On Optimal Pairwise Linear Classifiers for Normal Distributions: The d-Dimensional Case*

Luis Rueda† and **B. John Oommen**‡

Abstract

We consider the well-studied Pattern Recognition (PR) problem of designing linear classifiers. When dealing with normally distributed classes, it is well known that the optimal Bayes classifier is linear only when the covariance matrices are equal. This was the only known condition for classifier linearity. In a previous work, we presented the theoretical framework for optimal *pairwise* linear classifiers for two-dimensional normally distributed random vectors. We derived the necessary and sufficient conditions that the distributions have to satisfy so as to yield the *optimal* linear classifier as a pair of straight lines.

In this paper we extend the previous work to d-dimensional normally distributed random vectors. We provide the necessary and sufficient conditions needed so that the optimal Bayes classifier is a pair of hyperplanes. Various scenarios have been considered including one which resolves the multi-dimensional *Minsky's paradox* for the perceptron. We have also provided some three dimensional examples for all the cases, and tested the classification accuracy of the corresponding pairwise linear classifier. In all the cases, these linear classifiers achieve very good performance. To demonstrate that the current pairwise-linear philosophy yields superior discriminants on real life data, we have shown how linear classifiers determined using an MLE estimation applicable for this approach, yields better accuracy that the discriminants obtained by the traditional Fisher's classifier on a real life data set. The multi-dimensional generalization of the MLE estimate for these classifiers is currently being investigated.

1 Introduction

The problem of finding linear classifiers has been the study of many researchers in the field of Pattern Recognition (PR). Linear classifiers are very important because of their simplicity when it concerns implementation, and their classification speed. Various schemes to yield linear classifiers are reported in the literature such as *Fisher's approach* [3, 9, 19], the *perceptron algorithm* (the basis of the back propagation *neural network* learning algorithms) [7, 11, 14, 15], *piecewise recognition models* [12], *random search optimization* [13], *removal classification structures* [1], *adaptive linear dimensionality reduction* [8] (which outperforms Fisher's

* **Acknowledgments:** The authors are very grateful to the anonymous referee for his/her suggestions on enhancing our theorems to also include the necessary and sufficient conditions for the respective results. By utilizing these suggestions, we were also able to extend the previous results obtained for the 2-dimensional features [17], thus improving the quality of both the previous paper, and of this present one. We are truly grateful. A preliminary version of this paper can be found in the *Proceedings of AI'01, the 14th Australian Joint Conference on Artificial Intelligence*, December 2001, Adelaide, Australia.

† 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 for some data sets), and *linear constrained distance-based classifier analysis* [2] (an improvement to Fisher's approach designed for hyperspectral image classification). All of these approaches suffer from the lack of optimality, and thus, although they do determine linear classification functions, the classifier is not optimal.

Apart from the results reported in [17, 18], in *statistical* PR, the Bayesian linear classification for normally distributed classes involves a single case. This traditional case is when the covariance matrices are equal [5, 16, 20]. In this case, the classifier is a single straight line (or a hyperplane in the d-dimensional case) completely specified by a first-order equation.

In [17, 18], we showed that although the general classifier for two dimensional normally distributed random vectors is a second-degree polynomial, this polynomial degenerates to be either a single straight line or a pair of straight lines. Thus, we have found the necessary and sufficient conditions under which the classifier can be linear even when the covariance matrices are not equal. In this case, the classification function is a pair of first-order equations, which are factors of the second-order polynomial (i.e. the classification function). When the factors are equal, the classification function is given by a single straight line, which corresponds to the traditional case when the covariance matrices are equal.

Some examples of pairwise linear classifiers for two and three-dimensional normally distributed random vectors can be found in [3] pp. 42–43. By studying these, the reader should observe that the *existence* of such classifiers was known. The novelty of our results are the *conditions* for pairwise linear classifiers, and the demonstration that these, in their own right, lead to superior linear classifiers.

In this paper, we extend these conditions for d-dimensional normal random vectors, where $d > 2$. We assume that the features of an object to be recognized are represented as a d-dimensional vector which is an ordered tuple $X = [x_1 \dots x_d]^T$ characterized by a probability distribution function. We deal only with the case in which these random vectors have a jointly normal distribution, where class ω_i has a mean M_i and covariance matrix Σ_i, $i = 1, 2$.

Without loss of generality, we assume that the classes ω_1 and ω_2 have the same *a priori* probability, 0.5, in which case, the classifier is given by:

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

When $\Sigma_1 = \Sigma_2$, the classification function is linear [3, 4, 20]. For the case when Σ_1 and Σ_2 are arbitrary, the classifier results in a general equation of second degree which results in the classifier being a hyperparaboloid, a hyperellipsoid, a hypersphere, a hyperboloid, or a pair of hyperplanes. This latter case is the focus of our present study.

The results presented here have been rigorously tested. In particular, we present some empirical results for the cases in which the optimal Bayes classifier is a pair of hyperplanes. It is worth mentioning that we tested the case of Minsky's paradox on randomly generated samples, and we have found that the accuracy is very high even though the classes are significantly overlapping.

2 Linear Classifiers for Diagonalized Classes: The 2-D Case

The concept of *diagonalization* is quite fundamental to our study. Diagonalization is the process of transforming a space by performing linear and whitening transformations [4]. Consider a normally distributed random vector, $\mathbf{X}$, with any mean vector and covariance matrix. By performing diagonalization, $\mathbf{X}$ can be transformed into another normally distributed random vector, $\mathbf{Z}$, whose covariance is the identity matrix. This can be easily generalized to incorporate what is called "simultaneous diagonalization". By performing this process, two normally distributed random vectors, $\mathbf{X}_1$ and $\mathbf{X}_2$, can be transformed into two other normally distributed random vectors, $\mathbf{Z}_1$ and $\mathbf{Z}_2$, whose covariance matrices are the identity and a diagonal matrix, respectively. A more in-depth discussion of diagonalization can be found in [19, 4], and is omitted here as it is assumed to be fairly elementary. We discuss below the conditions for the mean vectors and covariance matrices of simultaneously diagonalized vectors in which the Bayes optimal classifier is pairwise linear.

In [17, 18], we presented the necessary and sufficient conditions required so that the optimal classifier is a pair of straight lines, for the two dimensional space. Using these results, we present here the cases for the d-dimensional case in which the optimal Bayes classifier is a pair of hyperplanes.

Since we repeatedly refer to the work of [17, 18], we state (without proof) the relevant results below.

One of the cases in which we evaluated the possibility of finding a pair of straight lines as the optimal classifier is when we have *inequality constraints*. This case is discussed below.

Theorem 1. *Let $\mathbf{X}_1 \sim N(M_1, \Sigma_1)$ and $\mathbf{X}_2 \sim N(M_2, \Sigma_2)$ be two normally distributed random vectors with parameters of the form:*

$$M_1 = \begin{bmatrix} r \\ s \end{bmatrix}, M_2 = \begin{bmatrix} -r \\ -s \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ and } \Sigma_2 = \begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix}. \tag{2}$$

There exist real numbers, r and s, such that the optimal Bayes classifier is a pair of straight lines if and only if one of the following conditions is satisfied:

(a) $0 < a < 1$ *and* $b > 1$,

(b) $a > 1$ *and* $0 < b < 1$.

Moreover, the optimal Bayes classifier is a pair of straight lines if and only if

$$a(1-b)r^2 + b(1-a)s^2 - \frac{1}{4}(ab - a - b + 1)\log ab = 0\,, \tag{3}$$

□

Another case evaluated in [17, 18] is when we have *equality constraints*. In this case, the optimal Bayes classifier is a pair of parallel straight lines. In particular, when $\Sigma_1 = \Sigma_2$, these lines are coincident.

Theorem 2. *Let $\mathbf{X}_1 \sim N(M_1, \Sigma_1)$ and $\mathbf{X}_2 \sim N(M_2, \Sigma_2)$ be two normally distributed random vectors with parameters of the form of (2). The optimal Bayes classifier is a pair of straight lines if and only if one of the following conditions is satisfied:*

(a) $a = 1$, $b \neq 1$, *and* $r = 0$, *(b)* $a \neq 1$, $b = 1$, *and* $s = 0$.

The classifier is a single straight line if and only if:

(c) $a = 1$ *and* $b = 1$. □

3 Multi-Dimensional Pairwise Hyperplane Classifiers

Let us consider now the more general case for $d > 2$. Using the results mentioned above, we derive the necessary and sufficient conditions for a pairwise-linear optimal Bayes classifier. From the inequality constraints (a) and (b) of Theorem 1, we state and prove that it is not possible to find the optimal Bayes classifier as a pair of hyperplanes for these conditions when $d > 2$. We modify the notation marginally. We use the symbols $(a_1^{-1}, a_2^{-1}, \ldots, a_d^{-1})$ to synonymously refer to the marginal variances $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$.

Theorem 3. *Let* $\mathbf{X}_1 \sim N(M_1, \Sigma_1)$ *and* $\mathbf{X}_2 \sim N(M_2, \Sigma_2)$ *be two normally distributed random vectors, such that*

$$M_1 = -M_2 = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_d \end{bmatrix}, \Sigma_1 = I, \text{ and } \Sigma_2 = \begin{bmatrix} a_1^{-1} & 0 & \ldots & 0 \\ 0 & a_2^{-1} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_d^{-1} \end{bmatrix} \tag{4}$$

where $a_i \neq 1$, $i = 1, \ldots, d$. *There are no real numbers* m_i, $i = 1, \ldots, d$, *such that the optimal Bayes classifier is a pair of hyperplanes.*

Proof. In order to obtain a pair of hyperplanes for the optimal classifier in the d-dimensional space, the projection of these hyperplanes onto the plane determined by the axes x_i and x_j must be a pair of straight lines for all $i, j = 1, \ldots, d$, $i \neq j$.

Without loss of generality[1], suppose that the conditions stated in Theorem 1 are satisfied for the axes x_i and x_j. Now, consider another axis, x_k, $k \neq i$, $k \neq j$, and let us analyze the possibility of finding a pair of straight lines, which is the optimal classifier between the axes x_k and x_i, and between the axes x_k and x_j.

- Suppose that $0 < a_i < 1$ and $a_j > 1$.

  In this case, if $0 < a_k < 1$, using the result of Theorem 1, we see that it is impossible to find the corresponding real numbers r_{ik} and s_{ik} such that the optimal classifier is a pair of straight lines between the axes x_k and x_i, since $0 < a_i < 1$.

  On the other hand, if $a_k > 1$, again, there are no real numbers r_{jk} and s_{jk} such that the optimal classifier is a pair of straight lines between the axes x_k and x_j, since $a_j > 1$.

- Suppose that $a_i > 1$ and $0 < a_j < 1$.

  Again, in this case, if $0 < a_k < 1$, it is not possible to find real numbers r_{jk} and s_{jk} so as to yield a pair of straight lines between the axes x_k and x_j, since $0 < a_j < 1$.

  Analogously, if $a_k > 1$, it is impossible to find a pair of straight lines between the axes x_k and x_i, since $a_i > 1$.

[1] We refer to the quantities r and s of Theorem 1 as r_{ij} and s_{ij} if the random variables considered are x_i and x_j.

We have thus shown that it is not possible to find a pair of hyperplanes when $a_i \neq 1, a_j \neq 1$, and $a_k \neq 1$. The result follows. $\square$

Informally speaking, the proof of Theorem 3 is accomplished by checking if there is an optimal pairwise linear classifier for all the pairs of axes. This is not possible since, if the condition has to be satisfied, when the first element on the diagonal is less than unity, the second one must be greater than unity. Consequently, there is no chance for a third element to satisfy this condition, pairwise, with the first two elements.

Using the results of Theorem 2, we now analyze the possibility of finding the optimal pairwise linear classifiers for the d-dimensional case when some of the entries in Σ_2 are unity.

Theorem 4. *Let $\mathbf{X}_1 \sim N(M_1, \Sigma_1)$ and $\mathbf{X}_2 \sim N(M_2, \Sigma_2)$ be two normally distributed random vectors with parameters of the form of (4). If there exists an index i such that $a_i \neq 1$, and $a_j = 1$, $m_j = 0$, for $j = 1, \ldots, d$, $i \neq j$, then the optimal Bayes classifier is a pair of hyperplanes.*

Proof. Consider two normally distributed random variable, $\mathbf{X}_1 \sim N(M_1, \Sigma_1)$ and $\mathbf{X}_2 \sim N(M_2, \Sigma_2)$, whose parameters are of the form given in (4).

Without loss of generality, suppose that there exist some i, $1 \leq i \leq d$, such that $a_i \neq 1$, and for all j, $1 \leq j \leq d$, $i \neq j$, $a_j = 1$ and $m_j = 0$.

To obtain a pair of hyperplanes as the optimal classifier in the d-dimensional space, the projection of these hyperplanes onto the plane determined by the axes x_i and x_j must be a pair of straight lines for i, $j = 1, \ldots, d$, $i \neq j$.

To consider all the pairs of axes, x_k and x_l, $k \neq l$, we analyze the following three mutually exclusive and exhaustive cases:

- **(a)** $k = i$, and $l = j$, where $j \neq i$. Since $a_k \neq 1$, $a_l = 1$, and $m_j = 0$, this case reduces to Case (b) of Theorem 2. Using the latter, we argue that the classifier between the axes x_k and x_l is a pair of straight lines.
- **(b)** $k \neq i$ and $l = j$, where $j \neq i$. In this case, we see that $a_k = 1$ and $a_l = 1$. Since $k \neq l$, by invoking Case (c) of Theorem 2, it is clear that the classifier between the axes x_k and x_l is a pair of straight lines because the corresponding sub-matrices are equal, leading to the linear (not pairwise linear) classifier determined by traditional methods.
- **(c)** $k = j$, where $j \neq i$, and $l = i$. Again, since $a_k = 1$, $m_k = 0$, and $a_l \neq 1$, by invoking Case (a) of Theorem 2, it follows that the classifier between the axes x_k and x_l is a pair of straight lines except that, as opposed to Case (a) above, k and l assume interchanged roles.

The result follows. $\square$

We now combine the results of Theorems 1 and 2, and state more general necessary and sufficient conditions to find a pair of hyperplanes as the optimal Bayes classifier. We achieve this using the inequality and equality constraints of these theorems.

The main difference between Theorem 4 and the theorem given below is that in the former, all the elements but one of the diagonal of Σ_2 are equal to unity. In the theorem presented below, there are two elements of the diagonal of Σ_2 which are not equal to unity, and therefore they must satisfy (3) and either condition (a) or (b) of Theorem 1.

Theorem 5. *Let $\mathbf{X}_1 \sim N(M_1, \Sigma_1)$ and $\mathbf{X}_2 \sim N(M_2, \Sigma_2)$ be two normally distributed random vectors with parameters of the form of (4). The optimal Bayes classifier is a pair of hyperplanes if there exist i and j such that any of the following conditions are satisfied:*

(a) $0 < a_i < 1$, $a_j > 1$, $a_k = 1$, $m_k = 0$, *for all* $k = 1, \ldots, d$, $i \neq j$, $k \neq i$, $k \neq j$, *with*

$$a_i(1-a_j)m_i^2 + a_j(1-a_i)m_j^2 - \frac{1}{4}(a_ia_j - a_i - a_j + 1)\log a_ia_j = 0\,. \tag{5}$$

(b) $a_i \neq 1$, $a_j = 1$, $m_j = 0$, *for all* $j \neq i$.

(c) $a_i = 1$, *for all* $i = 1, \ldots, d$.

Proof. As in Theorem 4, to obtain a pair of hyperplanes for the optimal classifier in the d-dimensional space, the projection of these hyperplanes onto the plane determined by the axes x_i and x_j must be a pair of straight lines for all $i, j = 1, \ldots, d$, $i \neq j$.

Without loss of generality, suppose that there exists an index i such that $0 < a_i < 1$. Consider another axis, x_j, $i \neq j$. We have to analyze the following cases for a_j:

(*i*) $0 < a_j < 1$: According to the conditions stated in Theorem 1, we cannot have a pairwise linear optimal classifier.

(*ii*) $a_j > 1$: If (5) is satisfied between x_i and x_j, for all the x_k, $k = 1, \ldots, d$, $i \neq j$, $k \neq i$, $k \neq j$, the optimal classifier between the axes x_i and x_k, x_j and x_k, is a pair of straight lines only if $a_k = 1$ and $m_k = 0$, using the equality constraint (b) of Theorem 2. Hence condition (a) is satisfied.

(*iii*) $a_j = 1$: which is the case of condition (b). Here, using the result of Theorem 4, an optimal pairwise linear classifier results if $a_j = 1$ and $m_j = 0$, for all $j = 1, \ldots, d$, $j \neq i$.

The final case considered in condition (c) corresponds to the traditional case in which the optimal Bayes classifier is a single hyperplane when both the covariance matrices are identical.

Hence the theorem. □

4 Linear Classifiers with Different Means

In [17], we have shown that given two normally distributed random vectors, $\mathbf{X}_1$ and $\mathbf{X}_2$, with mean vectors and covariance matrices of the form:

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

the optimal Bayes classifier is a pair of straight lines when $r^2 = s^2$, where a and b are any positive real numbers. The classifier for this case is given by:

$$a(x-r)^2 + b(y-s)^2 - b(x+r)^2 - a(y+s)^2 = 0\,. \tag{7}$$

We consider now the more general case for $d > 2$. We are interested in finding the conditions that guarantee a pairwise linear classification function. This is given in Theorem 6 below.

Theorem 6. *Let* $\mathbf{X}_1 \sim N(M_1, \Sigma_1)$ *and* $\mathbf{X}_2 \sim N(M_2, \Sigma_2)$ *be two normal random vectors such that*

$$\begin{aligned} M_1 &= [m_1, \ldots, m_i, \ldots, m_j, \ldots, m_d]^T \ , \\ M_2 &= [m_1, \ldots, m_{i-1}, -m_i, m_{i+1}, \ldots, m_{j-1}, -m_j, m_{j+1}, \ldots, m_d] \ , \end{aligned} \tag{8}$$

$$\Sigma_1 = \begin{bmatrix} a_1^{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_i^{-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_j^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_d^{-1} \end{bmatrix}, \ \textit{and}\ \Sigma_2 = \begin{bmatrix} a_1^{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_j^{-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_i^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_d^{-1} \end{bmatrix}. \tag{9}$$

The optimal classifier, obtained by Bayes classification, is a pair of hyperplanes when

$$m_i^2 = m_j^2 \,. \tag{10}$$

Proof. We shall prove the theorem by utilizing some of the results of the two dimensional case mentioned above. The classification function given in (1) can be written as follows:

$$(X - M_1)^T \Sigma_1^{-1} (X - M_1) - (X - M_2)^T \Sigma_2^{-1} (X - M_2) = 0 \,. \tag{11}$$

Note that $\log \frac{|\Sigma_2|}{|\Sigma_1|} = 0$, since $|\Sigma_1| = |\Sigma_2|$. After applying some simplifying operations to (11), the term $a_k(x_k - m_k)^2$ can be canceled along with its correspondent negative term $-a_k(x_k - m_k)^2$, for $k = 1, \ldots, d$, $i \neq j$, $k \neq i$, $k \neq j$. Equation (11) now reduces to:

$$a_i(x_i - m_i)^2 + a_j(x_j - m_j)^2 - a_j(x_i + m_i)^2 - a_i(x_j + m_j)^2 = 0 \,. \tag{12}$$

Equation (12) is the pairwise multi-dimensional version of (7). By choosing $a = a_i$, $b = a_j$, $r = m_i$, $s = m_j$, and after rather lengthy algebraic operations, it can be seen that the classification function is a pair of hyperplanes whenever:

$$m_i^2 = m_j^2 \,. \tag{13}$$

□

Theorem 6 can be interpreted geometrically as follows. Whenever we have two covariance matrices that differ only in two elements of their diagonal (namely a_i and a_j, where $i \neq j$); and whenever the two elements in the second covariance matrix are a permutation of the same rows in the first matrix, if the mean vectors differ only in positions[2] i and j, and (13) is satisfied, the resulting classifier is a pair of hyperplanes.

Indeed, by performing a projection of the space in the x_i and x_j axes, we observe that the classifier takes on exactly the same shape as that which is obtained from the distribution given in (6). Thus effectively, we

[2] Actually, the elements in positions i and j of M_1 are the negated values of those in positions i and j of M_2 respectively.

obtain a pair of straight lines in the two dimensional space from the projection of the pair of hyperplanes in the d-dimensional space.

5 Linear Classifiers with Equal Means

We consider now a particular instance of the problem discussed in Section 4, which leads to the resolution of the generalization of the d-dimensional Minsky's paradox. In this case, the covariance matrices have the form of (9), but the mean vectors are the same for both classes. We shall now show that, with these parameters, it is always possible to find a pair of hyperplanes, which resolves Minsky's paradox in the most general case.

Theorem 7. *Let $\mathbf{X}_1 \sim N(M_1, \Sigma_1)$ and $\mathbf{X}_2 \sim N(M_2, \Sigma_2)$ be two normal random vectors, where $M_1 = M_2 = [m_1, \ldots, m_d]^T$, and Σ_1 and Σ_2 have the form of (9). The optimal classifier, obtained by Bayes classification, is a pair of hyperplanes.*

Proof. From Theorem 6, we know that when the mean vectors and covariance matrices have the form of (8) and (9), respectively, then the optimal Bayes classifier is a pair of hyperplanes if $m_i^2 = m_j^2$. Since $M_1 = M_2$, then $m_i^2 = m_j^2$, for $j = 1, \ldots, d$, $i \neq j$. Hence the optimal Bayes classifier is a pair of hyperplanes. The theorem is thus proved. □

6 Simulation Results for Synthetic Data

In order to test the accuracy of the pairwise linear classifiers and to verify the results derived here, we have performed some simulations for the different cases discussed above. We have chosen the dimension $d = 3$, since it is easy to visualize and plot the corresponding hyperplanes. In all the simulations, we trained our classifier using 100 randomly generated training samples (which were three dimensional vectors from the corresponding classes). Using the *maximum likelihood* estimation (MLE) method [3], we then approximated the mean vectors and covariance matrices for each of the three cases.

In order to obtain a classifier that is a pair of hyperplanes, we have, in the various scenarios, selected parameters that satisfy the conditions given in (5), (12) and (13). Using these parameters, we have randomly generated 100 training samples from which the estimated parameters were obtained using a MLE method. These new parameters where constrained by forcing one value in a covariance matrix and a mean vector, so as to satisfy the required conditions. The corresponding classifiers were then tested on 100 random samples generated using the original parameters. In the next section, we discuss how we obtain pairwise linear classifiers for real life data, in which the parameters of the distributions do not necessarily satisfy the conditions given in (5), (12) and (13), and hence our classifiers are general enough for any data set.

We considered two classes, ω_1 and ω_2, which are represented by two normal random vectors, $\mathbf{X}_1 \sim N(M_1, \Sigma_1)$ and $\mathbf{X}_2 \sim N(M_2, \Sigma_2)$, respectively. For each class, we used two sets of 100 normal random points to test the accuracy of the classifiers.

In all the cases, to display the distribution, we plotted the ellipsoid of equi-probable points instead of the training points. This was because the plot of the three dimensional points caused too much cluttering, making the shape of the classes and the classifiers indistinguishable.

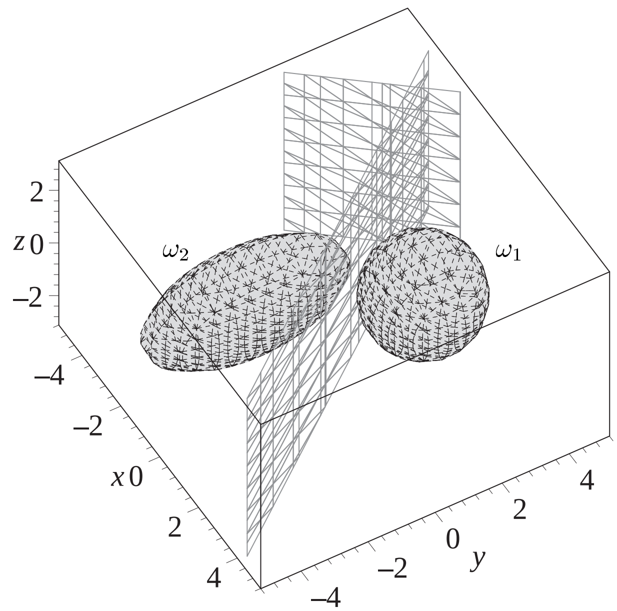

Figure 1: Example of a pairwise linear classifier for diagonalized normally distributed classes. This example corresponds to the data set DD-1.

6.1 Linear Classifiers for Two Diagonalized Classes

In the first test, DD-1, we considered the pairwise linear classification function for two diagonalized classes. These classes are normally distributed with covariance matrices being the identity matrix and another matrix in which two elements of the diagonal are not equal to unity and the remaining are unity. This is indeed, the case in which the optimal Bayes classifier is shown to be a pair of hyperplanes, stated and proven in Theorem 5. The following mean vectors and covariance matrices were estimated from 100 training samples to yield the respective classifier:

DD-1: $M_1 = -M_2 \approx \begin{bmatrix} 1.037 \\ 2.049 \\ 0 \end{bmatrix}, \Sigma_1 \approx I, \Sigma_2 \approx \begin{bmatrix} .481 & 0 & 0 \\ 0 & 3.131 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

The plot of the ellipsoid simulating the points and the linear classifier hyperplanes in the three dimensional space are depicted in Figure 1. The accuracy of the classifier was 96% for ω_1 and 97% for ω_2.

6.2 Pairwise Linear Classifiers with Different Means

To demonstrate the properties of the classifier satisfying the conditions of Theorem 6, we considered the pairwise linear classifier with different means. In this case, the diagonal covariance matrices differ only in two elements. These two elements in the first matrix have switched positions in the second covariance matrix. The remaining elements are identical in both covariance matrices. The mean vectors and covariance matrices estimated from 100 training samples are given below.

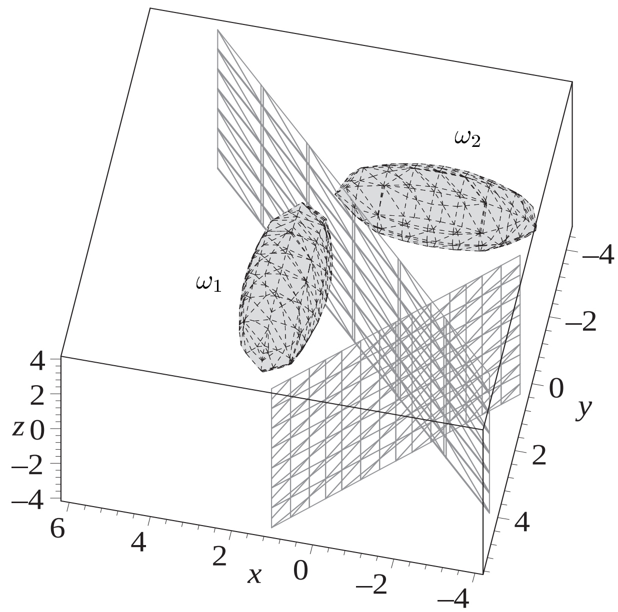

Figure 2: Example of a pairwise linear classifier with different means for the case described in Section 4. These classes corresponds to the data set DM-1.

$$\textbf{DM-1: } M_1 \approx \begin{bmatrix} 1.542 \\ 1.542 \\ 1.983 \end{bmatrix}, M_2 \approx \begin{bmatrix} -1.542 \\ -1.542 \\ 1.983 \end{bmatrix}, \Sigma_1 \approx \begin{bmatrix} .384 & 0 & 0 \\ 0 & 2.121 & 0 \\ 0 & 0 & .475 \end{bmatrix}, \Sigma_2 \approx \begin{bmatrix} 2.121 & 0 & 0 \\ 0 & .384 & 0 \\ 0 & 0 & .475 \end{bmatrix}$$

Using these parameters, the pairwise linear classifier was derived. The plot of the ellipsoid simulating the points and the linear classification hyperplanes are shown in Figure 2. With this classifier, we obtained an accuracy of 94% for ω_1 and 97% for ω_2.

6.3 Pairwise Linear Classifiers with Equal Means

We also tested our scheme for the case of the pairwise linear classifier with equal means, EM-1, for the generalized multi-dimensional Minsky's Paradox. This is the case in which we have coincident mean vectors, but covariance matrices as in the the case of DM-1. Two classes having parameters like these are proven in Theorem 7 to be optimally classified by a pair of hyperplanes. We obtained the following estimated mean vectors and covariance matrices from 100 training samples:

$$\textbf{EM-1: } M_1 = M_2 \approx \begin{bmatrix} -1.95 \\ 4.067 \\ 1.988 \end{bmatrix}, \Sigma_1 \approx \begin{bmatrix} 5.327 & 0 & 0 \\ 0 & .171 & 0 \\ 0 & 0 & .238 \end{bmatrix}, \Sigma_2 \approx \begin{bmatrix} .171 & 0 & 0 \\ 0 & 5.327 & 0 \\ 0 & 0 & .238 \end{bmatrix}$$

The shape of the overlapping classes and the linear classification function from these estimates are given in Figure 3. We evaluated the classifier with 100 randomly generated test points, and the accuracy was 82% for ω_1 and 85% for ω_2.

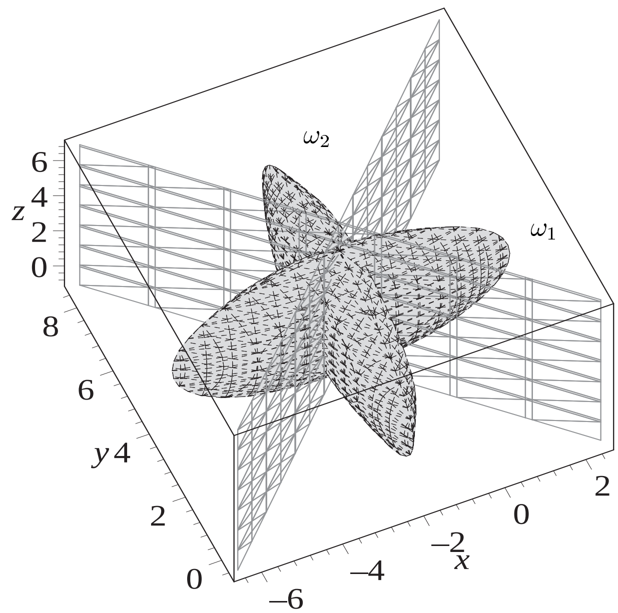

Figure 3: Example of a pairwise linear classifier with equal means for the case described in Section 5. The data set is EM-1. This resolves the generalized multi-dimensional Minsky's paradox.

Example	Accuracy for ω_1	Accuracy for ω_2
DD-1	96 %	97 %
DM-1	94 %	97 %
EM-1	83 %	88 %

Table 1: Accuracy of classification of 100 three dimensional random test points generated with the parameters of the examples presented above. The accuracy is given in percentage of points correctly classified.

6.4 Discussion of Results

Finally, we analyze the accuracy of the classifiers for the different cases discussed above. The accuracy of classification for the three cases is 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. Observe that the accuracy of DD-1 is very high. This case corresponds to the pairwise linear classifier when dealing with covariance matrices being the identity and another diagonal matrix in which two elements are not equal to unity, as shown in Theorem 5. The accuracy of the case in which the means are different and the covariance matrices as given in (8) and (9) (third row) is still very high. The fourth row corresponds to the case where the means are identical, referred to as EM-1. The accuracy is lower than that of the other cases but still very high, even though the classes overlap and the classification function is pairwise linear. This demonstrates the power of our scheme to resolve Minsky's Paradox in three dimensions.

7 Pairwise Linear Classifiers on Real Life Data

Real life data sets that satisfy the constraints given in (5), (12) and (13) are not very common, and in general, a classification scheme should be applicable to any data set, or eventually to a particular domain. Having demonstrated how our results are applicable to synthetic data sets, we now propose a method that substitutes the actual parameters of the data sets by approximated parameters for which the required constraints are satisfied. As we shall presently show, these parameters, in turn, are obtained by solving a constrained optimization problem. Using these principles, we present the empirical results of the experiments that we have conducted on real life data.

The approximation method and the empirical results are discussed for two-dimensional normally distributed random vectors. The evaluation of our optimal pairwise linear classifiers on real life data for the multi-dimensional case constitutes an open problem that is currently being investigated.

7.1 Parameter Approximation for 2-D Features

To solve the constrained optimization problem alluded to above, we propose a method that finds approximate parameters for normal distributions that satisfy the constraints of (4) and (5). The problem and its solution for the two-dimensional case are presented below.

Suppose that we are given two data sets containing labeled samples, $\mathcal{D}_1 = \{X_{1_1}, X_{1_2}, \ldots, X_{1_{N_1}}\}$ and $\mathcal{D}_2 = \{X_{2_1}, X_{2_2}, \ldots, X_{2_{N_2}}\}$, where X_{1_j} and X_{2_j} are drawn independently from their respective classes. We assume that the two classes corresponding to these samples are represented by two normally distributed random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, whose parameters, $\theta_1 = \begin{bmatrix} M_1 \\ \Sigma_1 \end{bmatrix}$ and $\theta_2 = \begin{bmatrix} M_2 \\ \Sigma_2 \end{bmatrix}$, are to be estimated. Our aim is to find the *maximum-likelihood* estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ that satisfy the constraints stated in (4) and (5).

We, first of all, use the maximum-likelihood method to estimate the parameters from $\mathcal{D}_1$ and $\mathcal{D}_2$, i.e. M_1, M_2, Σ_1 and Σ_2 [3]. By shifting the origin such that $M_1 = -M_2$, and performing the simultaneous diagonalization process discussed earlier, we obtain two transformed data sets, $\mathcal{D}'_1$ and $\mathcal{D}'_2$, as follows:

$$X'_{i_j} = \Psi_2^T \Lambda_1^{-\frac{1}{2}} \Phi_1^T \left\{ X_{i_j} - \left[M_2 + \frac{1}{2}(M_1 - M_2) \right] \right\}, \; i = 1, 2, \tag{14}$$

where Φ_1 is the eigenvector matrix of Σ_1, Λ_1 is the eigenvalue matrix of Σ_1, and Ψ_2 is the eigenvector matrix of Σ_{2_Z} obtained after performing the transformation $\mathbf{Z}_i = \Lambda_1^{-\frac{1}{2}} \Phi_1^T \mathbf{X}_i$, for $i = 1, 2$.

To clarify issues regarding the notation used here, we use "primed" variables with the "$'$" symbol to denote parameters or samples after the transformation of (14).

Once the samples are transformed into the new space, we proceed with the *constrained* maximum-likelihood estimation problem. The aim is to find the parameters $\{\hat{\theta}'_i\}$ that maximize the likelihood of $\{\theta_i\}$ with respect to the samples $\{\mathcal{D}'_i\}$, for $i = 1, 2$,

$$P(\mathcal{D}'_i | \theta'_i) = \prod_{j=1}^{N_i} P(X'_{i_j} | \theta'_i)\,, \tag{15}$$

while satisfying (4) and (5). This is equivalent to maximizing the *log-likelihood* function

$$l(\theta'_i) = \log P(\mathcal{D}'_i | \theta'_i)\,. \tag{16}$$

From (14), it can be seen that if $M_1 = -M_2$, then for $i = 1, 2$,

$$M_i' = \Psi_2^T \Lambda_1^{-\frac{1}{2}} \Phi_1^T \left\{ M_i - \left[M_2 + \frac{1}{2}(M_1 - M_2) \right] \right\} \qquad (17)$$

have the form $M_1' = -M_2'$, and hence our problem is to find Σ_1' and Σ_2' that maximize the likelihood function. Since, after the transformation, $\Sigma_1' = I$, we are left with the task of determining Σ_2' which we assume is diagonal, and which has the form $\Sigma_2' = \begin{bmatrix} a^{-1} & 0 \\ 0 & b^{-1} \end{bmatrix}$. Thus, (16) can be written as follows:

$$l(\theta_2') = \sum_{j=1}^{N_2} \log P(X_{2_j}' | \theta_2') \,. \qquad (18)$$

Substituting $P(X_{2_j}'|\theta_2')$ for the normal density function, and assuming that M_1' and M_2' have the form $M_1' = -M_2' = [r, s]^T$, we have the following constrained optimization problem:

Determine the value of $\hat{\Sigma}_2'$ that maximizes

$$l(\theta_2') = \sum_{j=1}^{N_2} \left\{ -\frac{1}{2} \log \left[(2\pi)^2 |\Sigma_2'| \right] - \frac{1}{2} (X_{2_j}' - M_2')^T (\Sigma_2')^{-1} (X_{2_j}' - M_2') \right\} , \qquad (19)$$

subject to the constraint

$$g(\theta_2') = a(1-b)r^2 + b(1-a)s^2 - \frac{1}{4}(ab - a - b + 1)\log(ab) = 0 \,. \qquad (20)$$

Using the Lagrange multiplier, this *constrained* optimization problem can be transformed into an *unconstrained* optimization problem, for which we have to find the solutions to the following system of equations. After the corresponding differentiation and some algebraic manipulations, and assuming that X_{2_j}' has the form $X_{2_j}' = \begin{bmatrix} x_{2_j}' \\ y_{2_j}' \end{bmatrix}$, we get:

$$\begin{aligned}
\frac{1}{2}\left[\frac{N_2}{a} - \sum_{j=1}^{N_2}(x_{2_j}' + r)^2\right] + \lambda\left[(1-b)r^2 - bs^2 - \frac{1}{4}(b-1)\log(ab) - \frac{1}{4a}(ab - a - b + 1)\right] &= 0 \qquad (21)\\
\frac{1}{2}\left[\frac{N_2}{b} - \sum_{j=1}^{N_2}(y_{2_j}' + s)^2\right] + \lambda\left[-ar^2 + (1-a)s^2 - \frac{1}{4}(a-1)\log(ab) - \frac{1}{4b}(ab - a - b + 1)\right] &= 0 \qquad (22)\\
a(1-b)r^2 + b(1-a)s^2 - \frac{1}{4}(ab - a - b + 1)\log(ab) &= 0 \qquad (23)
\end{aligned}$$

The system of equations given above has no closed-form algebraic solution for a, b and λ. Instead, it can be solved numerically for $a > 0$, $b > 0$, and where λ is a real number. The numerical solution that satisfies the constraints given in (4) and (5) are indeed $\hat{a}$ and $\hat{b}$. Using r, s, $\hat{a}$ and $\hat{b}$, the linear classifier can be easily derived. The corresponding equation to find this classifier can be found in [17].

7.2 Empirical Results

In this section, we discuss the empirical results that we obtained after performing classification tasks using the approximated optimal (pairwise) linear classifier on a real life data set drawn from the UCI machine learning

Classifier→ Class ↓	Fisher's	Pairwise
Benign	96%	98%
Malignant	91%	93%

Table 2: Accuracy in classification obtained from the Fisher's classification approach and the approximated optimal pairwise linear classifier on the WDBC data set.

repository[3], namely the Wisconsin Diagnostic Breast Cancer (WDBC) data set. It contains two data sets, one for the "benign" class and the other for the "malignant" class. Each sample contains 30 real-valued features representing the radius, texture, perimeter, area, smoothness, compactness, etc., computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Details of these features and how they are obtained can be found in [10].

The benign class and the malignant class data sets contain 357 and 212 samples respectively. From these data sets, we have randomly selected 100 samples for training, and 100 samples for the testing of each class. The training samples are different from those used in the testing. Thus, we have used the *two-folded cross validation* approach [3].

For the training and classification tasks we have composed 15 data subsets with all possible pairs of features obtained from the first six features. Our classifier was trained by following the procedure described in Subsection 7.1, thus yielding the approximated pairwise linear classifier for each subset. The classification task was performed using these linear classifiers, and a voting scheme was invoked. This scheme assigned a class to a testing sample for each of the 15 data subsets, and subsequently classified the sample to the class that yielded positive classification for eight or more voters.

To compare our result with a well-known scheme, we have also trained and performed classification tasks using the Fisher's classifier [3], and the same voting scheme on the same 15 data subsets.

The empirical results corresponding to the accuracy obtained in classification are shown in Table 2. Observe the superiority of the approximated pairwise linear classifier over Fisher's classifier. This again demonstrates the power of our scheme. This is further clarified from another perspective in Figure 4, where we plot the "area" and "smoothness" features.

The symbol '$\circ$' and '$+$' correspond to the benign class and the malignant class testing points respectively. These points were obtained after performing the transformation of (14). The circle and the ellipses correspond to the points with the same Mahalanobis distance (unity in this case), for the benign and the malignant classes respectively. Σ_2 is the covariance matrix obtained by using the *standard* maximum-likelihood estimate, and Σ_2' is the one obtained using the *constrained* maximum-likelihood introduced here. The parabola represents the optimal quadratic classifier obtained from the parameters estimated using the standard maximum-likelihood estimate. Observe that some points of the malignant class are misclassified when using Fisher's classifier. Although this is only one of the 15 classification tasks that we have performed, the superiority of our scheme over traditional linear classifiers (such as the Fisher's classification approach) is again corroborated. Observe that our linear classifier is closer to the optimal than Fisher's classifier in the area in which the samples are more likely to occur. Observe also that although our classifier is of the form of a pair of straight lines (one of them is not shown because it is outside the ranges of the graph), only the

[3] Available electronically at http://www.ics.uci.edu/~mlearn/MLRepository.html.

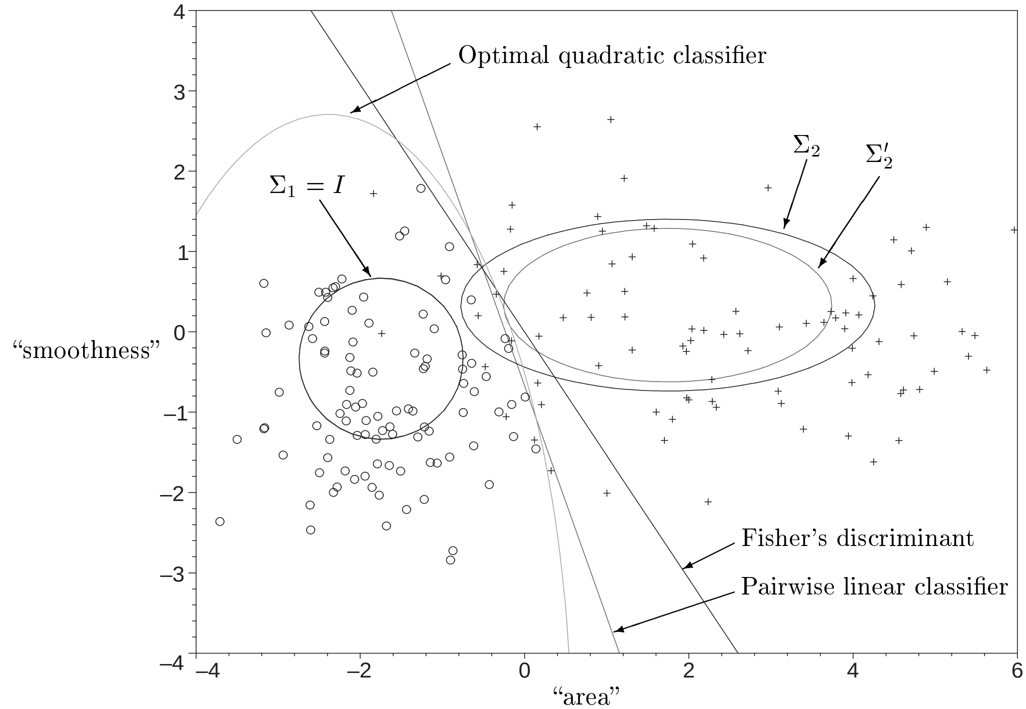

Figure 4: Optimal quadratic classifier, Fisher's classification function, and the approximated pairwise linear classifier used to classify the "area" and "smoothness" features from the WDBC data set.

one shown in the figure can be used. In spite of not utilizing the full capabilities of the pairwise classifier, we still maintain the superior classification accuracy.

We conclude this section by observing that although we have performed the classification tasks in the transformed space, this can be done in the original space by avoiding the burden of transforming each individual sample. In fact, by performing the inverse transformation on the classification functions, the classification can be done in the original space, and the classifiers are still linear. The details of this are omitted as they are straightforward.

7.3 Parameter Approximation for d-D Features

As seen above, the problem of approximating the parameters of the distributions so that they satisfy the constraints (4) and (5) for 2-dimensional normally distributed random vectors, is solved by formulating the problem as a *constrained* optimization problem. The same philosophy can be applied for approximating the parameters of the distributions so that *they* satisfy the d-dimensional versions of (4) and (5). It can be easily argued that the optimization for d-dimensional normally distributed random vectors can be solved in terms of the corresponding multi-dimensional *constrained* optimization problem.

By extending the concepts introduced for the 2-dimensional case, it is easy to see that the variables involved in the corresponding *unconstrained* optimization problem are the d elements in the diagonal of Σ_2' and a Lagrangian *vector*, namely, the $(d-1)$-dimensional vector $\boldsymbol{\lambda}= [\lambda_1, \ldots, \lambda_{d-1}]^T$. The issue now is one of solving for the optimal $2d-1$ variables, which constitute the non-zero diagonal elements of the covariance matrix, and the components of the Lagrangian, $\boldsymbol{\lambda}= [\lambda_1, \ldots, \lambda_{d-1}]^T$.

The solution to this problem is far from trivial. Indeed, we anticipate that various convergence problems will be encountered even if a numerical solution is attempted. This problem remains open and is currently being investigated.

8 Conclusions

In this paper, we have extended the theoretical framework of obtaining optimal pairwise linear classifiers for normally distributed classes to d-dimensional normally distributed random vectors, where $d > 2$.

We have determined the necessary and sufficient conditions for an optimal pairwise linear classifier when the covariance matrices are the identity and a diagonal matrix. In this case, we have formally shown that it is possible to find the optimal linear classifier by satisfying certain conditions specified in the planar projections of the various components.

In the second case, we have dealt with normally distributed classes having different mean vectors and with some special forms for the covariance matrices. When the covariance matrices differ only in two elements of the diagonal, and these elements are inverted in positions in the second covariance matrix, we have shown that the optimal classifier is a pair of hyperplanes only if the mean vectors differ in the two elements of these positions. The conditions for this have been formalized too.

The last case that we have considered is the generalized Minsky's paradox for multi-dimensional normally distributed random vectors. By a formal procedure, we have found that when the classes are overlapping and the mean vectors are coincident, under certain conditions on the covariance matrices, the optimal classifier is a pair of hyperplanes. This resolves the multi-dimensional Minsky's paradox.

We have provided some examples for each of the cases discussed above, and we have tested our classifier on some three dimensional normally distributed features. The classification accuracy obtained is very high, which is reasonable as the classifier is optimal in the Bayesian context. The degree of accuracy for the third case is not as high as that of the other cases, but is still impressive given the fact that we are dealing with significantly overlapping classes.

We have also presented a scheme from which the maximum likelihood pairwise linear classifier for two-dimensional random vectors can be estimated. The superiority of this approach over the traditional Fisher's classifier has been experimentally verified on a real life data set, namely the Wisconsin Diagnostic Breast Cancer data set. This superiority has been demonstrated numerically and graphically.

The extension of the approximation approach for multi-dimensional normal random vectors has been alluded to. Indeed, we have argued that this reduces to solving a constrained optimization problem in which the variables are the non-zero diagonal elements of the covariance matrix, and the components of the Lagrangian, which now becomes a *vector*, namely, $\boldsymbol{\lambda}= [\lambda_1, \ldots, \lambda_{d-1}]^T$.

The solution to this problem is far from trivial, and remains open. We anticipate that various convergence problems will be encountered even if a numerical solution is attempted.

References

[1] M. Aladjem. Linear Discriminant Analysis for Two Classes Via Removal of Classification Structure.

IEEE Trans. Pattern Analysis and Machine Intelligence, 19(2):187–192, 1997.

[2] Q. Du and C. Chang. A Linear Constrained Distance-based Discriminant Analysis for Hyperspectral Image Classification. *Pattern Recognition*, 34(2):361–373, 2001.

[3] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, Inc., New York, NY, 2nd edition, 2000.

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

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

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

[7] R. Lippman. An Introduction to Computing with Neural Nets. In Lau [6], pages 5–24.

[8] R. Lotlikar and R. Kothari. Adaptive Linear Dinensionality Reduction for Classification. *Pattern Recognition*, 33(2):185–194, 2000.

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

[10] O. Mangasarian, W. Street, and W. Wolberg. Breast Cancer Diagnosis and Prognosis via Linear Programming. *Operations Research*, 43(4):570–577, 1995.

[11] O. Murphy. Nearest Neighbor Pattern Classification Perceptrons. In Lau [6], pages 263–266.

[12] 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*, 21(2):159–173, 1999.

[13] S. Raudys. On Dimensionality, Sample Size, and Classification Error of Nonparametric Linear Classification. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(6):667–671, 1997.

[14] S. Raudys. Evolution and Generalization of a Single Neurone: I. Single-layer Perception as Seven Statistical Classifiers. *Neural Networks*, 11(2):283–296, 1998.

[15] S. Raudys. Evolution and Generalization of a Single Neurone: II. Complexity of Statistical Classifiers and Sample Size Considerations. *Neural Networks*, 11(2):297–313, 1998.

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

[17] Luis G. Rueda and B. John Oommen. On Optimal Pairwise Linear Classifiers for Normal Distributions: The Two-Dimensional Case. To appear in *IEEE Transations on Pattern Analysis and Machine Intelligence*. Also available as a Technical Report TR-01-01, School of Computer Science, Carleton University, Ottawa, Canada, 2001.

[18] Luis G. Rueda and B. John Oommen. The Foundational Theory of Optimal Bayesian Pairwise Linear Classifiers. In *Proceedings of the Joint IAPR International Workshops SSPR 2000 and SPR 2000*, pages 581–590, Alicante, Spain, August/September 2000. Springer.

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

[20] A. Webb. *Statistical Pattern Recognition*. Oxford University Press Inc., New York, 1999.