Non-iterative Heteroscedastic Linear Dimension Reduction for Two-Class Data
From Fisher to Chernoff

Marco Loog[1] and Robert P. W. Duin[2]

[1] Image Sciences Institute, University Medical Center Utrecht
Utrecht, The Netherlands
`marco@isi.uu.nl`
[2] Pattern Recognition Group, Department of Applied Physics
Delft University of Technology, Delft, The Netherlands

Abstract. Linear discriminant analysis (LDA) is a traditional solution to the linear dimension reduction (LDR) problem, which is based on the maximization of the between-class scatter over the within-class scatter. This solution is incapable of dealing with heteroscedastic data in a proper way, because of the implicit assumption that the covariance matrices for all the classes are equal. Hence, discriminatory information in the difference between the covariance matrices is not used and, as a consequence, we can only reduce the data to a single dimension in the two-class case. We propose a fast non-iterative eigenvector-based LDR technique for heteroscedastic two-class data, which generalizes, and improves upon LDA by dealing with the aforementioned problem. For this purpose, we use the concept of directed distance matrices, which generalizes the between-class covariance matrix such that it captures the differences in (co)variances.

1 Introduction

Probably the most well-known approach to supervised linear dimension reduction (LDR), or feature extraction, is *linear discriminant analysis* (LDA). This traditional and simple technique was developed by Fisher [6] for the two-class case, and extended by Rao [16] to handle the multi-class case. In LDA, a $d \times n$ transformation matrix that maximizes the *Fisher criterion* is determined. This criterion gives, for a certain linear transformation, a measure of the between-class scatter over the within-class scatter (cf. [7,9]). An attractive feature of LDA is the fast and easy way to determine this optimal linear transformation, merely requiring simple matrix arithmetics like addition, multiplication, and eigenvalue decomposition. A limitation of LDA is its incapability of dealing with heteroscedastic data, i.e., data in which classes do not have equal covariance matrices.

This paper focusses on the generalization of the Fisher criterion to the heteroscedastic case in order to come to heteroscedastic linear dimension reduction

(HLDR). We restrict our attention to two-class data, e.g. where pattern classes can typically be divided into good or bad, 0 or 1, benign or malignant, on or off, foreground or background, yin or yang, etc. With this kind of data the limitation of LDA is very obvious: A reduction to only a single dimension is possible (cf. [7]).

Our generalization takes into account the discriminatory information that is present in the difference of the covariance matrices. This is done by means of directed distance matrices (DDMs) [12], which are generalizations of the between-class covariance matrix. This between-class covariance matrix, as used in LDA, merely takes into account the discriminatory information that is present in the differences between class means and can be associated with the Euclidean distance.

The specific heteroscedastic generalization of the Fisher criterion, we study more closely in Section 2, is based on the *Chernoff distance* [2,3]. This measure of affinity of two densities considers mean differences as well as covariance differences—as opposed to the Euclidean distance—and can be used to extend LDA, while retaining the attractive feature of fast and easily determining a dimension reducing transformation. Furthermore, we are able to reduce the data to any dimension d smaller than n and not only to a single dimension. We call our HLDR criterion the *Chernoff criterion*.

Several alternative approaches to HLDR are known, of which we mention the following ones. See also [14].

In the two-class case, under the assumptions that both classes are normally distributed and that one wants a reduction to one dimension, Kazakos [10] reduces the LDR problem to a one-dimensional search problem. Finding the optimal solution for this search problem, is equivalent to finding the optimal linear feature. The work of Kazakos is closely related to [1].

Three other HLDR approaches for two-class problems, that generalize upon Fisher, were proposed in [13], [4], and [5], of which the latter is also applicable in the multi-class case. [13] uses scatter measures different to the one used in LDA. In [4] and [5] the criterions to be optimized utilize the Bhattacharyya distance (cf. [7]) and the Kullback divergence, respectively. The drawback of these criteria is that the maximization of them needs complex or iterative optimization procedures.

Another iterative multi-class HLDR procedure, which is based on a maximum likelihood formulation of LDA, is studied in [11]. Here LDA is generalized by dropping the assumption that all classes have equal within-class covariance matrices, and maximizing the likelihood for this model.

A fast HLDR method based on a singular value decomposition (svd) was developed in [19] by Tubbs et al. We discuss this method in more detail in Section 3, where we also compare our non-iterative method to theirs. The comparison is done on three artificial and seven real-world data sets.

Section 4 presents the discussion and the conclusions.

2   From Fisher to Chernoff

2.1   The Fisher Criterion

LDR is concerned with the search for a linear transformation that reduces the dimension of a given n-dimensional statistical model to d $(d < n)$ dimensions, while maximally preserving the discriminatory information for the several classes within the model. Due to the complexity of utilizing the Bayes error as the criterion to optimize, one resorts to suboptimal criteria. LDA is such a suboptimal approach. It determines a linear mapping $\mathbf{L}$, a $d \times n$-matrix, that maximizes the so-called *Fisher criterion* J_F [7,9,12,13,16]:

$$J_\mathrm{F}(\mathbf{A}) = \mathrm{tr}((\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_B\mathbf{A}^t)). \tag{1}$$

Here $\mathbf{S}_B := \sum_{i=1}^{K} p_i(\mathbf{m}_i - \bar{\mathbf{m}})(\mathbf{m}_i - \bar{\mathbf{m}})^T$ and $\mathbf{S}_W := \sum_{i=1}^{K} p_i\mathbf{S}_i$ are the between-class and the average within-class scatter matrix, respectively; K is the number of classes, $\mathbf{m}_i$ is the mean vector of class i, p_i is its a priori probability, and the overall mean $\bar{\mathbf{m}}$ equals $\sum_{i=1}^{K} p_i\mathbf{m}_i$. Furthermore, $\mathbf{S}_i$ is the within-class covariance matrix of class i.

From Equation (1) we see that LDA maximizes the ratio of between-class scatter to average within-class scatter in the lower-dimensional space.

Our focus is on the two-class case, in which case we have $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$ [7,12], $\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2$, and $p_1 = 1 - p_2$. Optimizing (1) comes down to determining an eigenvalue decomposition of $\mathbf{S}_W^{-1}\mathbf{S}_B$, and taking the rows of $\mathbf{L}$ to equal the d eigenvectors corresponding to the d largest eigenvalues [7].

Note that the rank of $\mathbf{S}_B$ is 1 in the two-class case, assuming unequal class means, and so we can only reduce the dimension to 1: According to the Fisher criterion there is no discriminatory information in the features, apart from this single dimension.

2.2   Directed Distance Matrices

We now turn to the concept of *directed distance matrices* (DDMs) [12], by which means, we are able to generalize LDA in a proper way.

Assume that the data is linearly transformed such that the within-class covariance matrix $\mathbf{S}_W$ equals the identity matrix, then $J_\mathrm{F}(\mathbf{A})$ equals $\mathrm{tr}((\mathbf{A}\mathbf{A}^t)^{-1}$ $(\mathbf{A}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{A}^t))$, which is maximized by taking the eigenvector $\mathbf{v}$ associated with the largest eigenvalue λ of the matrix $(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$. As pointed out earlier, this matrix has only one nonzero eigenvalue and we can show that $\mathbf{v} = \mathbf{m}_1 - \mathbf{m}_2$ and $\lambda = \mathrm{tr}((\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t) = (\mathbf{m}_1 - \mathbf{m}_2)^t(\mathbf{m}_1 - \mathbf{m}_2)$. The latter equals the squared Euclidean distance between the two *class means*, which we denote by ∂_E.

The matrix $(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$, which we call $\mathbf{S}_E$ from now on, not only gives us the distance between two distributions, but it also provides the direction, by means of the eigenvectors, in which this specific distance can be found. As a matter of a fact, if both classes are normally distributed and have

equal covariance matrix, there is only distance between them in the direction $\mathbf{v}$ and this distance equals λ. All other eigenvectors have eigenvalue 0, indicating that there is no distance between the two classes in these directions. Indeed, reducing the dimension using one of these latter eigenvectors results in a complete overlap of the classes: There is no discriminatory information in these directions, the distance equals 0.

The idea behind DDMs is to give a generalization of $\mathbf{S}_E$. If there is discriminatory information present because of the heteroscedasticity of the data, then this should become apparent in the DDM. This extra distance because of the heteroscedasticity, is, in general, in different directions then the vector $\mathbf{v}$, which separates the means, and so DDMs have more than one nonzero eigenvalue.

The specific DDM we propose is based on the Chernoff distance ∂_C. For two normally distributed densities, it is defined as[1] [2,3]

$$
\begin{aligned}
\partial_C =&(\mathbf{m}_1 - \mathbf{m}_2)^t (\alpha \mathbf{S}_1 + (1-\alpha)\mathbf{S}_2)^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \\
&+ \frac{1}{\alpha(1-\alpha)} \log \frac{|(\alpha \mathbf{S}_1 + (1-\alpha)\mathbf{S}_2)|}{|\mathbf{S}_1|^\alpha |\mathbf{S}_2|^{1-\alpha}} .
\end{aligned}
\tag{2}
$$

Like ∂_E, we can obtain ∂_C as the trace of a positive semi-definite matrix $\mathbf{S}_C$. Simple matrix manipulation [18] shows that this matrix equals[2] (cf. [12])

$$
\begin{aligned}
\mathbf{S}_C :=& \mathbf{S}^{-\frac{1}{2}}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t \mathbf{S}^{-\frac{1}{2}} \\
&+ \frac{1}{\alpha(1-\alpha)}(\log \mathbf{S} - \alpha \log \mathbf{S}_1 - (1-\alpha) \log \mathbf{S}_2) ,
\end{aligned}
\tag{3}
$$

where $\mathbf{S} := \alpha \mathbf{S}_1 + (1-\alpha)\mathbf{S}_2$. Now, before we get to our HLDR criterion, we make the following remarks. (Still assume that $\mathbf{S}_W$ equals the identity matrix.) We want our criterion to be a generalization of Fisher's, so if the data is homoscedastic, i.e., $\mathbf{S}_1 = \mathbf{S}_2$, we want $\mathbf{S}_C$ to equal $\mathbf{S}_E$. This suggests to set α equal to p_1, from which it directly follows that $1 - \alpha$ equals p_2. In this case $\mathbf{S}_C = \mathbf{S}_E$, and we obtain the same dimension reducing linear transform via an eigenvalue decomposition on either.

Now assume that $\mathbf{S}_1$ and $\mathbf{S}_2$ are diagonal—$\mathrm{diag}(a_1, \ldots, a_n)$ and $\mathrm{diag}(b_1, \ldots, b_n)$, respectively—but not necessarily equal. Furthermore, let $\mathbf{m}_1 = \mathbf{m}_2$. Now because $\alpha = p_1$, and hence $\alpha \mathbf{S}_1 + (1-\alpha)\mathbf{S}_2 = \mathbf{I}$, we have

$$
\mathbf{S}_C = \frac{1}{p_1 p_2} \mathrm{diag}(\log \frac{1}{a_1^{p_1} b_1^{p_2}}, \ldots, \log \frac{1}{a_n^{p_1} b_n^{p_2}}) .
\tag{4}
$$

---

[1] Often, the Chernoff distance is defined as $\frac{\alpha(1-\alpha)}{2}\partial_C$, this constant factor, however, is of no essential influence on the rest of our discussion.

[2] We define the function f, e.g. some power or the logarithm, of a symmetric positive definite matrix $\mathbf{A}$, by means of its eigenvalue decomposition $\mathbf{R}\mathbf{V}\mathbf{R}^{-1}$, with eigenvalue matrix $\mathbf{V} = \mathrm{diag}(v_1, \ldots, v_n)$. We let $f(\mathbf{A})$ equal $\mathbf{R}\,\mathrm{diag}(f(v_1), \ldots, f(v_n))\,\mathbf{R}^{-1} = \mathbf{R}(f(\mathbf{V}))\mathbf{R}^{-1}$. Although generally $\mathbf{A}$ is nonsingular, determining $f(\mathbf{A})$ might cause problems, because the matrix is close to singular. Most of the times, alleviation of this problem is possible by using the svd instead of an eigenvalue decomposition, or by properly regularizing $\mathbf{A}$.

On the diagonal of $\mathbf{S}_C$ are the Chernoff distances of the two densities if the the dimension is reduced to one in the associated direction, e.g., linearly transforming the data by the n-vector $(0, \ldots, 0, 1, 0, \ldots, 0)$, where only the dth entry is 1 and all the other equal 0, would give us a Chernoff distance of $\frac{1}{p_1 p_2} \log \frac{1}{a_d^{p_1} b_d^{p_2}}$ in the one-dimensional space. Hence, determining a LDR transformation by an eigenvalue decomposition of the DDM $\mathbf{S}_C$, means that we determine a transform which preserves as much of the Chernoff distance in the lower dimensional space as possible.

In the two cases above, where in fact, we considered $\mathbf{S}_1 = \mathbf{S}_2$ and $\mathbf{m}_1 = \mathbf{m}_2$ respectively, we argued that our criterion gives eligible results. We also expect reasonable results if we do not necessarily have equality of means or covariance matrices, because in this case we obtain a solution that is approximately optimal with respect to the Chernoff distance. In conclusion: The DDM $\mathbf{S}_C$, captures differences in covariance matrices in a certain way and indeed generalizes the homoscedastic DDM $\mathbf{S}_E$.

2.3   Heteroscedasticization of Fisher: The *Chernoff Criterion*

If $\mathbf{S}_W = \mathbf{I}$, $J_{\mathrm{F}}(\mathbf{A})$ equals $\mathrm{tr}((\mathbf{AA}^t)^{-1}(\mathbf{AS}_E\mathbf{A}^t))$. Hence in this case, regarding the discussion in the foregoing subsection, we simply substitute $\mathbf{S}_C$ for $\mathbf{S}_E$, to obtain a heteroscedastic generalization of LDA, because optimizing this criterion is similar to optimizing J_{F}: Determine an eigenvalue decomposition of $\mathbf{S}_C$, and take the rows of the transform $\mathbf{L}$ to equal the d eigenvectors corresponding to the d largest eigenvalues.

In case $\mathbf{S}_W \neq \mathbf{I}$, note that we can first transform our data by $\mathbf{S}_W^{-\frac{1}{2}}$, so we *do* have $\mathbf{S}_W = \mathbf{I}$. In this space we then determine our criterion, which for LDA equals $\mathrm{tr}((\mathbf{AA}^t)^{-1}(\mathbf{AS}_W^{-\frac{1}{2}}\mathbf{S}_B\mathbf{S}_W^{-\frac{1}{2}}\mathbf{A}^t))$, and then transform it back to our original space using $\mathbf{S}_W^{\frac{1}{2}}$, giving the criterion $\mathrm{tr}((\mathbf{AS}_W^{\frac{1}{2}}\mathbf{S}_W^{\frac{1}{2}}\mathbf{A}^t)^{-1}(\mathbf{AS}_B\mathbf{A}^t))$. Hence for LDA, this procedure gives us just Criterion (1), as if it was determined directly in the original space. For our heteroscedastic *Chernoff criterion* J_{C} the procedure above gives the following:

$$
\begin{aligned}
J_{\mathrm{C}}(\mathbf{A}) :={}& \mathrm{tr}((\mathbf{AS}_W\mathbf{A}^t)^{-1}(\mathbf{A}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{A}^t \\
& - \mathbf{AS}_W^{\frac{1}{2}} \frac{p_1 \log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_1\mathbf{S}_W^{-\frac{1}{2}}) + p_2 \log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_2\mathbf{S}_W^{-\frac{1}{2}})}{p_1 p_2} \mathbf{S}_W^{\frac{1}{2}}\mathbf{A}^t)) \, .
\end{aligned}
\tag{5}
$$

This is maximized by determining an eigenvalue decomposition of

$$
\mathbf{S}_W^{-1}(\mathbf{S}_B - \mathbf{S}_W^{\frac{1}{2}} \frac{p_1 \log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_1\mathbf{S}_W^{-\frac{1}{2}}) + p_2 \log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_2\mathbf{S}_W^{-\frac{1}{2}})}{p_1 p_2} \mathbf{S}_W^{\frac{1}{2}}) \, ,
\tag{6}
$$

and taking the rows of the transform $\mathbf{L}$ to equal the d eigenvectors corresponding to the d largest eigenvalues.

3 Experimental Results: Comparing Chernoff to Fisher and Svd

This section compares the performance of the HLDR transformations obtained by means of the Chernoff criterion with transformations obtained by the traditional Fisher criterion, and by the svd method as discussed in, e.g., [19]. The latter method determines a dimension reducing transform—in the two-class case— by constructing an $n \times (n + 1)$-matrix $\mathbf{T}$ that equals $(\mathbf{m}_2 - \mathbf{m}_1, \mathbf{S}_2 - \mathbf{S}_1)$, then performing an svd on $\mathbf{TT}^t = \mathbf{USV}^t$, and finally choosing the row vectors from $\mathbf{U}$ associated with the largest d singular values as the HLDR transformation.

Tests were performed on three artificial [7] (cf. [12])—labelled (a) to (c)— and seven real-world data sets [8,15]—labelled (d) to (j). To be able to see what discriminatory information is retained in using a HLDR, classification is done with a quadratic classifier assuming the underlying distributions to be normal. Results obtained with the svd-based approach, and the Chernoff criterion are presented in the Figures 1(a) to 1(j), and indicated by gray and black lines, respectively. Figures 1(a) to (j) are associated to data sets (a) to (j). The dimension of the subspace is plotted horizontally and the classification error vertically. Results of reduction to a single dimension, mainly for comparison with LDA, are in Table 1.

In presenting the results on the real-world data sets, we restricted ourselves to discussing the main results, and to the most interesting observations. The p-values stated in this part are obtained by comparing the data via a signed rank test [17].

3.1 Fukunaga's Heteroscedastic Two-Class Data and Two Variations

Fukunaga [7] describes a heteroscedastic model consisting of two classes in eight dimensions. The classes are normally distributed with $\mathbf{m}_1 = (0, \ldots, 0)^t$, $\mathbf{S}_1 = \mathbf{I}$, and

$$\mathbf{m}_2 = (3.86, 3.10, 0.84, 0.84, 1.64, 1.08, 0.26, 0.01)^t, \tag{7}$$

$$\mathbf{S}_2 = \operatorname{diag}(8.41, 12.06, 0.12, 0.22, 1.49, 1.77, 0.35, 2.73). \tag{8}$$

Furthermore, $p_1 = p_2 = \frac{1}{2}$.

The first test (a) on artificial data uses these parameters. Two variants are also considers. In the first variant (b), the two means are taken closer to each other to elucidate the performance of the Chernoff criterion, when most of the discriminatory information is in the difference in covariances. For this variant we take the mean of the second class to equal $\frac{1}{10}\mathbf{m}_2$ (cf. [12]). The second variant (c) is a variation on the first, where we additionally set $p_1 = \frac{1}{4}$ and $p_2 = \frac{3}{4}$. This is to elucidate the effect of a difference in class priors, something the svd approach does not account for.

Tests are carried out using Monte Carlo simulation in which we take a total of 1,000,000 instances from the two classes, proportional to the values p_1 and p_2, and determine the error by classifying these instances.

The results from Figures 1(a)–(c) are clear: For a reduction to one dimension, the LDR by means of the Chernoff criterion is as good as, or unmistakably better than LDR by the Fisher criterion or the svd approach. Furthermore, when reducing the data to several dimensions, our approach is also preferable to the svd approach, which merely outperforms our approach in reducing the dimension to 2 in experiment (b).

Table 1. Information w.r.t. the 10 data sets including (mean) classification errors for comparison of the three considered LDR techniques for reduction to one dimension (last three columns). Best results are in boldface

data set		n	size	N	Fisher	Chernoff	svd
Fukunaga's two-class data	(a)	8	-	-	**0.054**	**0.054**	0.140
Variation one	(b)	8	-	-	0.415	**0.231**	**0.231**
Variation two	(c)	8	-	-	0.245	**0.159**	0.240
Wisconsin breast cancer	(d)	9	682	350	0.028	**0.027**	0.031
Wisconsin diagnostic breast cancer	(e)	30	569	500	0.035	**0.029**	0.086
Bupa liver disorders	(f)	6	345	200	**0.364**	0.407	0.466
Cleveland heart-disease	(g)	13	297	200	0.174	**0.172**	0.463
Pima indian diabetes	(h)	8	768	576	0.230	**0.229**	0.342
Musk "Clean2" database	(i)	166	6598	6268	**0.056**	0.061	0.152
Lung/non-lung classification	(j)	11	72000	36000	0.223	0.225	**0.217**

3.2   Real-World Data

Six tests in this subsection are on data sets from the UCI Repository of machine learning databases [15], a seventh test is on the chest radiograph database used in [8]. For a description of the first six data sets refer to [15]. The seventh database consists of 115 chest radiograph images and the classification of their pixels in lung or non-lung. For our purpose, we took 20 images and sub-sampled them from 256×256 to 64×64 pixels. To come to a lung/non-lung classification of a pixel, we used its gray value, its eight neighboring gray values, and its x- and y-coordinate as features, which finally gives us 72000 instances[3] in an 11-dimensional feature space.

The seven tests are carried out by randomly drawing N instances from the data for training and using the remaining instances for testing. (If provided, the value N is taken from the UCI Repository.) This procedure is repeated 100 times and the mean classification error is given. Most of the time, the 100 repetitions give us enough measurements to reliably decide whether one approach consistently outperform the other. This decision is based on the signed rank test [17] for which the p-values are provided.

---

[3] There are only 72000 instances, because in building the feature vector, we excluded pixels that were too close to the border of the image resulting in $60 \times 60 \times 20$ instances.
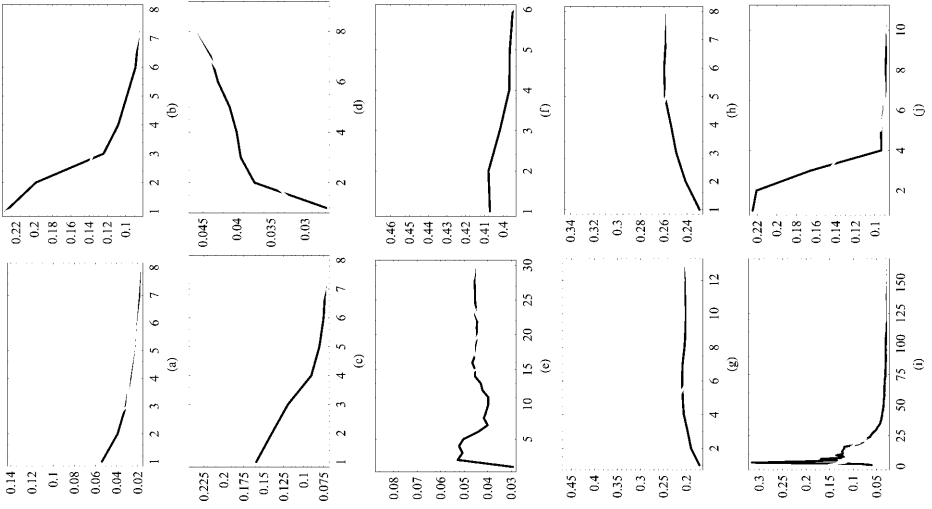
Fig. 1. Plots of feature dimension (vertically) versus classification error (horizontally) for comparison of HLDR via the Chernoff criterion and via the svd-based approach. The grey lines give the results obtained by svd, the black lines provide results obtained by using the Chernoff criterion

Considering Tables 1 and 2, and Figures 1(d) to 1(j), we can generally conclude that the Chernoff criterion improves upon the Fisher criterion. Even though the Chernoff criterion clearly needs more than one dimension, about 25, to outperform LDA in case of data set (i), it dramatically improves upon LDA in case of a dimension greater than, say, 50, with its best result at $d = 106$. Fisher is clearly better for data set (f), however for all other data sets Chernoff is, in general, the one to be preferred although its improvement w.r.t. Fisher for data sets (d) and (h) are considered insignificant.

Concerning the comparison of Chernoff with the svd approach, we can be brief: In case of reducing data set (j) to one or two dimensions, the svd approach is clearly preferable, however, in all other cases the use of the Chernoff criterion is preferable to the svd approach.

See also the captions of Tables 1 and 2.

4   Discussion and Conclusions

We proposed a new heteroscedastic linear dimension reduction (HLDR) criterion for two-class data, which generalizes the well-known Fisher criterion used in LDA. After noting that the Fisher criterion can be related to the Euclidean distance between class means, we used the concept of directed distance matrices (DDMs) to replace the matrix that incorporates the Euclidean distance by one

Table 2. Results w.r.t. the 7 real-world data sets—(d) to (j). Included are the best results over all dimensions ($< n$) for the three approaches (for LDA this, of course, equals $d = 1$). The dimension for which this result is obtained is denoted by d. For comparison of our approach (Chernoff) to both other approaches, the p-values are provided, which are indicated between the compared approaches. Best overall results are in boldface

data set	Fisher	p-value	Chernoff	d	p-value	svd	d
(d)	0.028	*0.070*	**0.027**	1	*0.000*	0.031	1
(e)	0.035	*0.006*	**0.029**	1	*0.000*	0.043	16
(f)	**0.364**	*0.000*	0.396	5	*0.000*	0.416	5
(g)	0.174	*0.033*	**0.172**	1	*0.000*	0.195	7
(h)	0.230	*0.721*	**0.229**	1	*0.000*	0.256	6
(i)	0.056	*0.000*	**0.030**	106	*0.035*	0.031	165
(j)	0.223	*0.000*	**0.089**	9	*0.000*	0.090	10

incorporating the *Chernoff distance*. This distance takes into account the difference in the covariance matrices of both groups, which, by means of a DDM, can be used to find an LDR transformation that takes such differences into account. In addition, it enables us to reduce the dimension of the two-class data to more than a single one.

An other important property of our *Chernoff criterion* is that it is computed in a simple and efficient way, merely using standard matrix arithmetics and not using complex or iterative procedures. Hence its computation is almost as easy as determining an LDR transform using LDA. Furthermore, it should be noted that, although it was used in the derivation of our criterion, it is not necessary that both classes are normally distributed. The Chernoff criterion only uses the first and second order central moments of the class distribution in a way that is plausible for dimension reduction, whether the data is normally distributed or not.

In ten experiments, we compared the performance of the Chernoff criterion to that of LDA and an other simple and efficient approach based on the svd. Three of these experiments were on artificial data and seven on real-world data. The experiments clearly showed the improvement that is possible when utilizing the Chernoff criterion instead of the Fisher criterion or the svd-based approach and we can generally conclude that our method is clearly preferable to the others.

Finally, it is of course interesting to look at the possibility of extending our criterion to the multi-class case. [7] and [12] offer ideas for doing so. They suggest a certain averaged criterion that takes into account all multi-class discriminatory information at once. Future investigations will be on these kinds of extensions of the Chernoff criterion.

References

1. T. W. Anderson and R. R. Bahadur. Classification into two multivariate normal distributions with different covariance matrices. *Annals of Mathematical Statistics*, 33:420–431, 1962. 509
2. C. H. Chen. On information and distance measures, error bounds, and feature selection. *The information scientist*, 10:159–173, 1979. 509, 511
3. J. K. Chung, P. L. Kannappan, C. T. Ng, and P. K. Sahoo. Measures of distance between probability distributions. *Journal of mathematical analysis and applications*, 138:280–292, 1989. 509, 511
4. H. P. Decell and S. K. Marani. Feature combinations and the Bhattacharyya criterion. *Communications in Statistics. Part A. Theory and Methods*, 5:1143–1152, 1976. 509
5. H. P. Decell and S. M. Mayekar. Feature combinations and the divergence criterion. *Computers and Mathematics with Applications*, 3:71–76, 1977. 509
6. R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936. 508
7. K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 1990. 508, 509, 510, 513, 516
8. B. van Ginneken and B. M. ter Haar Romeny. Automatic segmentation of lung fields in chest radiographs. *Medical Physics*, 27(10):2445–2455, 2000. 513, 514
9. A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000. 508, 510
10. D. Kazakos. On the optimal linear feature. *IEEE Transactions on Information Theory*, 24:651–652, 1978. 509
11. N. Kumar and A. G. Andreou. Generalization of linear discriminant analysis in a maximum likelihood framework. In *Proceedings of the Joint Meeting of the American Statistical Association*, 1996. 509
12. M. Loog. *Approximate Pairwise Accuracy Criteria for Multiclass Linear Dimension Reduction: Generalisations of the Fisher Criterion*. Number 44 in WBBM Report Series. Delft University Press, Delft, 1999. 509, 510, 511, 513, 516
13. W. Malina. On an extended Fisher criterion for feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3:611–614, 1981. 509, 510
14. G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, New York, 1992. 509
15. P. M. Murphy and D. W. Aha. UCI Repository of machine learning databases. [http://www.ics.uci.edu/~mlearn/mlrepository.html]. 513, 514
16. C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B*, 10:159–203, 1948. 508, 510
17. J. A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, second edition, 1995. 513, 514
18. G. Strang. *Linear algebra and its applications*. Harcourt Brace Jovanovich, third edition, 1988. 511
19. J. D. Tubbs, W. A. Coberly, and D. M. Young. Linear dimension reduction and Bayes classification. *Pattern Recognition*, 15:167–172, 1982. 509, 513