The tangent classifier

José R. Berrendero and Javier Cárcamo *

Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid

June 18, 2012

Abstract

Given a classifier, we describe a general method to construct a simple linear classification rule. This rule, called the *tangent classifier*, is obtained by computing the tangent hyperplane to the separation boundary of the groups (generated by the initial classifier) at a certain point. When applied to a quadratic region, the tangent classifier has a neat closed-form expression. We discuss various examples and the application of this new linear classifier in two situations under which standard rules may fail: When there is a fraction of outliers in the training sample, and, when the dimension of the data is large in comparison with the sample size.

Keywords: Discrimination; Supervised classification; Fisher linear discriminant analysis; Quadratic discriminant analysis; Regularized discriminant analysis; Robust discriminant analysis; Multivariate normal distribution; Exponential distribution.

*This research was supported by the Spanish MCyT grants MTM2007-66632, MTM2010-17366 and MTM2011-27248, and Comunidad de Madrid-UAM grant CCG10-UAM/ESP-5494.

E-mail addresses: joser.berrendero@uam.es and javier.carcamo@uam.es

1 Introduction

We consider the classic problem of discriminating between two groups. We observe the pair $(\mathbf{X}, G)$, where $\mathbf{X}$ is a predictor vector taking values in $\mathbb{R}^d$ and $G \in \{0, 1\}$ is a categorical response variable representing the class memberships. We want to predict G based on the d variables in $\mathbf{X}$. Therefore, the goal is to obtain a classifier of the form $\eta(\mathbf{x}) = \mathrm{I}_A(\mathbf{x})$, where $A \subset \mathbb{R}^d$ and I_A stands for the indicator function of the set A. The notation means that, if $\mathbf{x}$ is an observation with unknown membership, we assign $G = \eta(\mathbf{x})$. Throughout this work, we denote by $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ the expectation and the covariance matrix of $\mathbf{X}$ in the group i $(i = 0, 1)$. The covariance matrices are assumed to exist and be positive definite. We also consider column vectors and $\mathbf{v}^\top$ stands for the transpose of the vector $\mathbf{v}$.

Many classification rules can be expressed as

$$\eta_g(\mathbf{x}) = \mathrm{I}_{\{\mathbf{y} \,:\, g(\mathbf{y}) > 0\}}(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^d, \tag{1}$$

where $g : \mathbb{R}^d \to \mathbb{R}$ is a *classification function*. That is, the separation boundary of the two groups determined by the classifier η_g is the level set $\{\mathbf{x} \in \mathbb{R}^d \,:\, g(\mathbf{x}) = 0\}$. For instance, one of the first and most applied methods to obtain a classifier is the celebrated *Fisher's linear discriminant analysis* (see Fisher (1936) [5]). Under homoscedasticity ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$) this approach leads to the classifier η_f, where f is the linear function

$$f(\mathbf{x}) = \mathbf{w}_F^\top \left(\mathbf{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} \right), \tag{2}$$

and the *Fisher's linear discriminant vector* is

$$\mathbf{w}_F = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0). \tag{3}$$

When the random vector $\mathbf{X}|\{G = i\}$ has probability density function f_i, ano-

ther important example is the *Bayes classifier*, η_b with

$$b(\mathbf{x}) = c(0|1)\pi_1 f_1(\mathbf{x}) - c(1|0)\pi_0 f_0(\mathbf{x}), \tag{4}$$

where $\pi_i = \mathrm{P}(G = i)$ is the prior probability of the group i and $c(1-i|i)$ is the cost of misclassifying an observation with membership $G = i$ as $G = 1 - i$ $(i = 0, 1)$.

It is well-known that the Bayes classifier is theoretically optimal (see for example Devroye *et al.* (1996) [4]). However, the (possibly complicated) function b in (4) depends on the conditional densities f_0 and f_1 which in practice are not specified and have to be estimated. Further, the generated rule could be difficult to interpret. In contrast, linear classifiers are seldom optimal, but they have other advantages. They have a clear interpretation, which is an important feature of a classifier. Indeed, in some areas like credit scoring interpretability is a legal requirement. Linear classifiers also allow us to identify the most relevant variables in the discrimination procedure, so they can be used as tools to obtain discriminative information and reduce dimensionality. Moreover, the performance of a linear classifier can be reasonably good. In fact, as pointed out by Hand (2006) [8], Fisher's method is often very competitive against more sophisticated classifiers.

In many important situations the classification function is quadratic. The implementation of both linear and quadratic classifiers is direct and fairly straightforward. However, it is important to note that there is a tremendous difference in the interpretation of the generated classifiers. In linear rules, we obtain a classification function $g(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i\, x_i$ $(\mathbf{x} = (x_1, \ldots, x_d)^\top)$ whose weights w_i help the user to understand the role played by each marginal variable in the procedure. If $g(\mathbf{x}) = w_0 + \sum_{1 \le i \le j \le d} w_{i,j}\, x_i x_j$, it is not common to understand the meaning of the weights in terms of the original variables, specially when d is large. Although we can always embed our measurement space in a higher dimensional space (the

space of the product variables) to transform the quadratic rule into a linear one, the interpretation of the new variables is not always clear.

In practice, the Fisher's classifier is usually applied even when the homoscedastic assumption is clearly not satisfied. In the two groups setting, this essentially means that the underlying Bayes classifier is being approximated by the Fisher's linear rule. However, the hyperplane obtained by the Fisher's approach can be far away from the separation boundary defined by the theoretically optimal rule. To overcome this problem in the multivariate normal setting, Anderson and Bahadur (1962) [1] developed a method to obtain a linear classifier under heteroscedasticity. This classifier is found computationally to minimize the probability of misclassification and strongly depends on the normality assumption.

There are many other classifiers beyond those commented before. New proposals arise not only in the statistical literature but also in other research areas such as machine learning, data mining, bioinformatics and other applied fields. A non-exhaustive enumeration of methods would include neural networks, nearest neighbors and kernel non parametric methods, tree classifiers and support vector machines. Ensemble learning methods generate many classifiers and aggregate their results. For instance, random forests and other procedures based on bagging and boosting. The interested reader can consult Bishop (2006) [2], Devroye *et al.* (1996) [4], and Hastie *et al.* (2001) [9] as useful and general references on supervised classification.

Given the classifier η_g defined in (1) by a smooth function g, the aim of this work is to obtain a simple and easy to implement a linear classification rule close to η_g. This new rule shares the main advantages of the linear classifiers, as simplicity and interpretability, and its computation is straightforward. In some situations it

can be seen as a correction of the Fisher's rule when the covariance structures of the predictor vector in the two groups are different. An interesting feature of this new linear classifier is that, in contrast with the Fisher's rule, it can be computed even when d is higher than the sample size by using a regularized approach (see Subsection 5.2).

The key idea to obtain such a classifier is very simple. As a natural linear approximation, we consider a tangent hyperplane to the separation boundary of the groups defined by η_g. We call this classifier the *tangent classifier to* η_g. This procedure is explained in Section 2. In Section 3, we discuss in depth the case in which the classification function is quadratic. We derive a simple and easy to handle expression for the tangent classifier. Two illustrative examples are considered in Section 4. It is shown in some situations the tangent classifier can be nearly optimal and extremely different from the Fisher's method. When the estimated underlying rule η_g has good properties regarding the misclassification errors, we expect that the associated tangent classifier inherits this good behavior. This is analyzed in Section 5 by means of some simulations. Two situations under which standard estimators may fail are also discussed: When there is a fraction of outliers in the training sample, and, when we have a high-dimensional data set in which the number of variables is large in comparison with the sample size. In Section 6 a real data example is analyzed. Finally, some conclusions and final remarks close the paper.

2 The tangent classifier

Given $g : \mathbb{R}^d \to \mathbb{R}$ a smooth enough classification function, we want to construct a linear classifier close to η_g. We consider the tangent hyperplane to the level set

$g(\mathbf{x}) = 0$ at a certain point $\boldsymbol{\mu}$ as a linear approximation of the separation boundary of the region defined by η_g. The derived classifier is η_{t_g}, where

$$t_g(\mathbf{x}) = \mathbf{w}_g^\top(\mathbf{x} - \boldsymbol{\mu}), \quad \text{with} \quad \mathbf{w}_g = a\nabla g(\boldsymbol{\mu}) \quad (a > 0 \text{ constant}). \tag{5}$$

Hence, $\mathbf{w}_g$ is a normal vector to the (hyper)surface $g(\mathbf{x}) = 0$ at the point $\boldsymbol{\mu}$. We call η_{t_g} the *tangent classifier (to η_g at the point $\boldsymbol{\mu}$)*. Therefore, η_{t_g} can be viewed as a linearization, and thus a simplification, of η_g.

Though we can select any point $\boldsymbol{\mu}$ (such that $g(\boldsymbol{\mu}) = 0$) to approximate the surface $g(\mathbf{x}) = 0$, it is sensible to choose a relevant point for the discrimination procedure. Since in many occasions $g(\boldsymbol{\mu}_0) < 0$ and $g(\boldsymbol{\mu}_1) > 0$ (i.e., the centers of the groups are correctly classified), a natural choice is a point of the form $\boldsymbol{\mu} = \alpha\boldsymbol{\mu}_0 + (1 - \alpha)\boldsymbol{\mu}_1$, for some $\alpha \in [0, 1]$ (see Figure 1). That is, $\boldsymbol{\mu}$ lies in the segment joining $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$. However, there are situations in which it may happen that $\alpha \notin [0, 1]$. Of course, the value of α also depends on the prior probabilities and the misclassification costs.

3 Quadratic classification functions

Under heteroscedasticity ($\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$), several well-known classifiers adopts a quadratic form. The classification rule is η_q, where q is a (hyper)quadratic form usually expressed as

$$q(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\mathbf{x} - \boldsymbol{\mu}_0) - (\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + c, \tag{6}$$

and c is a real constant.

Since the quantity $(\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$ $(i = 0, 1)$ is the square of the Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}_i$, η_q corresponds to the Mahalanobis classifier

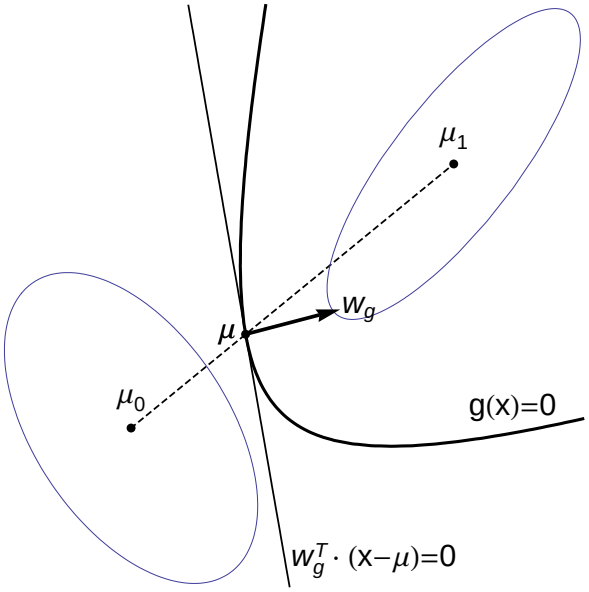

Figure 1: The tangent classifier to the rule generated by g.

when $c = 0$. In other words, $\mathbf{x}$ is assigned to the group whose center is closer, according to the Mahalanobis distance. If additionally $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$, then $q(\mathbf{x}) = 2f(\mathbf{x})$, where f is defined in (2). Hence, the Fisher's classifier can be seen as a particular case of η_q when $c = 0$ and under homoscedasticity.

In general, the constant c in (6) may depend on the total variability of the vector $\mathbf{X}$ in the groups (usually accounted by the determinants $|\boldsymbol{\Sigma}_0|$ and $|\boldsymbol{\Sigma}_1|$), on the prior probabilities of the groups and on the misclassification costs. For instance, if the distribution of $\mathbf{X}$ in both groups is d-variate normal, the Bayes rule (4) is η_q with

$$c = 2 \log \left(\frac{c(0|1)\pi_1 |\boldsymbol{\Sigma}_0|^{1/2}}{c(1|0)\pi_0 |\boldsymbol{\Sigma}_1|^{1/2}} \right). \tag{7}$$

The following result provides a closed-form expression for the tangent classifier to the rule η_q at a point in the line containing $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$.

Theorem 1. *Assume that* $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$ *and* $\Delta_0^2\Delta_1^2 + c(\Delta_1^2 - \Delta_0^2) \geq 0$*, where*

$$\Delta_i^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \qquad i = 0, 1. \tag{8}$$

We consider the point $\boldsymbol{\mu} = \alpha\boldsymbol{\mu}_0 + (1-\alpha)\boldsymbol{\mu}_1$*, with*

$$\alpha = \frac{\Delta_0^2 + c}{\Delta_0^2 + \sqrt{\Delta_0^2\Delta_1^2 + c(\Delta_1^2 - \Delta_0^2)}}. \tag{9}$$

We have that $q(\boldsymbol{\mu}) = 0$ *and the tangent classifier to* η_q *at the point* $\boldsymbol{\mu}$*,* η_{t_q}*, is determined by the classification function*

$$t_q(\mathbf{x}) = \mathbf{w}_q^\top(\mathbf{x} - \boldsymbol{\mu}), \tag{10}$$

where

$$\mathbf{w}_q = \left[(1-\alpha)\boldsymbol{\Sigma}_0^{-1} + \alpha\boldsymbol{\Sigma}_1^{-1}\right](\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0). \tag{11}$$

Proof. First, if $\mathbf{x} = \alpha\boldsymbol{\mu}_0 + (1-\alpha)\boldsymbol{\mu}_1$, $\alpha \in \mathbb{R}$, it is easy to see that $q(\mathbf{x}) = 0$ if and only if α is a solution of the second order equation

$$(\Delta_0^2 - \Delta_1^2)\alpha^2 - 2\Delta_0^2\alpha + (\Delta_0^2 + c) = 0, \tag{12}$$

where Δ_0^2 and Δ_1^2 are defined in (8). It is straightforward to check that α in (9) is a solution of the previous equation, so we conclude that $q(\boldsymbol{\mu}) = 0$.

On the other hand, q can be rewritten as

$$q(\mathbf{x}) = \mathbf{x}^\top\left[\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1}\right]\mathbf{x} + 2\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)^\top\mathbf{x} + d,$$

where $d = \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + c$ is a constant. Then, the normal vector $n(\mathbf{x}) = \nabla q(\mathbf{x})/2$ is given by

$$n(\mathbf{x}) = \left[\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1}\right]\mathbf{x} + \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0.$$

Finally, for any $\alpha \in \mathbb{R}$, it can be readily checked that

$$n(\alpha\boldsymbol{\mu}_0 + (1-\alpha)\boldsymbol{\mu}_1) = \left[(1-\alpha)\boldsymbol{\Sigma}_0^{-1} + \alpha\boldsymbol{\Sigma}_1^{-1}\right](\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0),$$

which finishes the proof. □

Under the assumptions of Theorem 1 and when $\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$, there are two values of α such that $q(\alpha\boldsymbol{\mu}_0 + (1-\alpha)\boldsymbol{\mu}_1) = 0$, i.e., equation (12) has two solutions. One solution is given in (9) and the other one is

$$\alpha' = \frac{\Delta_0^2 + \sqrt{\Delta_0^2\Delta_1^2 + c(\Delta_1^2 - \Delta_0^2)}}{\Delta_1^2 - \Delta_0^2}.$$

However, in general, the point $\boldsymbol{\mu}' = \alpha'\boldsymbol{\mu}_0 + (1-\alpha')\boldsymbol{\mu}_1$ does not lie between $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, so it is not relevant (for the discrimination purpose) to approximate q around the point $\boldsymbol{\mu}'$. Moreover, the value α' makes no sense when $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$.

Remark. From (9), and after some simple computations, we obtain

$$\frac{\partial\alpha}{\partial c} = \frac{1}{2\sqrt{\Delta_0^2\Delta_1^2 + c(\Delta_1^2 - \Delta_0^2)}} > 0.$$

Therefore, α is an increasing function of c. As c increases, the point $\boldsymbol{\mu} = \alpha\boldsymbol{\mu}_0 + (1-\alpha)\boldsymbol{\mu}_1$ (and hence η_q and η_{t_q}) gets closer to $\boldsymbol{\mu}_0$. For example, if c is given by (7), the greater $c(0|1)$ is with respect to $c(1|0)$ (or π_1 with respect to π_0, or $|\boldsymbol{\Sigma}_0|$ with respect to $|\boldsymbol{\Sigma}_1|$), the bigger α is.

The quantity Δ_i^2 defined in (8) is the square of the Mahalanobis distance between $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ with respect to $\boldsymbol{\Sigma}_i$. Note that in almost all important situations, q fulfills $\Delta_0^2\Delta_1^2 + c(\Delta_1^2 - \Delta_0^2) \geq 0$. Indeed, since

$$q(\boldsymbol{\mu}_0)q(\boldsymbol{\mu}_1) = (c + \Delta_0^2)(c - \Delta_1^2) = c^2 - [\Delta_0^2\Delta_1^2 + c(\Delta_1^2 - \Delta_0^2)],$$

it follows that $\Delta_0^2\Delta_1^2 + c(\Delta_1^2 - \Delta_0^2) \geq 0$ when $q(\boldsymbol{\mu}_0)q(\boldsymbol{\mu}_1) \leq 0$. This holds when q assigns the centers $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ to different groups.

Actually, if $-\Delta_0^2 \leq c \leq \Delta_1^2$, then α in (9) belongs to the interval $[0, 1]$. For instance, this happens if the centers are correctly classified. In particular, when $c = 0$ (i.e., η_q is the Mahalanobis classifier), we have the following corollary:

Corollary 1. *Assume that* $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$ *and* $c = 0$. *Let us consider the point*

$$\boldsymbol{\mu} = \left(\frac{\Delta_0}{\Delta_0 + \Delta_1}\right)\boldsymbol{\mu}_0 + \left(\frac{\Delta_1}{\Delta_0 + \Delta_1}\right)\boldsymbol{\mu}_1.$$

We have that $q(\boldsymbol{\mu}) = 0$ *and the tangent classifier to* η_q *at the point* $\boldsymbol{\mu}$, η_{t_q}, *is determined by the classification function* $t_q(\mathbf{x}) = \mathbf{w}_q^\top(\mathbf{x} - \boldsymbol{\mu})$, *where*

$$\mathbf{w}_q = \left[\left(\frac{\Delta_1}{\Delta_0 + \Delta_1}\right)\boldsymbol{\Sigma}_0^{-1} + \left(\frac{\Delta_0}{\Delta_0 + \Delta_1}\right)\boldsymbol{\Sigma}_1^{-1}\right](\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0).$$

Furthermore, if additionally $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$, *then* η_{t_q} *coincides with the Fisher's rule given in (2).*

Let us consider the Bayes classifier under normality η_q, with c given in (7). Under homoscedasticity, the tangent classifier is generated by the function

$$t_q(\mathbf{x}) = \mathbf{w}_F^\top(\mathbf{x} - \boldsymbol{\mu}),$$

where $\mathbf{w}_F$ is the Fisher's discriminant vector (3) and $\boldsymbol{\mu} = \alpha\boldsymbol{\mu}_0 + (1-\alpha)\boldsymbol{\mu}_1$ with

$$\alpha = \frac{1}{2} + \frac{1}{\Delta^2}\log\left(\frac{c(0|1)\pi_1}{c(1|0)\pi_0}\right) \quad \text{and} \quad \Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0).$$

Some simple computations show that

$$t_q(\mathbf{x}) = \mathbf{w}_F^\top\left(\mathbf{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2}\right) + \log\left(\frac{c(0|1)\pi_1}{c(1|0)\pi_0}\right),$$

which, of course, is the well-known Bayes classifier under normality and homoscedasticity.

When dealing with multivariate normal distributions with different covariance matrices, Anderson and Bahadur (1962) [1] developed a method to obtain a linear classifier called the *best linear classification rule*. The normal vector of the associated hyperplane is

$$\mathbf{w} = [t_1\boldsymbol{\Sigma}_0 + t_2\boldsymbol{\Sigma}_1]^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0),$$

where t_1 and t_2 are scalars chosen by systematic trial and error to minimize the probability of misclassification.

4 Two illustrative examples

The aim of this section is twofold. On the one hand, we want to identify some situations in which the tangent classifier is nearly optimal. On the other hand, we also show that the tangent and the Fisher's classifiers can be spectacularly different. This is carried out by two illustrative examples. For the sake of simplicity, we only consider equal costs and prior probabilities.

When $\mathbf{X}|\{G = i\}$ has density $f_i(\mathbf{x})$ $(i = 0, 1)$, the theoretically optimal classifier is defined by the classification function $b(\mathbf{x}) = f_1(\mathbf{x}) - f_0(\mathbf{x})$. The tangent classifier will be almost optimal if the curve $f_1(\mathbf{x}) = f_0(\mathbf{x})$, or, equivalently, the likelihood ratio $f_1(\mathbf{x})/f_0(\mathbf{x}) = 1$, is almost linear in regions in which $f_1(\mathbf{x})$ and $f_0(\mathbf{x})$ are large (the regions of interest for the discrimination purpose). In the first example below we consider such a case. In the second example, we discuss a similar situation in which the Fisher and the tangent classifier can be extremely different.

4.1 Exponential marginal distributions

For $i = 0, 1$, we assume $\mathbf{X}|\{G = i\}$ has exponential distribution with parameter $\mathbf{a}_i = (a_1^i, \ldots, a_d^i)^\top \in (0, \infty)^d$. That is, for $\mathbf{x} \in [0, \infty)^d$, its density is

$$f_i(\mathbf{x}) = c_i \exp\left(-\mathbf{a}_i^\top \mathbf{x}\right), \qquad c_i = a_1^i \cdots a_d^i.$$

It is easy to check that the Bayes classifier (4) can be expressed as η_b, where

$$b(\mathbf{x}) = (\mathbf{a}_0 - \mathbf{a}_1)^\top \mathbf{x} - \log(c_0/c_1), \qquad \mathbf{x} \in [0, \infty)^d.$$

It should be observed that whenever $\mathbf{a}_0 \neq \mathbf{a_1}$, η_b correctly classifies $\boldsymbol{\mu}_i = (1/a_1^i, \ldots, 1/a_d^i)^\top$ $(i = 0, 1)$. That is, $b(\boldsymbol{\mu}_0) < 0$ and $b(\boldsymbol{\mu}_1) > 0$. For instance, we

have

$$b(\boldsymbol{\mu}_1) = \sum_{i=1}^{d} \xi(a_i^0/a_i^1),$$

where $\xi(x) = x - \log(x) - 1$ $(x > 0)$. The function ξ is strictly convex on $(0, \infty)$, attains its global minimum at the point $x = 1$ and $\xi(1) = 0$. Therefore, $b(\boldsymbol{\mu}_1) > 0$ if and only if $\mathbf{a}_0 \neq \mathbf{a}_1$. An analogous reasoning shows that $b(\boldsymbol{\mu}_0) < 0$. In particular, this implies that there exists a point $\boldsymbol{\mu} = \alpha\boldsymbol{\mu}_0 + (1-\alpha)\boldsymbol{\mu}_1$ with $\alpha \in (0,1)$ such that $b(\boldsymbol{\mu}) = 0$. It can be readily seen that

$$\alpha = \frac{\sum_{i=1}^{d}[a_i^0/a_i^1 - \log(a_i^0/a_i^1) - 1]}{\sum_{i=1}^{d}[a_i^0/a_i^1 + a_i^1/a_i^0 - 2]},$$

and therefore the optimal rule coincides with the tangent classifier.

If we introduce a correlation between the variables of $\mathbf{X}$, in general the Bayes classifier will not be linear (not even quadratic) but the behavior in the area of interest for the discrimination purpose could be close to linearity. Thus, the tangent classifier will provide a simple linear classification rule with a nearly optimal behavior.

As illustration, we consider the Gumbel's bivariate exponential distribution (see Kotz *et al.* [12]) with parameters $\mathbf{a} = (a_1, a_2)^\top \in (0,\infty)^2$ and $\theta \in [0,1]$. For $x, y \in [0, \infty)$, the density function is

$$f(x,y) = a_1 a_2[(1 + a_1\theta x)(1 + a_2\theta y) - \theta] \exp\left[-(a_1 x + a_2 y + a_1 a_2 \theta x y)\right].$$

The vector $\mathbf{a}$ is a scale parameter and the marginal distributions are univariate exponentials. The parameter θ introduces a (negative) correlation given by

$$\frac{e^{1/\theta}}{\theta} \int_1^\infty \frac{e^{-t/\theta}}{t}\, \mathrm{d}t - 1.$$

This correlation decreases from 0 to $-0.403653\cdots$ as θ goes from 0 to 1. Obviously, when $\theta = 0$ the marginal distributions are independent.

For $i = 0, 1$, we assume $\mathbf{X}|\{G = i\}$ has the Gumbel's exponential distribution with parameters $\mathbf{a}_i$ and θ_i. In this case, the associated Bayes classifier is *not* quadratic unless $\theta_0 = \theta_1 = 0$ and it has an unpleasant expression. However, the tangent classifier can be easily computed numerically and it is often a sharp linear approximation to the optimal rule. Some examples are considered in Figure 2 in which the Bayes, tangent and Fisher's classifiers corresponding to different parameter values are plotted. Although the homoscedastic assumption is not fulfilled, Fisher's rule can be computed using the normal vector $\mathbf{w}_F = [(\mathbf{\Sigma}_0 + \mathbf{\Sigma}_1)/2]^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. As it can be noticed, the tangent classifier provides a nice linear approximation of the theoretically optimal rule. Further, the tangent classifier can be very different from the Fisher's rule.

We have carried out a small simulation study to compare the performance of the three classification rules (Fisher, tangent and Bayes) under the six models displayed in Figure 2. The Bayes rule has been computed using the true parameters so it is a landmark for the best possible result. For the other two rules the parameters are estimated from a training sample using the sample means and covariance matrices. The dependence parameter θ is estimated with the method of moments. We have carried out 1000 replications of the following procedure. First, training samples are drawn from each group with sizes $n_0 = n_1 = 100$. These training samples are used to compute the Fisher's and tangent classifiers (note that the Bayes classifier does not depend on the training sample). Then, two additional test samples of size 1000 are drawn from each group, and classified with the three rules, providing an estimation of the corresponding probabilities of misclassification. In Figure 3 we display the boxplots corresponding to the misclassification proportions for the three rules across the 1000 replications of the experiment. Notice that the

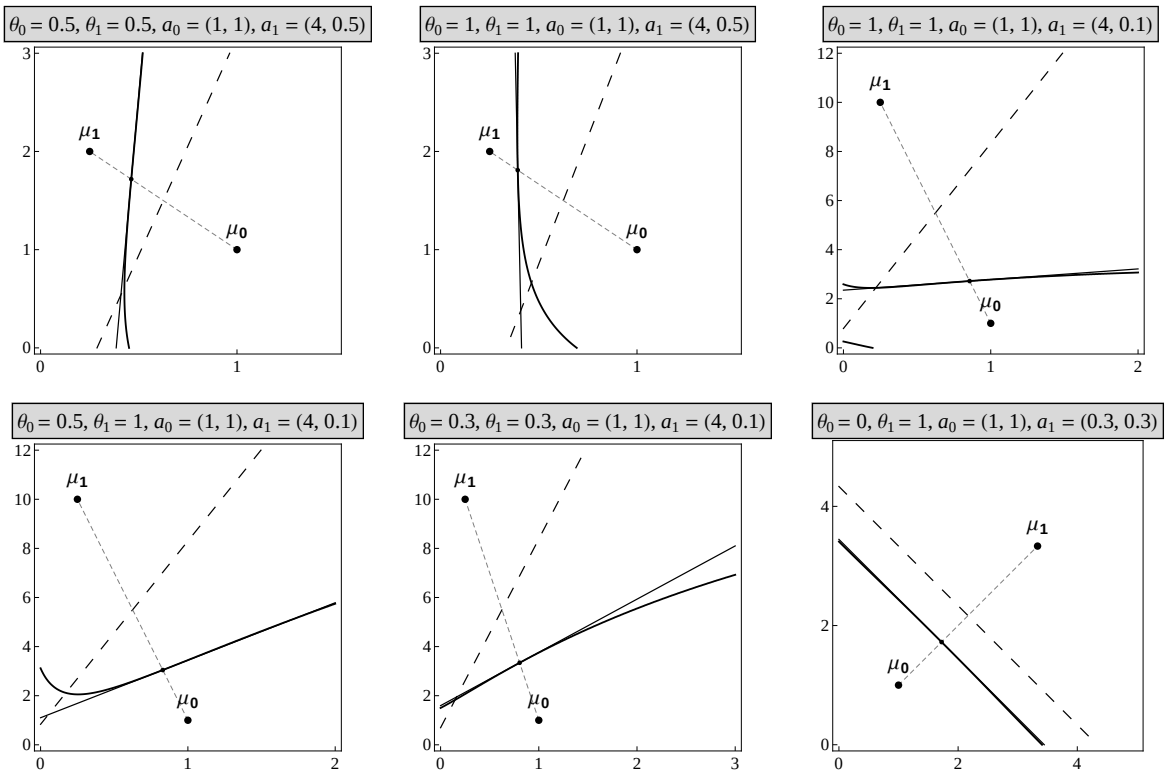

Figure 2: The Bayes classifier (solid curve), the tangent classifier (solid line) and the Fisher's classifier (dashed line) for different values of the parameters.

variation corresponding to the Bayes rule is only due to the test sample. We observe that, as expected, the tangent classifier gives better results than the Fisher's one across all the considered models. In fact, for some models the performance of the tangent rule is remarkably close to the optimal one.

4.2 Exponential versus normal class distributions

Let us assume that the vector $\mathbf{X}|\{G = 0\}$ has the exponential density of parameter $\mathbf{a} = (a_1, \ldots, a_d)^\top$, i.e.,

$$f_0(\mathbf{x}) = c_0 \, \exp\left(-\mathbf{a}^\top \mathbf{x}\right), \qquad \mathbf{x} \in [0, \infty)^d, \quad c_0 = a_1 \cdots a_d,$$

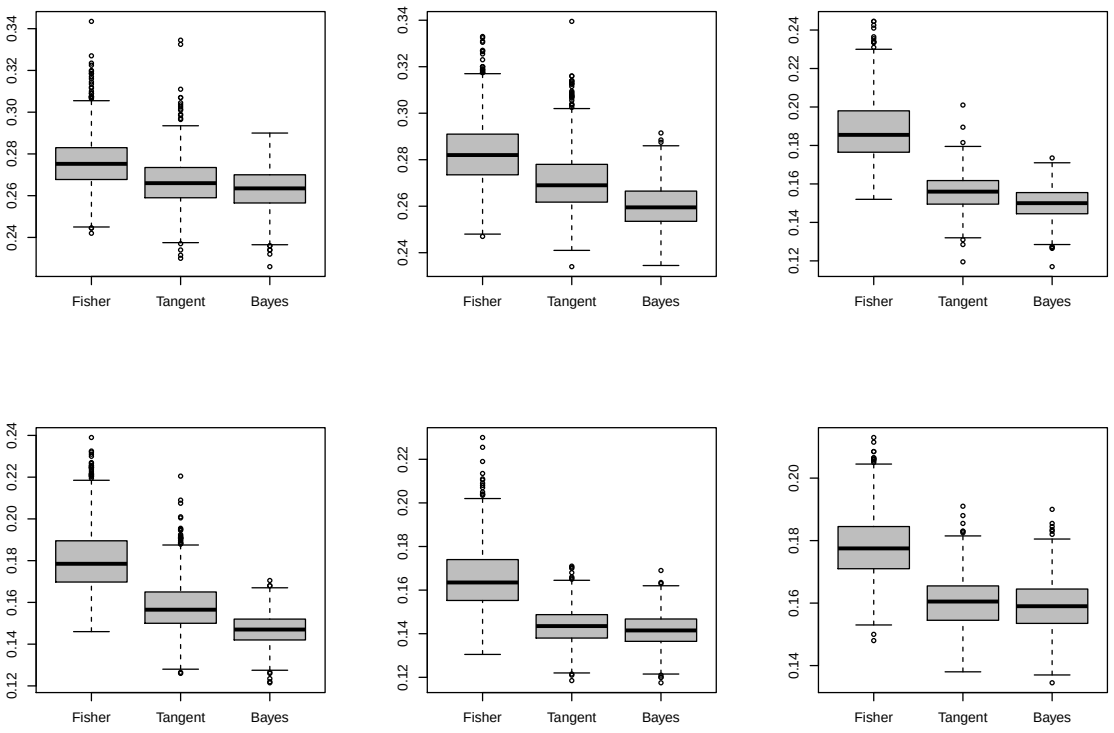

Figure 3: Misclassification proportions for the Fisher's, tangent and Bayes classifiers for the six models in Figure 2.

(and hence $\boldsymbol{\mu}_0 = (1/a_1, \ldots, 1/a_d)^\top$ and $\boldsymbol{\Sigma}_0 = \mathrm{diag}(1/a_1^2, \ldots, 1/a_d^2)$) and that $\mathbf{X}|\{G = 1\}$ has normal density

$$f_1(\mathbf{x}) = |2\pi\boldsymbol{\Sigma}_1|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right], \qquad \mathbf{x} \in \mathbb{R}^d.$$

The Bayes rule satisfies $\eta_b(\mathbf{x}) = 1$ whenever $\mathbf{x} \in \mathbb{R}^d - [0, \infty)^d$ and on $[0, \infty)^d$,

$$b(\mathbf{x}) = 2\mathbf{a}^\top \mathbf{x} - (\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + 2\log\left[\frac{|2\pi\boldsymbol{\Sigma}_1|^{-1/2}}{c_0}\right]. \tag{13}$$

We can follow the same steps as in the proof of Theorem 1 to show the following result. Details are left to the reader.

Theorem 2. *Let us assume that* $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$ *and* $[\mathbf{a}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)]^2 + c\,\Delta_1^2 \geq 0$, *where* Δ_1^2 *is defined in (8) and*

$$c = 2\mathbf{a}^\top \boldsymbol{\mu}_1 + 2\log\left[\frac{|2\pi\boldsymbol{\Sigma}_1|^{-1/2}}{c_0}\right].$$

We consider the point $\boldsymbol{\mu} = \alpha\boldsymbol{\mu}_0 + (1-\alpha)\boldsymbol{\mu}_1$ *with*

$$\alpha = \frac{1}{\Delta_1^2}\left[\mathbf{a}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \pm \sqrt{[\mathbf{a}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)]^2 + c\,\Delta_1^2}\right]. \tag{14}$$

We have that $b(\boldsymbol{\mu}) = 0$ *and the tangent classifier to* η_b *at the point* $\boldsymbol{\mu}$, η_{t_b}, *is determined by the classification function*

$$t_b(\mathbf{x}) = \mathbf{w}_b^\top(\mathbf{x} - \boldsymbol{\mu}), \qquad \textit{where} \qquad \mathbf{w}_b = \alpha\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + \mathbf{a}.$$

In this example, the right sign in the square root in (14) depends on the particular choice of the parameters $\mathbf{a}$, $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$. In practice, α can be selected as the closest value to 0.5 so that $\boldsymbol{\mu}$ lies on an important region for the discriminant analysis.

Since $\mathbf{a} = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$, under homoscedasticity $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$, we have that

$$\mathbf{w}_b = \boldsymbol{\Sigma}^{-1}[(1-\alpha)\boldsymbol{\mu}_0 + \alpha\boldsymbol{\mu}_1].$$

This example is fairly rich in the sense that a wide casuistry arises by selecting different parameter values. The centers are not necessarily well classified (even under equal priors and costs) and the tangent rule often provides a good approximation to the Bayes classifier. The tangent classifier can be extremely different from the Fisher's rule as the following curious example shows.

Example. Let $\mathbf{a} = (1,1)^\top$, $\boldsymbol{\mu}_1 = (1/2, 3/2)^\top$ and $\boldsymbol{\Sigma}_1 = \frac{e^2}{2\pi}\mathbf{I} \approx (1.176)\,\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Since $\boldsymbol{\Sigma}_0 = \mathbf{I}$, it is interesting to note that this example is not

far from being homoscedastic. The Bayes rule determined by (13) assigns $G = 1$ if $\mathbf{x} \in \mathbb{R}^2 - [0, \infty)^2$ or $\mathbf{x}$ is in the circle centered at the point $\mathbf{c}_b$ and radius r_d, where

$$\mathbf{c}_b = \frac{1}{2}\left(1 + \frac{e^2}{\pi}, 3 + \frac{e^2}{\pi}\right)^\top \qquad \text{and} \qquad r_b = \frac{e^2}{\pi\sqrt{2}}.$$

For $\mathbf{x} \in [0, \infty)^2$, the tangent classifier (to η_b at the point $\boldsymbol{\mu} = \boldsymbol{\mu}_1$) is η_t with

$$t(\mathbf{x}) = \mathbf{w}_t^\top \mathbf{x} - 2, \qquad \mathbf{w}_t = (1, 1)^\top,$$

and the Fisher's rule is η_f with

$$f(\mathbf{x}) = \mathbf{w}_f^\top \mathbf{x} - \frac{1}{2}, \qquad \mathbf{w}_f = (-1, 1)^\top.$$

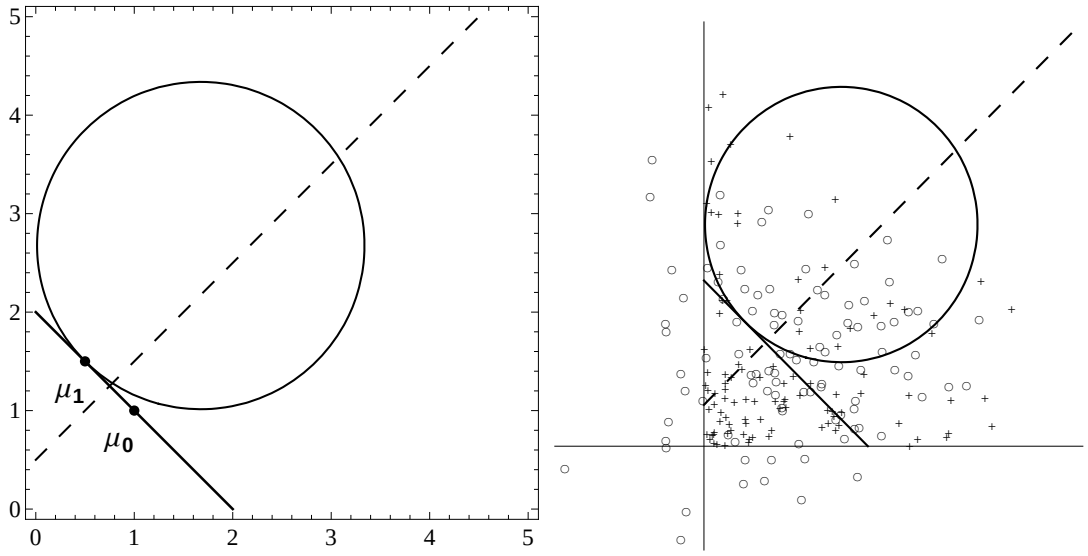

Figure 4: The Bayes classifier (circle), the tangent classifier (solid line) and the Fisher's classifier (dashed line). On the right, 100 realizations of the exponential (+ symbol) and the normal (∘ symbol).

In Figure 4, we see these three classifiers. The interesting characteristic of this example is that the separation boundaries of the tangent and the Fisher's rule are *orthogonal* ($\mathbf{w}_t^\top \mathbf{w}_f = 0$).

5 Further applications and numerical results

In practice, the classification function g has to be estimated with a training sample of independent copies of the pair $(\mathbf{X}, G)$. For each estimate $\hat{g}$ of g, we can compute the estimated classifier $\eta_{\hat{g}}$ and the associated tangent classifier, $\eta_{t_{\hat{g}}}$. For the Fisher's classifier defined in (2)–(3) we need to estimate the expectations in both groups and their common covariance matrix. If these quantities are estimated by the sample means, $\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{x}}_i$, and the pooled within-groups covariance matrix $\mathbf{S}_w$, we obtain an estimate $\hat{f}$ that leads to what is known in the literature as *linear discriminant analysis* (LDA). For the quadratic classifier η_q determined by (6), the expectations and covariances in both groups and the constant c have to be estimated. Estimating these quantities by the sample means, $\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{x}}_i$, and sample covariances, $\hat{\boldsymbol{\Sigma}}_i = \mathbf{S}_i$ $(i = 0, 1)$, we obtain an estimate $\hat{q}$ (whenever the constant c only depends on these quantities) that leads to the so-called *quadratic discriminant analysis* (QDA). The classifier $\eta_{\hat{q}}$ is expected to work well if the class conditional densities are elliptically symmetric (see Velilla and Hernández (2005) [17]), but we need to estimate correctly the covariance matrices of both groups. The procedure described in the previous sections allows us to define a third rule, different from LDA and QDA, by plugging in the sample means and covariances in (8)–(11). We call this resulting approach *tangent discriminant analysis* (TDA).

In the remainder of this section we point out two further instances in which sample means and covariances are not reliable estimators, and therefore it is helpful to replace them with other alternatives. The first situation corresponds to the presence of outliers in the training sample. The second one concerns the case in which the dimension of the data is large in comparison with the sample size. In both cases, the methodology described in the previous sections yields new linear

classifiers, easy to implement and interpret, which may be useful.

Regarding the simulations with Gaussian vectors considered in this section, it should be observed there exists a linear transformation of the data that simultaneously reduces $\mathbf{\Sigma}_0$ to the identity matrix and diagonalices $\mathbf{\Sigma}_1$ (see e.g. Gilbert (1969) [7], p. 506). As a consequence, considering models with diagonal covariance matrices is not too restrictive for comparing the behavior of the classifiers.

5.1 Robust tangent discriminant analysis

The potentially harming influence of outliers on sample means and covariance matrices is a well studied subject. Atypical observations may drastically change the location of the sample means or artificially inflate the covariance estimators. As a consequence, when these estimates are employed we may obtain distorted classifiers with a poor performance. In the literature there exist many robust alternative estimators of the population parameters in order to avoid this problem.

In this section, we consider the MCD-estimators of location and covariance introduced by Rousseeuw (1985) [14], tuned to have a 50% breakdown point. These estimators are based on the subset of the data of size $\lfloor n+d+1 \rfloor/2$ ($\lfloor \cdot \rfloor$ being the floor function) for which the determinant of the covariance matrix is minimal. The MCD-estimators of location and covariance are the mean and covariance of these $\lfloor n+d+1 \rfloor/2$ observations. Several authors have proposed the use of these estimators to obtain both linear and quadratic robust discriminant rules (see Hubert and van Driessen (2004) [10] or Croux *et al.* (2008) [3] and the references therein). We propose to use MCD-estimators instead of sample means and covariance matrices in (8)–(11). We call the resulting classification rule *robust tangent discriminant analysis* (RTDA).

We have carried out a simulation study to compare the performance of LDA, QDA, TDA and their robust counterparts RLDA, RQDA, and RTDA. We have assumed that the prior probabilities and misclassification costs are equal in both groups. The simulation layout is based on Joossens and Croux (2004) [11]. We consider four different models without outliers and the same models contaminated with 10% of outliers. For each case, class conditional distributions are 3-variate normal with different centers and covariance matrices (see below). We have carried out 1000 replications of the following procedure. First, training samples are drawn from each group with sizes $n_0 = n_1 = 1000$. These training samples are used to compute the six classification rules. Then, two additional test samples of size 5000 are drawn from each group, and classified with the six rules, providing an estimation of the misclassification probability. Test data are never contaminated with outliers. We report the test misclassification proportions averaged over the 1000 replications together with the corresponding standard deviations. In order to compute the MCD-estimators we have used the function *covMcd* of the R package *robustbase* (see Rousseeuw *et al.* (2011) [15]) which in turn uses the Fast MCD algorithm of Rousseeuw and Van Driessen (1999) [16].

If $\mathbf{a}$ stands for the vector $(a, a, a)^\top$ and $\mathbf{I}$ denotes the 3×3 identity matrix, the considered models are:

M1: $\boldsymbol{\mu}_0 = -\mathbf{1}$, $\boldsymbol{\mu}_1 = \mathbf{1}$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$ and $\boldsymbol{\Sigma}_1 = 0.25\mathbf{I}$. In the contaminated version (M1out), the parameters for the outliers distributions are $\boldsymbol{\mu}_0 = \mathbf{9}$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$, $\boldsymbol{\mu}_1 = -\mathbf{9}$ and $\boldsymbol{\Sigma}_1 = 0.25\mathbf{I}$.

M2: $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\mu}_1 = \mathbf{2}$, $\boldsymbol{\Sigma}_0 = 2.25\mathbf{I}$ and $\boldsymbol{\Sigma}_1 = 0.25\mathbf{I}$. In the contaminated version (M2out), the parameters for the outliers distributions are $\boldsymbol{\mu}_0 = \mathbf{3}$, $\boldsymbol{\Sigma}_0 = 9\mathbf{I}$, $\boldsymbol{\mu}_1 = -\mathbf{1}$ and $\boldsymbol{\Sigma}_1 = \mathbf{I}$.

M3: $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\mu}_1 = \mathbf{1}$, $\boldsymbol{\Sigma}_0 = 4\mathbf{I}$ and $\boldsymbol{\Sigma}_1 = 16\mathbf{I}$. In the contaminated version (M2out) the parameters for the outliers distributions are $\boldsymbol{\mu}_0 = \mathbf{4}$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$, $\boldsymbol{\mu}_1 = -\mathbf{16}$ and $\boldsymbol{\Sigma}_1 = \mathbf{I}$.

M4: $\boldsymbol{\mu}_0 = -\mathbf{1}$, $\boldsymbol{\mu}_1 = \mathbf{1}$, $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \mathbf{I}$. In the contaminated version (M4out) the parameters for the outliers distributions are $\boldsymbol{\mu}_0 = \mathbf{9}$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$, $\boldsymbol{\mu}_1 = -\mathbf{9}$ and $\boldsymbol{\Sigma}_1 = \mathbf{I}$.

Model	LDA	QDA	TDA	RLDA	RQDA	RTDA
M1	2.10 (0.16)	0.76 (0.09)	1.17 (0.12)	2.11 (0.17)	0.77 (0.09)	1.18 (0.12)
M1out	49.75 (4.03)	26.41 (0.42)	42.84 (0.39)	2.11 (0.17)	0.78 (0.09)	1.31 (0.14)
M2	6.23 (0.29)	1.82 (0.13)	4.57 (0.25)	6.24 (0.30)	1.83 (0.13)	4.59 (0.27)
M2out	6.34 (0.33)	8.05 (0.36)	35.84 (1.21)	6.08 (0.31)	1.89 (0.14)	5.07 (0.29)
M3	37.41 (0.58)	20.13 (0.40)	37.69 (0.50)	37.42 (0.59)	20.16 (0.40)	37.70 (0.56)
M3.out	56.06 (0.87)	23.67 (0.43)	48.60 (0.17)	37.09 (0.60)	21.02 (0.47)	41.16 (1.42)
M4	4.20 (0.19)	4.20 (0.19)	4.20 (0.19)	4.20 (0.19)	4.21 (0.19)	4.21 (0.20)
M4.out	49.93 (4.15)	49.78 (3.17)	49.79 (4.69)	4.18 (0.21)	4.19 (0.21)	4.19 (0.20)

Table 1: Average misclassification proportions of six classifiers over 1000 runs (standard deviations between parenthesis).

From Table 1 we see that robust rules behave similarly to classical ones when there are no outliers. This is reasonable for large samples, in which case the lower efficiency of robust methods should not have much effect on the non-outlier data. However, in the presence of outliers, robust versions clearly outperform the classical ones. Under homoscedasticity (M4 and M4out) the behavior of linear and quadratic rules is similar. It seems that large variances (M3 and M3out) affect negatively the behavior of the tangent classifiers. However, if we compare the two linear classifiers (LDA and TDA, both in the robust and the classical versions)

we see that TDA tends to behave better under heteroscedasticity, that is, when a quadratic rule is more suitable. The conclusion is that when quadratic rules are better than linear ones, but, for the sake of simplicity, a linear rule is preferable then the classical and robust versions (depending on the absence of presence of outliers) of TDA may be an appropriate choice.

5.2 Regularized tangent discriminant analysis

When the class sample sizes are relatively small in comparison with the dimension of the measurement space, the covariance matrices estimates are more variable and the quadratic classifier may degrade quickly. In this situation, it is convenient to apply a regularization method to estimate the covariances (see Friedman (1989) [6] for more details). Friedman's regularized classifier depends on two tuning parameters, but it turns out to be quadratic so we denote it RegQDA. The fact that Friedman's rule is quadratic also allows us to apply Theorem 1 to obtain, in the obvious way, the corresponding regularized tangent rule, denoted RegTDA.

We have carried out a simulation study to compare the performance of LDA, QDA, TDA and the regularized rules RegQDA and RegTDA. We have assumed that the prior probabilities and misclassification costs are equal in both groups. The simulation layout is based on Friedman (1989) [6]. We consider four different models. For each one, the class conditional distributions are 40-variate normal with different centers and covariance matrices (see below). We have carried out 1000 replications of the same experiment described in Subsection 5.1. In this case, training samples are drawn from each group with sizes $n_0 = n_1 = 50$ (observe the sample sizes are only slightly greater than the dimension) and test samples of size 5000 are drawn from each of the groups. In order to compute the regularized rules

we have used the function *rda* of the R package *klaR* using the default options (see Weihs *et al.* (2005) [19]). Tuning parameters are determined numerically by a Nelder-Mead optimization algorithm. The goal function to be minimized is the misclassification rate, estimated by cross-validation.

If $\mathbf{a}$ stands for $(a, \ldots, a)^\top \in \mathbb{R}^{40}$ and $\mathbf{I}$ denotes the 40×40 identity matrix, the considered models are:

M1: $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\mu}_1 = (3, 0, \ldots, 0)^\top$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \mathbf{I}$.

M2: $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\mu}_1 = (3, 0, \ldots, 0)^\top$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$ and $\boldsymbol{\Sigma}_1 = 2\mathbf{I}$.

M3.1: For $i = 1, \ldots, 40$ define $\lambda_i = (9(i-1)/39 + 1)^2$ and $m_i = 2.5\sqrt{\lambda_i/40}(40 - i)/19$. Then $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\mu}_1 = (m_1, \ldots, m_{40})^\top$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_{40})$.

M3.2: For $i = 1, \ldots, 40$ define $m_i = 2.5\sqrt{\lambda_i/40}(i-1)/19$, with λ_i as in M3.1. Then, $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\mu}_1 = (m_1, \ldots, m_{40})^\top$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_{40})$.

M4: For $i = 1, \ldots, 40$ define $\mu_i = (9(i - 39/2)/39)^2$ and λ_i as in M3.1. Then $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\mu}_1 = \mathbf{14}/\sqrt{40}$, $\boldsymbol{\Sigma}_0 = \mathrm{diag}(\mu_1, \ldots, \mu_{40})$ and $\boldsymbol{\Sigma}_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_{40})$.

Model	LDA	QDA	RegQDA	TDA	RegTDA
M1	14.26 (2.01)	38.06 (3.32)	8.97 (1.18)	21.97 (3.99)	9.77 (1.51)
M2	19.67 (2.26)	41.07 (3.03)	3.80 (1.20)	28.61 (4.06)	14.34 (1.25)
M3.1	14.57 (2.02)	38.19 (2.32)	16.72 (3.64)	22.11 (3.84)	16.58 (3.10)
M3.2	14.50 (2.08)	38.20 (3.29)	9.51 (1.05)	22.23 (4.08)	10.81 (1.50)
M4	9.41 (1.84)	7.73 (4.03)	0.21 (0.36)	16.24 (2.73)	5.31 (0.93)

Table 2: Average misclassification proportions of five classifiers over 1000 runs (standard deviations between parenthesis).

From Table 2 we see that RegQDA yields the best results. However if we are interested in simpler and more easily interpretable rules, we should look at the performance of the three linear rules included in the study (LDA, TDA and RegTDA). In this respect, RegTDA gives the best results among the three linear rules across all the considered models with the exception of M3.1, for which the results are not very different. An additional advantage of RegTDA is that it is possible to compute this linear classifier even when the dimension of the data is greater than the sample size and LDA is no longer applicable.

6 Real data example

In this section, we apply the tangent classifier to a real data set. Data are a subset of a coronary risk-factor study carried out in Western Cape, South Africa (see e.g. Hastie *et al.* (2001) [9], p. 100, and references therein). The two classes correspond to the presence or absence of myocardial infarction at the time of the survey. There are 160 cases and 302 controls in the sample. In order to classify an observation in one of the two classes we use the following eight variables: *sbp* (systolic blood pressure), *tobacco* (total lifetime usage), *ldl* (low density lipoprotein cholesterol), *adiposity*, *typea* (type-a behavior), *obesity*, *alcohol* (current alcohol consumption) and *age* (age at onset).

We have used several versions of cross-validation in order to compare the performance of five different classification rules: LDA, TDA, QDA, RegTDA and RegQDA (here, we follow the nomenclature of the previous section). Data are split into two parts, the training sample and the test sample. We use the training sample to compute the five classifiers and then classify the test sample. Misclassification proportions are recorded for the five classifiers and the process is repeated

1000 times. We investigate the behaviour of the classifiers when the size of the training samples ranges from a value very close to the number of explanatory variables ($n_0 = n_1 = 9$) to larger values ($n_0 = n_1 = 100$). The average and standard deviations of the misclassification proportions are reported in Table 3, together with the results corresponding to leave-one-out cross-validation.

Sample size	LDA	TDA	QDA	RegTDA	RegQDA
$n_0 = n_1 = 9$	39.69 (5.55)	44.16 (7.57)	47.81 (10.24)	**37.06** (5.05)	38.46 (5.46)
$n_0 = n_1 = 20$	35.73 (3.28)	35.77 (6.00)	39.65 (3.59)	**35.27** (2.89)	36.30 (3.16)
$n_0 = n_1 = 50$	33.65 (2.42)	**32.50** (7.41)	35.35 (2.50)	34.85 (2.76)	34.85 (2.51)
$n_0 = n_1 = 100$	33.40 (2.73)	**27.69** (8.42)	33.21 (2.71)	35.72 (3.50)	34.22 (2.97)
Leave-one-out	31.82	**30.52**	31.60	33.55	32.68

Table 3: Average misclassification errors for five classifiers (standard deviations between parenthesis). Lowest averages for each sample size in bold.

In this example, tangent rules give fairly good results when compared with their quadratic counterparts so in this case the additional simplicity comes for free. Notice also that there are no big differences in the average classification errors but some version of the tangent classifier (the regularized one for small sample sizes and the usual one for large ones) yields always the lowest values. On the other hand, TDA presents more variability except at the smallest sample size.

The weights assigned to the explanatory variables by a linear classifier represent the relevance of each variable in determining the overall score of each observation. In Figure 5 we display the weights (normalized so that they have unit square norm) for the three linear classifiers used in this example. In this case, LDA and RegTDA weights are fairly similar whereas those corresponding to TDA are

slightly different. Of course, this kind of pictures cannot be produced for quadratic and other non-linear rules.

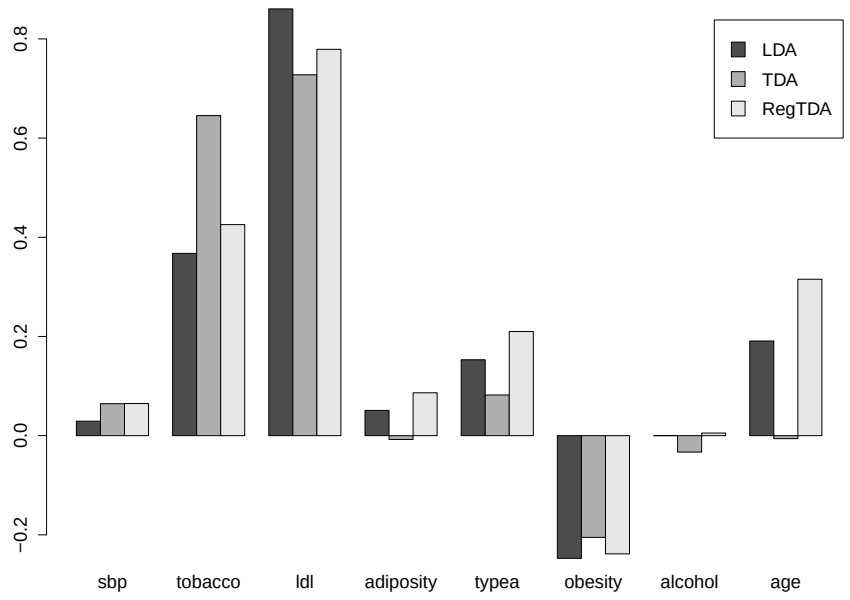

Figure 5: Weights assigned to the explanatory variables by LDA, TDA and RegTDA.

7 Conclusions

There exists an extensive literature comparing the properties of LDA and QDA (see for instance the discussion in Hastie *et al.* (2001) [9], p. 89, and the references therein). In this article we introduce a very simple linear rule, TDA, whose properties typically lie between those of LDA and QDA. In particular, in those situations where QDA is preferable to LDA, the tangent classifier provides a simple, easily interpretable classification rule whose behavior tends to be better than that

of LDA.

The method can be used in different contexts as long as a quadratic classifier is suitable. In particular, it allows us to define new robust linear classification rules, and also new linear rules when the dimension of the data is larger than the sample size.

When we are willing to assume a parametric model for the class conditional distributions, the tangent classifier can be almost optimal in some situations and it might be really different from the Fisher's classifier.

From a pedagogical point of view, we believe the derivations leading to TDA could be a valuable exercise in a course on multivariate analysis. They are fairly simple but, at the same time, they require to handle the main expressions that appear in LDA and QDA. The applications in Section 5 may also be useful for students to understand the difference between the population and the sample versions of a classifier.

Acknowledgments

We are grateful to the Editor, the Associate Editor and two referees for their helpful and constructive comments on this article. In particular, an Editor's remark led to the examples in Section 4.

References

[1] Anderson, W., and Bahadur, R. R. (1962). Classification into two multivariate normal distributions with different covariance matrices. *Ann. Math. Statist.*, **33**, 420–431.

[2] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Springer, New York.

[3] Croux, C., Filzmoser, P., and Joossens, K. (2008). Classification efficiencies for robust linear discriminant analysis. *Statistica Sinica*, **18**, 581–599.

[4] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York.

[5] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, **7**, 179–188.

[6] Friedman, J. H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.*, **84**, 165–175.

[7] Gilbert, E. S. (1969). The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. *Biometrics*, **25**, 505–515.

[8] Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statist. Sci.*, **21**, 1–14.

[9] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data mining, Inference and Prediction.* Springer, New York.

[10] Hubert, M., and van Driessen, K. (2004). Fast and robust discriminant analysis. *Computational Statistics and Data Analysis*, **45**, 301–320.

[11] Joosseens, K., and Croux, C. (2004). Empirical comparison of the classification performance of robust linear and quadratic discriminant analysis. In Hubert, M. *et al.* (eds.) *Theory and applications of recent robust methods*, 131–140, Birkhäuser, Basel.

[12] Kotz, S., Balakrishnan, N., and Johnson, N.L. *Continuous Multivariate Distributions, Volume 1, Models and Applications*, 2nd Edition, Wiley.

[13] Marks, S., and Dunn, O. J. (1974). Discriminant functions when covariance matrics are unequal. *J. Amer. Statist. Assoc.*, **69**, 555–559.

[14] Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In Grossmann, W *et al.* (eds.) *Mathematical Statistics and Applications*, vol. B., 283–297, Reidel, Dordrecht.

[15] Rousseeuw, P. J., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., and Maechler, M. (2011). robustbase: Basic Robust Statistics. R package version 0.7-6. URL `http://CRAN.R-project.org/package=robustbase`

[16] Rousseeuw, P. J., and van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.

[17] Velilla, S., and Hernández, A. (2005). On the consistency of linear and quadratic discriminant analysis. *J. Multivariate Anal.*, **96** 219–236.

[18] Wald, P. W., and Kronmal, R.A. (1977). Discriminant functions when covariances are unequal and sample sizes are moderate. *Biometrics*, **33**, 479–484.

[19] Weihs, C., Ligges, U., Luebke, K., and Raabe, N. (2005). klaR Analyzing German Business Cycles. In Baier, D., Decker, R. and Schmidt-Thieme, L. (eds.). *Data Analysis and Decision Support*, 335-343, Springer-Verlag, Berlin.