-th group (class) $C_k$ depends on some parameters denoted generally by $\boldsymbol{\theta_k} = (\theta_1^{(k)}, \ldots, \theta_s^{(k)})$ characterizing, e.g., the location and shape of that distribution.

Apart from the class specific parameters $\boldsymbol{\theta_k}$ the distribution $p(\mathbf{x}|C_k)$ may depend from some other constants, denoted by us generally by $\boldsymbol{\phi}$, which are the same for all groups (classes) of the analyzed data. We might emphasize this directly by writing

$$p(\mathbf{x}|C_k) = p(\mathbf{x}|C_k; \boldsymbol{\theta_k}, \boldsymbol{\phi}).$$

In the following we will assume that $K = 2$, thus we will consider only two groups of data, and our interest will be focused on the first group indexed by $k = 1$. The other group, indexed as $k = 2$ will serve as the reference group.

An interesting and useful family of distributions considered frequently in statistics is the *exponential* family which can be written down as (see e.g. [12, 10])

$$p(\mathbf{x}; \boldsymbol{\vartheta}) = h(\mathbf{x}) exp[\sum_{i=1}^{s} \boldsymbol{\eta}_i(\boldsymbol{\vartheta})\mathbf{T}_i(\mathbf{x}) - U(\boldsymbol{\vartheta})]. \tag{8}$$

Here $\boldsymbol{\eta}(.)$ and $\mathbf{T}(.)$ denote real functions, which may be vectorized; the vector $\boldsymbol{\vartheta}$ denotes generally the parameters characterizing the probability distribution $p(\mathbf{x}; \boldsymbol{\vartheta})$.

In particular the considered above probability distribution function $(p\mathbf{x}|C_k)$ describing the probability distribution for the $k$th group of data could also for some data belong to the exponential family of distributions, what would then be indicated by writing explicitly

$$p(\mathbf{x}|C_k) = p(\mathbf{x}|C_k, \boldsymbol{\vartheta}).$$

Let us substitute

$$\begin{aligned}
\boldsymbol{\vartheta} &= (\boldsymbol{\theta}_k, \boldsymbol{\phi}), \\
\boldsymbol{\eta}_1(\boldsymbol{\vartheta}) &= \boldsymbol{\theta_k}, \quad \mathbf{T}_1(\mathbf{x}) = \mathbf{x}, \\
\boldsymbol{\eta}_i(\boldsymbol{\vartheta}) &= \boldsymbol{\eta}_i(\boldsymbol{\phi}), \ (i = 2, \ldots, s), \\
\sum_{i=2}^{s} \boldsymbol{\eta}_i(\boldsymbol{\phi})\mathbf{T}_i(\mathbf{x}) &= B(\mathbf{x}, \boldsymbol{\phi}), \\
-U(\boldsymbol{\theta}_k, \boldsymbol{\phi}) &= A(\boldsymbol{\theta}_k, \boldsymbol{\phi}).
\end{aligned}$$

Then the formula for the exponential family of distributions reduces to the equation:

$$p(\mathbf{x} \mid C_k; \boldsymbol{\theta_k}, \boldsymbol{\phi}) = h(\mathbf{x}) exp\{A(\boldsymbol{\theta}_k, \boldsymbol{\phi}) + B(\mathbf{x}, \boldsymbol{\phi}) + \mathbf{x}\boldsymbol{\theta}_k\} \tag{9}$$

The family of distributions given by the equation above will be in the following called "specially reduced exponential family" of distributions.

It can be proved that the following distributions belong to the specially reduced exponential family (9):

(a) Multivariate normal $N_p(\boldsymbol{\mu_k}, \boldsymbol{\Sigma})$ with the same covariance matrix ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$) in both groups of data,
(b) Poisson distribution $P(\lambda_k)$,
(c) Binomial distribution $b(n, p_k)$,
(d) Multinomial distribution $w(n, s, \mathbf{p}_{(k)})$,
(e) Negative binomial distribution $NegBin(r, p_k)$,
(f) Gamma distribution $f_\Gamma(x; \alpha, \beta_k)$, with $\alpha$ denoting the scale parameter, and $\beta_k$ the shape parameter,
(g) Any mixture of (a)–(f), however assuming *independent* components.

Bishop [3] notified that if the probability distribution function under consideration (i.e. the $p(\mathbf{x} \mid C_k; \boldsymbol{\theta_k}, \boldsymbol{\phi})$ $(k = 1, 2)$ follows (9), then we may further reduce the expression $a(\mathbf{x})$ given by (6). Substituting (9) into (6) we obtain:

$$\begin{aligned} a(\mathbf{x}) &= ln \left\{ \frac{h(\mathbf{x}) exp\{A(\boldsymbol{\theta}_1, \boldsymbol{\phi}) + B(\mathbf{x}, \boldsymbol{\phi}) + \mathbf{x}\boldsymbol{\theta}_1\} P(C_1)}{h(\mathbf{x}) exp\{A(\boldsymbol{\theta}_2, \boldsymbol{\phi}) + B(\mathbf{x}, \boldsymbol{\phi}) + \mathbf{x}\boldsymbol{\theta}_2\} P(C_2)} \right\} = \\ &= ln \left\{ exp\{A(\boldsymbol{\theta}_1, \boldsymbol{\phi}) - A(\boldsymbol{\theta}_2, \boldsymbol{\phi}) + \mathbf{x}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\} \frac{P(C_1)}{P(C_2)} \right\} \\ &= \mathbf{xb} + b_0 , \end{aligned}$$

where

$$\mathbf{b} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \quad b_0 = A(\boldsymbol{\theta}_1, \boldsymbol{\phi}) - A(\boldsymbol{\theta}_2, \boldsymbol{\phi}) + ln \, \frac{P(C_1)}{P(C_2)} \, .$$

Thus we see that (5) reduces – in case of p.d.f. of the exponential type (9) – to

$$P(C_1 \mid \mathbf{x}) = \frac{1}{1 + exp\{-(b_0 + \mathbf{xb})\}}, \tag{10}$$

which may be rewritten as

$$P(C_1 \mid \mathbf{x}) = \frac{1}{1 + exp\{-l(\mathbf{x})\}}, \quad \text{with } l(\mathbf{x}) = b_0 + \mathbf{xb}. \tag{11}$$

The formula above presents the famous **logistic discriminant function**. It can be seen that the probability *á posteriori* of belonging to class $C_1$ depends only on $(p + 1)$ parameters $b_0, b_1, \ldots, b_p$ and does not need explicit a detailed knowledge of the distributions $p(\mathbf{x} \mid C_k; \boldsymbol{\theta_k}, \boldsymbol{\phi})$ and their parameters $\boldsymbol{\theta_k}$ $(k = 1, 2)$ and $\boldsymbol{\phi}$. The formula (10) permits to evaluate the posterior probability $P(C_1 \mid \mathbf{x})$ as a simple (although nonlinear) function of the linear score

$$l = l(\mathbf{x}) = b_0 + b_1 x_1 + \ldots + b_p x_p \, .$$

In other words: in the representation by the logistic discriminant function the posterior probability $P(C_1 \mid \mathbf{x})$ is expressed as a (nonlinear) function of one summary variable $l$ evaluated as a linear form of the observed values $x_1, \ldots, x_p$.

However, let us say, this does not mean, that the coefficients of the linear score $l(\mathbf{x})$ are exactly equal to the coefficients of the Fisherian linear discriminant function. This is only true, when the underlying probability density functions are multivariate normal with the same covariance matrix.

# 5 Practice of using the logistic discriminant function

## 5.1 Computational issues

The computation of the *Fisherian linear discriminant function* by use of the explicit formula (2) is straightforward and very quick. Also, the search for the variables with the greatest discriminative power – stepwise search or all subset search – can be carried out quite quickly.

Computation of the parameters of the *logistic discriminant function* can be done only in an iterative way with all drawbacks of this method (much longer computing time, possible problems with convergence, local maxima, colinearity of computed variables). Usually this is done by the maximum likelihood approach employing a kind of Newton-Raphson algorithm. The widely used algorithm is that proposed by Jennrich and Moore [9]. There is also known and still in use a classical sequential algorithm described by Walker and Duncan [15] permitting to adjust the solutions sequentially when adding new observational vectors to the data base.

The goals and the immediate results furnished by the linear Fisherian and the logistic discriminant function are different.

The goal of the Fisherian discriminant analysis is to provide a tool called discriminant function – which permits for each new observation vector $\mathbf{x}$ to classify it either to group 1 or group 2.

The logistic function is oriented rather to computing probability of belonging to group 1 – provided that the observation vector $\mathbf{x}$ is given. In some problems we are rather concerned with evaluation of the probabilities, and not with classification (see, e.g. [17]).

Of course, in the case of computing the Fisherian discriminant function after calculation of the coefficients $b_0, b_1, \ldots, b_p$ we may evaluate easily $p(C_2 \mid \mathbf{x})$ – as indicated by formula (7).

## 5.2 What can be expected for some distributions

Considering the specific exponential family of distributions described by formula (9) in Section 4 we have seen that one special case is the multivariate distribution with the same covariance matrix in both groups of data. Thus in the case, when we have to do with continuous variables with nearly normal shape but with unequal covariance matrices in the two considered groups of data – then the linear and also the logistic function will be inappropriate for discriminant analysis unless the considered groups are really far apart each from the other. It can be expected that in such cases the *quadratic discriminant function* will be much more appropriate: it will give lower error rates (for the definition of the quadratic discriminant function see, e.g., Lachenbruch [11]). Thus there is no reason to expect that the logistic discriminant function will be superior to the Fisherian discriminant function in such cases.

On the other hand, the exponential family of distributions contains some long-tailed distributions (like the *Gamma* and the *Negative Binomial* distributions) which may give different covariances in the considered two groups of data: in such cases the logistic discriminant function may be more appropriate then the linear one.

## 5.3 Reports in the statistical literature

In the statistical literature one may find reports from special investigations on assessing the fit and performance of the logistic model in various circumstances and trying to elucidate that topic. Especially the fit for the binary and mixed explanatory variables, the robustness and some diagnostics have been considered. The apparent error rates and asymptotic efficiency of the logistic discriminant rule as compared to the ordinary linear discriminant rule were quite often investigated; a.o. Lachenbruch [11], Fatti, Hawkins, and Raath [8], Dietz [5] and McLachlan [13] report such studies. The conclusions are that the linear and logistic discriminant functions show nearly the same performance in discriminating between two groups of data, at least when the analysed data do not differ much in shape from the multivariate normal distribution with equal covariance matrices in both groups of data.

What concerns the quadratic discriminant function, several studies (reported e.g. by Dietz [5]) have found that even in the case of continuous variables and obviously different covariance matrices it happened, that the linear or logistic discriminant functions yielded smaller error rates than the quadratic one. This happened especially when the sample sizes were small, and the number of variates (also the number of the estimated covariances between the variables) quite considerable. Dietz [5] shows also results from 3 real data sets in several (4–11) variables, for which the linear and logistic discriminant functions yielded practically the same result, while the quadratic discriminant function was in one set slightly better and in the two remaining data sets slightly worse.

Bartkowiak [2] reported an epidemiological study in Coronary Heart Disease risk factors (5 variables) in which the so called *risk deciles* were calculated by use of the linear, logistic and quadratic discriminant function. Independently, the bounds of the deciles were established by a bootstrap method. The risk deciles obtained by the three discriminant functions considered in the study were located within the confidence bounds established by the bootstrap method.

## 5.4 Our simulation studies for data with different covariance matrices

All the formerly reported studies have considered in principle some particular multivariate data and the obtained results were also particular: it seems that the results with indications on the preference of the individual discriminant functions were much dependent on the data on which the analysis was performed.

To find out exactly how the individual discriminant functions act we have planned a series of very simple experiments for which we could see exactly, what is going on.

We have performed a simulation study generating bivariate data from three distributions: (A) *Banana* shape, (B) *Higleyman's* shape, (C) *Lithuania* (sausage) shape.

The simulation study was carried out using the package *PRT2.1* (Pattern Recognition Tools version 2.1) developed by Robert Duin [7, 6] at the Delft University of Technol-

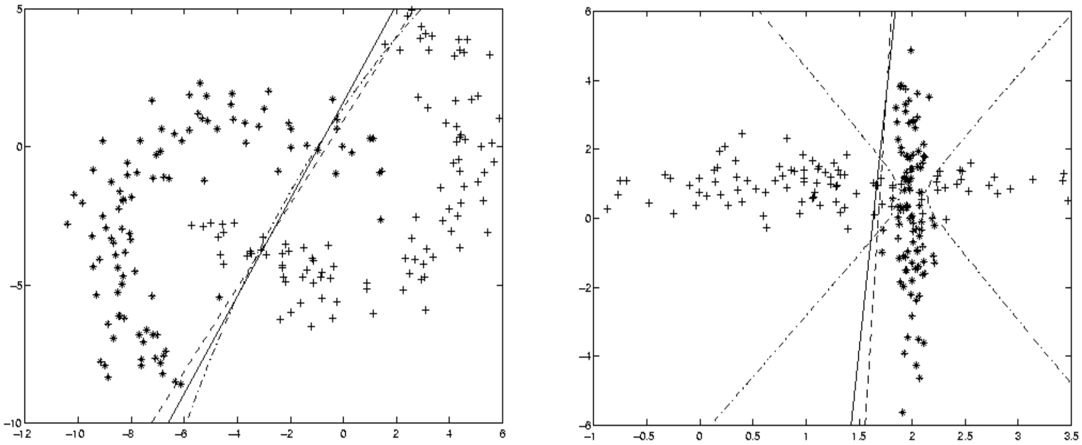


*Figure 2. Scatterplots showing two groups of of two-dimensional data with a linear, logistic and quadratic discrimination boundary overlaid.*
*Boundaries: linedraw for linear, - - - - logistic, -·-·- quadratic.*
Left: *Banana shape with ss=0.8.*
*Error rates: Lin.: 0.1250, Logist.: 0.1313, Quadr.: 0.1313.*
Right: *Higleyman shape.*
*Error rates: Linear: 0.1625, Logistic: 0.1500, Quadratic: 0.0312.*

ogy. The package needs the *MATLAB* and the *MATLAB Neural Network Toolbox* [4] environment.

For each distribution (A), (B) and (C) mentioned above we generated 2 groups of data, which were differentiated in the location and shape parameters. Then we have evaluated discriminant functions (discriminant boundaries between the 2 groups) using the linear, quadratic and logistic discriminant functions. Tho experiment was repeated several times varying the parameters (shapes) of the generated distributions.

Since the generated data vectors were two-dimensional, we could make a scatterplot for each investigated data set, with the discriminant boundaries overlaid.

In that situation, along with calculating the error of misclassification (defined as the ratio of the misclassified points to the total number of points in both groups of data using the resubstitution method) we could also judge 'by eye', which discriminant function was the most appropriate.

Exemplary scatterplots are shown in Fig. 2 (banana and Higleyman shapes) and Fig. 3 (Lithuania sausage).

Let us point out that while the banana shape of distribution occurs perhaps rarely in medical studies – the other two shapes may occur quite frequently.

Higleyman shape conveys the fact that investigating features in two diseases we may find, that in one disease one feature is contracted (has smaller variance) while the other feature gets deregularized and we notice much increased variance. In the other disease the opposite may happen: the variance increases for the first feature and decreases for the second feature. E.g. such situation happens when investigating variables characterizing some respiratory diseases [1].

The Lithuania or sausage shape is characterized by a parameter which we have denoted as $ss$. This parameter characterizes the thickness and curvature of the sausage.

The error rates for the data exhibited in the scatterplots (and some others for which the scatterplots are not shown here) are presented in Table 1:

Generally we have stated in that part of our simulation study, that

*Table 1. Error rates (resubstitution method) when applying the linear, the logistic and the quadratic discriminant functions to various data sets*

| Experiment and its parameters | Discriminant function Linear | Logistic | Quadratic |
|---|---|---|---|
| Banana shape with $ss = 0.8$ | 0.1250 | 0.1313 | 0.1313 |
| Higleyman shape | 0.1500 | 0.1250 | 0.0375 |
| Higleyman shape | 0.1625 | 0.1500 | 0.0312 |
| Higleyman shape | 0.1688 | 0.1688 | 0.0437 |
| Lithuania shape with $ss = 0.5$ | 0.1625 | 0.1313 | 0.1000 |
| Lithuania shape with $ss = 0.5$ | 0.1187 | 0.1125 | 0.0688 |
| Lithuania shape with $ss = 0.5$ | 0.1187 | 0.1125 | 0.1062 |
| Lithuania shape with $ss = 0.5$ | 0.1625 | 0.1938 | 0.1187 |
| Lithuania shape with $ss = 0.5$ | 0.1100 | 0.1350 | 0.0600 |
| Lithuania shape with $ss = 0.5$ | 0.1100 | 0.1100 | 0.0450 |
| Lithuania shape with $ss = 1$ | 0.0938 | 0.0938 | 0.0312 |
| Lithuania shape with $ss = 1$ | 0.1750 | 0.1750 | 0.1250 |
| Lithuania shape with $ss = 1.2$ | 0.3000 | 0.3063 | 0.2750 |
| Lithuania shape with $ss = 1.2$ | 0.2188 | 0.2313 | 0.1688 |
| Lithuania shape with $ss = 1.2$ | 0.1550 | 0.1550 | 0.1000 |
| Lithuania shape with $ss = 2$ | 0.2062 | 0.2062 | 0.1437 |
| Lithuania shape with $ss = 2$ | 0.2750 | 0.2812 | 0.2125 |
| Lithuania shape with $ss = 2$ | 0.2875 | 0.2812 | 0.2750 |
| Lithuania shape with $ss = 3$ | 0.3312 | 0.3312 | 0.3375 |
| Lithuania shape with $ss = 3$ | 0.3438 | 0.3563 | 0.3500 |

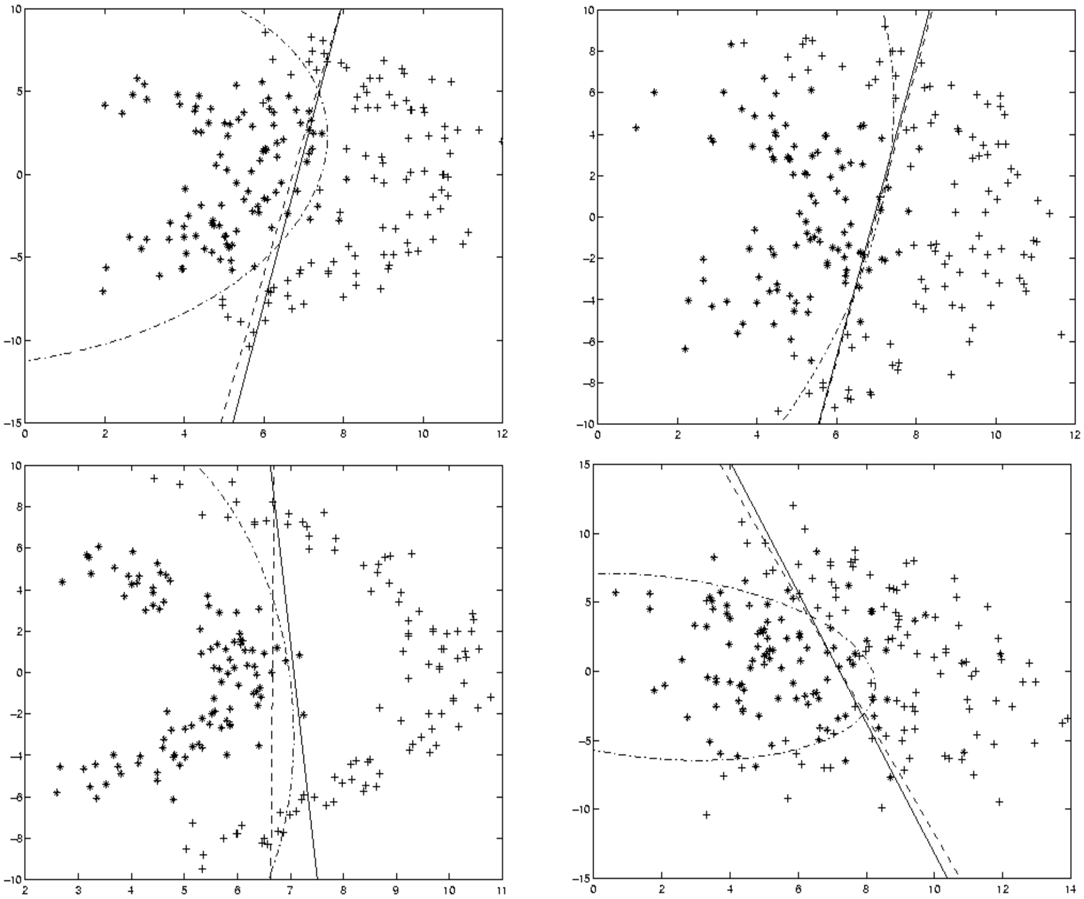


*Figure 3. Scatterplots showing 4 groups of two-dimensional data generated as 'Lithuania sausage' with varying parameters ss. Linear (linedraw), logistic ( - - - - ) and quadratic ( -·-·- quadratic)* discrimination boundaries are overlaid.
Top Left: *ss=1.0. Error rates: Linear: 0.0938, Logistic 0.0938, Quadratic: 0.0312..*
Top Right: *ss=1.0. Error rates: Linear: 0.1750, Logistic 0.1750, Quadratic: 0.1250.*
Bottom Left: *ss=0.5. Error rates: Linear: 0.1625, Logistic 0.1313, Quadratic: 0.1000.*
Bottom Right: *ss=2.0. Error rates: Linear: 0.2062, Logistic 0.2062, Quadratic: 0.1437.*

- In case of banana shape distributions all the three methods yielded approximately the same error rates.
- In case of Higleyman's and Lithuania sausage distributions the linear and logistic discriminant functions acted very similarly, while the quadratic discriminant function proved to be often much better, yielding in almost all cases considerably lower error rates as those stated for the linear or logistic discriminant function.

## 5.5 Our simulation studies for data sampled from the Gamma and the Negative Binomial distribution

It was shown in section 4, that for distributions being a member of the specific exponential family described by equation (9), the discriminant function depends on a score $l(\mathbf{x}) = b_0 + b_1x_1 + \ldots + b_px_p$ which is a linear function of the observed variables. For the multivariate normal distribution with the same covariance matrix the coefficients $b_1, \ldots, b_p$ are exactly the same as those appearing in the Fisherian linear discriminant function. For non–normal distributions this is difficult to prove. To elucidate the problem, we performed a series of simulation studies generating two–dimensional distributions $(X_1, X_2)$ from the negative binomial $NegBin(r, p_k)$ and the $Gamma(a, b_k)$ distributions. These distributions are defined as follows:

`The negative binomial distribution` $NegBin(r, p_k)$:

$$P_k(X = x; r, p_k) = \binom{r + x - 1}{x} (1 - p_k)^r p_k^x,$$

where $r \in \{1, 2, \ldots\}$, $0 < p_k < 1$, $x = 0, 1, 2, \ldots$. We have for that variable: $EX = rp_k/(1 - p_k)$, $VarX = rp_k/(1 - p_k)^2$.

`The Gamma distribution` $f_\Gamma(x; a, b_k)$:

$$f_\Gamma(x; a, b_k) = \frac{x^{a-1} e^{-\frac{x}{b_k}}}{\Gamma(a) b_k^a},$$

where $x > 0$, $a > 0$ denotes the shape parameter, and $b_k > 0$ the scale parameter, $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$, $EX = ab_k$, $VarX = ab_k^2$.

Usually we have been fixing the sample size as $n_1 = 50$ for the first group and $n_2 = 70$ for the second group of the generated data. The variables $X_1$ and $X_2$ were supposed to be independent, with varying parameters in the generated two groups of data. To satisfy (9) the appropriate parameters $r$ and $a$ were considered as fixed, and the parameters $p_k$ and $b_k$ as group-specific.

The generated 2-groups data were exhibited in a scatterplot, where also the corresponding Fisherian linear and the logistic linear discriminant functions were added.

Two exemplary scatterplots are shown in Fig. 4. Points from the two generated groups of data are marked in the plots by the characters + and o.

Looking at the plots exhibited in Fig. 4. one may see that the two exhibited groups of data have different variances. The linear discriminant function is certainly inappropriate for the exhibited data; none the less it could be adjusted by a shift towards to the origin of the coordinate system (this would be achieved by changing the intercept $b_0$).

Both the gamma and the negative binomial distributions are long-tailed. When generating observations from these distributions we have obtained in some experiments also points from the tails of the respective distributions, which in fact are some extreme points. These points look like outliers – which they are not. One such situation is shown in the upper plot of Fig. 4. One can see there one point marked by 'o', thus generated from the second distribution, which is located among the points from the first distribution.

Such situations were observed also in some other graphs, not shown here. It would be interesting to investigate, how much these extreme points are influential for the coefficients of the linear and the logistic discriminant function.

Generally we have performed several series of simulations, varying the parameters of the generated pseudo–random observations. Details for four series of experiments are given in Table 2. The symbols '*Av. ERR* lin' and '*Av. ERR* logi' in that table denote the average error rate observed for the Fisherian linear and the logistic discriminant functions; 'logi<lin' denotes the percentage of the experiments (in the series of simulations with the same parameters) for which the logistic discriminant function produced lower error rates; *nb runs* denotes length of the series, i.e. the number of repetitions using data generated with the same parameters. The error rates were calculated by the resubstitution method.

Generally we have observed the following facts:

*Table 2. Average error rates (Av. ERR) obtained when applying the linear and the logistic discriminant functions to four series of experiments of simulated data*

| Experiment & parameters | *Av. ERR* lin | *Av. ERR* logi | logi<lin | *nb runs* |
|---|---|---|---|---|
| NegBin$(r = 10, p_1 = 0.8, p_2 = 0.3)$ | 0.0242 | 0.0091 | 54.5% | 22 |
| NegBin$(r = 5, p_1 = 0.1, p_2 = 0.9)$ | 0.0256 | 0.0241 | 89.5% | 19 |
| Gamma $(a = 4.0, b_1 = 0.5, b_2 = 5.0)$ | 0.0504 | 0.0312 | 90.0% | 20 |
| Gamma $(a = 10, b_1 = 1.0, b_2 = 5.0)$ | 0.0452 | 0.0373 | 61.9% | 21 |

- There was *in the average* a slight reduction in the error rates when applying the logistic discriminant function.
- It happened quite often that among the generated data (which were generated from long-tailed distributions) some 'outliers' could be observed, i.e. single observations which being generated according to the formula describing the probability distribution in group 1 appeared among the observations belonging to group 2; or vice-versa.

  For the moment we can not say decidedly whether the logistic discriminant function proved to be more robust against such outliers.
- It was observed quite frequently, that after adjusting the intercept of the Fisherian linear discriminant function – the results of the discrimination by the logistic and Fisherian linear discriminant functions would be very similar.

# 6 Discussion and Summary of the Results

It has been stated that the logistic discriminant function with a linear score $a(\mathbf{x})$ in the observed variables may give a good classification rule when the distributions of the observed variables follow a distribution which is a special class of the exponential class of distributions and was specified by (9) . Although this formula allows to account to this class, a.o., the univariate gamma, the binomial, the negative binomial and the Poisson distributions, these should be introduced into the model as mutually independent variables, what is a serious restriction because in real data it happens usually that the considered variables are mutually dependent. It may also happen that in one of the groups the categorial variables are somehow interdependent and exhibit interactions of higher terms, while in the other group they are not. In such case the logistic discriminant function will be inappropriate.

Our main conclusions are the following:

1. The logistic discriminant function may give in some special cases, especially when the data follow the special exponential family distribution (9) – slightly better results than the Fisherian linear discriminant function.

2. It is always worth to investigate the covariance matrices in the considered groups of data. If they are apparently different and the number of variables dealt with is not very high as compared to the size of the data, then it is worth to try to apply the quadratic discriminant function.

**Acknowledgements**

The research was partially sponsored from the Polish State Research Committee KBN grant no 8T11 C031 16.

The authors express thanks to prof. R. Duin from the Technical University Delft, the Netherlands, for lending the elaborated by him *PRT* software package for Pattern Recognition.

# References

[1] Bartkowiak A., and Liebhart J. (1995). Estimation of the Spirometric Residual Volume by a regression built from Gower Distances. *Biometrical Journal* 37, 2, 131-149.

[2] Bartkowiak A. (1988). Logistic and Fisherian discrimination as applied to a coronary heart disease study. *Int. Workshop on Theory and Practice in Data Analysis, Proceedings, August 19-21, 1988*, Berlin GDR. Edition: Report R-MATH-01/89, Akademie der Wissenschaften der DDR, Karl-Weierstrass-Institut für Mathematik, Berlin, p.184-193.

[3] Bishop Ch. M. (1995). *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford (U.K.).

[4] Demuth H., and Beale M. (1997). *MATLAB, Neural Network Toolbox, User's Guide, version 2, Computation, Visualization, Programming.* Copyright 1992–1997, MathWorks, Natick, MA.

[5] Dietz E. (1987). Application of logistic regression and logistic discrimination in medical decision making. *Biometrical Journal* **29**, 6, 747–751.

[6] Duin R.P.W., and Kröse B.J. (1997). *Advanced Pattern Recognition.* Ph.D. Course, 23–31. October 1997, TU Delft.

[7] Duin R.P.W. (1997). *Introduction to Statistical Pattern Recognition.* Pattern Recognition Group. Delft University of Technology. The Netherlands.

[8] Fatti P., Hawkins D.M., and Raath E.L. (1982). Discriminant Analysis. In: D.M. Hawkins (Ed), *Topics in Applied Multivariate Analysis*, Cambridge University Press, Cambridge, London, 1–71.

[9] Jennrich R.I., and Moore R.H. (1975). Maximum likelihood estimation by means of nonlinear least squares. *Am. Statistical Association, Proceedings of the Statistical Computing Section*, 57–65.

[10] Krzyśko M. (1996). *Mathematical Statistics.* In Polish. Wydawnictwo Naukowe UAM.

[11] Lachenbruch P.A. (1975). *Discriminant analysis.* Hafner Press, London.

[12] Lehman E. L. (1983). *The Theory of Point Estimation.* Wiley N.Y. 1983. Polish translation: *Teoria estymacji punktowej,* PWN Warszawa 1991.

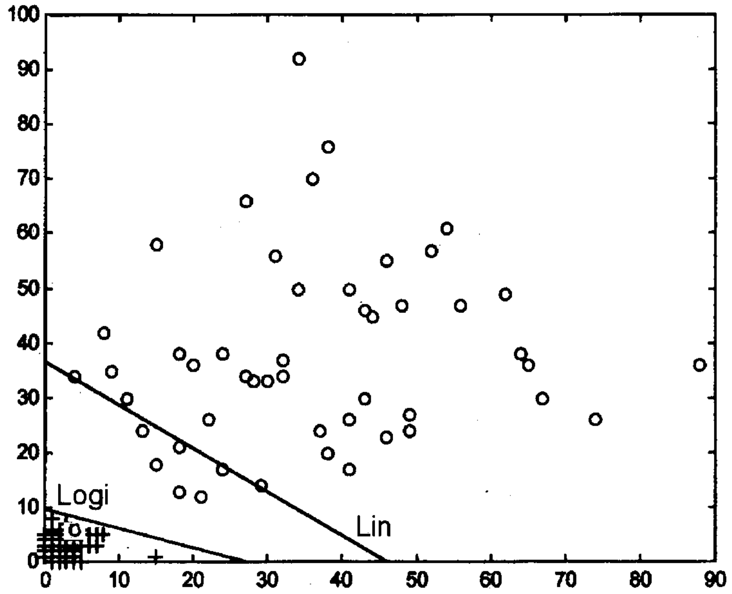
[13] McLachlan G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* Wiley New York.

[14] Steyerberg E.W., Eijkemans M.J.C., van Houwelingen J. C., Lee K.L. and Habbema J.D.F. (2000). Prognostic models based on literature and individual patient data in logistic regression analysis. *Statistics in Medicine* **19**, 2, 141–160.

[15] Walker S.B., and Duncan D.B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika* **54**, 167–179.

[16] van Houwelingen J.C., and le Cessie S. (1988). Logistic regression, a review. *Statistica Neerlandica* **52**, 4, 215–232.

[17] Wooff D.A., Scheult A.H. and Coolen F.P.A. (1999). Bayesian discrimination with uncertain covariates for pesticide contamination. In: V. Barnett, A. Stein and K.F. Turkman (Eds), *Statistical Aspects of Health and the Environment.* Wiley, Chichester, 337–353.



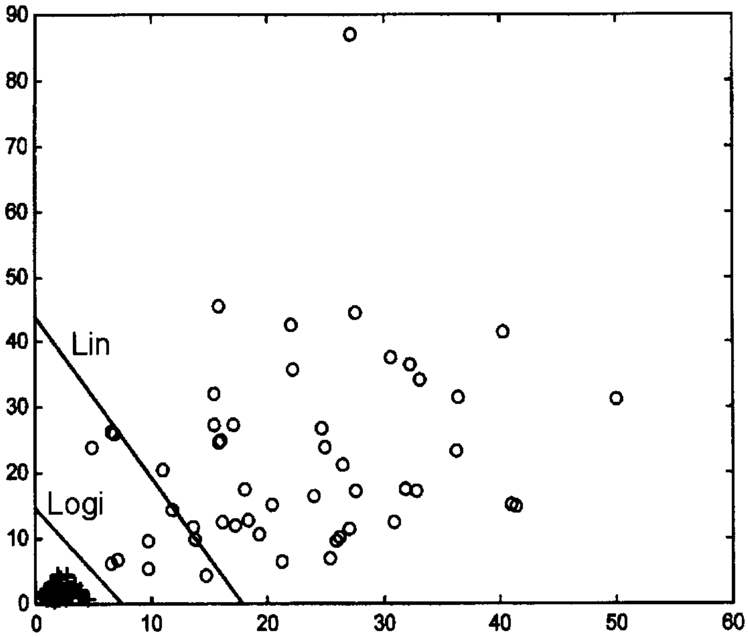




*Figure 4. Boundaries between two groups of data established by the linear and the logistic discriminant functions*

Top: *Generated from the negative binomial distribution with parameters* $p_1 = 0.4$ *and* $p_2 = 0.9$ *keeping* $r = 4$ *as constant. Error rate: 0.05 for the linear and 0.0083 for the logistic discriminant function.*

Bottom: *Generated from the gamma distribution with parameters* $b_1 = 0.4$ *and* $b_2 = 5$ *keeping* $a = 10$ *as constant. Error rate: 0.0833 for the linear and 0.0000 for the logistic discriminant function.*