

THE GEOMETRY OF LINEAR SEPARABILITY IN DATA SETS

ADI BEN-ISRAEL AND YURI LEVIN

ABSTRACT. We study the geometry of datasets, using an extension of the Fisher linear discriminant to the case of singular covariance, and a new regularization procedure. A dataset is called **linearly separable** if its different clusters can be reliably separated by a linear hyperplane. We propose a measure of linear separability, easily computed as an angle that arises naturally in our analysis. This **angle of separability** assumes values between 0 and $\pi/2$, with high [resp. low] values corresponding to datasets that are linearly separable, resp. inseparable.

1. INTRODUCTION

1.1. **Background.** A variable y (*dependent variable, class membership, or output*) is assumed to depend in some fashion on p variables $\mathbf{x} = (x_1, \dots, x_p)$ (*independent variables, predictors, attributes, or inputs*). The variables \mathbf{x} and y take values in sets $X = X_1 \times X_2 \times \dots \times X_p$ and Y , respectively, where the X_i are real intervals or finite sets, and Y is a finite set, in particular $\{-1, 1\}$.

The relation between y and \mathbf{x} is known through an empirical *dataset*

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$$

consisting of N previously observed points $(\mathbf{x}, y) \in X \times Y$.

The problem is to determine a rule, say

$$y = f(\mathbf{x}), \tag{1}$$

for the value of y corresponding to an observed $\mathbf{x} \in X$.

This problem appears in many areas and contexts, including statistical estimation, regression, learning theory, and artificial intelligence. In typical applications, the values of \mathbf{x} can be observed or measured cheaply, but the exact determination of y is complicated and costly, hence the need to predict y given \mathbf{x} .

For example, in typical medical applications y takes two values (e.g. $\{-1, 1\}$), denoting respectively the absence or presence of disease. The values $\mathbf{x} = (x_1, \dots, x_p)$ come from diagnostic tests. The determination of y dictates the course of treatment, in particular, $y = 1$ may result in additional tests or even surgery. In general, the two possible errors:

type 1 (false positive): declaring $y = 1$ when it is $= -1$, and

type 2 (false negative): declaring $y = -1$ when it is $= 1$,

differ in their consequences, with type 2 more serious.

A good repository of machine learning datasets is available from the University of California–Irvine (UCI), see [6].

1.2. **Previous work.** In [1] we described a method for determining a metric rule f in (1), using a classification of a training set \mathcal{D} into clusters, and estimating y according to a cluster, nearest in some sense.

Date: April 29, 2005.

Key words and phrases. Classification, cluster analysis, Tikhonov regularization, linear discriminant, separability of datasets.

Presented at the Haifa Conference on Matrix Theory, January 3–5, 2005.

For example, in the binary case, if \mathcal{D} is partitioned into two clusters \mathcal{C}_{-1} and \mathcal{C}_1 , with means $(\bar{\mathbf{x}}_{-1}, \bar{y}_{-1})$ and $(\bar{\mathbf{x}}_1, \bar{y}_1)$, respectively, then the rule is

$$y = \begin{cases} -1, & \text{if } d(\mathbf{x}, \bar{\mathbf{x}}_{-1}) < d(\mathbf{x}, \bar{\mathbf{x}}_1); \\ 1, & \text{otherwise,} \end{cases} \quad (2)$$

where d is a metric on X .

This method, outlined in § 1.3, uses the nearest mean reclassification algorithm, see [2].

1.3. Classification using Metric Clustering of Data. We assume that a suitable *distance function*, $d(\cdot, \cdot)$, is defined on $X \times Y$. This distance can be constructed from distances d_X and d_Y , defined on X and Y , respectively¹, for example,

$$d((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)) = \sqrt{d_X^2(\mathbf{x}_1, \mathbf{x}_2) + \alpha d_Y^2(y_1, y_2)}, \quad (3)$$

where the parameter $\alpha \geq 0$ measures the relative importance of the y -component, see [1].

For $X \subset \mathbb{R}^p$ we can use the Euclidean distance,

$$d_X(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T (\mathbf{x}_1 - \mathbf{x}_2)}, \quad (4)$$

or the Mahalanobis distance, [3],

$$d_X(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T S^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}, \quad (5)$$

where S is a pooled covariance matrix, and for $Y \subset \mathbb{R}$ the distance,

$$d_Y(y_1, y_2) = |y_1 - y_2|. \quad (6)$$

The method of [1] for predicting y given $\mathbf{x} \in X$, uses a classification of the data \mathcal{D} into clusters $\{\mathcal{C}_1, \dots, \mathcal{C}_m\}$. The i^{th} -cluster \mathcal{C}_i has a centroid $\mathbf{c}_i = (\bar{\mathbf{x}}_i, \bar{y}_i)$ of its \mathbf{x} and y components, respectively. Each cluster \mathcal{C}_i is computed, recursively, as all points (\mathbf{x}, y) in \mathcal{D} that are closer to \mathbf{c}_i than to any other centroid \mathbf{c}_j , see [1] for details. The X -projection of a cluster \mathcal{C}_i is the set $\mathcal{C}_i^X \subset X$ consisting of all vectors \mathbf{x} with $(\mathbf{x}, y) \in \mathcal{C}_i$.

Having thus classified the data \mathcal{D} , any point (\mathbf{x}, y) with y unknown can be assigned to a cluster \mathcal{C}_i such that the X -centroid $\bar{\mathbf{x}}_i$ of its projected cluster \mathcal{C}_i^X is closest to \mathbf{x} . The corresponding Y -centroid, \bar{y}_i , is then used as prediction of y . If Y is a discrete set, the values of \bar{y}_i need discretization. In particular, if $Y = \{-1, 1\}$, a cut-off value p is used to infer

$$y = \begin{cases} 1, & \text{if } \bar{y}_i > p; \\ -1, & \text{if } \bar{y}_i \leq p. \end{cases} \quad (7)$$

In spite of its simplicity, the proposed algorithm, even in its most elementary form (e.g., using the Euclidean distance (4) which ignores statistical information), performed very well on some datasets in [6], notably *Breast Cancer* and *Hepatitis*, and performed credibly on others, see [1]. In fact, our method compared favorably to other, better-justified, methods.

A natural question is what makes certain datasets amenable to metric clustering, as in [1].

¹In the absence of linear structure on X , Y and $X \times Y$, the distance functions d_X , d_Y and d are not associated with norms.

1.4. **Current paper.** In the attempt to answer the above question, we study here the geometry of datasets, using an extension of the Fisher linear discriminant to the case of singular covariance, and a new regularization procedure, §§ 2–4. We call a dataset **linearly separable** if its different clusters can be reliably separated by a linear hyperplane, **linearly inseparable** otherwise.

Our definition is vague on purpose, and allows misclassification, i.e. some points falling on the wrong side of the hyperplane. In the literature, linear separability often means an absolute property, with no misclassification, but we consider it to be a relative property of a dataset.

We propose in § 5 a new measure for linear separability of datasets, easily computed as an angle that arises naturally in our analysis. This **angle of separability** assumes values between 0 and $\pi/2$, with high [resp. low] values corresponding to datasets that are linearly separable, resp. inseparable.

2. THE LINEAR DISCRIMINANT

Let random vectors $\mathbf{x} \in \mathbb{R}^p$ belong to one of two populations distributed with equal covariance matrix Σ . Samples are taken from these two populations, and the sample means $\bar{\mathbf{x}}_i$ and the (pooled) sample covariance matrix S are computed. The matrix S is assumed nonsingular.

The problem is to find $\mathbf{u} \in \mathbb{R}^p$ maximizing

$$\frac{(\mathbf{u}^T \bar{\mathbf{x}}_1 - \mathbf{u}^T \bar{\mathbf{x}}_2)^2}{\mathbf{u}^T S \mathbf{u}}. \quad (8)$$

2.1. **Rationale.** Let $y = \mathbf{u}^T \mathbf{x}$. Then

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\mathbf{u}^T \bar{\mathbf{x}}_1 - \mathbf{u}^T \bar{\mathbf{x}}_2)^2}{\mathbf{u}^T S \mathbf{u}},$$

showing (8) to be the ratio of variances between and within the y -values corresponding to the two populations.

2.2. **Solution.** We solve the problem in the form

$$\max \{(\mathbf{u}^T \mathbf{d})^2 : \mathbf{u}^T S \mathbf{u} = 1\}, \quad (\text{P})$$

where

$$\mathbf{d} := \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2.$$

The problem (P) has the optimal solution,

$$\mathbf{u} = \frac{1}{\sqrt{\mathbf{d}^T S^{-1} \mathbf{d}}} S^{-1} \mathbf{d} \quad (9)$$

and the optimal value

$$\max \frac{(\mathbf{u}^T \mathbf{d})^2}{\mathbf{u}^T S \mathbf{u}} = \mathbf{d}^T S^{-1} \mathbf{d}. \quad (10)$$

The vector \mathbf{u} in (9) is the normal to the hyperplane separating the two samples, called the **Fisher linear discriminant**. It is given by the hyperplane

$$\mathbf{d}^T S^{-1} \mathbf{x} = \alpha, \quad (11)$$

where

$$\alpha = \frac{1}{2} \mathbf{d}^T S^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2). \quad (12)$$

The linear discriminant is illustrated in Figure 1, where the samples are represented by ellipses.

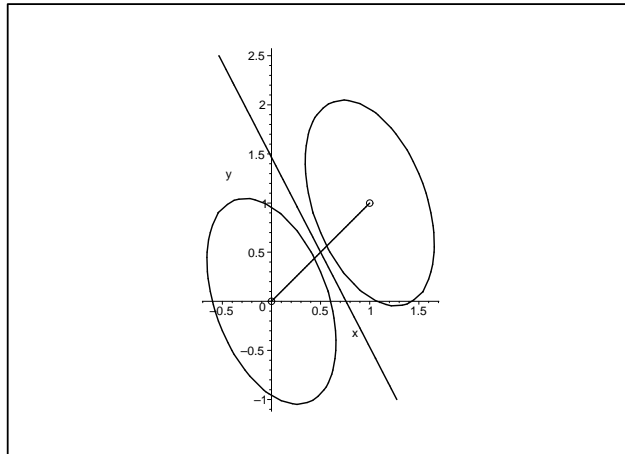


FIGURE 1. Illustration of the linear discriminant

2.3. Classification using Fisher's Discriminant. Let $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, \mathbf{d} , S be as above. Assign an observation \mathbf{x} to population 1 if

$$\mathbf{d}^T S^{-1} \mathbf{x} > \frac{1}{2} \mathbf{d}^T S^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2),$$

to population 2 otherwise, see [5, p. 320].

3. EXTENSION OF THE LINEAR DISCRIMINANT TO THE CASE OF POSSIBLY SINGULAR COVARIANCE

Given a matrix S we denote by $R(S)$, $N(S)$ and S^\dagger respectively the range, nullspace and Moore-Penrose inverse of S . By P_L we denote the orthogonal projector on a subspace L .

Given a vector $\mathbf{d} \in \mathbb{R}^p$, and a positive semi-definite matrix $S \in \mathbb{R}^{p \times p}$, consider the problem:

$$\max \{(\mathbf{d}^T \mathbf{u})^2 : \mathbf{u}^T S \mathbf{u} = 1\}. \quad (\text{P})$$

The Lagrangian of this problem is

$$L(\mathbf{u}, \lambda) = (\mathbf{d}^T \mathbf{u})^2 - \lambda(\mathbf{u}^T S \mathbf{u} - 1).$$

An optimal solution of problem (P) must satisfy

$$\frac{1}{2} \nabla L(\mathbf{u}, \lambda) = (\mathbf{d}^T \mathbf{u}) \mathbf{d} - \lambda S \mathbf{u} = \mathbf{0}.$$

Therefore,

$$S \mathbf{u} = \left(\frac{\mathbf{d}^T \mathbf{u}}{\lambda} \right) \mathbf{d}. \quad (13)$$

Consider equation (13) in two cases: when $\mathbf{d} \in R(S)$ and $\mathbf{d} \notin R(S)$.

Case 1: $\mathbf{d} \in R(S)$.

In this case

$$\mathbf{u} = \left(\frac{\mathbf{d}^T \mathbf{u}}{\lambda} \right) S^\dagger \mathbf{d} = \alpha S^\dagger \mathbf{d},$$

where $\alpha = \frac{\mathbf{d}^T \mathbf{u}}{\lambda}$,

$$\begin{aligned} \therefore \mathbf{u}^T S \mathbf{u} &= \alpha^2 \mathbf{d}^T S^\dagger S S^\dagger \mathbf{d} = \alpha^2 \mathbf{d}^T S^\dagger \mathbf{d} = 1, \\ \therefore \alpha^2 &= \frac{1}{\mathbf{d}^T S^\dagger \mathbf{d}}, \end{aligned}$$

and

$$\mathbf{u} = \frac{1}{\sqrt{\mathbf{d}^T S^\dagger \mathbf{d}}} S^\dagger \mathbf{d}, \quad (14)$$

analogously to (9). The optimal value of (P) is

$$(\mathbf{d}^T \mathbf{u})^2 = \mathbf{d}^T S^\dagger \mathbf{d}. \quad (15)$$

Case 2: $\mathbf{d} \notin R(S)$.

Let $\mathbf{z} = P_{N(S)} \mathbf{d}$, $\mathbf{z} \neq \mathbf{0}$, and let \mathbf{u}_0 satisfy the equation $\mathbf{u}_0^T S \mathbf{u}_0 = 1$. Also introduce

$$\mathbf{u}(t) := \mathbf{u}_0 + t\mathbf{z}. \quad (16)$$

Then, $\mathbf{u}(t)^T S \mathbf{u}(t) = 1$, for all t .

But

$$\begin{aligned} \mathbf{d}^T \mathbf{u}(t) &= \mathbf{d}^T \mathbf{u}_0 + t \mathbf{d}^T \mathbf{z} \\ &= \mathbf{d}^T \mathbf{u}_0 + t \mathbf{d}^T P_{N(S)} \mathbf{d} \\ &= \mathbf{d}^T \mathbf{u}_0 + t \|P_{N(S)} \mathbf{d}\|^2 \\ &= \mathbf{d}^T \mathbf{u}_0 + t \|\mathbf{z}\|^2, \\ \therefore |\mathbf{d}^T \mathbf{u}(t)|^2 &= O(t^2) \rightarrow \infty \text{ with } t. \end{aligned} \quad (17)$$

Thus, problem (P) has no optimal solution, as the values for this problem are unbounded.

Remark. Since the normal \mathbf{n} of a hyperplane is determined up to a sign ($-\mathbf{n}$ is also a normal), it follows that the problem (P) of § 2.2 can be written as

$$\min \{\mathbf{u}^T \mathbf{d} : \mathbf{u}^T S \mathbf{u} \leq 1\}, \quad (Q)$$

which is a convex programming problem². The conclusions of this section can then be shown to follow from the duality theorem of convex programming.

4. REGULARIZATION IN CASE $\mathbf{d} \notin R(S)$

We saw that problem (P) has no solution if $\mathbf{d} \notin R(S)$. In this case we can regularize (P), replacing it by a problem $(P(\kappa))$, where κ is the regularization parameter. The problem $(P(\kappa))$ has a nonsingular matrix $S(\kappa)$, and therefore an optimal solution $\mathbf{u}(\kappa)$ as in (9). The limit of these solutions, as $\kappa \rightarrow \infty$, is the least squares solution (14) of equation (13).

Recall the problem,

$$\max \{(\mathbf{d}^T \mathbf{u})^2 : \mathbf{u}^T S \mathbf{u} = 1\}. \quad (P)$$

Denote

$$Q = P_{N(S)} = I - S^\dagger S,$$

and define the regularized covariance matrix

$$S(\kappa) = S + \kappa Q. \quad (18)$$

²We thank Professor A. Ben-Tal for this observation.

Its inverse, for $\kappa \neq 0$, can be shown to equal

$$S(\kappa)^{-1} = S^\dagger + \frac{1}{\kappa} Q. \quad (19)$$

Consider the regularized problem

$$\max \{(\mathbf{d}^T \mathbf{u})^2 : \mathbf{u}^T S(\kappa) \mathbf{u} = 1\} \quad (\text{P}(\kappa))$$

with the optimal solution

$$\begin{aligned} \mathbf{u}(\kappa) &= \frac{1}{\sqrt{\mathbf{d}^T S(\kappa)^{-1} \mathbf{d}}} S(\kappa)^{-1} \mathbf{d} \\ &= \frac{1}{\sqrt{\mathbf{d}^T (S^\dagger + \frac{1}{\kappa} Q) \mathbf{d}}} \left(S^\dagger + \frac{1}{\kappa} Q \right) \mathbf{d} \end{aligned} \quad (20)$$

and the optimal value

$$\begin{aligned} (\mathbf{d}^T \mathbf{u}(\kappa))^2 &= \frac{A^2 + \frac{2AB}{\kappa} + \frac{B^2}{\kappa^2}}{A + \frac{B}{\kappa}} \\ &\text{where } A = (\mathbf{d}^T S^\dagger \mathbf{d}), \quad B = \|P_{N(S)} \mathbf{d}\|^2. \end{aligned}$$

In the limit, as $\kappa \rightarrow \infty$, we get,

$$\lim_{\kappa \rightarrow \infty} \mathbf{u}(\kappa) = \frac{1}{\sqrt{\mathbf{d}^T S^\dagger \mathbf{d}}} S^\dagger \mathbf{d}, \quad \text{as in (14),}$$

and,

$$\lim_{\kappa \rightarrow \infty} (\mathbf{d}^T \mathbf{u}(\kappa))^2 = \mathbf{d}^T S^\dagger \mathbf{d},$$

in agreement with (15).

The regularization (18) is a Tikhonov-type regularization, with the advantage that the inverse (19) is readily available.

5. A MEASURE OF LINEAR SEPARABILITY

5.1. Geometry. We denote the vectors of \mathbb{R}^{p+1} by (\mathbf{x}, z) , with $\mathbf{x} \in \mathbb{R}^p$, $z \in \mathbb{R}$. The **standard inner product** in \mathbb{R}^{p+1} is denoted by

$$(\boldsymbol{\xi}, \zeta) \cdot (\mathbf{x}, z) = \sum_{i=1}^p \xi_i x_i + \zeta z,$$

and the **Euclidean norm** of (\mathbf{x}, z) is

$$\|(\mathbf{x}, z)\| = \sqrt{(\mathbf{x}, z) \cdot (\mathbf{x}, z)}.$$

A hyperplane \mathcal{H} in \mathbb{R}^{p+1} is given by its **normal** $\mathbf{n} = (\boldsymbol{\xi}, \zeta)$ and z -**intercept** β as

$$\mathcal{H} = \{(\mathbf{x}, z) : (\boldsymbol{\xi}, \zeta) \cdot (\mathbf{x}, z) = \beta\}, \quad (21)$$

where the normal $\mathbf{n} = (\boldsymbol{\xi}, \zeta)$ is normalized, i.e., $\|\mathbf{n}\| = 1$.

The hyperplane \mathcal{H} is called **horizontal** if z is constant for all $(\mathbf{x}, z) \in \mathcal{H}$, i.e. if $\mathbf{n} = (\mathbf{0}, 1)$ is its normal. In particular, \mathbb{R}^p is identified with the horizontal hyperplane $z = 0$.

Given a hyperplane \mathcal{H} with a normal $\mathbf{n} = (\boldsymbol{\xi}, \zeta)$, $\|\mathbf{n}\| = 1$, the **angle of inclination** θ of \mathcal{H} is defined as the angle between \mathbf{n} and the vector $(\mathbf{0}, 1)$, i.e.,

$$\cos \theta = \frac{(\boldsymbol{\xi}, \zeta) \cdot (\mathbf{0}, 1)}{\|(\boldsymbol{\xi}, \zeta)\| \|(\mathbf{0}, 1)\|} = \zeta. \quad (22)$$

A horizontal hyperplane has $\theta = 0$. A **vertical** hyperplane is similarly defined by $\theta = \pi/2$.

5.2. A measure of separability. Consider a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\} \subset \mathbb{R}^{p+1}$ consisting of two clusters

$$\mathcal{D} = \mathcal{C}_{-1} \cup \mathcal{C}_1,$$

where

$$\mathcal{C}_{-1} = \{(\mathbf{x}_i, -1) \in \mathcal{D}\},$$

and

$$\mathcal{C}_1 = \{(\mathbf{x}_i, 1) \in \mathcal{D}\}.$$

We identify \mathbb{R}^p with the horizontal hyperplane $z = 0$ of

$$\mathbb{R}^{p+1} = \{(\mathbf{x}, z) : \mathbf{x} \in \mathbb{R}^p, z \in \mathbb{R}\},$$

and consider the clusters \mathcal{C}_{-1} and \mathcal{C}_1 to lie in the horizontal hyperplanes $z = -1$ and $z = 1$, respectively.

The two clusters \mathcal{C}_{-1} and \mathcal{C}_1 are linearly separable in \mathbb{R}^{p+1} , in particular, the horizontal hyperplane $z = 0$ separates \mathcal{C}_{-1} and \mathcal{C}_1 , as does any horizontal hyperplane with $-1 < z < 1$.

We denote the orthogonal projections of the clusters \mathcal{C}_{-1} and \mathcal{C}_1 on the hyperplane $z = 0$ by $\widehat{\mathcal{C}}_{-1}$ and $\widehat{\mathcal{C}}_1$, respectively. $\widehat{\mathcal{C}}_{-1}$ and $\widehat{\mathcal{C}}_1$ may be considered subsets of \mathbb{R}^p .

We assume that the clusters $\widehat{\mathcal{C}}_{-1}$ and $\widehat{\mathcal{C}}_1$ are samples from normal distributions on \mathbb{R}^p with the same nonsingular covariance $p \times p$ matrix $\widehat{\Sigma}$. Then the original clusters \mathcal{C}_{-1} and \mathcal{C}_1 in \mathbb{R}^{p+1} come from a singular covariance matrix

$$\Sigma = \begin{bmatrix} \widehat{\Sigma} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix}, \quad (23)$$

and the pooled covariance matrix of the dataset \mathcal{D} is therefore of the form

$$S = \begin{bmatrix} \widehat{S} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix}, \quad (24)$$

where \widehat{S} is nonsingular with probability 1.

Let the means of the clusters \mathcal{C}_{-1} and \mathcal{C}_1 be $(\bar{\mathbf{x}}_{-1}, -1)$ and $(\bar{\mathbf{x}}_1, 1)$, respectively. Their difference

$$\mathbf{d} = (\bar{\mathbf{x}}_1, 1) - (\bar{\mathbf{x}}_{-1}, -1) = (\widehat{\mathbf{d}}, 2), \quad (25)$$

where $\widehat{\mathbf{d}}$ is the difference of means of the clusters $\widehat{\mathcal{C}}_1$ and $\widehat{\mathcal{C}}_{-1}$,

$$\widehat{\mathbf{d}} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_{-1}. \quad (26)$$

By (24), the range of S lies in the hyperplane $z = 0$. Therefore, the difference \mathbf{d} in (25) does not lie in $R(S)$. We regularize S as in (18),

$$S(\kappa) = \begin{bmatrix} \widehat{S} & \mathbf{0} \\ \mathbf{0}^T & \kappa \end{bmatrix}. \quad (27)$$

By (19) and (9), the linear discriminant separating \mathcal{C}_{-1} and \mathcal{C}_1 is the hyperplane $\mathcal{H}(\kappa)$ in \mathbb{R}^{p+1} with normal

$$\mathbf{n} = \begin{bmatrix} \widehat{S}^{-1}\widehat{\mathbf{d}} \\ \frac{2}{\kappa} \end{bmatrix}. \quad (28)$$

The angle $\theta(\kappa)$ between the normal \mathbf{n} and the z -axis is defined by its cosine,

$$\cos \theta(\kappa) = \frac{\frac{2}{\kappa}}{\sqrt{\|\widehat{S}^{-1}\widehat{\mathbf{d}}\|^2 + \frac{4}{\kappa^2}}}. \quad (29)$$

We call $\theta(\kappa)$ the **angle of separability** between the clusters $\widehat{\mathcal{C}}_i$, $i = \pm 1$.

Equivalently,

$$\theta(\kappa) = \arccos \frac{\frac{2}{\kappa}}{\sqrt{\|\widehat{S}^{-1}\widehat{\mathbf{d}}\|^2 + \frac{4}{\kappa^2}}} = \arctan \frac{\kappa \|\widehat{S}^{-1}\widehat{\mathbf{d}}\|}{2}. \quad (30)$$

As $\kappa \rightarrow \infty$, the hyperplane $\mathcal{H}(\kappa)$ tends to be vertical, and its intersection with \mathbb{R}^p (i.e., the horizontal hyperplane $z = 0$) tends to the Fisher linear discriminant of $\widehat{\mathcal{C}}_{-1}$ and $\widehat{\mathcal{C}}_1$ in \mathbb{R}^p . It is however more interesting to observe $\theta(\kappa)$ for small fixed values of κ , say $\kappa = 1$, in which case we write θ for $\theta(1)$. If the clusters $\widehat{\mathcal{C}}_{-1}$ and $\widehat{\mathcal{C}}_1$ are not well separated, i.e. if $\|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_{-1}\|$ is small, then θ is small, i.e., $\mathcal{H}(\kappa)$ is nearly horizontal. On the other extreme, if the clusters are well separated so that $\|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_{-1}\|$ is large, then θ is large, and the hyperplane $\mathcal{H}(\kappa)$ is nearly vertical.

We propose θ as a measure of the linear separability of the dataset in question.

6. DISCUSSION

The angle of separability θ is a good measure of the linear separability of the dataset $\widehat{\mathcal{D}} = \widehat{\mathcal{C}}_{-1} \cup \widehat{\mathcal{C}}_1$. We illustrate this in §§ 6.1–6.2.

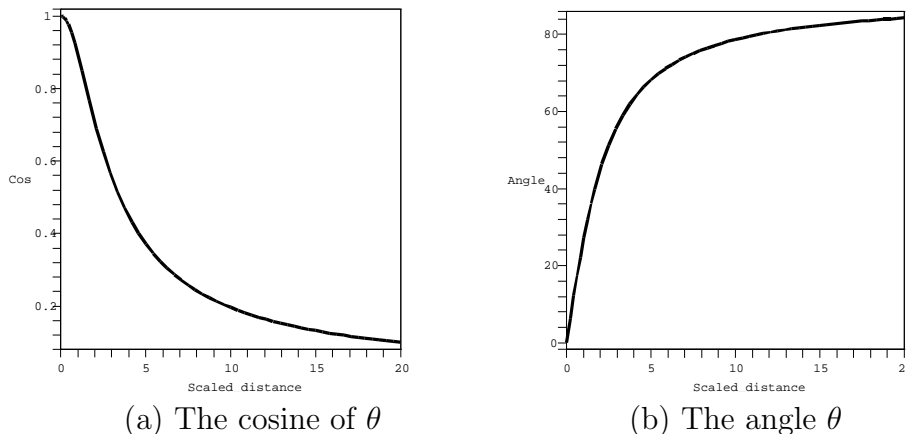
6.1. A numerical experiment. We study experimentally the dependence of θ on the scaled distance $\|\widehat{S}^{-1}\widehat{\mathbf{d}}\|$. For simplicity of notation we drop all the $\widehat{\cdot}$ symbols, writing $S, \mathbf{d}, \mathcal{C}_{-1}$ instead of $\widehat{S}, \widehat{\mathbf{d}}, \widehat{\mathcal{C}}_{-1}$, etc.

Let $N_2(\Sigma, \boldsymbol{\mu})$ denote the normal distribution on \mathbb{R}^2 with covariance Σ and mean $\boldsymbol{\mu}$, and let

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

We simulate two random samples (or clusters), \mathcal{C}_{-1} from $N_2(\Sigma, \mathbf{0})$ and \mathcal{C}_1 from $N_2(\Sigma, \boldsymbol{\mu})$, each with 100 observations, calculate $\cos \theta$ and θ using (29) and (30), and repeat for different $\boldsymbol{\mu} = (\mu_x, 0)$, $\mu_x = 0, 0.5, \dots, 50$. As μ_x increases the two clusters become more separable. This is illustrated in Figure 2 which gives the (experimental values of the) angle of separability θ and its cosine as functions of the scaled distance $\|S^{-1}\mathbf{d}\|$. In particular, the angle of separability is 0 for $\mu_x = \mathbf{0}$ (i.e. for $\|S^{-1}\mathbf{d}\| = 0$), and it increases asymptotically to $\pi/2$ as $\|S^{-1}\mathbf{d}\|$ increases.

6.2. Some real datasets. We compute the angle of separability θ and its cosine for five of the datasets in [6]. The results are tabulated in Table 1. The last two columns give the best (Max) and the worst (Min) performances, in percentages of correct predictions, from among the 33 algorithms (22 decision tree, 9 statistical and 2 neural network algorithms) compared in [4]. The procedure was ten-fold cross validation, with 90% of the dataset in the training set, and 10% used for testing. There was no over-all champion; the winning algorithm in one dataset, may be an also-ran in another dataset.

FIGURE 2. The angle of separability θ as a function of the scaled distance $\|S^{-1}\mathbf{d}\|$

Name of Dataset in [6]	Angle of Separability		Results of [4]	
	$\cos \theta$	θ	max %	min %
Breast Cancer	0.74	43°	97	91
Liver	0.99	4°	72	57
Diabetes	0.99	3°	78	69
Voting	0.18	80°	96	94
Hepatitis	0.42	65°	83	N/A

TABLE 1. The angle of separability and the maximum and minimum percentage of correctly predicted observations among 33 prediction methods in [4] for five datasets in [6]

We see that for datasets with a larger angles of separability (Breast Cancer and Voting), all methods in [4] gave good predictions, while for datasets with smaller separability angles (Liver and Diabetes), all methods performed poorly.

6.3. Statistical questions. The Fisher linear discriminant requires that the two populations have equal covariances. However, in medical datasets there is no reason to expect this (the healthy and the sick populations may be too dissimilar.)

In the case of equal covariances the distribution of the angle of separability θ is available, allowing for a simple tests of hypotheses, e.g., testing the hypothesis $\theta = 0$ vs. the alternative $\theta \neq 0$. This will be reported in a sequel study.

Our experiments show that the angle θ is a good measure of separability even for clusters with unequal covariances, although the pooled covariance (24) cannot be justified in this case.

REFERENCES

- [1] A. Ben-Israel and Y. Levin, An Estimation Algorithm using Distance Clustering of Data, *OPSEARCH* **38**(2001), 443–455
- [2] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd edition, *Academic Press Inc.*, Boston, MA, 1990
- [3] R. Gnanadesikan, J. Harvey and J. Kettenring, Mahalanobis metrics for cluster analysis, *The Indian Journal of Statistics. Series A* **55**(1993), 494–505
- [4] T. Lim, W. Loh, and Y. Shih, A comparison of prediction accuracy, complexity, and training time of thirty three old and new classification algorithms, *Machine Learning* **40**, 203–228

- [5] K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate Analysis*, Academic Press, 1979
- [6] C. Merz and P. Murphy, UCI Repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA, 1996. (<http://www.ics.uci.edu/mllearn/MLRepository.html>)

ADI BEN-ISRAEL, RUTCOR–RUTGERS CENTER FOR OPERATIONS RESEARCH, RUTGERS UNIVERSITY, 640 BARTHOLOMEW RD., PISCATAWAY, NJ 08854-8003, USA

E-mail address: `benisrael@rbsmail.rutgers.edu`

YURI LEVIN, SCHOOL OF BUSINESS, QUEEN'S UNIVERSITY, 143 UNION STR., KINGSTON, ON, K7L 3N6, CANADA

E-mail address: `ylevin@business.queensu.ca`