On Optimal Pairwise Linear Classifiers for Normal Distributions: The Two-Dimensional Case[*]

Luis Rueda[†] and **B. John Oommen**[‡]

Revised September 17, 2001

Abstract

Computing linear classifiers is a very important problem in statistical Pattern Recognition (PR). These classifiers have been investigated by the PR community extensively since they are the ones which are both easy to implement and comprehend. It is well known that when dealing with normally distributed classes, the *optimal* discriminant function for two-classes is linear only when the covariance matrices are equal. Other approaches, such as the Fisher's discriminant, the perceptron algorithm, minimum square distance classifiers, etc., have solved this problem by generating a linear classifier in normal and non-normal distributions, but these classifiers are typically *suboptimal*.

In this paper we shall focus on some special cases of the normal distribution with non-equal covariance matrices. We present a complete analysis of the case when the classifier is pairwise linear, and to our knowledge this is a pioneering work for the use of such classifiers in any are of statistical PR. We shall determine the conditions that the mean vectors and covariance matrices have to satisfy in order to obtain the *optimal* linear classifier. However, as opposed to the state of the art, in all the cases discussed here, the linear classifier is given by a *pair* of straight lines, which is a particular case of the general equation of second degree. One of these cases is when we have two overlapping classes with equal means, which is a general case of the Minsky's Paradox for the Perceptron. We present a general linear classifier for this particular case which can be obtained directly from the parameters of the distribution. Numerous other analytic results for two dimensional normal random vectors have been derived. Finally, we have also provided some empirical results in all the cases, and demonstrated that these linear classifiers achieve very good performance.

1 Introduction

The aim of statistical Pattern Recognition (PR) is to find a discriminant function which can be used to classify an object, represented by its features, which belongs to a certain class. In most cases, this function is linear or quadratic. When the classes are normally distributed, it is not always possible to find the optimal linear

[*] To appear in *IEEE Pattern Analysis and Machine Intelligence.*

[†] School of Computer Science, Carleton University, 1125 Colonel By Dr., Ottawa, ON, K1S 5B6, Canada. E-mail: lrueda@scs.carleton.ca. Partially supported by Departamento de Informática, Universidad Nacional de San Juan, Argentina.

[‡] Senior Member, IEEE. School of Computer Science, Carleton University, 1125 Colonel By Dr., Ottawa, ON, K1S 5B6, Canada. E-mail: oommen@scs.carleton.ca. Partially supported by NSERC, the National Science and Engineering Research Council of Canada.

classifier. In all the known results in this field, determining a linear function to achieve Bayes classification for normally distributed class-conditional distributions, has only been reported when the covariance matrices are equal [1], [2].

As opposed to optimal linear classifiers, many attempts have been made to yield linear classifiers, using Fisher's approach [3], [4], [5], the perceptron algorithm (the basis of the back propagation Neural Network learning algorithms) [6], [7], [8], [9], Piecewise Recognition Models [10], Random Search Optimization [11], and Removal Classification Structures [12]. All of these approaches suffer from the lack of optimality, and thus although they find linear discriminant functions, the classifier is not optimal.

In this paper, we show that there are other cases for normal distributions and non-equal covariance matrices in which the discriminant function is linear and the classifier is *optimal.* One of these cases is when we have two overlapping classes with equal means, as depicted in Figure 1. But as opposed to all the previously studied linear classifiers, the new techniques introduced here yield *pairwise* linear classifiers, which emerge as degenerate cases of the general quadratic classifier.

Minsky showed that it is not possible to find a single linear classifier for the simple case of Figure 1 in which the features of one class are the Exclusive-OR of a 2-bit binary vector and the features of the second class are the negated features. This paradox, also called the Minsky's Paradox [13], demonstrated that a single perceptron could not correctly classify in this simple scenario.

As opposed to this, we show that it is possible to find two optimal linear discriminant functions, given as a pair of straight lines, which is a particular case of the quadratic discriminant function. These classifiers have some advantages over the traditional linear discriminant approaches, such as Fisher's, perceptron learning, and other ones, because the classifier that we obtain is both linear and optimal. Finally, we conclude this introductory section by observing that, to the best of our knowledge, the results of this paper are pioneering. We are unaware of any work that has been done in statistical PR, which investigates the design and use of optimal pairwise linear classifiers.

2 Pattern Classification

2.1 Bayes Decision Theory

The main goal of PR is to find the class that an object belongs to given its features[1]. In statistical models, the features are represented as random vectors in the domain of the real numbers. A *random vector* is an ordered tuple $X = [x_1, \ldots, x_d]^T$that is characterized by a probability distribution function, where d represents the dimension of the problem[2]. In particular, the probability distribution function for a random vector X which is normally distributed is

$$p(X) = \frac{1}{2\pi^{\frac{d}{2}}|\Sigma|^{\frac{d}{2}}} e^{-\frac{1}{2}(X-M)^T\Sigma^{-1}(X-M)}, \tag{1}$$

where M is the mean vector and Σ is the covariance matrix [14].

[1] This introductory section is included primarily to lay the foundation for the arguments presented later. It also demonstrates why quadratic classifiers can be treated as a starting point in our analysis.

[2] In this report, we consider only two dimensional normal random vectors, $X = [x_1, x_2]^T$. The multi-dimensional case is currently being prepared.

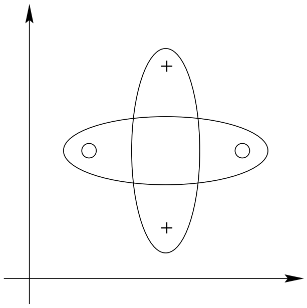

Figure 1: Two overlapping classes, ω_1 and ω_2, with equal means.

In general, the PR recognition problem deals with c classes, $\omega_1, \ldots, \omega_c$, with the *a priori* probability of ω_i, $P(\omega_i)$. Consider the problem of classifying an object whose features are given by the random vector X. The aim of Bayesian classification is to decide and choose the class that maximizes the *a posteriori* probability [15], [16], given by:

$$p(\omega_i|X) = \frac{p(X|\omega_i)P(\omega_i)}{p(X)} \tag{2}$$

Suppose we have two classes, ω_1 and ω_2, with *a priori* probability, $P(\omega_1)$ and $P(\omega_2)$. From Equation (2), we can write the general inequality specifying the Bayesian classification between two classes as follows:

$$p(X|\omega_1)P(\omega_1) \underset{\omega_1}{\overset{\omega_2}{\lessgtr}} p(X|\omega_2)P(\omega_2) \tag{3}$$

Equality in (3) represents the discriminant function. Assuming that ω_1 and ω_2 are represented by normal random vectors with covariance matrices Σ_1, Σ_2, and mean vectors M_1, M_2, respectively, the discriminant function is given by:

$$\left[\frac{1}{\sqrt{2\pi}|\Sigma_1|^{1/2}} e^{-\frac{1}{2}[(X-M_1)^T\Sigma_1^{-1}(X-M_1)]}\right] P(\omega_1) = \left[\frac{1}{\sqrt{2\pi}|\Sigma_2|^{1/2}} e^{-\frac{1}{2}[(X-M_2)^T\Sigma_2^{-1}(X-M_2)]}\right] P(\omega_2) \tag{4}$$

Without loss of generality, we assume that ω_1 and ω_2 have the same *a priori* probability, 0.5. Taking the logarithm of both sides of (4), we have:

$$\log\frac{|\Sigma_2|}{|\Sigma_1|} - [(X-M_1)^T\Sigma_1^{-1}(X-M_1)] + [(X-M_2)^T\Sigma_2^{-1}(X-M_2)] = 0 \tag{5}$$

Consider the two cases in (5). When $\Sigma_1 = \Sigma_2$ the discriminant function is linear [17]. For the case when Σ_1 and Σ_2 are arbitrary, the classifier results in a general equation of second degree in the form of a hyperparaboloid, hyperellipsoid, hypersphere, hyperboloid, or a pair of hyperplanes. Indeed, in our discussion, we are interested in the case when the classifier is a pair of straight lines for $d = 2$.

2.2 Diagonalization

Diagonalization is the process of transforming a space by performing linear and whitening transformations [18]. As our linear classifier depend primarily on a preprocessing involving whitening, we present below a

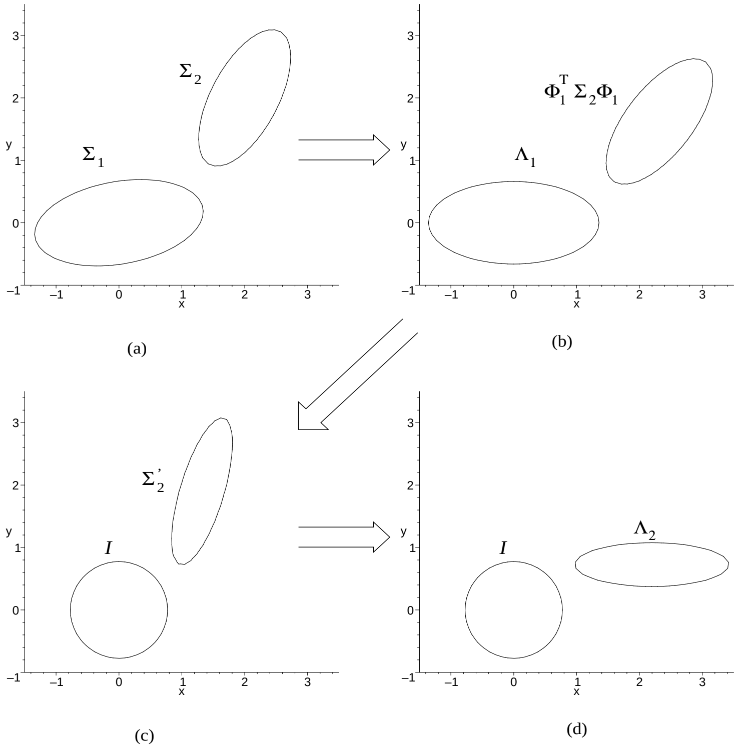

Figure 2: Simultaneous diagonalization: orthonormal and whitening transformations.

brief summary of this strategy.

Suppose we have a normal random vector $X \sim N(M_X, \Sigma_X)$. The linear transformation consists of transforming the random vector X into another vector Y as follows: $Y = \Phi^T X$, where Φ is a dxd matrix composed of the d eigenvectors of Σ: $[\phi_1|\ldots|\phi_d]$. The distribution of Y in the transformed space results in $Y \sim N(\Phi^T M_X, \Phi^T\Sigma_X\Phi)$.

The whitening transformation is a non-orthonormal transformation consisting of $Z = \Lambda^{-\frac{1}{2}}Y$, where Λ is a diagonal matrix whose elements are the eigenvalues of Σ: $\lambda_1, \ldots \lambda_d$. The distribution of Z in the transformed space results in $Z \sim N(\Lambda^{-\frac{1}{2}}\Phi^T M_X, \Lambda^{-\frac{1}{2}}\Phi^T\Sigma_X\Phi\Lambda^{-\frac{1}{2}}) = N(\Lambda^{-\frac{1}{2}}\Phi^T M_X, I)$, since

$$\Lambda^{-\frac{1}{2}}\Phi^T\Sigma_X\Phi\Lambda^{-\frac{1}{2}} = \Lambda^{-\frac{1}{2}}\Lambda\Lambda^{-\frac{1}{2}} = I.$$

Suppose we have two normal random vectors, $X_1 \sim N(M_1, \Sigma_1)$ and $X_2 \sim N(M_2, \Sigma_2)$. The transformation of these two random vectors is called *simultaneous diagonalization.*

The first transformation consists of $Z_i = \Lambda_1^{-\frac{1}{2}}\Phi_1^T(X_i - M_1)$, $i = 1, 2$, where $\Lambda^{-\frac{1}{2}}$ is the eigenvalue matrix of Σ_1, and Φ_1 is the eigenvector matrix of Σ_1. We subtract the mean vector of X_1 so that the origin of our system is M_1. After this transformation, the resulting random vectors are $Z_1 \sim N(\underline{0}, I)$ and $Z_2 \sim N(\Lambda_1^{-\frac{1}{2}}\Phi_1^T[M_2 - M_1], \Lambda_1^{-\frac{1}{2}}\Phi_1^T\Sigma_2\Phi_1\Lambda_1^{-\frac{1}{2}})$.

In the second transformation, we set the axes of the coordinate system to be in the direction of the eigenvectors of $\Lambda_1^{-\frac{1}{2}}\Phi_1^T\Sigma_2\Phi_1\Lambda_1^{-\frac{1}{2}} = \Sigma_2'$. The process consists of the transformation $V_i = \Phi_2^T Z_i$, $i = 1, 2$, where Φ_2 is the eigenvector matrix of Σ_2'. After the transformation, the resulting random vectors are $V_1 \sim N(\underline{0}, I)$ and $V_2 \sim N(\Phi_2^T\Lambda_1^{-\frac{1}{2}}\Phi_1^T[M_2 - M_1], \Phi_2^T\Sigma_2'\Phi_2)$. After this transformation, the covariance matrix of V_2 is Λ_2, a diagonal matrix whose diagonal elements are the eigenvalues of Σ'_2.

One of the implications of the simultaneous diagonalization is that Z_1 is in the origin (i.e. its mean vector is $\underline{0}$) and its covariance matrix is the identity. This means that any orthonormal transformation, such as $V_i = \Phi_2^T Z_i$, does not change the distribution of Z_1. In addition, both covariance matrices are diagonal, their eigenvectors are parallel to the axes, and the random variables, composing the random vectors, are uncorrelated.

The simultaneous diagonalization procedure for $d = 2$ is shown in Figure 2. The original distributions are depicted in Figure 2(a). Figure 2(b) represents the orthonormal transformation $Y_i = \Phi_1^T(X_i - M_1)$ with the axes of the system in the direction of the eigenvectors of Σ_1. The whitening transformation $Z_i = \Lambda_1^{-\frac{1}{2}}$ is depicted in Figure 2(c); the new value of Σ_1 is the identity matrix drawn as a circle. The orthonormal transformation $V_i = \Phi_2^T Z_i$ is depicted in Figure 2(d). The axes of the system are in the direction of the eigenvectors of Σ'_2.

3 Linear Discriminants in Diagonalization

As discussed in Section 2.2, given any arbitrary normal random vectors, X_1 and X_2, whose covariance matrices are Σ_1 and Σ_2, we can perform simultaneous diagonalization to obtain normal random vectors, V_1 and V_2, whose covariance matrices are diagonal, namely I and Λ, respectively. In what follows, we assume that the dimension of our problem is $d = 2$. Also, to be more specific, we assume that after simultaneous diagonalization, the mean vectors and covariance matrices have the form:

$$M_1 = \begin{bmatrix} p \\ q \end{bmatrix}, M_2 = \begin{bmatrix} r \\ s \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ and } \Sigma_2 = \begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix}. \tag{6}$$

Since we will be, for the present, consistently dealing with two-dimensional vectors we shall assume that the feature vector has the form $X = \begin{bmatrix} x \\ y \end{bmatrix}$.

For our discussion, we let $X \sim N(M, \Sigma)$ denote a normal random vector, X, with covariance matrix Σ and mean vector M. We will now present a linear transformation that will later prove useful in simplifying complex expressions. The transformation is stated more formally in Theorem 1.

Theorem 1. *Let $X_1 \sim N(M_{1_X}, \Sigma_1)$ and $X_2 \sim N(M_{2_X}, \Sigma_2)$ be two normal random vectors with parameters as in (6). Vectors X_1 and X_2 can be transformed into $Z_1 \sim N(M_{1_Z}, \Sigma_1)$ and $Z_2 \sim N(M_{2_Z}, \Sigma_2)$, where*

$$M_{1_Z} = \begin{bmatrix} t \\ u \end{bmatrix}, \text{and} M_{2_Z} = \begin{bmatrix} -t \\ -u \end{bmatrix}.$$

Proof. Suppose that we perform the following linear transformation:

$$Z_1 = X_1 - \left\{ M_{2_X} + \frac{M_{1_X} - M_{2_X}}{2} \right\}, \text{ and } Z_2 = X_2 - \left\{ M_{2_X} + \frac{M_{1_X} - M_{2_X}}{2} \right\}$$

As a result of this transformation, Σ_1 and Σ_2 will remain unchanged. The new mean vector of Z_1 results in

$$M_{1_Z} = M_{1_X} - \left\{ M_{2_X} + \frac{M_{1_X} - M_{2_X}}{2} \right\} = \begin{bmatrix} p \\ q \end{bmatrix} - \begin{bmatrix} r \\ s \end{bmatrix} - \frac{\begin{bmatrix} p \\ q \end{bmatrix} - \begin{bmatrix} r \\ s \end{bmatrix}}{2} = \begin{bmatrix} \frac{p-r}{2} \\ \frac{q-s}{2} \end{bmatrix}$$

In the same way, the new mean vector of Z_2 results in

$$M_{2_Z} = M_{2_X} - \left\{ M_{2_X} + \frac{M_{1_X} - M_{2_X}}{2} \right\} = \begin{bmatrix} r \\ s \end{bmatrix} - \begin{bmatrix} r \\ s \end{bmatrix} - \frac{\begin{bmatrix} p \\ q \end{bmatrix} - \begin{bmatrix} r \\ s \end{bmatrix}}{2} = \begin{bmatrix} -\frac{p-r}{2} \\ -\frac{q-s}{2} \end{bmatrix}$$

Therefore, the new random vectors are $Z_1 \sim N(M_{1_Z}, \Sigma_1)$ and $Z_2 \sim N(M_{2_Z}, \Sigma_2)$, where

$$M_{1_Z} = \begin{bmatrix} \frac{p-r}{2} \\ \frac{q-s}{2} \end{bmatrix}, \text{ and } M_{2_Z} = \begin{bmatrix} -\frac{p-r}{2} \\ -\frac{q-s}{2} \end{bmatrix}.$$

□

Theorem 2. *Let $X_1 \sim N(M_1, \Sigma_1)$ and $X_2 \sim N(M_2, \Sigma_2)$ be two random vectors, such that:*

$$M_1 = \begin{bmatrix} r \\ s \end{bmatrix}, M_2 = \begin{bmatrix} -r \\ -s \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ and } \Sigma_2 = \begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix} \tag{7}$$

then the optimal classifier obtained by Bayes classification is a pair of straight lines if and only if there exist positive real numbers, a and b, such that:

$$a(1-b)r^2 - \frac{1}{4}(ab - a - b + 1)\log ab = (a-1)bs^2\,. \tag{8}$$

Proof. The discriminant function given in Equation (5) is now:

$$\log \left| \begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix} \right| - \begin{bmatrix} x-r \\ y-s \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x-r \\ y-s \end{bmatrix} + \begin{bmatrix} x+r \\ y+s \end{bmatrix}^T \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} x+r \\ y+s \end{bmatrix} = 0, \tag{9}$$

After performing rather lengthy matrix operations, Equation (9) results in:

$$a(x+r)^2 + b(y+s)^2 - (x-r)^2 - (y-s)^2 - \log ab = 0. \tag{10}$$

Expanding the quadratic terms and grouping by x and y, we have:

$$(a-1)x^2 + (b-1)y^2 + 2(a+1)rx + 2(b+1)sy + ar^2 + bs^2 - r^2 - s^2 - \log ab = 0 \tag{11}$$

Equation (11) is a general equation of second degree, and can represent either a circle, an ellipse, a parabola, a hyperbola, or a pair of straight lines. Indeed, we are interested in the latter. Equation (11) represents a pair of straight lines if and only if the following condition is satisfied [19]:

$$(a-1)(b-1)(ar^2 + bs^2 - r^2 - s^2 - \log ab) - (a-1)[(b+1)s]^2 - (b-1)[(a+1)r]^2 = 0 \tag{12}$$

Equation (12) is equivalent to:

$$-4abr^2 - 4abs^2 + 4ar^2 + 4bs^2 - (ab - a - b + 1)\log ab = 0 \tag{13}$$

By multiplying both sides of (13) by $\frac{1}{4}$ and grouping some terms, we get:

$$a(1-b)r^2 - \frac{1}{4}(ab - a - b + 1)\log ab = (a-1)bs^2\,. \tag{14}$$

We have proved that (12) ≡ (13) ≡ (14), all of which are *iff* assertions. The result follows. □

Equation (14) is the necessary and sufficient condition that real numbers $a > 0, b > 0, r$, and s, must satisfy in order to yield the optimal linear classifier between two classes represented by normal random vectors with parameters of the form given in (7).

Consider the following: Given positive real numbers $a > 0$ and $b > 0$, we would like to find real numbers, r and s, that satisfy (14). When neither a nor b equals unity, we have four possible cases:

$$\left.\begin{array}{ll} \text{Case I}: & 0 < a < 1,\, b > 1, \\ \text{Case II}: & a > 1,\, 0 < b < 1, \\ \text{Case III}: & 0 < a < 1,\, 0 < b < 1, \text{ and} \\ \text{Case IV}: & a > 1,\, b > 1 \end{array}\right\} \tag{15}$$

The cases for which it is possible to find real numbers, r and s, satisfying Equation (14) are stated and proved in Theorem 3 below. These constitutes the necessary and sufficient conditions as will be explained presently.

Theorem 3. *Let $X_1 \sim N(M_1, \Sigma_1)$ and $X_2 \sim N(M_2, \Sigma_2)$ be two normal random vectors, such that*

$$M_1 = \begin{bmatrix} r \\ s \end{bmatrix}, M_2 = \begin{bmatrix} -r \\ -s \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ and } \Sigma_2 = \begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix} \tag{16}$$

For any positive real numbers a and b ($a \neq 1$, $b \neq 1$), there exist real numbers r and s permitting pairwise linear classification if and only if $a > 1$ and $0 < b < 1$, OR *$0 < a < 1$ and $b > 1$.*

Proof. IF: We have to prove that if a and b are as in Cases I and II of (15), there exist real numbers, r and s, permitting pairwise linear classification.

Consider the non-quadratic term of Equation (14), $l = \frac{1}{4}(ab - a - b + 1)\log ab$.

Case I: $0 < a < 1$ and $b > 1$.

We know that

$0 < a < 1 \Rightarrow -1 < a - 1 < 0 \Rightarrow (a-1)b < 0$ since $b > 1$, and

$b > 1 \Rightarrow -b < -1 \Rightarrow 1 - b < 0 \Rightarrow a(1-b) < 0$ because $a > 0$.

Since $s^2 \geq 0$, (14) can be written as follows:

$$s^2 = \frac{a(1-b)r^2 - l}{(a-1)b} \geq 0. \tag{17}$$

We are now required to find a real number r such that (17) is satisfied.

Since $(a-1)b < 0$, we must find r such that the following inequality is satisfied:

$$a(1-b)r^2 \leq l \;\equiv\; r^2 \geq \frac{l}{a(1-b)} \tag{18}$$

Utilizing the fact that $r^2 \geq 0$, we can argue that we can choose r such that r^2 is sufficiently large so as to satisfy the inequality given in (18).

Case II: $a > 1$ and $0 < b < 1$.

We know that

$0 < b < 1 \Rightarrow 0 > -b > -1 \Rightarrow 1 - b > 0 \Rightarrow a(1-b) > 0$ since $a > 1$, and

$a > 1 \Rightarrow a - 1 > 0 \Rightarrow b(a-1) > 0$ because $b > 0$.

Since $s^2 \geq 0$, (14) can be written as follows:

$$s^2 = \frac{a(1-b)r^2 - l}{(a-1)b} \geq 0. \tag{19}$$

We are now required to find a real number r such that (19) is satisfied.

Since $(a-1)b > 0$, we must find r such that the following inequality is satisfied:

$$r^2 \geq \frac{l}{a(1-b)}. \tag{20}$$

But we know that r^2 is always positive, and so we can choose r such that r^2 is sufficiently large so as to satisfy (20).

Cases I and II prove the sufficiency conditions.

ONLY IF: We are to prove that if the classifier is pairwise linear, a and b fall in Case either I or II. We prove this by showing that the negation of the implication is true, or equivalently if a and b do not fall in Cases I and II (i.e. thy fall in cases III or IV), it is not possible to find real numbers r and s permitting pairwise linear classification.

Case III: $0 < a < 1$ and $0 < b < 1$.

In this case we know that

$0 < a < 1 \Rightarrow (a-1)b < 0$ and $0 < b < 1 \Rightarrow a(1-b) > 0$.

Since $r^2 \geq 0$ and $s^2 \geq 0$ then $(a-1)br^2 \leq 0$ and $a(1-b)s^2 \geq 0$.

We shall show that this implies that if $l > 0$, it is not possible to find real numbers r and s.

Suppose that it were possible to find real numbers, r, s, and $l > 0$. This implies that

$$(ab - a - b + 1)\log ab > 0, \text{and}$$

$$ab - a - b + 1 < 0 \tag{21}$$

since $\log ab < 0$.

By our hypothesis, we know that

$0 < a < 1$ and $0 < b < 1 \Rightarrow 0 < a + b - ab < 1 \Rightarrow ab - a - b + 1 > 0$,

which contradicts (21). Therefore, there are no real numbers r and s that satisfy Equation (14) for $0 < a < 1$ and $0 < b < 1$.

Case IV: $a > 1$ and $b > 1$.

In this case we know that

$a > 1 \Rightarrow (a-1)b > 0$ and $b > 1 \Rightarrow a(1-b) < 0$.

Since $r^2 \geq 0$ and $s^2 \geq 0$, then $(a-1)br^2 \geq 0$ and $a(1-b)s^2 \leq 0$.

We again argue that this implies that if $l < 0$, it is not possible to find real number r and s.

Suppose that we are able to find real numbers, r, s, and $l < 0$. This implies that

$$(ab - a - b + 1)\log ab < 0, \text{and}$$

$$ab - a - b + 1 > 0, \tag{22}$$

since $\log ab > 0$.

By our hypothesis, we know that

$a > 1$ and $b > 1 \Rightarrow a + b - ab > 1 \Rightarrow ab - a - b + 1 < 0$,

which contradicts (22). Therefore, there are no real numbers r and s that satisfy Equation (14) for $a > 1$ and $b > 1$.

Hence the necessary conditions.

The theorem is thus proved. □

In the above theorem, we considered the cases only when both $a \neq 1$ and $b \neq 1$. The case when either a or b, or both, is unity is given below.

Theorem 4. *Let $X_1 \sim N(M_1, \Sigma_1)$ and $X_2 \sim N(M_2, \Sigma_2)$ be two normal random vectors with parameters of the form (16). When either a or b can have the value of unity, the optimal Bayesian classifier is a pair of straight lines if and only if:*

(i) $a = 1$, $b \neq 1$, *and* $r = 0$, *or*

(ii) $a \neq 1$, $b = 1$, *and* $s = 0$, *or*

(iii) $a = 1$ *and* $b = 1$.

Proof. The proof of the above cases can be derived by substituting these conditions into the relevant expressions of Theorem 3. In Case (iii), we have the situation in which $\Sigma_1 = \Sigma_2 = I$, which is the only scenario known to have a linear discriminant function as the classifier (i.e. a *single* straight line[3]) [18]. □

4 Special Cases of Linear Discriminant

In this section we analyze two special cases of diagonal covariance matrices that lead to the optimal linear discriminant function. The necessary and sufficient conditions to achieve a linear classifier are discussed in both cases. The second, and more specific case, is that where the mean vector is the same for the two classes under consideration.

4.1 Linear Discriminant with Different Means

Consider two normal random vectors of dimension $d = 2$. Using the diagonalization process discussed in Section 2.2, any covariance matrices and mean vectors can be converted into the form of (6). Starting with normal random vectors and these parameters, we are interested in analyzing linear classifiers for a more particular case.

Theorem 5. *Let X_1 and X_2 be two normal random vectors with covariance matrices and mean vectors as in (6). It is possible to transform X_1, X_2 into $Z_1 = A^T X_1$, $Z_2 = A^T X_2$, respectively, where $A = \begin{bmatrix} a^{\frac{1}{4}} & 0 \\ 0 & a^{-\frac{1}{4}} \end{bmatrix}$, and the new covariance matrices and mean vectors have the form:*

[3] Note that a single straight line means that the quadratic polynomial of (11) has a single root, or in the richer context of this paper, it has *two* coincident roots.

$$M_{1_Z} = \begin{bmatrix} p' \\ q' \end{bmatrix}, M_{2_Z} = \begin{bmatrix} r' \\ s' \end{bmatrix}, \Sigma_{1_Z} = \begin{bmatrix} a' & 0 \\ 0 & b' \end{bmatrix}, \textit{ and } \Sigma_{2_Z} = \begin{bmatrix} b' & 0 \\ 0 & a' \end{bmatrix} \tag{23}$$

if and only if $b = a^{-1}$.

Proof. Consider a matrix $A = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$ where α and β are any real numbers. The covariance matrices in the transformed space are:

$$\Sigma_{1_Z} = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} = \begin{bmatrix} \alpha^2 & 0 \\ 0 & \beta^2 \end{bmatrix}, \qquad \text{and}$$

$$\Sigma_{2_Z} = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}^T \begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix} \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} = \begin{bmatrix} a^{-1}\alpha^2 & 0 \\ 0 & b^{-1}\beta^2 \end{bmatrix}.$$

In order to obtain the form of (23), we need that $\alpha^2 = b^{-1}\beta^2$ and $a^{-1}\alpha^2 = \beta^2$. Substituting α^2 into the second expression, we have $a^{-1}b^{-1}\beta^2 = \beta^2 \Rightarrow b = a^{-1}$. Therefore, the covariance matrix of X_2 must be of the form $\Sigma_{2_X} = \begin{bmatrix} a^{-1} & 0 \\ 0 & a \end{bmatrix}$.

By choosing $A = \begin{bmatrix} a^{\frac{1}{4}} & 0 \\ 0 & a^{-\frac{1}{4}} \end{bmatrix}$, the covariance matrices in the transformed space are:

$$\Sigma_{1_Z} = \begin{bmatrix} a^{\frac{1}{4}} & 0 \\ 0 & a^{-\frac{1}{4}} \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a^{\frac{1}{4}} & 0 \\ 0 & a^{-\frac{1}{4}} \end{bmatrix} = \begin{bmatrix} a^{\frac{1}{2}} & 0 \\ 0 & a^{-\frac{1}{2}} \end{bmatrix}, \text{ and}$$

$$\Sigma_{2_Z} = \begin{bmatrix} a^{\frac{1}{4}} & 0 \\ 0 & a^{-\frac{1}{4}} \end{bmatrix}^T \begin{bmatrix} a^{-1} & 0 \\ 0 & a \end{bmatrix} \begin{bmatrix} a^{\frac{1}{4}} & 0 \\ 0 & a^{-\frac{1}{4}} \end{bmatrix} = \begin{bmatrix} a^{-\frac{1}{2}} & 0 \\ 0 & a^{\frac{1}{2}} \end{bmatrix}.$$

The mean vectors in the transformed space are $M_{1_Z} = \begin{bmatrix} a^{\frac{1}{4}}p \\ a^{-\frac{1}{4}}q \end{bmatrix}$ and $M_{2_Z} = \begin{bmatrix} a^{\frac{1}{4}}r \\ a^{-\frac{1}{4}}s \end{bmatrix}$.

The "only if" part follows by traversing the algebraic steps in a reverse manner. □

We now state the conditions necessary to obtain a pair of straight lines when we have the form of (23).

Theorem 6. *Let $X_1 \sim N(M_1, \Sigma_1)$ and $X_2 \sim N(M_2, \Sigma_2)$ be two normal random vectors such that*

$$M_1 = \begin{bmatrix} p \\ q \end{bmatrix}, M_2 = \begin{bmatrix} r \\ s \end{bmatrix}, \Sigma_1 = \begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix}, \textit{ and } \Sigma_2 = \begin{bmatrix} b^{-1} & 0 \\ 0 & a^{-1} \end{bmatrix}, \tag{24}$$

The optimal Bayes classifier is a pair of straight lines if and only if $(p - r)^2 = (q - s)^2$, for a, b any positive real numbers. Moreover, if $M_1 = \begin{bmatrix} r \\ s \end{bmatrix}$ and $M_2 = \begin{bmatrix} -r \\ -s \end{bmatrix}$, the condition simplifies to $r^2 = s^2$.

Proof. Multiplying both sides of (5) by -1, the discriminant function can be expressed as follows:

$$\begin{bmatrix} x-p \\ y-q \end{bmatrix}^T \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} x-p \\ y-q \end{bmatrix} - \begin{bmatrix} x-r \\ y-s \end{bmatrix}^T \begin{bmatrix} b & 0 \\ 0 & a \end{bmatrix} \begin{bmatrix} x-r \\ y-s \end{bmatrix} = 0 \tag{25}$$

Note that $\log \frac{|\Sigma_2|}{|\Sigma_1|} = 0$, since $|\Sigma_1| = |\Sigma_2|$.

After performing matrix operations, (25) reduces to:

$$a(x-p)^2 + b(y-q)^2 - b(x-r)^2 - a(y-s)^2 = 0 \tag{26}$$

By expanding the quadratic terms and applying the distributive law we obtain:

$$ax^2 - 2apx + ap^2 + by^2 - 2bqy + bq^2 - bx^2 + 2brx - br^2 - ay^2 + 2asy - as^2 = 0 \tag{27}$$

Taking common factors x and y from the linear and quadratic terms, we obtain:

$$(a-b)x^2 + (b-a)y^2 + 2(-ap+br)x + 2(-bq+as)y + (ap^2 + bq^2 - br^2 - as^2) = 0 \tag{28}$$

Equation (28) is a pair of straight lines *if and only if* the following condition is satisfied [19]:

$$(a-b)(b-a)(ap^2 + bq^2 - br^2 - as^2) - (a-b)(-bq+as)^2 + (a-b)(-ap+br)^2 = 0 \tag{29}$$

Dividing (29) by $(a-b)$ yields:

$$(b-a)(ap^2 + bq^2 - br^2 - as^2) - (-bq+as)^2 + (-ap+br)^2 = 0 \tag{30}$$

By expanding the quadratic terms and applying the distributive law again, we obtain:

$$abp^2 + b^2q^2 - b^2r^2 - abs^2 - a^2p^2 - abq^2 + abr^2 + a^2s^2 - b^2q^2 + 2abqs - a^2s^2 + a^2p^2 - 2abpr + b^2r^2 = 0 \tag{31}$$

After canceling some terms and dividing by ab, (31) can be simplified to:

$$(p^2 - 2pr + r^2) - (q^2 - 2qs + s^2) = 0 \tag{32}$$

We can rewrite (32) in quadratic terms as follows:

$$(p-r)^2 = (q-s)^2 \tag{33}$$

This proves the first assertion.

By setting the means to $M_1 = \begin{bmatrix} r \\ s \end{bmatrix}$ and $M_2 = \begin{bmatrix} -r \\ -s \end{bmatrix}$, (33) can be rewritten as:

$$r^2 = s^2. \tag{34}$$

The “only if” part of the proof is achieved by following the algebraic steps, in a reverse manner, from (33) to (25).

Hence the theorem. □

Theorem 6 states the necessary and sufficient condition for a pairwise linear classifier between two normal random vectors, with means and covariances of the form given in (24).

4.2 Linear Discriminant with Equal Means

In this section, we discuss a particular instance of the problem discussed in Section 4.1. Let us consider the generalization of Minsky's paradox, that is, when $M_1 = M_2$. We shall now show that it is always possible to find a pair of straight lines when $M_1 = M_2$, $\Sigma_1 = \begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix}$, and $\Sigma_2 = \begin{bmatrix} b^{-1} & 0 \\ 0 & a^{-1} \end{bmatrix}$, thus resolving the paradox in the most general case.

Theorem 7. *Let $X_1 \sim N(M_1, \Sigma_1)$ and $X_2 \sim N(M_2, \Sigma_2)$ be two random vectors such that:*

$$M_1 = M_2 = \begin{bmatrix} r \\ s \end{bmatrix}, \Sigma_1 = \begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix}, \text{ and } \Sigma_2 = \begin{bmatrix} b^{-1} & 0 \\ 0 & a^{-1} \end{bmatrix},$$

The optimal classifier obtained by Bayes classification is a pair of straight lines for positive real numbers a and b, where r and s are any real numbers.

Proof. The proof of this theorem is straightforward using the result of Theorem 6.

We know that for Σ_1 and Σ_2, it is possible to find p, q, r, and s, such that $(p-s)^2 = (q-s)^2$. Substituting r for p and s for q, in (33), we have $(r-r)^2 = (s-s)^2$. Therefore, for positive real numbers a and b, where r and s are any real numbers, the optimal classifier is a pair of straight lines. □

Clearly, Theorem 7 states the necessary and sufficient condition for a pairwise linear classifier between two normal random vectors with covariance matrices of the form given in (24), where the means are equal. The power of this will be obvious when the classification results are discussed in a subsequent section.

5 Classification

5.1 The Discriminant Function

In this section we discuss classification with the linear discriminant functions determined in Section 4.2, for dimension $d = 2$.

Equations (11) and (28) are the discriminant functions that represent a pair of straight lines for the cases discussed in Sections 3 and 4.1, respectively. For the purpose of classification, we need to find one equation for each straight line. This is done by inspection or by solving the quadratic equation in terms of y [19]. These second degree polynomial equations have the following roots:

$$y_+ = A_1 x + B_1, \text{ and } y_- = A_2 x + B_2 \tag{35}$$

Let us consider now the third case discussed in Section 4.2. The equation for each straight line can be found as per the following theorem.

Theorem 8. *Let $X_1 \sim N(M_1, \Sigma_1)$ and $X_2 \sim N(M_2, \Sigma_2)$ be two random vectors such that*

$$M_1 = M_2 = \begin{bmatrix} r \\ s \end{bmatrix}, \Sigma_1 = \begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix}, \text{ and } \Sigma_2 = \begin{bmatrix} b^{-1} & 0 \\ 0 & a^{-1} \end{bmatrix}.$$

The equations of the linear discriminant functions, i.e. the optimal classifiers, are

$$y_+ = -x + (r + s), \text{ and}$$

$$y_- = x + (s - r)\ .$$

Proof. Consider Equation (28), which specifies the discriminant function between ω_1 and ω_2 with the distributions as stated in Theorem 6.

Since $M_1 = M_2 = \begin{bmatrix} r \\ s \end{bmatrix}$, (28) can be expressed as follows:

$$(b-a)y^2 + 2(a-b)sy + [(a-b)x^2 + 2(b-a)rx + (a-b)r^2 + (b-a)s^2] = 0. \tag{36}$$

Being a polynomial of second degree in y, the roots of this equations are given by

$$\frac{-2(a-b)s \pm \sqrt{4(a-b)^2s^2 - 4(b-a)[(a-b)x^2 + 2(b-a)rx + (a-b)r^2 + (b-a)s^2]}}{2(b-a)}. \tag{37}$$

After applying the distributive law, and canceling some terms and lengthy manipulations, we can rewrite (37) as follows:

$$\frac{-2(a-b)s \pm \sqrt{4(a-b)^2(x-r)^2}}{2(b-a)}. \tag{38}$$

After solving the square root expression and canceling the terms $(a-b)$ and $(b-a)$, we have a resulting equation that has two solutions:

$$y_+ = -x + (r+s) \text{ and } y_- = x + (s-r). \tag{39}$$

□

To conclude this section, we give the discriminant functions for the distributions discussed in Section 4.1.

(a) Suppose also that

$$M_1 = \begin{bmatrix} r \\ s \end{bmatrix} \text{ and } M_2 = \begin{bmatrix} -r \\ -s \end{bmatrix}\ .$$

We earlier showed that we obtain a pairwise linear discriminant function when $r^2 = s^2$. It can be shown that the equations of the linear discriminant functions are[4]:

$$\begin{array}{lll} y_+ = -x, & y_- = x - \frac{2(a+b)r}{a-b} & \text{if } s = r, \text{ and} \\ y_+ = x, & y_- = -x + \frac{2(a+b)r}{a-b} & \text{if } s = -r. \end{array} \tag{40}$$

(b) Consider now the distributions discussed in Section 3, in which

$$M_1 = \begin{bmatrix} r \\ s \end{bmatrix},\ M_2 = \begin{bmatrix} -r \\ -s \end{bmatrix},\ \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ \text{and } \Sigma_2 = \begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix}.$$

By solving for the roots of (11), a quadratic polynomial of y, we have[4]:

[4] This can be verified by using the symbolic computation package Maple V [20], [21].

$$\begin{aligned} y_+ &= \alpha(a-1)x + \alpha(a+1)r - \beta s \quad \text{and} \\ y_- &= \alpha(1-a)x - \alpha(a+1)r - \beta s, \end{aligned} \tag{41}$$

where $\alpha = \frac{\sqrt{\frac{1-b}{a-1}}}{b-1}$ and $\beta = \frac{b+1}{b-1}$.

5.2 Pairwise Linear Classification

The general scheme of a linear classifier is the following [22]: Given a vector, X, the discriminant function as a linear combination of X is:

$$g(X) = W^T X + w, \tag{42}$$

where W is the *weight vector* that gives the direction of the hyperplane and w is the *threshold weight* that represents the distance from the origin to the hyperplane. We decide ω_1, if $g(x) > 0$, and ω_2, if $g(x) < 0$. The main problem in this scheme is to find W and w, given the training feature vectors. One of the most popular approaches uses the *perceptron* algorithm that is the basis of the learning algorithm in *neural networks*. One of the disadvantages of this approach is that it uses sub-optimal optimization techniques, such as the *gradient descent* procedure. This procedure can become "stuck" in a local optimum.
In our case, we have the optimal linear classifier, represented by a pair of straight lines, as given by the following equations

$$g_1(X) = A_1 y + B_1 x + C_1 \text{ and } g_2(X) = A_2 y + B_2 x + C_2 \tag{43}$$

The weight vectors and the threshold weights are:

$$W_1 = \begin{bmatrix} B_1 \\ A_1 \end{bmatrix}, W_2 = \begin{bmatrix} B_2 \\ A_2 \end{bmatrix}, w_1 = C_1, \text{ and } w_2 = C_2 \tag{44}$$

We can write the general inequality for classification with a pair of straight lines as follows:

$$g_1(X) g_2(X) \underset{\omega_2}{\overset{\omega_1}{\lesseqgtr}} 0 \tag{45}$$

where $g_1(X) = W_1^T X + w_1$ and $g_2(X) = W_2^T X + w_2$.
The classification is done using the following scheme:

- $g_1(X) < 0$ and $g_2(X) < 0 \;\Rightarrow\; X \in \omega_2$
- $g_1(X) < 0$ and $g_2(X) > 0 \;\Rightarrow\; X \in \omega_1$
- $g_1(X) > 0$ and $g_2(X) < 0 \;\Rightarrow\; X \in \omega_1$
- $g_1(X) > 0$ and $g_2(X) > 0 \;\Rightarrow\; X \in \omega_2$

Consider the second equation of (39). This can be written as follows:

$$y - x + (r - s) = 0 \tag{46}$$

Dividing Equation (36) by Equation (46) yields:

$$(b - a)[y + x - (r + s)] = 0 \tag{47}$$

Note that we do not cancel the term $b - a$, since we are looking for a classifier that can be obtained by replacing the equality symbol in the discriminant function by the symbol $\lesseqgtr$, and $b - a$ can be positive or negative. Next we divide (47) by $|b - a|$ and multiply it by $sgn(b - a)$, where $sgn(b - a)$ is 1 if $b - a$ is positive and -1 otherwise.

The discriminant functions in vectorial form are:

$$\begin{aligned} g_1(X) = &\quad sgn(b-a)\left\{ \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + (-r - s) \right\}, \quad \text{and} \\ g_2(X) = &\quad \begin{bmatrix} -1 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + (r - s) \end{aligned}$$

In an analogous manner, the discriminant functions, $g_1(X)$ and $g_2(X)$, and the classification scheme for equations in (40) and (41) can be obtained.

6 Simulation Results

In this section we present some examples illustrating the different cases discussed in previous sections. In all of the examples we have chosen the dimension $d = 2$ and two classes, ω_1 and ω_2. We also discuss the empirical results obtained after testing the linear classifier with 100 points for each class generated randomly using the *maximum likelihood* approach in estimating the parameters [5], assuming that they are of the form found in the respective cases.

The two classes, ω_1 and ω_2, are represented by two normal random vectors, $X_1 \sim N(M_1, \Sigma_1)$ and $X_2 \sim N(M_2, \Sigma_2)$, respectively. We used two instances of each of the three cases and generated a set of 100 normal random points, in order to test the accuracy of the classifiers.

6.1 Pairwise Linear Discriminant in Diagonalization

In the first test (referred to as DD-1 and DD-2) we considered the pairwise linear discriminant function in diagonalization. We used the following covariance matrices and mean vectors (estimated from 100 training samples) to yield the respective classifier:

DD-1: $M_1 \approx \begin{bmatrix} 1.0342 \\ 1.8686 \end{bmatrix}, M_2 \approx \begin{bmatrix} -1.0342 \\ -1.8686 \end{bmatrix}, \Sigma_1 \approx \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \Sigma_2 \approx \begin{bmatrix} .4599 & 0 \\ 0 & 2.8232 \end{bmatrix}$

DD-2: $M_1 \approx \begin{bmatrix} -1.4602 \\ -1.1913 \end{bmatrix}, M_2 \approx \begin{bmatrix} 1.4602 \\ 1.1913 \end{bmatrix}, \Sigma_1 \approx \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \Sigma_2 \approx \begin{bmatrix} 1.8395 & 0 \\ 0 & .4278 \end{bmatrix}$

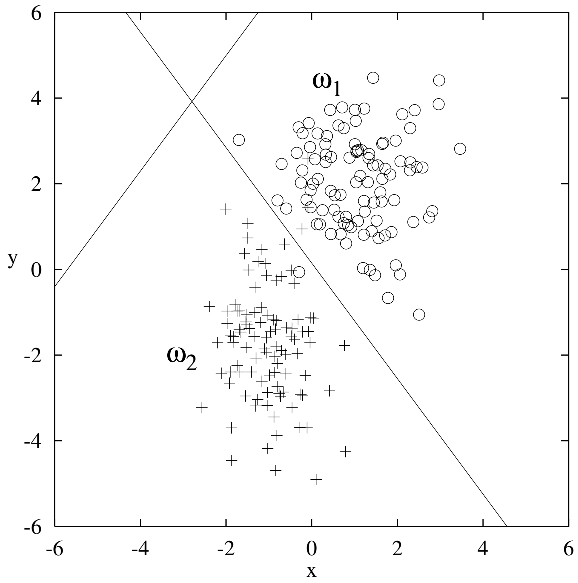

Figure 3: Example of pairwise linear discriminant in diagonalization for the case described in Theorem 2. The data set is DD-1.

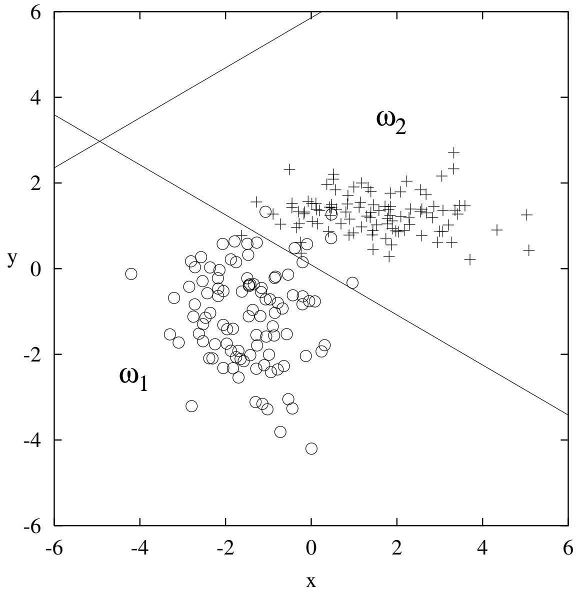

Figure 4: Example of pairwise linear discriminant in diagonalization for the case described in Theorem 2. The data set is DD-2.

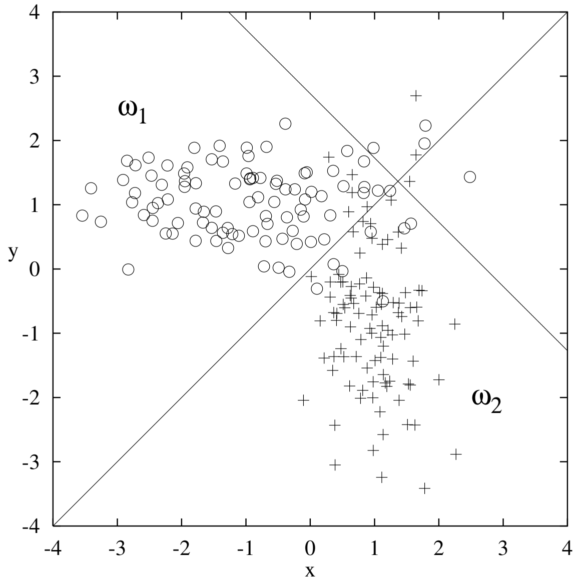

Figure 5: Example of pairwise linear discriminant with different means for the case described in Theorem 6. The data set is DM-1.

The plot of the points and the linear discriminant function are depicted in Figures 3 and 4. The accuracy of the classifier was 99% for ω_1 and ω_2 in the first test set, DD-1, and 94% for ω_1 and 93% for ω_2 in the second test set, DD-2. The power of the scheme is obvious!

6.2 Pairwise Linear Discriminant with Different Means

In the second test (referred to as DM-1 and DM-2) we considered the pairwise linear discriminant with different means. The following estimated covariance matrices and mean vectors were obtained by using 100 training samples.

$$\textbf{DM-1: } M_1 \approx \begin{bmatrix} -.9555 \\ .9555 \end{bmatrix}, M_2 \approx \begin{bmatrix} .9555 \\ -.9555 \end{bmatrix}, \Sigma_1 \approx \begin{bmatrix} 1.8077 & 0 \\ 0 & .3188 \end{bmatrix}, \Sigma_2 \approx \begin{bmatrix} .3188 & 0 \\ 0 & 1.8077 \end{bmatrix}$$

$$\textbf{DM-2: } M_1 \approx \begin{bmatrix} .8293 \\ .8293 \end{bmatrix}, M_2 \approx \begin{bmatrix} -.8293 \\ -.8293 \end{bmatrix}, \Sigma_1 \approx \begin{bmatrix} .3615 & 0 \\ 0 & 2.5503 \end{bmatrix}, \Sigma_2 \approx \begin{bmatrix} 2.5503 & 0 \\ 0 & .3615 \end{bmatrix}$$

Using the above parameters, the pairwise linear classifier was derived. The plot of the points and the linear discriminant function are shown in Figures 5 and 6. The accuracy of the classifier was 91% for ω_1 and 95% for ω_2 in the first test set, DM-1, and 88% for ω_1 and ω_2 in the second test set, DM-2.

6.3 Pairwise Linear Discriminant with Equal Means

To show the power of the scheme, we also tested our results for the case of the pairwise linear classifier with equal means (EM-1 and EM-2) for the generalized Minsky's Paradox. By using 100 training samples

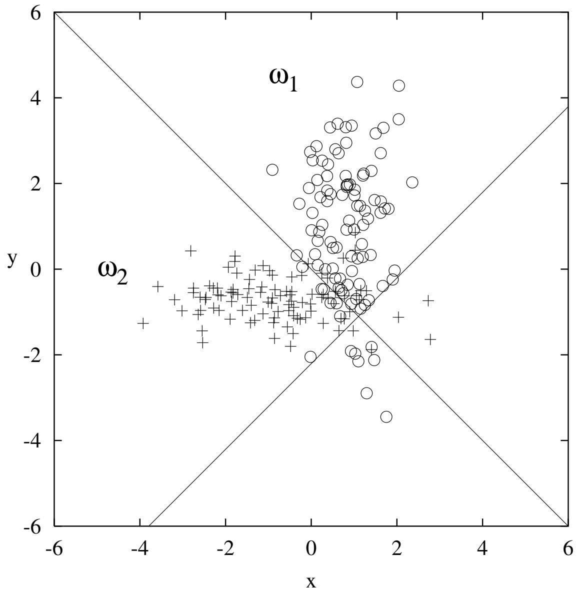

Figure 6: Example of pairwise linear discriminant with different means for the case described in Theorem 6. The data set is DM-2.

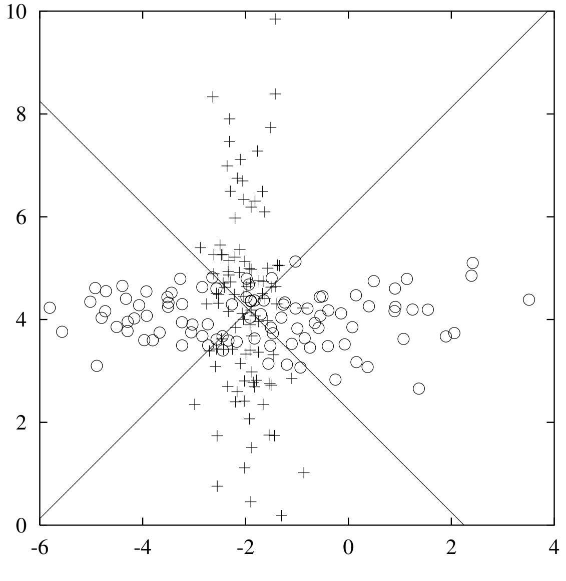

Figure 7: Example of pairwise linear discriminant with equal means for the case described in Theorem 7. The data set is EM-1.

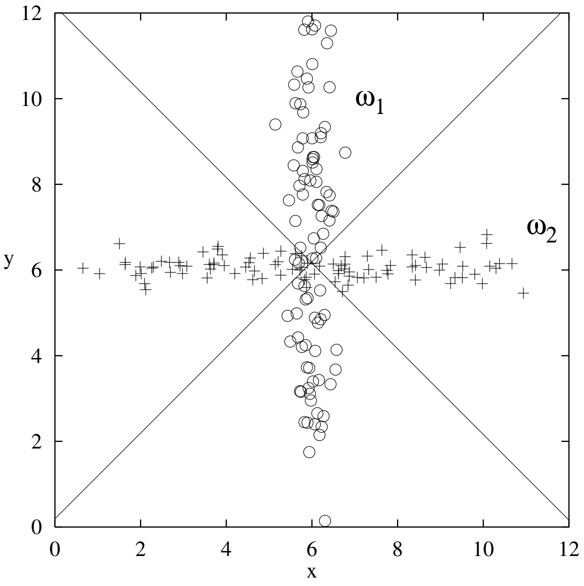

Figure 8: Example of pairwise linear discriminant with equal means for the case described in Theorem 7. The data set is EM-2.

generated with equal means but mirrored covariances, we obtained the following estimated covariance matrices and mean vectors:

EM-1: $M_1 \approx \begin{bmatrix} -1.9394 \\ 4.1875 \end{bmatrix}, M_2 \approx \begin{bmatrix} -1.9394 \\ 4.1875 \end{bmatrix}, \Sigma_1 \approx \begin{bmatrix} 4.5193 & 0 \\ 0 & .255 \end{bmatrix}, \Sigma_2 \approx \begin{bmatrix} .255 & 0 \\ 0 & 4.5193 \end{bmatrix}$

EM-2: $M_1 \approx \begin{bmatrix} 5.9841 \\ 6.1766 \end{bmatrix}, M_2 \approx \begin{bmatrix} 5.9841 \\ 6.1766 \end{bmatrix}, \Sigma_1 \approx \begin{bmatrix} .0904 & 0 \\ 0 & 12.7515 \end{bmatrix}, \Sigma_2 \approx \begin{bmatrix} 12.7515 & 0 \\ 0 & .0904 \end{bmatrix}$

The plot of the points and the linear discriminant function from these estimates are given in Figures 7 and 8. The accuracy of the classifier was 82% for ω_1 and 85% for ω_2 in the first test set, EM-1, and 96% for ω_1 and 90% for ω_2 in the second test set, EM-2.

6.4 Overall Observations

The empirical results of classification are given in Table 1. The first column corresponds to the test case. The second and third columns represent the percentage of correctly classified points belonging to ω_1 and ω_2, respectively. Note the very high accuracy in the first two cases, DD-1 and DD-2, which correspond to the pairwise linear discriminant in diagonalization case. DM-1 and DM-2 are “worse” in accuracy, but it is still high. These correspond to the pairwise linear discriminant with different means. The sixth and seventh rows correspond to the case where the means are identical, referred to as EM-1 and EM-2. The accuracy is very high in this case, despite the fact that the classes overlap and the discriminant functions are pairwise linear. The power of the results presented are again obvious and the resolution of Minsky’s Paradox is clear.

Example	accuracy for ω_1	accuracy for ω_2
DD-1	98 %	99 %
DD-2	95 %	94 %
DM-1	91 %	95 %
DM-2	88 %	88 %
EM-1	82 %	85 %
EM-2	96 %	90 %

Table 1: Accuracy in classification of test points generated with the six examples presented above. The accuracy is given in percentage of points correctly classified.

7 Conclusions

In this paper we have shown the problem of determining pairwise linear classifiers for the case of normally distributed classes. We have shown that, contrary to what is known, it is possible to find the optimal linear discriminant function even though the covariance matrices are different. In all the cases discussed here, the functions obtained are pairs of straight lines, which is a particular case of the second degree general equation.

By a formal procedure, we have determined the conditions for these particular discriminant functions in three cases. The first case occurs after diagonalization. Having two classes with normal distribution and any covariance matrices, we can always perform diagonalization and obtain the required covariance matrices. We have explicitly derived the necessary and sufficient conditions for the covariance matrices and the mean vectors so as to yield a pair of straight lines for the optimal classifier.

The second case is when we have particular forms in the two diagonal covariance matrices. In this case, two terms of the diagonal of one matrix have to be a permutation of two terms of the other matrix and the remaining terms of both matrices must be identical. For this case we have again derived the necessary and sufficient conditions for an optimal pairwise linear discriminant function.

In the third case, assuming equal means, we have found that it is always possible to obtain a pair of straight lines when we have covariance matrices with the same form as found in the second case. This re-solves Minsky's paradox!

The results derived in the paper have also been experimentally verified. The empirical results obtained show that the accuracy of the classifier is very high. This is understandable since the classifier is optimal. The degree of this accuracy is even more amazing when we recognize that we are dealing with a linear discriminant function for classes which are significantly overlapping.

References

[1] B. Ripley, *Pattern Recognition and Neural Networks.* Cambridge Univ. Press, 1996.

[2] W. Krzanowski, P. Jonathan, W. McCarthy, and M. Thomas, "Discriminant Analysis with Singular Covariance Matrices: Methods and Applications to Spectroscopic Data," *Applied Statistics*, vol. 44, pp. 101–115, 1995.

[3] R. Duda and P. Hart, *Pattern Classification and Scene Analysis.* John Wiley and Sons, Inc., 1973.

[4] W. Malina, "On an Extended Fisher Criterion for Feature Selection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 3, pp. 611–614, 1981.

[5] R. Schalkoff, *Pattern Recognition: Statistical, Structural and Neural Approaches.* John Wiley and Sons, Inc., 1992.

[6] R. Lippman, "An Introduction to Computing with Neural Nets," in Lau [23], pp. 5–24.

[7] O. Murphy, "Nearest Neighbor Pattern Classification Perceptrons," in Lau [23], pp. 263–266.

[8] S. Raudys, "Evolution and Generalization of a Single Neurone: I. Single-layer Perception as Seven Statistical Classifiers," *Neural Networks*, vol. 11, no. 2, pp. 283–296, 1998.

[9] S. Raudys, "Evolution and Generalization of a Single Neurone: II. Complexity of Statistical Classifiers and Sample Size Considerations," *Neural Networks*, vol. 11, no. 2, pp. 297–313, 1998.

[10] A. Rao, D. Miller, K. Rose, , and A. Gersho, "A Deterministic Annealing Approach for Parsimonious Design of Piecewise Regression Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 2, pp. 159–173, 1999.

[11] S. Raudys, "On Dimensionality, Sample Size, and Classification Error of Nonparametric Linear Classification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 6, pp. 667–671, 1997.

[12] M. Aladjem, "Linear Discriminant Analysis for Two Classes Via Removal of Classification Structure," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 187–192, 1997.

[13] M. Minsky, *Perceptrons.* MIT Press, 2nd ed., 1988.

[14] C. Chatterjee and V. Roychowdhury, "An Adaptive Stochastic Approximation Algorithm for Simuntaneous Diagonalization of Matrix Sequences with Applications," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, pp. 282–287, 1997.

[15] T. M. Ha, "The Optimum Class-Selective Rejection Rule," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 6, pp. 608–615, 1997.

[16] T. M. Ha, "Optimum Decision Rules in Pattern Recognition," *Advances in Pattern Recognition, SSPR'98 - SPR'98*, pp. 726–735, 1998.

[17] K. Fukunaga, "Statistical Pattern Recognition," *Handbook of Pattern Recognition and Computer Vision*, pp. 33–60, 1993.

[18] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

[19] J. Brown and C. Manson, *The Elements of Analiytical Geometry.* McMillan and Co., Limited, 1950.

[20] B. Char, K. Geddes, G. Gonnet, B. Leong, M. Monagan, and S. Watt, *Maple V - Reference Manual.* Springer-Verlag, 1991.

[21] E. Deeba and A. Gunawardena, *Interactive Linear Algebra with MAPLE V.* Springer, 1997.

[22] P. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach.* Prentice-Hall, 1982.

[23] C. Lau, ed., *Neural Networks: Theoretical Foundations and Analysis*, IEEE Press, 1992.