Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion

Marco Loog and Robert P.W. Duin, *Member*, *IEEE*

Abstract—We propose an eigenvector-based *heteroscedastic* linear dimension reduction (LDR) technique for multiclass data. The technique is based on a heteroscedastic two-class technique which utilizes the so-called *Chernoff criterion*, and successfully extends the well-known *linear discriminant analysis* (LDA). The latter, which is based on the *Fisher criterion*, is incapable of dealing with heteroscedastic data in a proper way. For the two-class case, the between-class scatter is generalized so to capture differences in (co)variances. It is shown that the classical notion of between-class scatter can be associated with Euclidean distances between class means. From this viewpoint, the between-class scatter is generalized by employing the Chernoff distance measure, leading to our proposed heteroscedastic measure. Finally, using the results from the two-class case, a multiclass extension of the Chernoff criterion is proposed. This criterion combines separation information present in the class mean as well as the class covariance matrices. Extensive experiments and a comparison with similar dimension reduction techniques are presented.

Index Terms—Linear dimension reduction, linear discriminant analysis, Fisher criterion, Chernoff distance, Chernoff criterion.

1 Introduction

Linearly reducing the dimensionality of the features space, i.e., feature extraction, is a common technique in statistical pattern recognition typically used to lower the size of statistical models and overcome estimation problems, often resulting in an improved classifier accuracy in this lower-dimensional space. *Linear discriminant analysis* (LDA) is probably the most well-known approach to supervised *linear dimension reduction* (LDR). This classical technique was developed by Fisher [9] for the two-class case and extended by Rao [21] to handle the multiclass case.

In LDA, a transformation matrix from an n-dimensional feature space to a d-dimensional space is determined such that the Fisher criterion of between-class scatter over within-class scatter is maximized (cf. [8], [10], [12], [18]). An attractive feature of LDA is the fast and easy way to determine this optimal linear transformation, only requiring simple matrix arithmetics. A limitation of LDA is that it merely tries to separate class means as good as possible and it does not take the discriminatory information that is present in the difference of the covariance matrices into account. It is incapable of dealing explicitly with *heteroscedastic* data, i.e., data in which classes do not have equal covariance matrices. This limitation becomes very apparent in the two-class case, in which a reduction to only a single dimension is possible (cf. [10]), while the K-class case allows only for a reduction to at most $K-1$ dimensions.

When linearly reducing the dimensionality, the $K-1$ dimensions do not necessarily contain all the relevant data for the classification task and even if $K-1$ dimensions do so, it is not clear that LDA will discern them. Taking the heteroscedasticity of the data into account, we develop an LDR technique that extends and improves upon classical LDA. This extension is obtained via the use of directed distance matrices (DDMs) [15], which can be considered generalizations of the between-class scatter matrix. The between-class scatter matrix, as used in LDA, merely takes into account the discriminatory information that is present in the pairwise differences of class means and can be associated with the squared Euclidean distance between pairs of class means.

The specific heteroscedastic extension of the Fisher criterion, studied more closely in Sections 2 and 3, is based on the *Chernoff distance* [4], [5]. This measure of affinity of two densities considers mean differences as well as covariance differences—as opposed to the Euclidean distance—and is used to extend LDA. Section 2 discusses the LDA extension for two-class data as proposed in an earlier article [16]. In Section 3, we come to our heteroscedastic multiclass measure, which extends LDA, by comparing the K classes in a *pairwise* fashion and using the two-class measure as a building block. While doing so, we retain the attractive feature of quickly and easily determining a dimension reducing transformation, as with LDA. Furthermore, we are able to reduce the data to any dimension d smaller than n and not only to, at most, $K-1$ dimensions.

In Section 4.2 of [10], Fukunaga discusses several ways of extending the linear classifiers to unequal covariance matrices and nonnormal distributions. The criteria derived can also be used for the purpose of dimensionality reduction.

- *M. Loog is with the Image Sciences Institute, University Medical Center Utrecht, E.01.335, PO Box 85500, 3508 GA Utrecht, The Netherlands. E-mail: marco@isi.uu.nl.*
- *R.P.W. Duin is with the Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, HB 11.260, PO Box 5046, 2600 GA Delft, The Netherlands. E-mail: r.p.w.duin@ewi.tudelft.nl.*

Manuscript received 22 Apr. 2003; revised 11 Sept. 2003; accepted 12 Jan. 2004.
Recommended for acceptance by A. Yuille.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0047-0403.

However, besides the fact that they are merely derived for the two-class case and not readily extendible to the multiclass case, the criteria essentially give a single LDR vector that takes the difference between the class means into account. In addition, some of the approaches need an iterative optimization procedure.

Several alternative approaches to heteroscedastic LDR (HLDR) are known of which we mention the following ones. See also [8], [10], and [18], and references therein.

Under the assumption that all classes are normally distributed, [23] gives a computationally demanding approach to solving the LDR problem by minimizing the actual Bayes error in the linearly reduced space. This is done using simulated annealing in combination with an exact integration over the lower-dimensional feature space.

Straightforward extensions of the Fisher criterion were proposed in [7] and [20], the former of which is based on the Kullback divergence. As opposed to our criterion, their iterative optimization procedures are clearly more complex than optimizing the Fisher criterion. A broad overview of feature extraction techniques based on probabilistic separability and interclass distance measures—some of them related to the previous mentioned techniques—can be found in [8]. Again, mostly, time-consuming iterative procedures should be employed to optimize these criteria.

Different extensions of Fisher's LDA are given by Hastie et al., see [11]. We mention penalized discriminant analysis (PDA), which can also be used for the purpose of LDR. By means of regularization, PDA is able to deal with data in which one has many highly correlated features and LDA would suffer from overfitting. However, PDA does not explicitly use the discriminatory information present in the covariance terms as the Chernoff criterion does. We note that the regularizations suggested for PDA are readily applicable within our approach.

Another multiclass HLDR procedure, which is based on a maximum-likelihood formulation of LDA, is studied in [13]. Here, LDA is generalized by dropping the assumption that all classes have equal within-class covariance matrices and iteratively maximizing the likelihood for this model.

Of the computationally intensive methods, we finally mention the nonparametric approaches presented in [3] and [14]. These techniques work directly on the data and try to maintain as much of the separation information as possible in the lower-dimensional space. The amount of separability in the subspace is measured using a certain nearest-neighbor procedure, which accounts for a large part of the computational complexity. Comparable to these approaches is the one given in [8] based on Parzen estimates.

Two fast LDR methods based on the singular value decomposition (svd) were introduced in [24] and [2], respectively. The first one by Tubbs et al. presents an HLDR method while the latter is Mahalanobis distance-based and basically homoscedastic. We describe both methods in some more details in Section 4, where we also compare our noniterative method to theirs and to LDA on 12 real-world data sets from the UCI Repository [19].

Section 5 completes the paper with a discussion and the conclusions.

2 The Chernoff Criterion: Two-Class Case

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} [10], [12], [15], [21]:

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

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 estimated 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, and $\mathbf{A}$ is a $d \times n$-matrix. From (1), we see that LDA maximizes the ratio of between-class scatter to average within-class scatter in the lower-dimensional space. 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 [8], [10].

This section focuses on the two-class case, in which case we have $\mathbf{S}_B = p_1\,p_2(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$ [10], [15], [16], $\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2$, and $p_1 = 1 - p_2$. Note that, in this case, the rank of $\mathbf{S}_B$ is 1—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

For now, 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 $\mathrm{tr}((\mathbf{A}\mathbf{A}^t)^{-1}(\mathbf{A}p_1\,p_2\,(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{A}^t))$, which is maximized by taking the eigenvector $\mathbf{v}$ associated with the largest eigenvalue λ of the matrix $\mathbf{S}_E := (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$. (Note that $\mathbf{S}_B = p_1\,p_2\,\mathbf{S}_E$) This matrix has only one nonzero eigenvalue which equals $\lambda = \mathrm{tr}((\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t) = (\mathbf{m}_1 - \mathbf{m}_2)^t(\mathbf{m}_1 - \mathbf{m}_2)$, with associated eigenvector $\mathbf{v} = \mathbf{m}_1 - \mathbf{m}_2$. Note that the eigenvalue equals the squared *Euclidean* distance, denoted by ∂_E, between the two-class means.

The matrix $\mathbf{S}_E := (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$ 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 matrices, 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 directed distance matrices (DDMs) is to give a generalization of $\mathbf{S}_E$ and, hence, $\mathbf{S}_B$ [15]. If there is discriminatory information present because of the heteroscedasticity of the data, then this should become apparent in the DDM. This extra distance due to the heteroscedasticity is,

in general, in different directions than 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 between two probability density functions d_1 and d_2

$$\partial_C := -\log \int d_1^{\alpha}(x)\, d_2^{1-\alpha}(x)dx,$$

where $\alpha \in (0,1)$ is a constant.[1]

For two normally distributed densities, it equals[2] [4], [5]

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

Like ∂_E, we can obtain ∂_C as the trace of a positive semidefinite matrix $\mathbf{S}_C$ (cf. [15]):

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

where $\mathbf{S} := \alpha\mathbf{S}_1+(1-\alpha)\mathbf{S}_2$, $\mathbf{S}^{-\frac{1}{2}}$ is the inverted square root and $\log\mathbf{S}$ is the logarithm[3] of $\mathbf{S}$.

To see that the trace of $\mathbf{S}_C$ equals ∂_C, write out $\operatorname{tr}\mathbf{S}_C$:

$$\begin{aligned}\operatorname{tr}\mathbf{S}_C &= \operatorname{tr}(\mathbf{S}^{-\frac{1}{2}}(\mathbf{m}_1-\mathbf{m}_2)(\mathbf{m}_1-\mathbf{m}_2)^t\mathbf{S}^{-\frac{1}{2}})\\ &\quad+\operatorname{tr}\left(\frac{1}{\alpha(1-\alpha)}(\log\mathbf{S}-\alpha\log\mathbf{S}_1-(1-\alpha)\log\mathbf{S}_2)\right)\\ &= \operatorname{tr}((\mathbf{m}_1-\mathbf{m}_2)^t\mathbf{S}^{-1}(\mathbf{m}_1-\mathbf{m}_2))\\ &\quad+\frac{1}{\alpha(1-\alpha)}(\operatorname{tr}(\log\mathbf{S})-\alpha\operatorname{tr}(\log\mathbf{S}_1)\\ &\quad-(1-\alpha)\operatorname{tr}(\log\mathbf{S}_2))\\ &= (\mathbf{m}_1-\mathbf{m}_2)^t\mathbf{S}^{-1}(\mathbf{m}_1-\mathbf{m}_2)\\ &\quad+\frac{1}{\alpha(1-\alpha)}(\log|\mathbf{S}|-\alpha\log|\mathbf{S}_1|-(1-\alpha)\log|\mathbf{S}_2|).\end{aligned}$$

Finally, recalling that $\mathbf{S} := \alpha\mathbf{S}_1+(1-\alpha)\mathbf{S}_2$ and combining the three logarithms into a single one, we see that the resulting expression equals (2).

We want the final criterion to be an extension 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 setting α equal to p_1, from which it directly follows that $1-\alpha$ equals p_2. The link with homoscedastic LDA is clear from the foregoing.

To exemplify the behavior of the matrix $\mathbf{S}_C$ in the heteroscedastic case we consider the other extreme case in which the means are taken to be equal, i.e., $\mathbf{m}_1=\mathbf{m}_2$. In addition, assume that $\mathbf{S}_1$ and $\mathbf{S}_2$ are diagonal, $\operatorname{diag}(a_1,\ldots,a_n)$ and $\operatorname{diag}(b_1,\ldots,b_n)$, respectively, but not necessarily equal. Because $\alpha=p_1$, and $\alpha\mathbf{S}_1+(1-\alpha)\mathbf{S}_2=\mathbf{I}$ (by assumption), we have

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

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 others 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 view of the two cases above, we argue that our suggested DDM gives eligible results. In addition, we argue that this even holds if we do not have equality of means or covariance matrices because, also in this case, we obtain a solution that is based on the Chernoff distance, which is a certain weighted combination of both extreme cases above. In conclusion, the DDM $\mathbf{S}_C$ captures differences in covariance matrices and indeed gives an extension of the homoscedastic DDM $\mathbf{S}_E$.

2.3 The Two-Class Chernoff Criterion

If $\mathbf{S}_W=\mathbf{I}$, $J_{\mathrm{F}}(\mathbf{A})$ equals $\operatorname{tr}((\mathbf{A}\mathbf{A}^t)^{-1}(p_1\,p_2\,\mathbf{A}\mathbf{S}_E\mathbf{A}^t))$. Therefore, in this case, regarding the discussion in the foregoing section, we simply substitute $\mathbf{S}_C$ for $\mathbf{S}_E$, to obtain a heteroscedastic generalization of the Fisher criterion. In case $\mathbf{S}_W\neq\mathbf{I}$, we first transform the data by $\mathbf{S}_W^{-\frac{1}{2}}$, so we *do* have $\mathbf{S}_W=\mathbf{I}$. In this space, the criterion is determined—which for LDA equals

$$\operatorname{tr}((\mathbf{A}\mathbf{A}^t)^{-1}(p_1\,p_2\,\mathbf{A}\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_E\mathbf{S}_W^{-\frac{1}{2}}\mathbf{A}^t))$$

and then transformed back to the original space using $\mathbf{S}_W^{\frac{1}{2}}$. For the Fisher criterion this would finally result in

$$\operatorname{tr}((\mathbf{A}\mathbf{S}_W^{\frac{1}{2}}\mathbf{S}_W^{\frac{1}{2}}\mathbf{A}^t)^{-1}(p_1\,p_2\,\mathbf{A}\mathbf{S}_E\mathbf{A}^t)),$$

which equals (1), as if it was determined directly in the original space. Using $\mathbf{S}_C$ instead of $\mathbf{S}_E$, this procedure leads to the following heteroscedastic extension.

Definition. *The heteroscedastic two-class* Chernoff criterion J_{C} *is defined as*

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

3 The Multiclass Extension

In the previous section, we derived the Chernoff criterion for two-class data (see also [16]). In this section, we turn to the multiclass case. Based on a certain decomposition of the

1. In [6], the Chernoff distance is defined as the minimum of ∂_C over all $\alpha\in(0,1)$.

2. Although the Chernoff distance actually equals $\frac{\alpha(1-\alpha)}{2}\partial_C$ in this case, this constant factor is of no essential influence on the rest of our discussion.

3. 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 $\mathbf{V}=\operatorname{diag}(v_1,\ldots,v_n)$. We let $f(\mathbf{A})$ equal $\mathbf{R}\operatorname{diag}(f(v_1),\ldots,f(v_n))\,\mathbf{R}^{-1}=\mathbf{R}(f(\mathbf{V}))\mathbf{R}^{-1}$. Although, generally, $\mathbf{A}$ is nonsingular, determining $f(\mathbf{A})$ might cause numerical problems, if the matrix is close to singular. Alleviation of this computational problem is possible by using the svd instead of an eigenvalue decomposition, or by properly regularizing $\mathbf{A}$.

between-class scatter matrix, we construct a measure for HLDR using the two-class criterion as a building block.

3.1 Decomposing the Between-Class Scatter Matrix

The decomposition of the between-class scatter matrix $\mathbf{S}_B$ we use to generalize the Chernoff criterion to the multiclass case is as follows:

$$\begin{aligned}\mathbf{S}_B &= \sum_{i=1}^{K-1}\sum_{j=i+1}^{K} p_i\, p_j (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^t \\ &= \sum_{i=1}^{K-1}\sum_{j=i+1}^{K} p_i\, p_j\, \mathbf{S}_{Eij},\end{aligned} \tag{6}$$

where $\mathbf{S}_{Eij} := (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^t$. (See [15] for a proof of (6) above.) This decomposition shows how the scatter matrix captures the divergence of the class mean $\mathbf{m}_i$ from all other class means $\mathbf{m}_j$. For every pair of means the difference vector $\mathbf{m}_i - \mathbf{m}_j$ is determined and the sum of their outer products forms the between-class scatter.

Based on (6), J_{F} can be decomposed as

$$J_{\mathrm{F}}(\mathbf{A}) = \sum_{i=1}^{K-1}\sum_{j=i+1}^{K} p_i\, p_j \operatorname{tr}((\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_{Eij}\mathbf{A}^t)). \tag{7}$$

The foregoing expression allows a decomposition of the Fisher criterion into a sum of *pairwise* Fisher criteria. It consists of sums of Fisher criteria taken all class pairs into account separately (cf. [15]). Based on this pairwise decomposed Fisher criterion, we can now generalize the two-class Chernoff criterion to the multiclass case.

3.2 Weighted Two-Class Chernoff Criteria: The Heteroscedasticization of Fisher

Initially, as in Section 2, the within-class scatter $\mathbf{S}_W$ is assumed to equal the identity matrix. In this case, the Fisher criterion equals $\operatorname{tr}((\mathbf{A}\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_B\mathbf{A}^t))$, which can be optimized via an eigenvalue decomposition of the matrix $\mathbf{S}_B$. Decomposition (6) shows that $\mathbf{S}_B$ is a weighted sum of pairwise DDMs and as such can be considered a DDM itself: Its eigenvectors giving the direction in which there is distance, their eigenvalues giving the actual distance. Indeed, carrying out an LDA and assuming the within-class scatter matrix to be the identity, LDR is performed by taking those eigenvectors of $\mathbf{S}_B$ for which the associated eigenvalues are largest.

In light of Sections 2.2 and 2.3, and (7), foregoing considerations lead us to the Chernoff-based, multiclass extension of the two-class Chernoff criterion.

$$J_{\mathrm{C}}(\mathbf{A}) := \sum_{i=1}^{K-1}\sum_{j=i+1}^{K} p_i\, p_j \operatorname{tr}((\mathbf{A}\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_{Cij}\mathbf{A}^t)). \tag{8}$$

In this, $\mathbf{S}_{Cij}$ is the DDM capturing the Chernoff distance between class i and j, which is immediately determined by means of (3).

$$\begin{aligned}\mathbf{S}_{Cij} :=& \mathbf{S}_{ij}^{-\frac{1}{2}}(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^t\mathbf{S}_{ij}^{-\frac{1}{2}} \\ &+ \frac{1}{\pi_i\,\pi_j}(\log \mathbf{S}_{ij} - \pi_i \log \mathbf{S}_i - \pi_j \log \mathbf{S}_j).\end{aligned} \tag{9}$$

Here, $\pi_i := p_i/(p_i + p_j)$, and $\pi_j := p_j/(p_i + p_j)$ are *relative* priors, i.e., only taking the two classes into account that define the particular pairwise term. Furthermore, $\mathbf{S}_{ij}$ is the average *pairwise* within-class scatter matrix, defined as $\pi_i\,\mathbf{S}_i + \pi_j\,\mathbf{S}_j$.

Along the same line of reasoning as in Section 2.3, the final multiclass Chernoff criterion—in which the within-class scatter is not necessarily the identity matrix—can be obtained by first transforming the data such that the within-class scatter matrix is the identity, then determine the criterion J_{C} and, finally, do the inverse transformation, leading to

Definition. *For a $d \times n$-*matrix $\mathbf{A}$*, the multiclass measure of spread J_{C}, the* Chernoff criterion*, is defined as*

$$\begin{aligned}J_{\mathrm{C}}(\mathbf{A}) :=& \sum_{i=1}^{K-1}\sum_{j=i+1}^{K} p_i\, p_j\, \operatorname{tr}\Big((\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1} \\ &\times \mathbf{A}\mathbf{S}_W^{\frac{1}{2}}\big((\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_{ij}\mathbf{S}_W^{-\frac{1}{2}})^{-\frac{1}{2}}\mathbf{S}_W^{-\frac{1}{2}}(\mathbf{m}_i - \mathbf{m}_j) \\ &\times (\mathbf{m}_i - \mathbf{m}_j)^t\mathbf{S}_W^{-\frac{1}{2}}(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_{ij}\mathbf{S}_W^{-\frac{1}{2}})^{-\frac{1}{2}} \\ &+ \frac{1}{\pi_i\,\pi_j}((\log \mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_{ij}\mathbf{S}_W^{-\frac{1}{2}}) \\ &- \pi_i \log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_i\mathbf{S}_W^{-\frac{1}{2}}) \\ &- \pi_j \log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_j\mathbf{S}_W^{-\frac{1}{2}})))\mathbf{S}_W^{\frac{1}{2}}\mathbf{A}^t\Big),\end{aligned} \tag{10}$$

where $\pi_i := p_i/(p_i + p_j)$, $\pi_j := p_j/(p_i + p_j)$, *and* $\mathbf{S}_{ij} := \pi_i\,\mathbf{S}_i + \pi_j\,\mathbf{S}_j$.

The Chernoff criterion is maximized in a manner similar to optimizing the Fisher criterion: First, determine an eigenvalue decomposition of the $n \times n$-matrix

$$\begin{aligned}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} p_i\, p_j\, \mathbf{S}_W^{-1} &\times \mathbf{S}_W^{\frac{1}{2}}\big((\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_{ij}\mathbf{S}_W^{-\frac{1}{2}})^{-\frac{1}{2}} \\ &\times \mathbf{S}_W^{-\frac{1}{2}}(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^t \\ &\times \mathbf{S}_W^{-\frac{1}{2}}(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_{ij}\mathbf{S}_W^{-\frac{1}{2}})^{-\frac{1}{2}} \\ &+ \frac{1}{\pi_i\,\pi_j}(\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_{ij}\mathbf{S}_W^{-\frac{1}{2}}) \\ &- \pi_i \log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_i\mathbf{S}_W^{-\frac{1}{2}}) \\ &- \pi_j \log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_j\mathbf{S}_W^{-\frac{1}{2}})))\mathbf{S}_W^{\frac{1}{2}},\end{aligned} \tag{11}$$

then take the rows of the transformation matrix $\mathbf{L}$ to equal the d eigenvectors associated with the d largest eigenvalues [8], [10].

Note that, in the two-class case, $\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_{ij}\mathbf{S}_W^{-\frac{1}{2}} = \mathbf{I}$, hence, the foregoing weighted two-class Chernoff criterion boils down to the original two-class Chernoff criterion (5). Note also that, if all covariance matrices $\mathbf{S}_i$ are equal, the Chernoff criterion equals the Fisher criterion, i.e., $J_{\mathrm{C}} = J_{\mathrm{F}}$.

4 Experimental Results

This section compares the performance of the HLDR transformations obtained by means of the Chernoff criterion—based on an eigendecomposition of the matrix in (11)—with transformations obtained by the traditional Fisher criterion. In addition, the performances of the HLDR methods from [24] and [2], are also compared to the performance of the Chernoff criterion. For some other comparative studies, between several LDR techniques on several data sets, see, for example, [1] and [2].

TABLE 1

data set	label	n	PC	K	N
Wisconsin breast cancer	(a)	9	9	2	682
BUPA liver disorder	(b)	6	6	2	345
Pima indians diabetes	(c)	8	8	2	768
Wisconsin diagnostic breast cancer	(d)	30	7	2	569
Cleveland heart-disease	(e)	13	13	2	297
SPECTF heart	(f)	44	44	2	349
Iris plants	(g)	4	4	3	150
Thyroid gland	(h)	5	5	3	215
Vowel context	(i)	10	10	11	990
Landsat satellite	(j)	36	36	6	6435
Multifeature digit (Zernike moments)	(k)	47	33	10	2000
Glass identification	(l)	9	8	6	214

The 12 data sets taken from [19] used in the experiments. Information is provided on initial dimensionality n, dimensionality after principal component analysis PC, number of classes K, and number of total instances N.

The method in [24] determines a heteroscedastic dimension reducing transform by constructing an $n \times (n+1)(K-1)$-matrix $\mathbf{T}$ that equals $(\mathbf{m}_2 - \mathbf{m}_1, \ldots, \mathbf{m}_K - \mathbf{m}_1, \mathbf{S}_2 - \mathbf{S}_1, \ldots, \mathbf{S}_K - \mathbf{S}_1)$, then performing an svd on $\mathbf{T} = \mathbf{QSV}^t$ and, finally, choosing the column vectors from $\mathbf{Q}$ associated with the largest d singular values as the LDR transformation. As with our HLDR approach, this approach also allows for LDR to dimensions larger than $K-1$ (if $K-1 < n$) and up to n.

Similar to the foregoing method is the Mahalanobis distance-based method from [2], which determines an svd $\mathbf{QSV}^t$ of the $n \times \frac{1}{2}n(n-1)$-matrix

$$\mathbf{U} = ((\mathbf{S}_1 + \mathbf{S}_2)^{-1}(\mathbf{m}_1 - \mathbf{m}_2), (\mathbf{S}_1 + \mathbf{S}_3)^{-1} (\mathbf{m}_1 - \mathbf{m}_3), \ldots, (\mathbf{S}_{K-1} + \mathbf{S}_K)^{-1}(\mathbf{m}_{K-1} - \mathbf{m}_K)).$$

Again, the column vectors from $\mathbf{Q}$ associated with the largest d singular values are chosen as the LDR transformation. This technique can also be viewed as an extension to Fisher's LDA and allows for a reduction of dimensionality up to $d = \frac{1}{2}K(K-1)$ (if $\frac{1}{2}K(K-1) \leq n$, see [2]).

Tests were performed on 12 real-world data sets, labeled (a) to (l), taken from the UCI Repository of machine learning databases [19] (see Table 1). Instances with missing values were taken out of the data sets prior to the experiments.

The comparison is based on two different classifiers [8], [10], [12]:

- the linear classifier assuming all classes to be normally distributed with equal covariance matrix and
- the quadratic classifier assuming the underlying distributions to be normal with covariance matrices that are not necessarily equal.

These two classifiers are chosen because they stay close to the assumption that most of the relevant information is in the first and second order central moments, i.e., the means and the (co)variances. The first classifier merely takes means and average within-class covariances into account based upon which linear decision boundaries are constructed. The second can cope with all classes having different means and covariance matrices and allows the decision boundaries to be quadratic.

4.1 The Experimental Setup

For every of the 12 data sets and for every possible d to reduce the dimension to, the experiment described below is conducted a hundred times.

1. The data set is randomly split into a test and a train set. The test set contains (approximately) 10 percent of the data, while the train set contains the remaining 90 percent.
2. A PCA is performed on the train set after which all principal components with an eigenvalue smaller than one millionth of the total variance, i.e., the trace of the total covariance matrix, are discarded. In this way, problems related to (near) singular covariance matrices are avoided and all four transformations can be properly determined. See Table 1 for the data dimensionalities before and after PCA. Note that for most data sets all principal components are retained.
3. Using the transformed train data, we determine the four LDR transformations (or less, if a reduction to d dimensions is not possible with a certain transformation, i.e., the Fisher-based and the Mahalanobis distance-based transformations) and reduce the dimensionality of the train data to d.
4. In the d-dimensional reduced feature space, we determine the linear and the quadratic classifier using the train data and, subsequently, classify the test data after transforming its instances in the same way as the train instances. The classification error is estimated on the test data.

4.2 Analysis of Results

The per data set-performances of the several LDR techniques are compared. To this end, per classifier, data set and dimension d, the mean estimated classification error over the hundred runs is determined. This gives a final estimate of the classification error for the respective settings. For every LDR transform, only the optimal dimensionality to reduce the data to and the corresponding mean classification error (MCE) is reported. (Our method, as well as the other methods, give no direct means to determine an optimal

TABLE 2

	Full	Fisher		Chernoff		Mahalanobis		Tubbs	
label	MCE	MCE	(d)	MCE	(d)	MCE	(d)	MCE	(d)
(a)	0.063	**0.047**	(1)	**0.046***	(1)	**0.047**	(1)	**0.050**	(2)
(b)	**0.426**	0.427	(1)	**0.424***	(1)	0.427	(1)	0.428	(5)
(c)	**0.348**	**0.348***	(1)	**0.348***	(2/1)	**0.349**	(1)	**0.348***	(5/4)
(d)	0.142	0.177	(1)	**0.131***	(1)	0.208	(1)	0.140	(3)
(e)	**0.175**	**0.172**	(1)	**0.171***	(1)	**0.175**	(1)	**0.174**	(12/5)
(f)	0.279	0.272	(1)	0.266	(2)	0.239	(1)	**0.205***	(6)
(g)	0.051	0.039	(1)	**0.035**	(3)	**0.029***	(2)	0.046	(3)
(h)	**0.122**	**0.130**	(2)	**0.122***	(4/1)	**0.125**	(3/1)	**0.128**	(4)
(i)	0.636	**0.543***	(2)	**0.550**	(4)	0.595	(2)	0.620	(4)
(j)	0.217	**0.210***	(3)	0.212	(3)	0.219	(14)	0.217	(33)
(k)	0.539	**0.203***	(8)	0.226	(8)	0.270	(16)	0.404	(11)
(l)	0.568	**0.515***	(3)	**0.538**	(4)	0.552	(6)	0.571	(6)

*Observed MCE and optimal dimensionality (d) for the 12 data dets (a) to (l), Using the linear classifier and the four different LDR techniques indicated by "Fisher," "Chernoff, " "Tubbs" [24], and "Mahalanobis" [2]. Optimal observed MCE per data set is typeset in bold and a superscript * is added. In bold are the MCEs for transforms that also give, in comparison to the optimal transformation, indiscernible MCEs based on a signed rank test with significance level 0.01. An MCE in a lower-dimensional space indiscernible from the optimal one, is indicated by the second integer in parentheses on the right of the /. The estimated MCE using no LDR is below "Full."*

TABLE 3

	Full	Fisher		Chernoff		Mahalanobis		Tubbs	
label	MCE	MCE	(d)	MCE	(d)	MCE	(d)	MCE	(d)
(a)	0.050	**0.028**	(1)	**0.027***	(1)	**0.028**	(1)	**0.029**	(1)
(b)	0.402	**0.374***	(1)	**0.381**	(1)	**0.375**	(1)	0.421	(5)
(c)	0.260	**0.227**	(1)	**0.224***	(1)	0.229	(1)	0.254	(2)
(d)	0.062	0.059	(1)	**0.051***	(2)	0.063	(1)	0.059	(4)
(e)	**0.170**	**0.164**	(1)	**0.159***	(1)	**0.164**	(1)	**0.168**	(7)
(f)	**0.060**	0.256	(1)	**0.059***	(21)	0.245	(1)	**0.061**	(42)
(g)	**0.041**	**0.038**	(1)	**0.034***	(2/1)	**0.034***	(1)	0.041	(3)
(h)	**0.045**	**0.044**	(2/1)	**0.043**	(1)	**0.041***	(3/1)	**0.045**	(4)
(i)	**0.122**	0.169	(9)	**0.126***	(9)	0.148	(9)	0.136	(9)
(j)	0.145	0.141	(5)	0.143	(25)	**0.135***	(6)	0.139	(16)
(k)	0.175	0.178	(8)	**0.164***	(21)	**0.164***	(15)	**0.167**	(23)
(l)	0.750	**0.519**	(1)	**0.532**	(3)	**0.541**	(5)	**0.515***	(5/3)

*Observed MCE and optimal dimensionality (d) for the 12 data sets (a) to (l), using the quadratic classifier and the four different LDR techniques indicated by "Fisher," "Chernoff," "Tubbs" [24], and "Mahalanobis" [2]. Optimal observed MCE per data set is typeset in bold and a superscript * is added. In bold are the MCEs for transforms that also give, in comparison to the optimal transformation, indiscernible MCEs based on a signed rank test with significance level 0.01. An MCE in a lower-dimensional space indiscernible from the optimal one, is indicated by the second integer in parentheses on the right of the /. The estimated MCE using no LDR is below "Full."*

dimensionality to reduce to. However, the observed optimal MCEs give an indication of the attainable performance and can be used to compare the several approaches.) These numbers are presented in Table 2 and Table 3. The overall optimal MCE over all transforms is typeset in bold and a "*" is added in superscript. In bold are the transforms that also give, in comparison to the optimal transformation, statistically indiscernible classification errors. For this, results are compared using a signed rank test in which the desired level of significance is set to 0.01 (see [22]). If it is possible to attain an MCE not significantly different—again based on a signed rank test—from the optimal one in a lower-dimensional space, this is indicated by the second integer in parentheses on the right of the /. Tables 2 and 3 also give the MCE obtained when not performing an LDR.

We start with two general observations: First, the quadratic classifier performs, in general, better for most data sets. The two exceptions are data sets (g) and (l). This may indicate that in most data sets, there is indeed separation information present in the second order moments of the class distributions. Second, we see that LDR indeed can improve the accuracy of the classifier in most cases. Note, although, that this is not always the case (take, for example, data set (i) and the quadratic classifier) and if it does hold the improvements are sometimes not very convincing. However, even if the error rate does not drop considerably, the feature dimensionality often does and we can attain similar error rates in feature spaces having much lower dimensionality than the initial space. Very often even a reduction to a single dimension is possible.

In case of using the linear classifier (see Table 2), we see that in nine of the 12 data sets the Chernoff criterion was ranked among the best. In six cases, it provides the overall optimal LDR (indicated by the "*"s). The second best is LDR based on the Fisher criterion: In eight of the 12 cases, it is ranked among the best, and in five cases, it provides the optimal result. Both criteria produce in two cases an MCE that is significantly less in comparison to the other three MCEs: For the Fisher criterion these are data sets (j) and (k), for the Chernoff criterion (b) and (d). However, the performance improvement of Chernoff on data set (b) is, although significant in comparison to the other three, not very large. The same holds for the Fisher criterion on data set (j). The technique of Tubbs et al. provides the single optimal MCE on data set (f). The Mahalanobis distance-based approach is on none of the data sets the sole optimal technique. Note also that the Fisher criterion gives generally lower-dimensional data set representations as best solution.

For the classification results by a quadratic classifier (Table 3), the observations are different. The Mahalanobis distance-based technique performs relatively much better now. It ranks in eight of the 12 times among the best and provides in four cases the overall optimal results. In addition, for data set (j), it is significantly better compared to the three other transforms. However, again the Chernoff criterion scores best: In 11 of the 12 data sets, it ranks between the best performing LDR techniques, in eight of these cases it produces the optimal transform, and in two cases it provides the single optimal representation significantly better than the other three representations. Using the quadratic classifier, the results for the Fisher criterion get relatively worse.

Specifically comparing Chernoff to Fisher, the experiments show that, especially when using a quadratic classifier, Chernoff can improve significantly upon Fisher (in four out of 12 data sets). When using the linear classifier, Fisher can improve significantly upon Chernoff, which we see in two of the 12 instances. However, Chernoff now gives a significant improvement in three cases. In general, the Chernoff approach compares favorable to Fisher's LDA, giving only inferior results in very few cases.

5 Discussion and Conclusions

The linear dimension reduction (LDR) criterion presented in this paper extends the well-known Fisher criterion, as used in linear discriminant analysis (LDA), in a way that it can also deal with the heteroscedasticity of the data, i.e., it takes into account differences in within-class covariance matrices and the discriminatory information therein. After establishing the link between the squared Euclidean distance between classes and the Fisher criterion, the two-class heteroscedastic Chernoff criterion is defined by means of the Chernoff distance between two classes using the notion of directed distance matrices. Subsequently, the multiclass Chernoff criterion is constructed via a certain decomposition of the multiclass Fisher criterion in multiple two-class Fisher criteria. Substituting these two-class Fisher criteria by the two-class Chernoff criterion finally leads to our multiclass Chernoff criterion. Using the latter criterion, we can compute a LDR transform in a simple and efficient way comparable to LDA. It merely uses standard matrix arithmetics, avoiding complex or iterative procedures.

Using 12 data sets from the UCI Repository (Table 1), we compared our technique to Fisher's LDA and to two singular value decomposition-based methods for dimensionality reduction. One of these, the technique from [24], can also deal with heteroscedastic data. The other approach, which is Mahalanobis distance-based, is primarily homoscedastic and more directly related to the Fisher criterion (see [2]).

The experiments showed the clear improvements possible when using the Chernoff criterion instead of Fisher's. The improvements are slightly better in cases where the quadratic classifier is used. This may be due to the fact that the quadratic classifier takes second order information into account, as does the Chernoff criterion. In general, and not only compared to LDA, the Chernoff criterion gives better results in cases where the quadratic classifier is used. For the latter, Chernoff ranks among the best transforms in 11 of the 12 cases, while for the linear classifier this is nine out of 12.

The performance of the Chernoff-based technique is in both the linear and the quadratic case better than any of the three other tested LDR techniques. It significantly outperforms all other three transformations in only four of the 24 instances (using the linear classifier on data set (b) and (d), and using the quadratic classifier on data sets (d) and (i)). However, with respect to accuracy, the experiments indicate that doing Chernoff criterion-based LDR gives results better than, or at least comparable to, results obtained with any of the other three transforms. With respect to obtaining a lower-dimensional representation, there are few instances in which the Fisher or Mahalanobis-based transforms provide better representations, but also for these, the Chernoff criterion, in most cases, produces good results.

The main reason for the Chernoff criterion to work well for dimensionality reduction is that it, in a certain way, quantifies the amount of discrimination information in the several subspaces. The Chernoff distance is determined assuming the classes to be normally distributed; however, what is important is that it generally expresses discrimination information in terms of simple first and second order moments. In addition, there are only few parameters to be estimated in order to derive the criterion and obtain its associated eigenvectors and, therefore, it also allows for good generalization.

An improvement of the method may be possible by using some form of penalization [11], by weighting the relative contributions of the pairwise term [17] confining the influence of otherwise dominant terms on the final criterion, or by reweighting all eigenvalues of the individual terms [15]. All these techniques rely on a certain form of regularization of the covariance terms in (10). However, success is, of course, not necessarily guaranteed.

In conclusion, the multiclass Chernoff criterion provides a good alternative to the well-known Fisher criterion and extends its use to linear dimension reduction for heteroscedastic data. Although the number of data sets used for the tests is merely 12, these experiments clearly show the improvements possible when utilizing the Chernoff criterion, also in comparison with two other dimensionality reduction schemes.

ACKNOWLEDGMENTS

The authors would like to express their appreciation to the four anonymous reviewers for their critical and extensive appraisal and their thoughtful comments and suggestions.

REFERENCES

[1] S. Aeberhard, O. de Vel, and D. Coomans, "Comparative Analysis of Statistical Pattern Recognition Methods in High Dimensional Settings," *Pattern Recognition,* vol. 27, pp. 1065-1077, 1994.
[2] H. Brunzell and J. Eriksson, "Feature Reduction for Classification of Multidimensional Data," *Pattern Recognition,* vol. 33 pp. 1741-1748, 2000.
[3] L.J. Buturovic, "Toward Bayes-Optimal Linear Dimension Reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, pp. 420-424, 1994.
[4] C.H. Chen, "On Information and Distance Measures, Error Bounds, and Feature Selection," *The Information Scientist,* vol. 10, pp. 159-173, 1979.
[5] J.K. Chung, P.L. Kannappan, C.T. Ng, and P.K. Sahoo, "Measures of Distance between Probability distributions," *J. Math. Analysis and Applications,* vol. 138, pp. 280-292, 1989.
[6] T.M. Cover and J.A. Thomas, *Elements of Information Theory.* New York: Wiley Interscience 1991.
[7] H.P. Decell and S.M. Mayekar, "Feature Combinations and the Divergence Criterion," *Computers and Math. with Applications,* vol. 3, pp. 71-76, 1977.
[8] P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach.* London: Prentice-Hall, 1982.
[9] R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics,* vol. 7, pp. 179-188, 1936.
[10] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic Press, 1990.
[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* New York: Springer-Verlag, 2001.
[12] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 1, pp. 4-37, Jan. 2000.
[13] N. Kumar and A.G. Andreou, "Generalization of Linear Discriminant Analysis in a Maximum Likelihood Framework," *Proc. Joint Meeting of the Am. Statistical Assoc.,* 1996.
[14] X. Liu, A. Srivastava, and K. Gallivan, "Optimal Linear Representations of Images for Object Recognition," *Proc. 2003 Conf. Computer Vision and Pattern Recognition,* pp. 229-234, June 2003.
[15] M. Loog, *Approximate Pairwise Accuracy Criteria for Multiclass Linear Dimension Reduction: Generalisations of the Fisher Criterion,* Number 44 in WBBM Report Series. Delft, The Netherlands: Delft Univ. Press, 1999.
[16] M. Loog and R.P.W. Duin, "Non-Iterative Heteroscedastic Linear Dimension Reduction for Two-Class Data. From Fisher to Chernoff," *Proc. Fourth Int'l Workshop S+SSPR 2002,* pp. 508-517, 2002.
[17] M. Loog, R.P.W. Duin, and R. Haeb-Umbach, "Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, pp. 762-766, 2001.
[18] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition.* New York: John Wiley & Sons, 1992.
[19] P.M. Murphy and D.W. Aha UCI Repository of Machine Learning Databases, www.ics.uci.edu/mlearn/mlrepository.html, 2004.
[20] T. Okada and S. Tomita, "An Extended Fisher Criterion for Feature Extraction—Malina's Method and Its Problems," *Electronics and Comm. Japan,* vol. 67, pp. 10-17, 1984.
[21] C.R. Rao, "The Utilization of Multiple Measurements in Problems of Biological Classification," *J. Royal Statistical Soc., Series B,* vol. 10, pp. 159-203, 1948.
[22] J.A. Rice, *Mathematical Statistics and Data Analysis,* second ed. Belmont: Duxbury Press, 1995.
[23] M. Röhl and C. Weihs, "Optimal vs. Classical Linear Dimension Reduction," *Proc. 22nd Ann. GfKl Conf.,* pp. 252-259, 1998.
[24] J.D. Tubbs, W.A. Coberly, and D.M. Young, "Linear Dimension Reduction and Bayes Classification," *Pattern Recognition,* vol. 15, pp. 167-172, 1982.

Marco Loog graduated in mathematics from the Mathematical Institute, Utrecht University, The Netherlands, in 1997. In 1999, he finished the post-Master's program "Mathematical Support and Decision Models" in the Department of Mathematics and Computer Science at Delft University of Technology, The Netherlands. As part of this two year program, he did a one-year research project within the speech processing group at the Philips Research Laboratories Aachen, Germany. Since June 2000, he has been working as a PhD student at the Image Sciences Institute, Utrecht, The Netherlands, on the project entitled "Knowledge-Based filtering of X-Thorax Images," which is an IOP Project funded by the Dutch Ministry of Economic Affairs. The project involves the development and improvement of statistical pattern recognition methods, and their use in the processing and analysis of medical images. His principal research interests include scale space theory, folklore theorems, the back-transmutation of gold into lead, linear and nonlinear dimensionality reduction methods, and pattern analysis techniques for supervised image processing.

Robert P.W. Duin studied applied physics at Delft University of Technology in the Netherlands. In 1978, he received the PhD degree for a thesis on the accuracy of statistical pattern recognizers. In his research, he included various aspects of the automatic interpretation of measurements, learning systems, and classifiers. Between 1980 and 1990, he studied and developed hardware architectures and software configurations for interactive image analysis. After this period, his interest was redirected to neural networks and pattern recognition. At this moment, he is an associate professor in the Faculty of Electrical Engineering, Mathematics, and Computer Science at Delft University of Technology. His present research is in the design, evaluation, and application of algorithms that learn from examples. This includes neural network classifiers, support vector machines, and classifier combining strategies. Recently, he started to study alternative object representations for classification and became thereby interested in the use of relational methods for pattern recognition and in the possibilities to learn domain descriptions. He is a member of the IEEE and the IEEE Computer Society.