A New Approach to Multi-class Linear Dimensionality Reduction

Luis Rueda¹ and Myriam Herrera²

¹ Department of Computer Science and Center for Biotechnology
University of Concepción
Edmundo Larenas 215, Concepción, Chile
lrueda@inf.udec.cl

² Department and Institute of Informatics
National University of San Juan
Cereceto y Meglioli, San Juan, 5400, Argentina
mherrera@iinfo.unsj.edu.ar

Abstract. Linear dimensionality reduction (LDR) is quite important in pattern recognition due to its efficiency and low computational complexity. In this paper, we extend the two-class Chernoff-based LDR method to deal with multiple classes. We introduce the criterion, as well as the algorithm that maximizes such a criterion. The proof of convergence of the algorithm and a formal procedure to initialize the parameters of the algorithm are also given. We present empirical simulations on standard well-known multi-class datasets drawn from the UCI machine learning repository. The results show that the proposed LDR coupled with a quadratic classifier outperforms the traditional LDR schemes.

1 Introduction

The linear dimensionality reduction (LDR) problem is well-known in pattern recognition due to its linear time complexity and its simplicity to be implemented and understood. Various schemes that yield LDR to dimension one have been proposed, including the well known *Fisher's discriminant analysis* approach [5], *direct Fisher's discriminant analysis* [6], the *perceptron algorithm* (the basis of the back propagation neural network learning algorithms) [12], *piecewise recognition models* [11], *removal classification structures* [1], *adaptive linear dimensionality reduction* [9] (which outperforms Fisher's classifier for some data sets), *linear constrained distance-based classifier analysis* [4] (an improvement to Fisher's approach designed for hyperspectral image classification), *recursive Fisher's discriminant* [2], *pairwise linear classifiers* [16,17], and the *best hyperplane classifier* [13].

The problem we deal with consists of k classes, $\omega_1, \dots, \omega_k$, whose *a priori* probabilities are given by $p_1, \dots, p_k$, and which are represented by k n -dimensional normally distributed random vectors, $\mathbf{x}_1 \sim N(\mathbf{m}_1; \mathbf{S}_1), \dots, \mathbf{x}_k \sim N(\mathbf{m}_k; \mathbf{S}_k)$. The aim is to linearly transform $\mathbf{x}_1, \dots, \mathbf{x}_k$ into new normally distributed random vectors $\mathbf{y}_1, \dots, \mathbf{y}_k$ of dimension d , $d < n$, using a matrix $\mathbf{A}$ of order $d \times n$, in such a way that the classification error in the transformed space is as small as possible.

A typical approach to reduce the dimension of the data is principal component analysis (PCA) [5,18,21], but it better applies to unsupervised learning problems, since it takes the whole data as a “single” class, losing the discriminability power of labeled data. As opposed to this, linear discriminant analysis (LDA) aims to consider the class-conditional distributions to maximize the separability of the data in the transformed space. One of such approaches is Fisher’s discriminant (FD). Let $\mathbf{S}_E = \sum_{i=1}^k p_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t$, where $\mathbf{m} = \sum_{i=1}^k p_i \mathbf{m}_i$, be the *between-class* scatter matrix, and $\mathbf{S}_W = \sum_{i=1}^k p_i \mathbf{S}_i$ be the *within-class* scatter matrix. The FD approach aims to find a matrix $\mathbf{A}$ that maximizes the following criterion function [5]:

$$J_{FD}(\mathbf{A}) = \text{tr} \{ (\mathbf{A} \mathbf{S}_W \mathbf{A}^t)^{-1} (\mathbf{A} \mathbf{S}_E \mathbf{A}^t) \}, \quad (1)$$

and which is obtained by finding the d eigenvectors (whose eigenvalues are the largest ones) of $\mathbf{S}_W^{-1} \mathbf{S}_E$.

The Loog-Duin LD criterion for the multi-class problem aims to find the transformation $d \times n$ matrix $\mathbf{A}$ that maximizes the following function [8]:

$$J_{LD}(\mathbf{A}) = \sum_{i=1}^{k-1} \sum_{j=i+1}^k p_i p_j \text{tr} \left\{ (\mathbf{A} \mathbf{S}_W \mathbf{A}^t)^{-1} \mathbf{A} \mathbf{S}_W^{\frac{1}{2}} \left[(\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}})^{-\frac{1}{2}} \mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{E_{ij}} \mathbf{S}_W^{-\frac{1}{2}} (\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}})^{-\frac{1}{2}} + \frac{1}{\pi_i \pi_j} \left(\log(\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}}) - \pi_i \log(\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_i \mathbf{S}_W^{-\frac{1}{2}}) - \pi_j \log(\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_j \mathbf{S}_W^{-\frac{1}{2}}) \right) \right] \mathbf{S}_W^{\frac{1}{2}} \mathbf{A}^t \right\}, \quad (2)$$

where $\mathbf{S}_{E_{ij}} = (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^t$, $\pi_i = \frac{p_i}{p_i + p_j}$, $\pi_j = \frac{p_j}{p_i + p_j}$, and $\mathbf{S}_{ij} = \pi_i \mathbf{S}_i + \pi_j \mathbf{S}_j$. The LD criterion is maximized as it is done for the two-class case, by finding the matrix $\mathbf{A}$ composed of the d eigenvectors (whose eigenvalues are the largest ones) of the following matrix:

$$\mathbf{S}_{LD} = \sum_{i=1}^{k-1} \sum_{j=i+1}^k p_i p_j \mathbf{S}_W^{-1} \mathbf{S}_W^{\frac{1}{2}} \left[(\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}})^{-\frac{1}{2}} \mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{E_{ij}} \mathbf{S}_W^{-\frac{1}{2}} (\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}})^{-\frac{1}{2}} + \frac{1}{\pi_i \pi_j} \left(\log(\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_{ij} \mathbf{S}_W^{-\frac{1}{2}}) - \pi_i \log(\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_i \mathbf{S}_W^{-\frac{1}{2}}) - \pi_j \log(\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_j \mathbf{S}_W^{-\frac{1}{2}}) \right) \right] \mathbf{S}_W^{\frac{1}{2}}, \quad (3)$$

The traditional classification problem has usually been solved by maximizing the separability between the underlying distributions using different criteria. FD aims to minimize the error by maximizing the Mahalanobis distance between distributions, resulting in an optimal criterion when the covariance matrices are equal. In case the covariances are different, the optimal classifier is quadratic; the linear classification results in maximizing the separability between the distributions by generalizing the Mahalanobis distance [7]. On the other hand, the LD criterion utilizes, as pointed out above, a directed distance matrix, which is incorporated in Fisher’s criterion assuming the within-class scatter matrix is the identity.

In this paper, we extend the two-class LDR criterion as proposed in [14], and which takes advantage of the properties of the Chernoff distance to maximize the separability of the distributions in the transformed space. Since we are assuming the original distributions are normal, the distributions in the transformed space are also normal. Thus, the Bayes classifier in the transformed space is quadratic and deriving a closed-form expression for the classification error is not possible. However, the probability of classification error in the transformed space can be bounded in terms of the Chernoff distance between two distributions given by:

$$k(\beta, \mathbf{A}) = \frac{\beta(1-\beta)}{2} (\mathbf{A}\mathbf{m}_1 - \mathbf{A}\mathbf{m}_2)^t [\beta \mathbf{A}\mathbf{S}_1 \mathbf{A} + (1-\beta) \mathbf{A}\mathbf{S}_2 \mathbf{A}]^{-1} (\mathbf{A}\mathbf{m}_1 - \mathbf{A}\mathbf{m}_2) + \frac{1}{2} \log \frac{|\beta \mathbf{A}\mathbf{S}_1 \mathbf{A} + (1-\beta) \mathbf{A}\mathbf{S}_2 \mathbf{A}|}{|\mathbf{A}\mathbf{S}_1 \mathbf{A}|^\beta |\mathbf{A}\mathbf{S}_2 \mathbf{A}|^{1-\beta}}. \quad (4)$$

The larger the value of $k(\beta, \mathbf{A})$ is, the smaller the bound for the classification error is, and hence, in this paper, we propose to maximize (4). To clarify this, we note that the FD criterion also aims to maximize the separability between distributions in the transformed space, but coincides with the optimal classifier only when the latter is linear, i.e. when the covariance matrices are coincident, a rare case. As observed above, the LD criterion utilizes the Chernoff distance in its directed distance matrix but in the original space. This criterion, however, does not optimize such a distance in the *transformed* space, as it can be observed in the following example.

Consider two normally distributed random vectors, $\mathbf{x}_1 \sim N(\mathbf{m}_1, \mathbf{S}_1)$ and $\mathbf{x}_2 \sim N(\mathbf{m}_2, \mathbf{S}_2)$, whose parameters are: $\mathbf{m}_1 = [0.5001, 0.4947]^t$, $\mathbf{m}_2 = [2.1069, 1.4324]^t$, $\mathbf{S}_1 = [0.8205, 0.4177; 0.4177, 2.8910]$, $\mathbf{S}_2 = [5.1150, -4.3990; -4.3990, 5.7119]$, $p_1 = p_2 = 0.5$. Consider also a linear transformation $\mathbf{y} = \mathbf{A}\mathbf{x}$ to the one-dimensional space, i.e. $\mathbf{A}$ is of order 1×2 , namely a two-dimensional vector. As shown later in the paper, we can just “rotate” $\mathbf{A}$ and produce different values for the Chernoff distance in the transformed space, and *only one* value for each angle. Thus, in Fig. 1, we plot three different criteria for all possible values of the angle θ between $\mathbf{A}$ and $[0, 1]^t$, including $J_F(\mathbf{A})$ computed as in (1), $J_{LD}(\mathbf{A})$ computed as in (2), and the Chernoff distance in the transformed one-dimensional space computed as in (4), where $\beta = 1/2$. As we will also see later, we note that maximizing the criterion $J_F(\mathbf{A})$ or $J_{LD}(\mathbf{A})$ does not necessarily imply maximizing the Chernoff distance in the transformed space (as our criterion aims to), and hence minimizing the classification error. Also, we observe that the $k(\beta, \mathbf{A})$ function has more than one peak and so, as shown later, this makes it difficult to find a closed-form expression for the optimal solution.

2 The Proposed LDR Criterion

In this section, we formalize the proposed LDR criterion for the multi-class case, which aims to maximize the Chernoff distance in the transformed space. We suppose that we are dealing with k classes, $\{\omega_i\}$, whose *a priori* probabilities are given by $\{p_i\}$, with $i = 1, 2, \dots, k$. By following the notation used above, we define $\mathbf{S}_{E_{ij}} = (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^t$ and $\mathbf{S}_{W_{ij}} = \pi_i \mathbf{S}_i + \pi_j \mathbf{S}_j$ as the between-class and within-class scatter

Fig. 1. Plot of three different dimensionality reduction functions, namely Fisher's, Loog-Duin's, and the Chernoff distance in the transformed space, for a two-dimensional to one-dimensional reduction example. The x -axis represents the different angles of the transformation vector $\mathbf{A}$.

matrices respectively, where $\mathbf{m}_i$ and $\mathbf{S}_i$ are the mean and covariance matrices for ω_i respectively, $\pi_i = p_i/(p_i + p_j)$ and $\pi_j = p_j/(p_i + p_j)$.

The extension of the two-class classification is not straightforward, as there is no general formula for the Chernoff distance between more than two distributions. This is also the case of other classifiers, such as the well-known support vector machines or kernel-based classifiers, for which majority votes of $k(k-1)/2$ decisions are among the most efficient schemes [19], as opposed to other schemes like one-against-all or all-at-once, which suffer the problem of yielding unclassifiable regions [5,20]. In our case, however, it is natural to maximize the *weighted* sum of pairwise Chernoff distances between classes ω_i and ω_j , for all $i = 1, \dots, k-1, j = i+1, \dots, k$. The “weights” used for the pairwise class criterion are given by the normalized joint prior probabilities between classes ω_i and ω_j , $\pi_i \pi_j$. Thus, the criterion that we propose for multiple classes consists of finding the transformation $\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a matrix of order $d \times n$ that maximizes the function, where we have adopted a natural way to set the value of β , as $\beta = \pi_1$ and $1 - \beta = \pi_2$:

$$J_c^*(\mathbf{A}) = \sum_{i=1}^{k-1} \sum_{j=i+1}^k J_{c_{ij}}^*(\mathbf{A}), \quad (5)$$

with:

$$J_{c_{ij}}^*(\mathbf{A}) = \text{tr} \left\{ \pi_i \pi_j (\mathbf{A} \mathbf{S}_{W_{ij}} \mathbf{A}^t)^{-1} \mathbf{A} \mathbf{S}_{E_{ij}} \mathbf{A}^t \right. \\ \left. + \log(\mathbf{A} \mathbf{S}_{W_{ij}} \mathbf{A}^t) - \pi_i \log(\mathbf{A} \mathbf{S}_i \mathbf{A}^t) - \pi_j \log(\mathbf{A} \mathbf{S}_j \mathbf{A}^t) \right\} \quad (6)$$

Taking the Chernoff distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ as in (4), and doing some algebraic manipulations, it can be written in terms of $\text{tr}\{\dots\}$ as in (6) (cf. [15]). We show there always exists an orthogonal matrix $\mathbf{Q}$ for which the Chernoff distance in the new space is the same as that of using the matrix $\mathbf{A}$. The proof of the lemma can be found in [15].

Lemma 1. *Let $\mathbf{A}$ be any real $d \times n$ matrix, $d \leq n$, whose rows are linearly independent, and $J_c^*(\mathbf{A})$ be defined as in (5). Then:*

$$\max_{\{\mathbf{A}\}} J_c^*(\mathbf{A}) = \max_{\{\mathbf{A}: \mathbf{A}\mathbf{A}^t = \mathbf{I}_d\}} J_c^*(\mathbf{A}). \quad (7)$$

Proof (Sketch). Let $\mathbf{A}$ be any matrix of order $d \leq n$ whose rows are linearly independent. Then, by applying the QR decomposition [3], $\mathbf{A}$ can be decomposed as follows $\mathbf{A} = \mathbf{R}\mathbf{Q}$, and it follows that $J_c^*(\mathbf{A}) = J_c^*(\mathbf{Q})$. In this way, we ensure that $\max_{\{\mathbf{A}\}} J_c^*(\mathbf{A}) = \max_{\{\mathbf{A}: \mathbf{A}\mathbf{A}^t = \mathbf{I}_d\}} J_c^*(\mathbf{A})$. $\square$

To obtain $\mathbf{A}$, we first find the gradient matrix given by the first-order necessary condition. That is, we differentiate (5) with respect to $\mathbf{A}$, resulting in:

$$\nabla J_c^*(\mathbf{A}) = \frac{\partial J_c^*(\mathbf{A})}{\partial \mathbf{A}} = \frac{\partial}{\partial \mathbf{A}} \sum_{i=1}^{k-1} \sum_{j=i+1}^k J_{c_{ij}}^*(\mathbf{A}) = \sum_{i=1}^{k-1} \sum_{j=i+1}^k \nabla J_{c_{ij}}^*(\mathbf{A}), \quad (8)$$

where

$$\begin{aligned} \nabla J_{c_{ij}}^*(\mathbf{A}) = & 2\pi_i \pi_j [\mathbf{S}_{E_{ij}} \mathbf{A}^t (\mathbf{A} \mathbf{S}_{W_{ij}} \mathbf{A}^t)^{-1} \\ & - \mathbf{S}_{W_{ij}} \mathbf{A}^t (\mathbf{A} \mathbf{S}_{W_{ij}} \mathbf{A}^t)^{-1} (\mathbf{A} \mathbf{S}_{E_{ij}} \mathbf{A}^t) (\mathbf{A} \mathbf{S}_{W_{ij}} \mathbf{A}^t)^{-1}]^t \\ & + 2 [\mathbf{S}_{W_{ij}} \mathbf{A}^t (\mathbf{A} \mathbf{S}_{W_{ij}} \mathbf{A}^t)^{-1} - \pi_i \mathbf{S}_i \mathbf{A}^t (\mathbf{A} \mathbf{S}_i \mathbf{A}^t)^{-1} - \pi_j \mathbf{S}_j \mathbf{A}^t (\mathbf{A} \mathbf{S}_j \mathbf{A}^t)^{-1}]^t \end{aligned} \quad (9)$$

To maximize $J_c^*(\mathbf{A})$ we propose Algorithm **ChernoffLDA_Multi** given below.

Algorithm **ChernoffLDA_Multi**

Input: Threshold τ

begin

$\mathbf{A}^{(0)} \leftarrow \max_{\mathbf{A}} \{J_{c_{12}}^*(\mathbf{A}_{FD}), J_{c_{12}}^*(\mathbf{A}_{LD})\}$ // Maximum of FD or LD

$k \leftarrow 0$

repeat

$\eta_k \leftarrow \max_{\eta > 0} \phi_k(\eta)$

$\mathbf{B} \leftarrow \mathbf{A}^{(k)} + \eta_k \sum_{i=1}^{k-1} \sum_{j=i+1}^k \nabla J_{c_{ij}}^*(\mathbf{A}^{(k)})$

Decompose $\mathbf{B}$ into $\mathbf{R}$ and $\mathbf{Q}$

$\mathbf{A}^{(k+1)} \leftarrow \mathbf{Q}$

$k \leftarrow k + 1$

until $|J_{c_{ij}}^*(\mathbf{A}^{(k-1)}) - J_{c_{ij}}^*(\mathbf{A}^{(k)})| < \tau$

return $\mathbf{A}^{(k)}, \sum_{i=1}^{k-1} \sum_{j=i+1}^k J_{c_{ij}}^*(\mathbf{A}^{(k)})$

end

At each step, the value of η_k is obtained as the one that maximizes the following function:

$$\phi_k(\eta) = J_c^*(\mathbf{A}^{(k)} + \eta \nabla J_c^*(\mathbf{A}^{(k)})). \quad (10)$$

The algorithm initializes the matrix $\mathbf{A}$ by using the result of either FD or LD, depending on the one that gives the maximum value of the Chernoff distance in the transformed space. At each step, the matrix $\mathbf{A}$ is decomposed using the RQ decomposition, as shown in Lemma 1. Let us now see how the algorithm converges, using the convergence properties of the two-class case [15].

Theorem 1. *Let $\{\mathbf{A}^{(k)}\}_{k=1}^\infty$ be the sequence of matrices generated by Algorithm **Chernoff-LDA-Multi**. If $\nabla J_c^*(\mathbf{A}^{(k)}) \neq 0$, then $J_c^*(\mathbf{A}^{(k)}) < J_c^*(\mathbf{A}^{(k+1)})$. Otherwise, the algorithm terminates.*

Proof (Sketch). From (5), we have that $J_c^*(\mathbf{A}^{(k)}) = \sum_{i=1}^{k-1} \sum_{j=i+1}^k J_{c_{ij}}^*(\mathbf{A}^{(k)})$. By using the result of the two-class case [15], it is true that $J_{c_{ij}}^*(\mathbf{A}^{(k)}) < J_{c_{ij}}^*(\mathbf{A}^{(k+1)})$ if $\nabla J_{c_{ij}}^*(\mathbf{A}^{(k)}) \neq 0$. Also, since for all $i, j = 1, \dots, k$, $J_{c_{ij}}^*(\mathbf{A}^{(k+1)}) \geq 0$, we have that $J_c^*(\mathbf{A}^{(k)}) = \sum_{i=1}^{k-1} \sum_{j=i+1}^k J_{c_{ij}}^*(\mathbf{A}^{(k)}) < \sum_{i=1}^{k-1} \sum_{j=i+1}^k J_{c_{ij}}^*(\mathbf{A}^{(k+1)}) = J_c^*(\mathbf{A}^{(k+1)})$.

On the other hand, if $\nabla J_c^*(\mathbf{A}^{(k)}) = 0$, we have that $\mathbf{A}^{(k)} = \mathbf{A}^{(k+1)}$. Since we also know that $\tau > 0$, then we have that $|J_c^*(\mathbf{A}^{(k+1)}) - J_c^*(\mathbf{A}^{(k)})| < \tau$, and hence the algorithm terminates. The complete proof can be found in [15]. $\square$

We note that it is quite important to compute, at each step, a value of η_k that maximizes the function ϕ_k given in (10). Computing the first derivative of this function results in:

$$\frac{d\phi_k}{d\eta}(\eta) = [\nabla J_c^*(\mathbf{A}^{(k)} + \eta \nabla J_c^*(\mathbf{A}^{(k)}))] \cdot \nabla J_c^*(\mathbf{A}^{(k)}). \quad (11)$$

Using this expression, we compute $\eta^{(j+1)}$ as follows:

$$\eta^{(j+1)} = \eta^{(j)} - \frac{\eta^{(j)} - \eta^{(j-1)}}{\frac{d\phi_{k12}}{d\eta}(\eta^{(j)}) - \frac{d\phi_{k12}}{d\eta}(\eta^{(j-1)})} \frac{d\phi_{k12}}{d\eta}(\eta^{(j)}), \quad (12)$$

where the values of $\eta^{(0)}$ and $\eta^{(1)}$ are obtained using the following theorem.

Theorem 2. *Let $\phi_k : \mathbb{R}^{d \times n} \rightarrow \mathbb{R}$ be the continuously differentiable function defined in (10), where $J_c^*(\cdot)$ is defined in (5), and whose first derivative is given by (8).*

Then, the initial values of the secant method are given by $\eta_0 = 0$ and

$$\eta_1 = \frac{d^2\epsilon - d}{\text{tr}\{\mathbf{A}^{(k)}[\nabla J_c^*(\mathbf{A}^{(k)})]^t\}}, \quad (13)$$

where $\epsilon = \cos \theta$, with θ being the angle between $\mathbf{A}^{(k)}$ and $[\mathbf{A}^{(k)} + \eta_k \nabla J_c^*(\mathbf{A}^{(k)})]$.

Proof (Sketch). From Lemma 1, we have that $^1 [\mathbf{A}^{(k)}][\mathbf{A}^{(k)}]^t = [\mathbf{A}^{(k+1)}][\mathbf{A}^{(k+1)}]^t = \mathbf{I}_d$, and hence both matrices are located in the boundary of the environment of zero (null

¹ Note that we are considering that $\mathbf{A}^{(k)}$ is the matrix $\mathbf{Q}$, which is orthogonal and obtained by means of the RQ decomposition.

matrices) of radius d in the matrix space. The angle difference between the matrices at steps ' k ' and ' $k + 1$ ' are given by:

$$\cos \theta = \frac{1}{d} + \frac{\eta_k \text{tr}\{[\mathbf{A}^{(k)}][\nabla J_c^*(\mathbf{A}^{(k)})]^t\}}{d^2} \quad (14)$$

The result follows by setting $\cos \hat{\theta} = \epsilon$, and observing that the secant leads to a maximum. The complete proof can be found in [15]. $\square$

To conclude this section, we emphasize that the geometric interpretation of $\mathbf{A}^{(k)}$ that we did for the two-class case is also valid for the multi-class case [14]. Again, the rows of the matrix $\mathbf{A}^{(k)}$ are "rotated" independently using the same scalar, n_k . In the same way, we could also imagine a more general mechanism for updating $\mathbf{A}^{(k)}$ using a vector, $\boldsymbol{\eta}$, instead of a scalar. This problem is a possible future avenue for extending this work.

3 Empirical Results

In order to evaluate the classification performance of the proposed criterion, we present an empirical analysis of the classification accuracy and Chernoff distance in the transformed space on real-life data drawn from the UCI machine learning repository [10]. Three LDR techniques are compared, namely FD and LD as discussed in Section 1, and the proposed method, as presented in Section 2, namely RH. In order to analyze the classification power of the LDR techniques, two classifiers are used in the transformed space, the linear (L) and quadratic (Q) classifiers. The datasets involved in the experiments are Iris plants, Letter recognition, Pendigits, Thyrod gland, Wine, Glass identification, Landsat satellite, and Vowel context. In order to avoid ill-conditioned covariance matrices, we have applied principal component analysis (PCA) to Glass and reduced the data from dimension nine to eight, and removed class '6' to apply the 10-fold cross validation method. As in the two-class case, we trained the three LDR techniques, namely FD, LD and RH, followed by a quadratic or linear classifier, in a 10-fold cross-validation experiment. The average classification errors are given in Table 1, in which d^* indicates the dimension that yields the lowest error rate. For each classifier, quadratic and linear, the LDR method(s) that produce(s) the lowest error rate is(are) marked with a '*'. For the quadratic classifier, we note that the RH method yields to lower error rate in four times, while FD and LD reach the best error rate in three times. For the linear classifier, both FD and LD are superior to RH. This is as expected, since the RH aims to maximize the Chernoff distance in the transformed space, which is related to the Bayesian quadratic classifier, but not necessarily to the linear classifier. Also, the error rates obtained using RH and the quadratic classifier are in all cases (except in Iris and Glass) much smaller than the corresponding rate for the linear classifier, independently of the LDR technique coupled with the latter.

Fig. 2 (a) shows the plots for the error rates obtained after coupling the LDR techniques with the quadratic classifier on the Pendigits dataset. The plot shows the error rates for dimensions $d = 1, \dots, 15$. After dimension $d = 3$, the error rates corresponding to LD and RH become relatively lower than that of FD, and in general, the error rate of RH is lower than that of LD, showing the superiority of the former over the latter on

Table 1. Average error rates obtained from the three LDR techniques coupled with quadratic and linear classifiers on the multi-class datasets drawn from the UCI machine learning repository

Dataset	FD+Q	d^*	LD+Q	d^*	RH+Q	d^*	FD+L	d^*	LD+L	d^*	RH+L	d^*
Iris	0.0266	1	0.0200*	1	0.0200*	1	0.0200*	1	0.0200*	1	0.0200*	1
Letter	0.1186	15	0.1169*	15	0.1200	15	0.2973*	15	0.2997	14	0.3048	14
Pendigit	0.0493	9	0.0232	15	0.0223*	14	0.1240*	9	0.1291	15	0.1301	15
Thyroid	0.0329*	1	0.0422	1	0.0374	4	0.0935*	1	0.0935*	4	0.0935*	1
Wine	0.0111	2	0.0055*	2	0.0055*	2	0.0222	2	0.0111*	5	0.0164	2
Glass	0.4333*	2	0.4468	4	0.4496	4	0.3387*	4	0.3968	6	0.4106	6
Landsat	0.1518*	5	0.1597	18	0.1543	5	0.1722	5	0.1705*	10	0.1857	5
Vowel	0.3788	9	0.3222	6	0.3040*	6	0.4656	6	0.4444*	2	0.4444*	2

the quadratic classifier. For the linear classifier, the error rates for different dimensions are shown in Fig. 2 (b). After dimension $d = 4$, the error rate for FD is relatively lower than that of LD and RH, while the error rate for the latter is higher than that of LD. In all cases, the error rate tends to decrease as the reduced dimension becomes larger. Fig. 3 (a) shows the error rates for the quadratic classifier on the Vowel dataset. The error rate of FD is quite larger than that of the other two methods, while the error rate for RH starts decreasing to reach its minimum at $d = 6$, to increase a little bit on greater dimensions. Although LD exhibits a similar behavior, it does not reach the minimum as RH, which demonstrates that the former does not capture all the information for the Chernoff distance in the transformed space. For the linear classifier on the Vowel dataset, Fig. 3 (b), LD and RH lead to a minimum error rate for most of the values of d , but they show the minimum is reached early at dimension $d = 2$. Also, the peak shown by RH at $d = 3$ shows how maximizing the Chernoff distance in the transformed space does not necessarily imply minimizing the error rate for the linear classifier. Note, however, that the converse is not always true for the quadratic classifier, but the behavior exhibited in the figures shows how the RH criterion is more appropriate for a quadratic classifier in

(a) Quadratic classifier.

(b) Linear classifier.

Fig. 2. Error rates for all reduced dimensions on Pendigits

Fig. 3. Error rates for all reduced dimensions on the Vowel dataset

the transformed space, because it captures more information on the Chernoff distance which provides a good approximation of the error rate for that classifier.

4 Conclusion

We have introduced a new criterion for linear dimensionality reduction (LDR), which, unlike previous approaches such as Fisher's and Loog-Duin's, aims to maximize the Chernoff distance in the transformed space. We have extended the corresponding criterion for the multi-class case, and provided proofs for the convergence of the optimizing gradient-based algorithms. The empirical results on real-life datasets show the superiority of RH over the existing FD and LD criteria, mainly when the techniques are coupled with the *quadratic* classifier, demonstrating the importance of maximizing the Chernoff distance in the transformed space for such a classifier. One of the possible extensions for this work is to use a vector $\boldsymbol{\eta}$ to update the matrix $\mathbf{A}^{(k)}$, instead of a scalar. This is a problem that we are currently investigating.

Acknowledgments. This research has been supported by the Chilean National Council for Technological and Scientific Research, FONDECYT grant No. 1060904, and the Institute of Informatics, National University of San Juan.

References

1. M. Aladjem. Linear Discriminant Analysis for Two Classes Via Removal of Classification Structure. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(2):187–192, 1997.
2. T. Cooke. Two Variations on Fisher's Linear Discriminant for Pattern Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):268–273, 2002.
3. D. Harville. *Matrix Algebra from a Statisticians Perspective*. Springer-Verlag, New York, 1997.
4. Q. Du and C. Chang. A Linear Constrained Distance-based Discriminant Analysis for Hyperspectral Image Classification. *Pattern Recognition*, 34(2):361–373, 2001.

5. R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, Inc., New York, NY, 2nd edition, 2000.
6. H. Gao and J. Davis. Why Direct LDA is not Equivalent to LDA. *Pattern Recognition*, 39:1002–1006, 2006.
7. M. Herrera and R. Leiva. Generalización de la Distancia de Mahalanobis para el Análisis Discriminante Lineal en Poblaciones con Matrices de Covarianza Desiguales. *Revista de la Sociedad Argentina de Estadística*, 3(1-2):64–86, 1999.
8. M. Loog and P.W. Duin. Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):732–739, 2004.
9. R. Lotlikar and R. Kothari. Adaptive Linear Dimensionality Reduction for Classification. *Pattern Recognition*, 33(2):185–194, 2000.
10. D. Newman, S. Hettich, C. Blake, and C. Merz. UCI repository of machine learning databases, 1998. University of California, Irvine, Dept. of Computer Science.
11. A. Rao, D. Miller, K. Rose, , and A. Gersho. A Deterministic Annealing Approach for Parsimonious Design of Piecewise Regression Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(2):159–173, 1999.
12. S. Raudys. Evolution and Generalization of a Single Neurone: I. Single-layer Perception as Seven Statistical Classifiers. *Neural Networks*, 11(2):283–296, 1998.
13. L. Rueda. Selecting the Best Hyperplane in the Framework of Optimal Pairwise Linear Classifiers. *Pattern Recognition Letters*, 25(2):49–62, 2004.
14. L. Rueda and M. Herrera. A New Linear Dimensionality Reduction Technique based on Chernoff Distance. In *Proceedings of the 10th Iberoamerican Conference on Artificial Intelligence*, Ribeiro Prieto, Brazil, October 2006, pp. 299–308.
15. L. Rueda and M. Herrera. Linear Discriminant Analysis by Maximizing the Chernoff Distance in the Transformed Space. *Submitted for Publication*, 2006. Electronically available at <http://www.inf.udec.cl/~lrueda/papers/ChernoffLDAInl.pdf>.
16. L. Rueda and B. J. Oommen. On Optimal Pairwise Linear Classifiers for Normal Distributions: The Two-Dimensional Case. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):274–280, 2002.
17. L. Rueda and B. J. Oommen. On Optimal Pairwise Linear Classifiers for Normal Distributions: The d -Dimensional Case. *Pattern Recognition*, 36(1):13–23, January 2003.
18. S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Elsevier Academic Press, third edition, 2006.
19. D. Tsujinishi and S. Abe. Fuzzy Least Squares Support Vector Machines for Multi-class Problems. *Neural Networks*, 16:785–792, 2003.
20. V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
21. A. Webb. *Statistical Pattern Recognition*. John Wiley & Sons, N.York, second edition, 2002.