Optimal expected-distance separating halfspace

Emilio Carrizosa* Frank Plastria†

15th March 2004

Abstract

One recently proposed criterion to separate two datasets in discriminant analysis, is to use a hyperplane which minimises the sum of distances to it from all the misclassified data points. Here all distances are supposed to be measured by way of some fixed norm, while misclassification means lying on the wrong side of the hyperplane, or rather in the wrong halfspace. In this paper we study the problem of determining such an optimal halfspace when points are distributed according to an arbitrary random vector X in $\mathbb{R}^d$,. In the unconstrained case in dimension d, we prove that any optimal separating halfspace always balances the misclassified points. Moreover, under polyhedrality assumptions on the support of X, there always exists an optimal separating halfspace passing through d affinely independent points. It follows that the problem is polynomially solvable in fixed dimension by an algorithm of $O(n^{d+1})$ when the support of X consists of n points.

All these results are strengthened in the one-dimensional case, yielding an algorithm with complexity linear in the cardinality of the support of X.

If a different norm is used for each data set in order to measure distances to the hyperplane, or if all distances are measured by a

*Facultad de Matemáticas, Universidad de Sevilla, Tarfia s/n, 41012 Sevilla, Spain, e-mail: ecarrizosa@us.es

†MOSI - Department of Mathematics, Operational research, Statistics and Information systems for management, Vrije Universiteit Brussel, Pleinlaan 2, B 1050 Brussels, Belgium, e-mail: Frank.Plastria@vub.ac.be

fixed gauge, the balancing property still holds, and we show that, under polyhedrality assumptions on the support of X, there always exists an optimal separating halfspace passing through $d-1$ affinely independent data points.

These results extend in a natural way when we allow constraints modeling that certain points are forced to be correctly classified.

Keywords. Gauge-distance to hyperplane, separating halfspace, discriminant analysis.

1 Problem statement

In two-group probabilistic Discriminant Analysis models, one has two classes of objects, C_A, C_B, identified with points in $\mathbb{R}^d$, and the aim is to build, by minimizing a certain error function, a *classification rule*, i.e., a function that labels as member of C_A or member of C_B an entry $x \in \mathbb{R}^d$ whose class membership is unknown. See e.g. [3, 6, 8, 10, 16] for comprehensive surveys.

We assume full distributional knowledge: we are given a random vector (X, Y) in $\mathbb{R}^d \times \{-1, 1\}$; the d-variate random vector X models the coordinates representing objects and Y models class membership, where we identify the value 1 (respectively -1) with class C_A (respectively C_B), and assume the distribution of the vector (X, Y) to be known.

In practical settings, one usually has a data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_p, y_p)\} \subset \mathbb{R}^d \times \{-1, 1\}$, assumed to be a random sample of the random vector (X, Y), and then the distribution of (X, Y) is determined (estimated) from the sample by one of a number of procedures, as reviewed in e.g. [3], thus fulfilling this full-distributional knowledge assumption. For instance, in Bayesian models, a *prior* distribution for (X, Y) is assumed to be given and then updated after obtaining the sample $\mathcal{D}$; in nonparametric models, one assumes (X, Y) to be uniformly distributed over the finite set $\mathcal{D}$, or represented via a *kernel* as a mixture of densities (e.g. uniform distributions on hypercubes centered at the data points); in parametric approaches, (X, Y) is assumed to have a distribution known up to some parameters (e.g. gaussian distributions with unknown mean values and covariances), which are estimated by standard point-estimation methods once $\mathcal{D}$ becomes available.

Classification will be obtained according to a linear rule, defined via a hyperplane, as follows.

Let $\mathcal{H} \stackrel{\text{def}}{=} \left(\mathbb{R}^d \setminus \{0\}\right) \times \mathbb{R}$. For any pair $\sigma \stackrel{\text{def}}{=} (u, \beta) \in \mathcal{H}$, define the halfspaces and hyperplanes

$$H^{\#}(\sigma) = H^{\#}(u, \beta) \stackrel{\text{def}}{=} \{\ x \in \mathbb{R}^d \mid \langle\ u\ ;\ x\ \rangle \# \beta\ \}$$

where $\# \in \{\leq, <, =, \geq, >\}$. Note that these sets all remain the same when σ is multiplied by any strictly positive constant, but not when the sign is inversed. In fact we will use the halfspace $H^{<}(\sigma)$ to discriminate elements from C_A, as opposed to those of C_B. Therefore we say we separate by a halfspace (or oriented hyperplane), and the sign of σ is thus part of the information. We will also speak of the 'halfspace' σ by an abuse of terminology.

Given $\sigma = (u, \beta) \in \mathcal{H}$, $x \in \mathbb{R}^d$ and $y \in \{-1, 1\}$, we will say that (x, y) is *correctly classified* by σ iff $y\langle\ u\ ;\ x\ \rangle < y\beta$, which occurs iff $x \in H^{<}(y\sigma)$. Otherwise, (x, y) is said to be *misclassified* by σ.

When $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ is misclassified by σ, a penalty (misclassification cost) is incurred. This penalty is assumed to be proportional to the distance to the halfspace of correctly classified points, as detailed later.

Distances are measured according to gauges γ in $\mathbb{R}^d$, see e.g. [15, 7, 9] and Section 2, and the γ-distance of a point $a \in \mathbb{R}^d$ to $H^{\#}(\sigma)$ is then

$$d_\gamma(a, H^{\#}(\sigma)) \stackrel{\text{def}}{=} \inf\{\ \gamma(x - a) \mid x \in H^{\#}(\sigma)\ \}, \tag{1}$$

where this infimum is reached when $\# \in \{\leq, =, \geq\}$.

Gauges γ_A and γ_B are given to measure distances to halfspaces, and thus misclassification costs: for σ and (x, y) given, the misclassification cost $c((x, y), \sigma)$ associated with (x, y) when σ is used as classifier is defined as

$$c((x, y), \sigma) \stackrel{\text{def}}{=} \begin{cases} d_{\gamma_A}(x, H^{<}(\sigma)), & \text{if } y = 1 \\ d_{\gamma_B}(x, H^{>}(\sigma)), & \text{if } y = -1. \end{cases} \tag{2}$$

Hence, for a given $\sigma = (u, \beta) \in \mathcal{H}$, the expected misclassification cost of (X, Y) is then given by

$$f(\sigma) = \mathbb{E}\left[c((X, Y), \sigma)\right], \tag{3}$$

where $\mathbb{E}[\cdot]$ denotes expected value.

Classification constraints are allowed in the model: two subsets S_A^*, S_B^* of $\mathbb{R}^d$ are given such that the search for halfspaces is reduced to those which do

not classify as members of C_B points in S_A^* and do not classify as members of C_A points in S_B^*. In other words, σ must satisfy

$$S_A^* \subset H^{\leq}(\sigma),\ S_B^* \subset H^{\geq}(\sigma), \tag{4}$$

or equivalently,

$$\sup\{\ \langle\ u\ ;\ a\ \rangle \mid a \in S_A^*\ \} \leq \beta \leq \inf\{\ \langle\ u\ ;\ b\ \rangle \mid b \in S_B^*\ \},$$

where $\sup\{\ \langle\ u\ ;\ a\ \rangle \mid a \in S_A^*\ \} = -\infty$ (respectively $\inf\{\ \langle\ u\ ;\ b\ \rangle \mid b \in S_B^*\ \} = +\infty$) when $S_A^* = \emptyset$ (respectively $S_B^* = \emptyset$).

Since (4) is satisfied iff

$$\sup\{\ \langle\ u\ ;\ a\ \rangle \mid a \in cl(conv(S_A^*))\ \} \leq \beta \leq \inf\{\ \langle\ u\ ;\ b\ \rangle \mid b \in cl(conv(S_B^*))\ \},$$

where $cl(conv(Z))$ stands for the closure of the convex hull of the set Z, we can assume wlog that S_A^*, S_B^* are closed convex sets.

We are concerned in this paper with the existence of optimal hyperplanes passing through many points. The results are simplified if we can assert that $\sup\{\ \langle\ u\ ;\ a\ \rangle \mid a \in S_A^*\ \}$ and $\inf\{\ \langle\ u\ ;\ b\ \rangle \mid b \in S_B^*\ \}$, when finite, are attained. A sufficient condition for this to happen is to impose some structure on S_A^*, S_B^*. We make the following

Assumption 1 S_A^* and S_B^* are *asymptotically conical sets,* i.e., there exist compact convex sets M_A, M_B and closed convex cones K_A, K_B (possibly degenerated to $\{0\}$) such that

$$\begin{array}{rcl} S_A^* & = & M_A + K_A \\ S_B^* & = & M_B + K_B. \end{array} \tag{5}$$

See [4] for further results on this concept.

When S_A^* or S_B^* are non-empty, σ satisfying (4) may not exist. We exclude this by assuming throughout the paper linear separability of S_A^* and S_B^* :

Assumption 2 There exists $\sigma \in \mathcal{H}$ satisfying (4).

Moreover, for reasons that will become apparent later, we also impose the following:

Assumption 3 $0 < \tau < 1$, where $\tau \stackrel{\text{def}}{=} \mathbb{P}[Y = 1]$.

Assumption 4 Let $\|\cdot\|$ denote the Euclidean norm in $\mathbb{R}^d$. Then, $\mathbb{E}\left[\|X\||(Y=1)\right]$ and $\mathbb{E}\left[\|X\||(Y=-1)\right]$, the expected values of $\|X\|$ conditioned to the events $(Y=1)$ and $(Y=-1)$, exist and are finite.

Any optimal solution to

$$\begin{array}{ll} \min & f(\sigma) \\ \text{s.t.} & \sigma \in \mathcal{H},\ S_A^* \subset H^{\leq}(\sigma),\ S_B^* \subset H^{\geq}(\sigma) \end{array} \tag{6}$$

will be called hereafter an *optimal separating halfspace*.

Our aim is to study properties of optimal separating halfspaces in the general setting described above. With this, we provide deeper insight, and obtain a new algorithmic approach, to the work of [11], where problem (6) is addressed in the particular case that the distribution of (X, Y) is uniformly distributed on the finite set $\mathcal{D}$, and no constraints of type (4) are allowed.

In particular, it is shown in Theorem 2 below that optimal separating halfspaces exist. Moreover, under stronger assumptions on X and the gauges γ_A, γ_B, the search for an optimal separating halfspace may be reduced by Theorems 13 and 15, to halfspaces with boundary containing sufficiently many points of the support of X.

The remainder of the paper is structured as follows. In Section 2 we recall certain properties of gauge distances, which enable us to rewrite the objective function and view problem (6) as a nonlinear nonconvex optimisation problem in $d+1$ variables.

Section 3 discusses existence and balancing properties, extending those previously derived by the authors in [14] for the case of uniform distribution on $\mathcal{D}$. Section 4 then concerns the localization results for optimal separating halfspaces, first for the particular case in which the gauges are induced by a common norm, and then in the more general setting in which different norms (or gauges) are used for the different classes.

The following notation will be used. For a given set $S \subset \mathbb{R}^d$, $conv(S)$ will denote its convex hull. For a convex set S, let $dim(S)$ denote the dimension of S, and $relint(S)$ its relative interior. For an arbitrary S, $ext(S)$ will denote the set of extreme points of the closure of $conv(S)$.

For any random vector Z, let S_Z be its support, i.e., the smallest closed set S satisfying $\mathbb{P}\left[Z \in S\right] = 1$. Hence, $S_{X|Y=1}$ and $S_{X|Y=-1}$ represent the supports of the variable X conditioned to the values of Y, i.e., the distribution of X for classes C_A and C_B.

2 Misclassification costs evaluation

As announced in Section 1, misclassification costs are measured through distances, these being induced by gauges.

We recall that, given a compact convex set Γ containing the origin in its interior, the gauge γ with unit ball Γ is defined as

$$\gamma(x) \stackrel{\text{def}}{=} \min\{\ t \geq 0 \,|\, x \in t\Gamma\ \}.$$

The dual (or polar) γ° of γ is

$$\gamma^\circ(v) \stackrel{\text{def}}{=} \max\{\ \langle\ v\ ;\ y\ \rangle \,|\, \gamma(y) \leq 1\ \} = \max\{\ \langle\ v\ ;\ y\ \rangle \,|\, y \in \Gamma\ \},$$

which is well-defined and also a gauge on (the dual space of) $\mathbb{R}^d$, see e.g. [9, 15]. Note that this definition immediately implies the generalized Cauchy-Schwarz inequality $\langle\ v\ ;\ x\ \rangle \leq \gamma^\circ(v)\gamma(x)$.

Since the γ-distance from a point x to $H^{\#}(\sigma)$ equals 0 when $x \in H^{\#}(\sigma)$, and equals $d_\gamma(x, H^{=}(\sigma))$ otherwise, the formula in [13] for γ-distances to hyperplanes yields the following expression for the γ-distance to halfspaces:

$$d_\gamma(x, H^{<}(u,\beta)) = d_\gamma(x, H^{\leq}(u,\beta)) \;=\; \left(\frac{\langle\ u\ ;\ x\ \rangle - \beta}{\gamma^\circ(-u)}\right)^+ \tag{7}$$

$$d_\gamma(x, H^{>}(u,\beta)) = d_\gamma(x, H^{\geq}(u,\beta)) \;=\; \left(\frac{\beta - \langle\ u\ ;\ x\ \rangle}{\gamma^\circ(u)}\right)^+, \tag{8}$$

where $(\cdot)^+ = \max\{\cdot, 0\}$.

Expressions (7)-(8) enable us to rephrase in a more manageable form the expected misclassification cost $f(\sigma)$ defined in (3):

$$\begin{aligned} f(\sigma) &= \mathbb{E}\left[c((X,Y),\sigma)\right] \\ &= \tau\mathbb{E}\left[c((X,1),\sigma)|Y=1\right] + (1-\tau)\mathbb{E}\left[c((X,-1),\sigma)|Y=-1\right]. \end{aligned}$$

Hence, by (2),

$$f(\sigma) = \tau\mathbb{E}\left[d_{\gamma_A}(X, H^{<}(\sigma))|Y=1\right] + (1-\tau)\mathbb{E}\left[d_{\gamma_B}(X, H^{>}(\sigma))|Y=-1\right].$$

Denote respectively by F_A and F_B the distributions of the random vectors $X|(Y=1)$ and $X|(Y=-1)$. Then,

$$f(\sigma) = \tau \int \left(\frac{\langle\ u\ ;\ a\ \rangle - \beta}{\gamma_A^\circ(-u)}\right)^+ dF_A(a) + (1-\tau)\int \left(\frac{\beta - \langle\ u\ ;\ b\ \rangle}{\gamma_B^\circ(u)}\right)^+ dF_B(b). \tag{9}$$

Remark 1 The integrals in (9) are well defined and finite by Assumption 4. Indeed, the first integrand is non-negative and bounded above by $\frac{\|u\|\|a\|+|\beta|}{\gamma_A^\circ(-u)}$, thus the first integral is bounded above by $\frac{\|u\|\mathbb{E}[\|X\||Y=1]+|\beta|}{\gamma_A^\circ(-u)}$. Similarly, the second integrand is also non-negative, bounded above by $\frac{\|u\|\|b\|+|\beta|}{\gamma_B^\circ(u)}$, thus the second integral is bounded above by $\frac{\|u\|\mathbb{E}[\|X\||Y=-1]+|\beta|}{\gamma_B^\circ(u)}$. Hence,

$$f(\sigma) \leq \tau \frac{\|u\|\mathbb{E}\left[\|X\||Y=1\right]+|\beta|}{\gamma_A^\circ(-u)} + (1-\tau)\frac{\|u\|\mathbb{E}\left[\|X\||Y=-1\right]+|\beta|}{\gamma_B^\circ(u)} < +\infty,$$

as asserted □

Since for any $\lambda > 0$ we have $H^{\#}(\lambda\sigma) = H^{\#}(\sigma)$, the optimization problem (6) can be reformulated as

$$\begin{array}{ll} \min & \frac{\tau}{\gamma_A^\circ(-u)} \int (\langle\ u\ ;\ a\ \rangle - \beta)^+ \, dF_A(a) + \frac{1-\tau}{\gamma_B^\circ(u)} \int (\beta - \langle\ u\ ;\ b\ \rangle)^+ \, dF_B(b) \\ \text{s.t.} & \gamma(u) = 1 \\ & \sup\{\ \langle\ u\ ;\ a\ \rangle \,|\, a \in S_A^*\ \} \leq \beta \leq \inf\{\ \langle\ u\ ;\ b\ \rangle \,|\, b \in S_B^*\ \}, \end{array} \tag{10}$$

where γ is an arbitrary fixed gauge.

Some properties of this optimization problem are discussed in the next section.

3 General properties

3.1 Existence of optimal solutions

Problem (10) is, in general, non-differentiable, non-convex, and may have an unbounded feasible region (for instance, when S_A^* and S_B^* are empty, since β is then unconstrained). However, optimal solutions always exist. Indeed one has

Theorem 2 *Under assumptions 1-4 there always exists an optimal separating halfspace. Moreover, if $S_{X|Y=1}, S_{X|Y=-1}$ are compact sets, then there exists an optimal separating halfspace, the boundary of which intersects $S_{X|Y=1} \cup S_{X|Y=-1} \cup S_A^* \cup S_B^*$.*

Proof. Let us first show that an optimal separating halfspace exists. We start by constructing an $M > 0$ such that, if $|\beta| > M$, then no $(u, \beta) \in \mathcal{H}$ with $\gamma(u) = 1$ can be optimal to (10).

1. By Assumptions 3–4 and the equivalence of norms in $\mathbb{R}^d$, there exists $M_0 > 0$ such that

$$\mathbb{P}\left[\gamma^\circ(X) \leq M_0 | Y = -1\right] \geq \frac{1}{2}.$$

Then for any $\beta \geq M_0$ and u with $\gamma(u) = 1$, the Cauchy-Schwartz inequality yields

$$\begin{aligned}
\mathbb{P}\left[\langle\ u\ ;\ X\ \rangle \leq \beta | Y = -1\right] &\geq \mathbb{P}\left[\gamma(u)\gamma^\circ(X) \leq \beta | Y = -1\right] \\
&= \mathbb{P}\left[\gamma^\circ(X) \leq \beta | Y = -1\right] \\
&\geq \mathbb{P}\left[\gamma^\circ(X) \leq M_0 | Y = -1\right] \geq \frac{1}{2}.
\end{aligned}$$

In other words, for any $\beta \geq M_0$ and u with $\gamma(u) = 1$ one has

$$\mathbb{P}\left[X \in H^{\leq}(u, \beta) | Y = -1\right] \geq \frac{1}{2}$$

and thus

$$\begin{aligned}
f(u,\beta) &\geq \frac{1-\tau}{\gamma_B^\circ(u)} \int \left(\beta - \langle\ u\ ;\ b\ \rangle\right)^+ dF_B(b) \\
&\geq \frac{1-\tau}{\gamma_B^\circ(u)} \int_{H^{\leq}(u,M_0)} \left(\beta - \langle\ u\ ;\ b\ \rangle\right)^+ dF_B(b) \\
&= \frac{1-\tau}{\gamma_B^\circ(u)} \int_{H^{\leq}(u,M_0)} \left(\beta - \langle\ u\ ;\ b\ \rangle\right) dF_B(b) \\
&= \frac{1-\tau}{\gamma_B^\circ(u)} \beta \mathbb{P}\left[X \in H^{\leq}(u, M_0) | Y = -1\right] - \frac{1-\tau}{\gamma_B^\circ(u)} \int_{H^{\leq}(u,M_0)} \langle\ u\ ;\ b\ \rangle dF_B(b) \\
&\geq \frac{1-\tau}{\gamma_B^\circ(u)} \frac{\beta}{2} - \frac{1-\tau}{\gamma_B^\circ(u)} \int |\langle\ u\ ;\ b\ \rangle| dF_B(b).
\end{aligned}$$

Let now (u^0, β^0) be any feasible solution for (10). Then, a sufficient condition for $\beta \geq M_0$ to satisfy $f(u, \beta) > f(u^0, \beta^0)$ for all u with $\gamma(u) = 1$ is that

$$\frac{1-\tau}{\gamma_B^\circ(u)} \frac{\beta}{2} - \frac{1-\tau}{\gamma_B^\circ(u)} \int |\langle\ u\ ;\ b\ \rangle| dF_B(b) > f(u^0, \beta^0) \qquad \forall u, \gamma(u) = 1,$$

or, equivalently,

$$\beta > 2\int |\langle\ u\ ;\ b\ \rangle| dF_B(b) + \frac{2\gamma_B^\circ(u)}{1-\tau} f(u^0,\beta^0) \qquad \forall u, \gamma(u) = 1.$$

In particular, defining M_1 as

$$M_1 = \max_{\gamma(u)=1} \left[2\int |\langle\ u\ ;\ b\ \rangle| dF_B(b) + \frac{2\gamma_B^\circ(u)}{1-\tau} f(u^0,\beta^0) \right],$$

(which exists by continuity of the expression and compactness of the unit ball of γ) one has for any $\beta > M' \stackrel{\text{def}}{=} \max\{M_0, M_1\}$ that $f(u,\beta) > f(u^0,\beta^0)$ for all u with $\gamma(u) = 1$.

2. Similarly one can find $M'' > 0$ such that, for any $\beta < -M''$, $f(u,\beta) > f(u^0,\beta^0)$ for all u with $\gamma(u) = 1$.

3. defining $M \stackrel{\text{def}}{=} \max\{M', M''\}$, one has

$$f(u,\beta) > f(u^0,\beta^0) \text{ for all } u, \beta \text{ with } \gamma(u) = 1,\ |\beta| > M.$$

In particular, $|\beta^0| \leq M$. This implies that Problem (10) can be rephrased as

$$\begin{array}{ll} \min & \frac{\tau}{\gamma_A^\circ(-u)} \int (\langle\ u\ ;\ a\ \rangle - \beta)^+ dF_A(a) + \frac{1-\tau}{\gamma_B^\circ(u)} \int (\beta - \langle\ u\ ;\ b\ \rangle)^+ dF_B(b) \\ \text{s.t.} & \gamma(u) = 1 \\ & |\beta| \leq M \\ & \sup\{\ \langle\ u\ ;\ a\ \rangle \mid a \in S_A^*\ \} \leq \beta \leq \inf\{\ \langle\ u\ ;\ b\ \rangle \mid b \in S_B^*\ \}. \end{array} \tag{11}$$

The objective function of (11) is continuous on its feasible region R, which is compact and, since $(u^0,\beta^0) \in R$, it is non-empty. Hence, an optimal solution exists, as asserted.

Assume now that $S_{X|Y=1}$ and $S_{X|Y=-1}$ are compact sets. Let $\sigma^0 = (u^0,\beta^0)$ be an optimal separating halfspace, existence of which has just been shown, and consider the linear program in the one-dimensional variable β

$$\begin{array}{ll} \min & \frac{\tau}{\gamma_A^\circ(-u^0)} \int_{H^\geq(\sigma^0)} (\langle\ u^0\ ;\ a\ \rangle - \beta)\, dF_A(a) + \frac{1-\tau}{\gamma_B^\circ(u^0)} \int_{H^\leq(\sigma^0)} (\beta - \langle\ u^0\ ;\ b\ \rangle)\, dF_B(b) \\ \text{s.t.} & \sup\{\ \langle\ u^0\ ;\ x\ \rangle \mid x \in H^\leq(\sigma^0) \cap \left(S_{X|Y=1} \cup S_{X|Y=-1}\right)\ \} \leq \beta \\ & \inf\{\ \langle\ u^0\ ;\ x\ \rangle \mid x \in H^\geq(\sigma^0) \cap \left(S_{X|Y=1} \cup S_{X|Y=-1}\right)\ \} \geq \beta \\ & \sup\{\ \langle\ u^0\ ;\ a\ \rangle \mid a \in S_A^*\ \} \leq \beta \leq \inf\{\ \langle\ u^0\ ;\ b\ \rangle \mid b \in S_B^*\ \} \end{array} \tag{12}$$

One immediately has that β^0 is not only feasible but also optimal for (12). Moreover, since $H^{\leq}(\sigma^0) \cup H^{\geq}(\sigma^0) = \mathbb{R}^d$ and S_X is nonempty, the set of nontrivial constraints of problem (12) is nonempty. This implies that its feasible region is an interval with at least one finite endpoint. Since the objective is linear and bounded below, there exists an optimal solution β which is an endpoint, thus at least one of the constraints of (12) is binding at β. In other words, at least one of the four following conditions hold:

$$\sup\{ \langle u^0 ; x \rangle \,|\, x \in H^{\leq}(\sigma^0) \cap (S_{X|Y=1} \cup S_{X|Y=-1}) \} = \beta \quad (13)$$

$$\inf\{ \langle u^0 ; x \rangle \,|\, x \in H^{\geq}(\sigma^0) \cap (S_{X|Y=1} \cup S_{X|Y=-1}) \} = \beta \quad (14)$$

$$\sup\{ \langle u ; a \rangle \,|\, a \in S_A^* \} = \beta \quad (15)$$

$$\inf\{ \langle u ; b \rangle \,|\, b \in S_B^* \} = \beta. \quad (16)$$

Since $S_{X|Y=1}, S_{X|Y=-1}$ are compact, their intersection with $H^{\leq}(\sigma^0)$ or $H^{\geq}(\sigma^0)$ define compact sets. Hence, the supremum and infimum in (13)-(14) are attained. Moreover, since, by assumption, S_A^*, S_B^* are asymptotically conical, the same happens for the supremum and infimum of (15)-(16), if finite, must be attained. Indeed, by (5), one has

$$\begin{aligned}
\sup_{a \in S_A^*} \langle u ; a \rangle &= \sup_{a \in M_A + K_A} \langle u ; a \rangle \\
&= \sup_{m \in M_A} \langle u ; m \rangle + \sup_{k \in K_A} \langle u ; k \rangle \\
&= \begin{cases} \max_{m \in M_A} \langle u ; m \rangle, & \text{when } \langle u ; k \rangle \leq 0 \text{ for all } k \in K_A \\ +\infty, & \text{otherwise} \end{cases}
\end{aligned}$$

Similarly,

$$\inf_{b \in S_B^*} \langle u ; b \rangle = \begin{cases} \min_{m \in M_B} \langle u ; m \rangle, & \text{when } \langle u ; k \rangle \geq 0 \text{ for all } k \in K_B \\ -\infty, & \text{otherwise} \end{cases}$$

Moreover, the optimality of β for (12) implies that (u^0, β) is an optimal separating halfspace satisfying the conditions of the Theorem. □

Remark 3 Even when feasible, optimal solutions may not exist if Assumption 3 is dropped. Indeed, if we take X to be an arbitrary random vector with support $S_X = \mathbb{R}^d$, and Y degenerate to 1, then it is clear that, for any $u \in \mathcal{H}$, if $\{\beta_n\}$ is a sequence converging to $+\infty$, then $f(u, \beta_n)$ tends to 0, a value which is not attained at any β. Hence, no optimal separating hyperplane exists in this degenerate case. □

3.2 Balancing property

The search of an optimal separating halfspace, whose existence is shown in Theorem 2, can be further restricted to those halfspaces that balance the mass of misclassified points of populations C_A and C_B in the way specified below:

Definition 4 We say that $\sigma = (u, \beta) \in \mathcal{H}$ balances (X, Y) if

$$\frac{\tau}{\gamma_A^\circ(-u)}\mathbb{P}\left[X \in H^{\geq}(\sigma)|Y = 1\right] \geq \frac{1-\tau}{\gamma_B^\circ(u)}\mathbb{P}\left[X \in H^{<}(\sigma)|Y = -1\right] \quad (17)$$

$$\frac{\tau}{\gamma_A^\circ(-u)}\mathbb{P}\left[X \in H^{>}(\sigma)|Y = 1\right] \leq \frac{1-\tau}{\gamma_B^\circ(u)}\mathbb{P}\left[X \in H^{\leq}(\sigma)|Y = -1\right] \quad (18)$$

Theorem 5 *For a given $u \neq 0$, the set of β such that (u, β) is an optimal solution to (10) is a (possibly empty or degenerate) closed interval. Moreover, all $\sigma = (u, \beta)$ optimal to (10) satisfy*

1. *If $\sup_{a \in S_A^*} \langle\ u\ ;\ a\ \rangle < \beta < \inf_{b \in S_B^*} \langle\ u\ ;\ b\ \rangle$, then σ satisfies (17)-(18), i.e., σ balances X.*
2. *If $\beta = \inf_{b \in S_B^*} \langle\ u\ ;\ b\ \rangle$, then σ satisfies (17).*
3. *If $\beta = \sup_{a \in S_A^*} \langle\ u\ ;\ a\ \rangle$, then σ satisfies (18).*

Proof. Consider any fixed $u \neq 0$, and look for an optimal choice for a corresponding β. The objective for fixed u then looks like

$$f^u(\beta) = \frac{\tau}{\gamma_A^\circ(-u)} \int (\langle\ u\ ;\ a\ \rangle - \beta)^+ \, dF_A(a) + \frac{1-\tau}{\gamma_B^\circ(u)} \int (\beta - \langle\ u\ ;\ b\ \rangle)^+ \, dF_B(b).$$

Function f^u is convex since it is the sum of integrals of convex functions, finite everywhere, and having compact level sets. Hence, the set of optimal β is a closed interval.

Moreover, a necessary and sufficient condition for optimality at a given β^* is that 0 belongs to $\partial f^u(\beta^*) + N(\beta^*)$, where $\partial f^u(\beta^*)$ is the subdifferential of f^u at β^*, and $N(\beta^*)$ is the normal cone of $[\sup_{a \in S_A^*} \langle\ u\ ;\ a\ \rangle, \inf_{b \in S_B^*} \langle\ u\ ;\ b\ \rangle]$ at β^*.

The subdifferential is easily determined (see e.g. [9] p260-262) :

$$\partial f^u(\beta) = \left\{ \frac{\tau}{\gamma_A^\circ(-u)} \left(- \int_{H^{>}(\sigma)} dF_A(a) - \delta_A \mathbb{P}\left[X \in H^{=}(\sigma)|Y = 1\right] \right) \right.$$

$$
\begin{aligned}
&+ \frac{1-\tau}{\gamma_B^\circ(u)} \left(\int_{H^<(\sigma)} dF_B(b) + \delta_B \mathbb{P}\left[X \in H^=(\sigma) | Y=-1\right] \right) : \quad \delta_A, \delta_B \in [0,1] \Big\} \\
= \; & \Big\{ -\frac{\tau}{\gamma_A^\circ(-u)} \left(\mathbb{P}\left[X \in H^>(\sigma) | Y=1\right] + \delta_A \mathbb{P}\left[X \in H^=(\sigma) | Y=1\right] \right) \\
&+ \frac{1-\tau}{\gamma_B^\circ(u)} \left(\mathbb{P}\left[X \in H^<(\sigma) | Y=-1\right] + \delta_B \mathbb{P}\left[X \in H^=(\sigma) | Y=-1\right] \right) : \quad \delta_A, \delta_B \in [0,1] \Big\},
\end{aligned}
$$

where the minimum of this set is obtained for $\delta_A = 1, \delta_B = 0$, and the maximum for $\delta_A = 0, \delta_B = 1$. This implies that $\partial f^u(\beta)$ has the form $[l_1, l_2]$, with

$$
\begin{aligned}
l_1 &= -\tfrac{\tau}{\gamma_A^\circ(-u)} \mathbb{P}\left[X \in H^\geq(\sigma) | Y=1\right] + \tfrac{1-\tau}{\gamma_B^\circ(u)} \mathbb{P}\left[X \in H^<(\sigma) | Y=-1\right] \\
l_2 &= -\tfrac{\tau}{\gamma_A^\circ(-u)} \mathbb{P}\left[X \in H^>(\sigma) | Y=1\right] + \tfrac{1-\tau}{\gamma_B^\circ(u)} \mathbb{P}\left[X \in H^\leq(\sigma) | Y=-1\right].
\end{aligned} \tag{19}
$$

For part 1, $N(\beta) = \{0\}$, thus the optimality condition reads $0 \in \partial f^u(\beta) = [l_1, l_2]$, which holds iff σ balances X.

For part 2 (respectively part 3), $N(\beta) = [0, +\infty[$ (respectively $N(\beta) =]-\infty, 0]$), and the optimality condition is equivalent to (17) (respectively to (18)). □

3.3 Dimension one

The simplest case happens of course when the data points just consist of single values, i.e. $d = 1$. Note that all halfspaces are halflines, either of the form $\sigma = (1, \beta) =]-\infty, \beta]$ or $\sigma = (-1, \beta) = [\beta, +\infty[$. Theorem 5 yields:

Corollary 6 *In the one-dimensional case, if $(1, \beta^*)$ is optimal for (10) then $f(1, \beta^*) \leq f(-1, \beta^*)$ and one of the following conditions hold:*

1. $\max S_A^* < \beta^* < \min S_B^*$ *and*

$$
\begin{aligned}
\frac{\tau}{\gamma_A^\circ(-1)} \mathbb{P}\left[X \geq \beta^* | Y=1\right] &\geq \frac{1-\tau}{\gamma_B^\circ(1)} \mathbb{P}\left[X < \beta^* | Y=-1\right] \\
\frac{1-\tau}{\gamma_A^\circ(1)} \mathbb{P}\left[X \leq \beta^* | Y=-1\right] &\geq \frac{\tau}{\gamma_B^\circ(-1)} \mathbb{P}\left[X > \beta^* | Y=1\right]
\end{aligned}
$$

2. $\beta^* = \min S_A^*$ *and*

$$
\frac{\tau}{\gamma_A^\circ(-1)} \mathbb{P}\left[X \geq \beta^* | Y=1\right] \geq \frac{1-\tau}{\gamma_B^\circ(1)} \mathbb{P}\left[X < \beta^* | Y=-1\right]
$$

3. $\beta^* = \max S_B^*$ *and*

$$\frac{1-\tau}{\gamma_A^\circ(1)}\mathbb{P}\left[X \leq \beta^* | Y = -1\right] \geq \frac{\tau}{\gamma_B^\circ(-1)}\mathbb{P}\left[X > \beta^* | Y = 1\right]$$

□

In the proof of Theorem 5, the convex minimization problem in one variable

$$\begin{array}{ll} \min & f^u(\beta) \\ \text{s.t.} & S_A^* \subset H^{\leq}(u,\beta) \\ & S_B^* \subset H^{\geq}(u,\beta) \\ & \beta \in \mathbb{R} \end{array} \tag{20}$$

has been addressed for a fixed u. In fact, optimality conditions at β yield the balancing property of Theorem 5. In the particular case in which $S_{X|Y=1}$, $S_{X|Y=-1}$ are finite, it is shown below that (20) can be solved in time proportional to $|S_X| + T$, where T is the time needed to calculate $\sup_{a \in S_A^*} \langle\ u\ ;\ a\ \rangle$ and $\inf_{b \in S_B^*} \langle\ u\ ;\ b\ \rangle$. Indeed, one has

Theorem 7 *Let* $u \in \mathbb{R}^d$, $u \neq 0$. *Let* T *denote the time needed to determine* $\sup_{a \in S_A^*} \langle\ u\ ;\ a\ \rangle$ *and* $\inf_{b \in S_B^*} \langle\ u\ ;\ b\ \rangle$. *Then, (20) can be solved in* $O(|S_X| + T)$ *time.*

Proof. Function f^u is a convex piecewise linear function, having at most $|S_X|$ breakpoints, namely, those in the set $\{\langle\ u\ ;\ x\ \rangle : x \in S_X\}$.

As shown below, an adaptation of the algorithm of Balas-Zemel, [2] can be used to solve (20) in $O(|S_{X|Y=1}| + |S_{X|Y=-1}|)$ time in the unconstrained case in which $S_A^* = S_B^* = \emptyset$. Since f^u is convex, an optimal solution for the constrained problem is obtained by finding, in $O(T)$ time, the projection of the unconstrained optimal solution onto the feasible set $[\sup_{a \in S_A^*} \langle\ u\ ;\ a\ \rangle, \inf_{b \in S_B^*} \langle\ u\ ;\ b\ \rangle]$, yielding the result.

Hence we can assume that $S_A^* = S_B^* = \emptyset$, and we need to show that, in this case, $O(|S_X|)$ time suffices.

For each β and $\# \in \{\leq, <, =, \geq, >\}$, define

$$\begin{array}{rcl} \omega_A^{\#}(\beta) & \stackrel{\text{def}}{=} & \sum_{a \in S_{X|Y=1} : \langle\ u\ ;\ a\ \rangle \# \beta} \mathbb{P}\left[X = a | Y = 1\right] \\ \omega_B^{\#}(\beta) & \stackrel{\text{def}}{=} & \sum_{b \in S_{X|Y=-1} : \langle\ u\ ;\ b\ \rangle \# \beta} \mathbb{P}\left[X = b | Y = -1\right]. \end{array} \tag{21}$$

Observe that the coefficients in (21) at β^* can be calculated from those at β without evaluating *all* points in the sets $S_{X|Y=1}, S_{X|Y=-1}$. Indeed, for $\beta_1 \leq \beta_2$ one has

$$\begin{aligned}
\omega_A^{\geq}(\beta_2) &= \omega_A^{\geq}(\beta_1) - \sum_{a \in S_{X|Y=1}: \beta_1 \leq \langle\ u\ ;\ a\ \rangle < \beta_2} \mathbb{P}\left[X = a | Y = 1\right] \\
\omega_A^{>}(\beta_2) &= \omega_A^{>}(\beta_1) - \sum_{a \in S_{X|Y=1}: \beta_1 < \langle\ u\ ;\ a\ \rangle \leq \beta_2} \mathbb{P}\left[X = a | Y = 1\right] \\
\omega_B^{\leq}(\beta_2) &= \omega_B^{\leq}(\beta_1) + \sum_{b \in S_{X|Y=-1}: \beta_1 < \langle\ u\ ;\ b\ \rangle \leq \beta_2} \mathbb{P}\left[X = b | Y = -1\right] \\
\omega_B^{<}(\beta_2) &= \omega_B^{<}(\beta_1) + \sum_{b \in S_{X|Y=-1}: \beta_1 \leq \langle\ u\ ;\ b\ \rangle < \beta_2} \mathbb{P}\left[X = b | Y = -1\right].
\end{aligned}$$

In particular, the coefficients in β_2 (respect. β_1) are obtained from those in β_1 (respect. β_2) in time proportional to $|\{x \in S_X : \beta_1 \leq \langle\ u\ ;\ x\ \rangle \leq \beta_2\}|$.

From (19) we see that the left and right derivatives f_-^u and f_+^u of f^u at β are expressed in terms of the coefficients in (21) as

$$\begin{aligned}
f_-^u(\beta) &= \tfrac{-\tau}{\gamma_A^\circ(-u)}\omega_A^{\geq}(\beta) + \tfrac{1-\tau}{\gamma_B^\circ(u)}\omega_B^{<}(\beta) \\
f_+^u(\beta) &= \tfrac{-\tau}{\gamma_A^\circ(-u)}\omega_A^{>}(\beta) + \tfrac{1-\tau}{\gamma_B^\circ(u)}\omega_B^{\leq}(\beta).
\end{aligned} \tag{22}$$

Hence, a binary search can be implemented as follows: starting with the sets $S_A^0 \stackrel{\text{def}}{=} S_{X|Y=1}, S_B^0 \stackrel{\text{def}}{=} S_{X|Y=-1}$, at each stage k a median β^k of the set $\{\langle\ u\ ;\ x\ \rangle : x \in S_A^k \cup S_B^k\}$ is found (in time proportional to $|S_A^k| + |S_B^k|$); the directional derivatives $f_-^u(\beta^k), f_+^u(\beta^k)$ at β^k are obtained following (22), thus in time proportional to $|S_A^k| + |S_B^k|$. With this information, we check whether β^k is optimal (i.e., whether $f_-^u(\beta^k) \leq 0 \leq f_+^u(\beta^k)$). If it is the case we stop. Otherwise, if $f_-^u(\beta^k) > 0$, then the search can be reduced to those $\beta < \beta^k$, i.e., to the sets $S_A^{k+1} \stackrel{\text{def}}{=} \{a \in S_A^k : \langle\ u\ ;\ a\ \rangle < \beta^k\}$ and $S_B^{k+1} \stackrel{\text{def}}{=} \{b \in S_B^k : \langle\ u\ ;\ b\ \rangle < \beta^k\}$. Similarly, if $f_+^u(\beta^k) < 0$, then the search can be reduced to the sets $S_A^{k+1} \stackrel{\text{def}}{=} \{a \in S_A^k : \langle\ u\ ;\ a\ \rangle > \beta^k\}$ and $S_B^{k+1} \stackrel{\text{def}}{=} \{b \in S_B^k : \langle\ u\ ;\ b\ \rangle > \beta^k\}$. In any case, the cardinality of the candidate set (from which medians must be calculated) is halved at each stage. Hence, the total process can be done in time proportional to $|S_X|$, as asserted. □

This result immediately gives a linear-time algorithm for the one-dimensional case. Indeed, by using Theorem 7 to find β optimal for (20) with $u = 1$ and then $u = -1$, one obtains the following

Theorem 8 *In the one-dimensional case $d = 1$, with finite support, a separating hyperplane (point) is obtained in time $O(|S_X| + T,)$ where T denotes the time needed to calculate the maxima and minima of S_A^*, S_B^*.*

□

4 Localization properties

In this section we show that, when the support S_X of X and the sets S_A^*, S_B^* have a certain structure (which includes the case in which S_X is a finite union of possibly degenerate polytopes), an optimal hyperplane exists passing through many points, this number being related to the dimensionality of the convex hull of S_X, S_A^* and S_B^*.

We start by looking at the *single-norm case:* there exist a norm ν (i.e., ν is a gauge such that $\nu(x) = \nu(-x)$ for all x) and positive constants δ_A, δ_B such that

$$\gamma_A = \delta_A \nu, \quad \gamma_B = \delta_B \nu, \tag{23}$$

and generalize later the results to the case in which different norms or gauges are used as γ_A and γ_B.

4.1 Single-norm case

In the single-norm case (23), the cost structure defined in (2) admits the simpler form

$$c((x,y),\sigma) = \begin{cases} \delta_A d_\nu(x, H^<(\sigma)), & \text{if } y = 1 \\ \delta_B d_\nu(x, H^>(\sigma)), & \text{if } y = -1, \end{cases}$$

showing how constants δ_A, δ_B model the different importance of classifying in C_A an element of C_B and viceversa. Moreover, since

$$\gamma_A^\circ(-u) = \gamma_A^\circ(u) = \frac{1}{\delta_A}\nu^\circ(u), \qquad \gamma_B^\circ(u) = \frac{1}{\delta_B}\nu^\circ(u),$$

choosing $\gamma = \nu^\circ$, Problem (10) can be rephrased as

$$\begin{array}{ll} \min & \frac{\tau\delta_A}{\nu^\circ(u)} \int (\langle\ u\ ;\ a\ \rangle - \beta)^+ \, dF_A(a) + \frac{(1-\tau)\delta_B}{\nu^\circ(u)} \int (\beta - \langle\ u\ ;\ b\ \rangle)^+ \, dF_B(b), \\ \text{s.t.} & \nu^\circ(u) = 1 \\ & \sup\{\ \langle\ u\ ;\ a\ \rangle \,|\, a \in S_A^*\ \} \le \beta \le \inf\{\ \langle\ u\ ;\ b\ \rangle \,|\, b \in S_B^*\ \} \end{array}$$

or, equivalently,

$$\begin{array}{ll} \min & \tau\delta_A \mathbb{E}\left[(\langle\ u\ ;\ X\ \rangle - \beta)^+ \,|Y=1\right] + (1-\tau)\delta_B \mathbb{E}\left[(\beta - \langle\ u\ ;\ X\ \rangle)^+ \,|Y=-1\right] \\ \text{s.t.} & \nu^\circ(u)=1 \\ & \sup\{\ \langle\ u\ ;\ a\ \rangle \,|\, a \in S_A^*\ \} \leq \beta \leq \inf\{\ \langle\ u\ ;\ b\ \rangle \,|\, b \in S_B^*\ \} \end{array} \tag{24}$$

For the particular case in which X takes just a finite set of values, the objective function in (24) simplifies to

$$\begin{array}{l} \tau\delta_A \sum_{a\in S_{X|Y=1}} \mathbb{P}\left[X=a|Y=1\right](\langle\ u\ ;\ a\ \rangle - \beta)^+ + \\ +(1-\tau)\delta_B \sum_{b\in S_{X|Y=-1}} \mathbb{P}\left[X=b|Y=-1\right](\beta - \langle\ u\ ;\ b\ \rangle)^+, \end{array} \tag{25}$$

an expression equivalent to the following program introduced by Mangasarian, [11] and later addressed in [14]

$$\begin{array}{ll} \min & \tau\delta_A \sum_{a\in S_{X|Y=1}} \mathbb{P}\left[X=a|Y=1\right](\langle\ u\ ;\ a\ \rangle - \beta)^+ + \\ & +(1-\tau)\delta_B \sum_{b\in S_{X|Y=-1}} \mathbb{P}\left[X=b|Y=-1\right](\beta - \langle\ u\ ;\ b\ \rangle)^+ \\ \text{s.t.} & u \in \mathbb{R}^d,\ \beta \in \mathbb{R}. \end{array} \tag{26}$$

Theorem 2 shows that, for compact $S_{X|Y=1}, S_{X|Y=-1}$ an optimal halfspace exists with boundary intersecting $S_X \cup S_A^* \cup S_B^*$. As shown below, we can strengthen this result if we assume, roughly speaking, that the intersection of $S_{X|Y=1}$ and $S_{X|Y=-1}$ with any halfspace has a finite number of extreme points.

We formalize this by introducing the concept of *quasipolytopal* set.

Definition 9 A set $S \subset \mathbb{R}^d$ is said to be *quasipolytopal* if for any $\# \in \{\leq, =, \geq\}$ and $\sigma \in \mathcal{H}$, the set $conv(S \cap H^{\#}(\sigma))$ is a polytope.

Remark 10 For a convex set S to be quasipolytopal, it is necessary and sufficient to be a polytope. Indeed, if S is a polytope, then, by definition, S is quasipolytopal. Conversely, given a convex quasipolytopal set S, it follows that S is bounded, since else there would exist $\sigma \in \mathcal{H}$ such that $S \cap H^{\leq}(\sigma)$ would be an unbounded polyhedron, thus would not be a polytope. Hence, S is bounded, and we can take some $\sigma \in \mathcal{H}$ such that $S \subset H^{\leq}(\sigma)$. For such a σ we have by assumption that $S = conv(S \cap H^{\leq}(\sigma))$ is a polytope.

However, for S nonconvex, quasipolytopal sets which are not polytopes exist, as discussed in the following Remark. □

Remark 11 It follows from the definition that, if $S_1, \ldots, S_n$ are quasipolytopal sets, then $\bigcup_{i=1}^n S_i$ is also quasipolytopal. In particular, since a polytope is trivially quasipolytopal, the finite union of polytopes is also quasipolytopal. Taking polytopes reduced to singletons, we obtain that any finite set is quasipolytopal.

However, the intersection of two quasipolytopal sets may not be quasipolytopal: take e.g. in 2 dimensions, the quasipolytopal sets

$$\begin{aligned} S_1 &= \{(x_1,x_2)|x_1^2+x_2^2=1\} \cup \{(x_1,x_2)|\ \max\{|x_1|,|x_2|\}=2\} \\ S_2 &= \{(x_1,x_2)|x_1^2+x_2^2=1\} \cup \{(x_1,x_2)|\ \max\{|x_1|,|x_2|\}=3\}, \end{aligned}$$

whose intersection has as convex hull a disc, which is not polytopal. □

Before obtaining the announced property, a separation result is needed.

Lemma 12 *Let Γ_1, Γ_2 be non-empty polytopes in $\mathbb{R}^d$ with $relint(\Gamma_1) \cap relint(\Gamma_2) = \emptyset$. Then, there exists a hyperplane σ such that $\Gamma_1 \subset H^{\leq}(\sigma), \Gamma_2 \subset H^{\geq}(\sigma)$, and $H^{=}(\sigma)$ contains at least $\max\{dim(\Gamma_1), dim(\Gamma_2)\}$ affinely independent points of $ext(\Gamma_1) \cup ext(\Gamma_2)$.*

Proof. Without loss of generality, we can assume that $dim(\Gamma_1) \geq dim(\Gamma_2)$. Consider the polytope $\Gamma_1 - \Gamma_2$, and we will show the result separately for the cases in which $\Gamma_1 - \Gamma_2$ has not or has full dimension.

For $dim(\Gamma_1 - \Gamma_2) < d$, we can take $\sigma = (u, \beta) \in \mathcal{H}$, with $\beta \leq 0$, such that $\Gamma_1 - \Gamma_2 \subset H^{=}(\sigma)$. Hence,

$$\langle\ u\ ;\ x_1 - x_2\ \rangle = \beta \qquad \forall x_1 \in \Gamma_1, x_2 \in \Gamma_2.$$

Let $x_2^0 \in \Gamma_2$. Setting $\beta^* = \beta + \langle\ u\ ;\ x_2^0\ \rangle$, one has

$$\begin{aligned} \langle\ u\ ;\ x_1\ \rangle &= \beta^* \quad \forall x_1 \in \Gamma_1 \\ \langle\ u\ ;\ x_2\ \rangle &= \langle\ u\ ;\ x_2^0\ \rangle = \beta^* - \beta \geq \beta^* \quad \forall x_2 \in \Gamma_2. \end{aligned}$$

Hence,

$$\begin{aligned} \Gamma_1 &\subset H^{=}(u, \beta^*) \\ \Gamma_2 &\subset H^{\geq}(u, \beta^*), \end{aligned}$$

and $H^{=}(u, \beta^*)$ contains $1 + dim(\Gamma_1)$ affinely independent points of $ext(\Gamma_1) \subset ext(\Gamma_1) \cup ext(\Gamma_2)$, as asserted.

Now we consider the remaining case in which $dim(\Gamma_1 - \Gamma_2) = d$. By assumption,

$$0 \notin relint(\Gamma_1) - relint(\Gamma_2) = relint(\Gamma_1 - \Gamma_2),$$

see Corollary 6.6.2 of [15]. Hence, by Theorem 11.3 of [15], there must exist a hyperplane separating the origin and $\Gamma_1 - \Gamma_2$. In other words, there must exist $\overline{\sigma} = (\overline{u}, \overline{\beta}) \in \mathcal{H}$ such that

$$\begin{array}{rcl} 0 & \in & H^{\geq}(\overline{\sigma}) \\ \Gamma_1 - \Gamma_2 & \subset & H^{\leq}(\overline{\sigma}). \end{array}$$

This means

$$\overline{\beta} \leq 0 \tag{27}$$

$$\langle\ \overline{u}\ ;\ x_1\ \rangle \leq \langle\ \overline{u}\ ;\ x_2\ \rangle + \overline{\beta} \quad \forall x_1 \in \Gamma_1, x_2 \in \Gamma_2. \tag{28}$$

Set $\beta^* = \max_{x_1 \in \Gamma_1} \langle \overline{u}, x_1 \rangle - \overline{\beta}$. Then, (27)-(28) imply that

$$\langle \overline{u}, x_1 \rangle \leq \beta^* \quad \forall x_1 \in \Gamma_1 \tag{29}$$

$$\langle \overline{u}, x_2 \rangle \geq \beta^* \quad \forall x_2 \in \Gamma_2. \tag{30}$$

Define the set $\mathcal{T} \subset \mathbb{R}^d \times \mathbb{R}$

$$(u, \beta) \in \mathcal{T} \text{ iff } \left\{ \begin{array}{ll} \langle\ u\ ;\ x_1\ \rangle \leq \beta & \forall x_1 \in ext(\Gamma_1) \\ \langle\ u\ ;\ x_2\ \rangle \geq \beta & \forall x_2 \in ext(\Gamma_2) \end{array} \right.$$

Since Γ_1, Γ_2 are polytopes, $\mathcal{T}$ is polyhedral. Define also the affine function g,

$$g(u, \beta) = \sum_{x_1 \in ext(\Gamma_1)} (\beta - \langle\ u\ ;\ x_1\ \rangle) + \sum_{x_2 \in ext(\Gamma_2)} (\langle\ u\ ;\ x_2\ \rangle - \beta),$$

and the polyhedron P,

$$(u, \beta) \in P \text{ iff } \left\{ \begin{array}{lcl} (u, \beta) & \in & \mathcal{T} \\ g(u, \beta) & = & g(\overline{u}, \beta^*) \end{array} \right.$$

Clearly P is non-empty: by (29)-(30), $(\overline{u}, \beta^*) \in P$. Note also that $g(u, \beta) \geq 0$ for any $(u, \beta) \in P$, so $g(\overline{u}, \beta^*) \geq 0$. Now in case $g(\overline{u}, \beta^*) = 0$ we would have by (29)-(30) that

$$\begin{array}{rcll} \langle\ \overline{u}\ ;\ x_1\ \rangle & = & \beta^* & \forall x_1 \in ext(\Gamma_1) \\ \langle\ \overline{u}\ ;\ x_2\ \rangle & = & \beta^* & \forall x_2 \in ext(\Gamma_2), \end{array}$$

and thus $\Gamma_1 - \Gamma_2 \subset H^{=}(\overline{u}, 0)$, contradicting the assumption that $\Gamma_1 - \Gamma_2$ has full dimension. Hence,

$$g(\overline{u}, \beta^*) > 0. \tag{31}$$

Now if $(0, \beta) \in P$, we would have by definition of $\mathcal{T}$ that $0 \leq \beta \leq 0$, i.e., $\beta = 0$, or $(0, 0) \in P$, in contradiction with (31). It follows that

$$u \neq 0 \quad \forall (u, \beta) \in P. \tag{32}$$

Moreover, P is bounded. Indeed, if it were not the case, P would contain a ray. In other words, there would exist $(u, \beta) \neq (0, 0)$ such that

$$\begin{array}{l} \langle\ u\ ;\ x_1\ \rangle \leq \beta \quad \forall x_1 \in ext(\Gamma_1) \\ \langle\ u\ ;\ x_2\ \rangle \geq \beta \quad \forall x_2 \in ext(\Gamma_2) \\ \sum_{x_1 \in ext(\Gamma_1)} (\beta - \langle\ u\ ;\ x_1\ \rangle) + \sum_{x_2 \in ext(\Gamma_2)} (\langle\ u\ ;\ x_2\ \rangle - \beta) = 0, \end{array}$$

implying that

$$\begin{array}{rcll} \langle\ u\ ;\ x_1\ \rangle & = & \beta & \forall x_1 \in ext(\Gamma_1) \\ \langle\ u\ ;\ x_2\ \rangle & = & \beta & \forall x_2 \in ext(\Gamma_2), \end{array}$$

thus $\Gamma_1 - \Gamma_2 \subset H^{=}(u, 0)$, which contradicts the assumption that $\Gamma_1 - \Gamma_2$ has full dimension. Hence, P is bounded.

Let σ be an extreme point of P, which, by (32), defines a hyperplane. We then have by construction that

$$\begin{array}{rcl} \Gamma_1 & \subset & H^{\leq}(\sigma) \\ \Gamma_2 & \subset & H^{\geq}(\sigma), \end{array}$$

and, since at least d linearly independent inequalities defining P must be binding at σ, at least $d \geq \max\{dim(\Gamma_1), dim(\Gamma_2)\}$ affinely independent points of $ext(\Gamma_1) \cup ext(\Gamma_2)$ are contained in $H^{=}(\sigma)$, as asserted. □

Theorem 13 *In the single-norm case, when $S_{X|Y=1}, S_{X|Y=-1}, S_A^*, S_B^*$ are quasipolytopal sets, there exists an optimal σ such that $H^{=}(\sigma)$ passes through at least d^* affinely independent points of $ext(S_{X|Y=1}) \cup ext(S_{X|Y=-1}) \cup ext(S_A^*) \cup ext(S_B^*)$, where d^* is defined as*

$$d^* = \max\{dim(conv(S_{X|Y=1})), dim(conv(S_{X|Y=-1})), dim(S_A^*), dim(S_B^*)\}.$$

Proof. Let $\sigma^0 = (u^0, \beta^0)$ define an optimal solution to (24), existence of which has been guaranteed in Theorem 2. In particular,

$$S_A^* \subset H^{\leq}(\sigma^0),\ S_B^* \subset H^{\geq}(\sigma^0).$$

Let us consider separately the cases in which the optimal value $f(\sigma^0)$ is zero or it is strictly positive.

Case $f(\sigma^0) = 0$. This implies that

$$\begin{aligned} S_{X|Y=1} \cup S_A^* &\subset H^{\leq}(\sigma^0) \\ S_{X|Y=-1} \cup S_B^* &\subset H^{\geq}(\sigma^0). \end{aligned}$$

Since, by assumption, $S_{X|Y=1}, S_{X|Y=-1}, S_A^*, S_B^*$ are quasipolytopal sets, we have that

$$\begin{aligned} conv((S_{X|Y=1} \cup S_A^*) \cap H^{\leq}(\sigma^0)) &= conv(S_{X|Y=1} \cup S_A^*) \\ conv((S_{X|Y=-1} \cup S_B^*) \cap H^{\geq}(\sigma^0)) &= conv(S_{X|Y=-1} \cup S_B^*) \end{aligned}$$

and both are polytopes, separated by $H^{=}(\sigma^0)$.

In case $relint(conv(S_{X|Y=1} \cup S_A^*)) \cap relint(conv(S_{X|Y=-1} \cup S_B^*)) \neq \emptyset$, then

$$\begin{aligned} conv(S_{X|Y=1} \cup S_A^*) &\subset H^{=}(\sigma^0) \\ conv(S_{X|Y=-1} \cup S_B^*) &\subset H^{=}(\sigma^0). \end{aligned}$$

and so

$$ext(S_{X|Y=1}) \cup ext(S_{X|Y=-1}) \cup ext(S_A^*) \cup ext(S_B^*) \subset H^{=}(\sigma^0),$$

showing σ^0 satisfies the conditions.

On the other hand, in case $relint(conv(S_{X|Y=1} \cup S_A^*)) \cap relint(conv(S_{X|Y=-1} \cup S_B^*)) = \emptyset$, the hyperplane σ of Lemma 12 for $\Gamma_1 = S_{X|Y=1} \cup S_A^*$ and $\Gamma_2 = S_{X|Y=-1} \cup S_B^*$, satisfies the conditions.

Case $f(\sigma^0) > 0$. Define the set $\mathcal{T} \subset \mathbb{R}^d \times \mathbb{R}$ as

$$(u, \beta) \in \mathcal{T} \quad \text{iff} \quad \begin{cases} \langle\ u\ ;\ x\ \rangle \leq \beta & \forall x \in H^{\leq}(\sigma^0) \cap ext(S_{X|Y=1}) \\ \langle\ u\ ;\ x\ \rangle \leq \beta & \forall x \in H^{\leq}(\sigma^0) \cap ext(S_{X|Y=-1}) \\ \langle\ u\ ;\ x\ \rangle \leq \beta & \forall x \in ext(S_A^*) \\ \langle\ u\ ;\ x\ \rangle \geq \beta & \forall x \in H^{\geq}(\sigma^0) \cap ext(S_{X|Y=1}) \\ \langle\ u\ ;\ x\ \rangle \geq \beta & \forall x \in H^{\geq}(\sigma^0) \cap ext(S_{X|Y=-1}) \\ \langle\ u\ ;\ x\ \rangle \geq \beta & \forall x \in ext(S_B^*). \end{cases}$$

Since $(u^0, \beta^0) \in \mathcal{T}$, and by assumption, $S_{X|Y=1}$ and $S_{X|Y=-1}$ are quasipolytopal, $\mathcal{T}$ is a non-empty polyhedron. Moreover, the function $g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$,

$$\begin{aligned} g(u, \beta) &= \tau\delta_A \int_{a \in H^{>}(\sigma^0)} (\langle\ u\ ;\ a\ \rangle - \beta) dF_A(a) + (1-\tau)\delta_B \int_{b \in H^{<}(\sigma^0)} (\beta - \langle\ u\ ;\ b\ \rangle) dF_B(b) \\ &= \langle\ u\ ;\ \tau\delta_A \int_{a \in H^{>}(\sigma^0)} a\, dF_A(a) - (1-\tau)\delta_B \int_{b \in H^{<}(\sigma^0)} b\, dF_B(b)\ \rangle \\ &\quad + \beta(1-\tau)\delta_B \mathbb{P}\left[X \in H^{<}(\sigma^0) | Y = -1\right] - \beta\tau\delta_A \mathbb{P}\left[X \in H^{>}(\sigma^0) | Y = 1\right] \end{aligned}$$

is affine.

Consider the polyhedron P in $\mathbb{R}^d \times \mathbb{R}$ defined as

$$P = \{ (u,\beta) \in \mathcal{T} \mid g(u,\beta) = g(u^0,\beta^0) \},$$

which is non-empty since $(u^0,\beta^0) \in P$.

Moreover,

$$g(u^0,\beta^0) \neq 0, \tag{33}$$

since otherwise we would have by definition of g that

$$S_{X|Y=1} \subset H^{\leq}(\sigma^0), S_{X|Y=-1} \subset H^{\geq}(\sigma^0),$$

and thus $f(\sigma^0) = 0$, which is a contradiction. Hence, (33) holds, and thus,

$$g(u,\beta) \neq 0 \qquad \forall (u,\beta) \in P.$$

implying that

$$(0,0) \notin P. \tag{34}$$

Let us show now that P is bounded, and thus a polytope. Indeed, if unbounded, there would exist some $(u,\beta) \neq (0,0)$ such that $(\lambda u + u^0, \lambda\beta + \beta^0) \in P$, for all $\lambda \geq 0$. This would mean that we would have

$$\begin{aligned} \langle\ u\ ;\ a\ \rangle &\leq \beta \qquad \forall a \in (H^{\leq}(\sigma^0) \cap S_X) \cup S_A^* \\ \langle\ u\ ;\ b\ \rangle &\geq \beta \qquad \forall b \in (H^{\geq}(\sigma^0) \cap S_X) \cup S_B^* \\ g(u,\beta) &= 0 \end{aligned}$$

which, by the definition of g, implies that

$$S_{X|Y=1} \subset H^{\leq}(\sigma^0), S_{X|Y=-1} \subset H^{\geq}(\sigma^0),$$

which contradicts again the assumption that the optimal value is strictly positive.

Therefore P is compact, and hence a polytope. It contains the optimal solution (u^0,β^0), so any solution minimizing f on P is also an optimal solution for problem (24). Moreover,

$$\begin{aligned} \forall (u,\beta) \in P:\ f(u,\beta) &= \frac{g(u,\beta)}{\nu^{\circ}(u)} \\ &= \frac{g(u^0,\beta^0)}{\nu^{\circ}(u)} \end{aligned}$$

So minimizing f on P turns out to be equivalent to maximizing the function $(u, \beta) \mapsto \nu^\circ(u)$ on P. Since this function is convex, it attains its maximum on P at some extreme point $\sigma^1 = (u^1, \beta^1)$ of P. Moreover, by (34), $(u^1, \beta^1) \neq (0, 0)$.

Since σ^1 is obtained as the point common to d hyperplanes bounding linearly independent halfspaces defining P, and each such halfspace has as boundary a hyperplane passing through some point of either $ext(conv(S_{X|Y=1}))$, $ext(conv(S_{X|Y=-1}))$, $ext(S_A^*)$ or $ext(S_B^*)$, the hyperplane $H^=(\sigma^1)$ passes through d affinely independent points of the union of the sets above, as asserted. □

The result in the preceding Theorem is no longer true if the compactness of the supports, or the quasipolytopal structure of their convex hulls is removed, as shown in the following example:

Example 14 In $\mathbb{R}^d$, let $X|(Y=1), X|(Y=-1)$ have unbounded supports

$$\begin{aligned} S_{X|Y=1} &= \{(a_1, a_2, \ldots, a_d) : a_d \geq e^{a_1+\ldots+a_{d-1}}\} \\ S_{X|Y=-1} &= \{(b_1, b_2, \ldots, b_d) : b_d \leq -e^{b_1+\ldots+b_{d-1}}\} \end{aligned}$$

and let $S_A^* = S_B^* = \emptyset$. It is easy to check that the hyperplane $\sigma = (u_1, u_2, \ldots, u_d, \beta) = (0, \ldots, 0, 1, 0)$ is the unique hyperplane which separates $S_{X|Y=1}$ and $S_{X|Y=-1}$. Since $f(\sigma) = 0$, it is the (unique) optimal solution. However, it does not intersect $S_{X|Y=1} \cup S_{X|Y=-1}$.

Now consider in $\mathbb{R}^d$ random vectors whose compact (but not quasipolytopal) supports are the spheres centered at the points $(0, 0, \ldots, 0, 1)$ and $(0, 0, \ldots, 0, -1)$ and unit radius,

$$\begin{aligned} S_{X|Y=1} &= \{(a_1, a_2, \ldots, a_d) : a_1^2 + a_2^2 + \ldots + a_{d-1}^2 + (a_d - 1)^2 \leq 1\} \\ S_{X|Y=-1} &= \{(b_1, b_2, \ldots, b_d) : b_1^2 + b_2^2 + \ldots + b_{d-1}^2 + (b_d + 1)^2 \leq 1\} \end{aligned}$$

It is easy to check that the hyperplane $\sigma = (u_1, u_2, \ldots, u_d, \beta) = (0, 0, \ldots, 0, 1, 0)$ is the unique hyperplane which separates $S_{X|Y=1}$ and $S_{X|Y=-1}$. It has $f(\sigma) = 0$, and is the (unique) optimal solution. However, it intersects $S_{X|Y=1} \cup S_{X|Y=-1}$ at a unique single point, namely, the point of tangency of both spheres. □

4.2 Mixed norm or gauge distance

We extend now the previous localization result to the case in which γ_A, γ_B are (possibly different) gauges.

Theorem 15 *Let $S_{X|Y=1}$, $S_{X|Y=-1}$, S_A^*, and S_B^* be quasipolytopal sets. Then, there exists an optimal σ such that $H^=(\sigma)$ passes through at least d^*-1 affinely independent points of $ext(S_{X|Y=1}) \cup ext(S_{X|Y=-1}) \cup ext(S_A^*) \cup ext(S_B^*)$, where d^* is defined as*

$$d^* = \max\{dim(conv(S_{X|Y=1})), dim(conv(S_{X|Y=-1})), dim(S_A^*), dim(S_B^*)\}.$$

Proof.

Let $\sigma^0 = (u^0, \beta^0)$ be an optimal separating hyperplane.

The case $f(\sigma^0) = 0$ yields the stronger result that $H^=(\sigma)$ passes through at least d^* affinely independent points. The proof is identical to the first part of Theorem 13 and will not be repeated here. Hence, we assume in the sequel that $f(\sigma^0) > 0$. Define the functions f_A, f_B, f^0, g as

$$\begin{aligned}
f_A(u,\beta) &= \frac{\tau}{\gamma_A^\circ(-u)} \int_{H^\geq(\sigma^0)} (\langle\ u\ ;\ a\ \rangle - \beta)\, dF_A(a) \\
f_B(u,\beta) &= \frac{1-\tau}{\gamma_B^\circ(u)} \int_{H^\leq(\sigma^0)} (\beta - \langle\ u\ ;\ a\ \rangle)\, dF_B(b) \\
f^0(u,\beta) &= f_A(u,\beta) + f_B(u,\beta) \\
g(u,\beta) &= \frac{\tau}{\gamma_A^\circ(-u^0)} \int_{H^\geq(\sigma^0)} (\langle\ u\ ;\ a\ \rangle - \beta)\, dF_A(a) + \frac{1-\tau}{\gamma_B^\circ(u^0)} \int_{H^\leq(\sigma^0)} (\beta - \langle\ u\ ;\ a\ \rangle)\, dF_B(b).
\end{aligned}$$

Observe that

$$g(u^0,\beta^0) = f^0(u^0,\beta^0) = f(u^0,\beta^0),$$

while g is linear and f^0 is nonlinear.

Define also the set P as

$$(u,\beta) \in P \text{ iff } \begin{cases}
\langle\ u\ ;\ x\ \rangle \leq \beta & \forall x \in H^\leq(\sigma^0) \cap ext(S_{X|Y=1}) \\
\langle\ u\ ;\ x\ \rangle \leq \beta & \forall x \in H^\leq(\sigma^0) \cap ext(S_{X|Y=-1}) \\
\langle\ u\ ;\ x\ \rangle \leq \beta & \forall x \in ext(S_A^*) \\
\langle\ u\ ;\ x\ \rangle \geq \beta & \forall x \in H^\geq(\sigma^0) \cap ext(S_{X|Y=1}) \\
\langle\ u\ ;\ x\ \rangle \geq \beta & \forall x \in H^\geq(\sigma^0) \cap ext(S_{X|Y=-1}) \\
\langle\ u\ ;\ x\ \rangle \geq \beta & \forall x \in ext(S_B^*) \\
g(u,\beta) = f(u^0,\beta^0). &
\end{cases} \tag{35}$$

We have that

$$f^0(u,\beta) = f(u,\beta) \qquad \forall (u,\beta) \in P. \tag{36}$$

It is easy to see by similar arguments as used in Theorem 13 that P is a polytope, with $(0,0) \notin P$, $(u^0, \beta^0) \in P$.

Moreover, both f_A and f_B are quasiconcave on P, see e.g. [1]. Indeed, their upper level sets at level α are respectively given by inequalities of the form

$$\alpha\gamma_A^\circ(-u) - \int_{H^>(\sigma^0)} (\langle\ u\ ;\ a\ \rangle - \beta) dF_A(a) \leq 0$$

and

$$\alpha\gamma_B^\circ(u) - \int_{H^<(\sigma^0)} (-\langle\ u\ ;\ b\ \rangle + \beta) dF_B(b) \leq 0$$

which, by convexity of γ_A° and γ_B°, define convex sets.

Consider the biobjective quasiconcave minimization problem on the polytope P

$$\min_{(u,\beta)\in P} (f_A(u,\beta), f_B(u,\beta))\,. \tag{37}$$

Since (u^0, β^0) minimizes f on $\mathcal{H}$, (36) shows that (u^0, β^0) also minimizes $f^0 = f_A + f_B$ on P. Hence, (u^0, β^0) is an efficient solution for (37).

It was shown in [5] that in this case the set of *edges* (one-dimensional faces) of P constitutes a *dominator* of P, i.e. for any $(u,\beta) \in P$ there exists some (u', β') on some edge of P with $f_A(u', \beta') \leq f_A(u, \beta)$ and $f_B(u', \beta') \leq f_B(u, \beta)$. Since $(u^0, \beta^0) \in P$ there is some (u', β') belonging to an edge of P such that

$$\begin{aligned} f_A(u', \beta') &\leq f_A(u^0,\ \beta^0) \\ f_B(u', \beta') &\leq f_B(u^0,\ \beta^0), \end{aligned}$$

and thus, by (36),

$$f(u', \beta') \leq f(u^0, \beta^0).$$

Since (u^0, β^0) was assumed to minimize f on P, it follows that (u', β') also minimizes f on P and thus also on $\mathcal{H}$.

Any point in an edge of $P \subset \mathbb{R}^d \times \mathbb{R}$, satisfies at least d linearly independent equations among those defining P in (35). Since, in the definition (35) there is just one equality constraint, (u', β') corresponds to a halfspace with boundary hyperplane $H^=(u', \beta')$ passing through $d - 1 \geq d^* - 1$ affinely independent points of $ext(S_{X|Y=1}) \cup ext(S_{X|Y=-1}) \cup ext(S_A^*) \cup ext(S_B^*)$, as asserted. □

Particularizing Theorems 13 and 15 to Problem (26), we obtain the following results previously derived in [14].

Corollary 16 *Define* d^* *as*

$$d^* = \max\{dim(conv(S_{X|Y=1})), dim(conv(S_{X|Y=-1}))\}.$$

There exists σ*, optimal solution to (26), such that* $H^=(\sigma)$ *passes through at least* $d^* - 1$ *affinely independent points of* $ext(S_{X|Y=1}) \cup ext(S_{X|Y=-1})$*. Moreover, in the single-norm case, such* $H(\sigma)$ *passes through* d^* *affinely independent points.*

Acknowledgements. The research of the first author is supported in part by grants BFM2002-04525-C02-02 and BFM2002-11282-E, MCYT, Spain.

References

[1] Avriel, M., Diewert, W.E., Schaible, S., and Zhang, I., *Generalized Concavity*, Plenum Press, New York, New York, 1988.

[2] Balas, E., and Zemel, E. An algorithm for large zero-one knapsack problems, *Operations Research*, Vol. 28, pp. 1130–1154, 1980.

[3] Bock, H.H., Classification methodology, in *Handbook of data mining and knowledge discovery*, W. Klösgen and Żytkow (Eds.), Oxford University Press, 258–267, 2002.

[4] Carrizosa, E., and Fliege, J., Generalized Goal Programming: polynomial methods and applications, *Mathematical Programming*, Vol. 93, No. 2, pp. 281–303, 2002.

[5] Carrizosa, E., and Plastria, F., Dominators for Multiple-objective Quasiconvex Maximization Problems, *Journal of Global Optimization*, Vol. 18, No. 1, pp. 35–58, 2000.

[6] Devroye, L., Györfy, L. and Lugosy, G., *A probabilistic theory of Pattern Recognition*, Springer, New York, 1997.

[7] Durier, R., and Michelot, C., Geometrical Properties of the Fermat-Weber Problem, *European Journal of Operational Research*, Vol. 20, pp. 332–343, 1985.

[8] Hastie, T., Tibshirani, R., and Friedman, J., *The elements of Statistical Learning.* Springer, New York, 2001.

[9] Hiriart-Urruty, J.B, and Lemaréchal, C., *Convex analysis and optimisation algorithms*, Springer, New York, 1993.

[10] McLachlan, G.J., *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, 1992.

[11] Mangasarian, O.L., Arbitrary-Norm Separating Plane, *Operations Research Letters*, Vol. 24, pp. 15–23, 1999.

[12] Michelot, C., The Mathematics of Continuous Location, *Studies in Locational Analysis*, Vol. 5, pp. 59–83, 1993.

[13] Plastria, F. and Carrizosa, E., Gauge-distances and median hyperplanes, *Journal of Optimization Theory and Applications*, Vol. 110, pp.173-182, 2001.

[14] Plastria, F. and Carrizosa, E., Optimal distance separating halfspace. *Working Paper:* Report BEIF/124, Vrije Universiteit Brussel, 2002.

[15] Rockafellar, T., Convex Analysis, Princeton University Press, Princeton, New Jersey, 1970.

[16] Vapnik, V.N., *Statistical Learning Theory*. Wiley, Nueva York, 1998.