Extending Kernel Fisher Discriminant Analysis with the Weighted Pairwise Chernoff Criterion

Guang Dai, Dit-Yan Yeung, and Hong Chang

Department of Computer Science, Hong Kong University of Science and Technology,
Clear Water Bay, Kowloon, Hong Kong
{daiguang, dyyeung, hongch}@cs.ust.hk

Abstract. Many linear discriminant analysis (LDA) and kernel Fisher discriminant analysis (KFD) methods are based on the restrictive assumption that the data are homoscedastic. In this paper, we propose a new KFD method called heteroscedastic kernel weighted discriminant analysis (HKWDA) which has several appealing characteristics. First, like all kernel methods, it can handle nonlinearity efficiently in a disciplined manner. Second, by incorporating a weighting function that can capture heteroscedastic data distributions into the discriminant criterion, it can work under more realistic situations and hence can further enhance the classification accuracy in many real-world applications. Moreover, it can effectively deal with the small sample size problem. We have performed some face recognition experiments to compare HKWDA with several linear and nonlinear dimensionality reduction methods, showing that HKWDA consistently gives the best results.

1 Introduction

In many classification applications in machine learning and pattern recognition, dimensionality reduction of the input space often plays an important role in reducing the complexity of the classification model and possibly leading to higher classification accuracy in the lower-dimensional feature space. This process is typically referred to as feature extraction or feature selection[1]. Linear discriminant analysis (LDA) is a classical linear dimensionality reduction method for feature extraction that has been used successfully for many classification applications. However, traditional LDA suffers from at least two limitations. First, the solution of LDA is optimal only when the data distributions for different classes are homoscedastic. In particular, the probability density functions of all classes are assumed to be Gaussian with identical covariance matrix. Second, for multi-class problems involving more than two classes, the linear transformation of traditional LDA tends to preserve the inter-class distances of well-separated classes in the input space at the expense of classes that are close to each other leading to significant overlap between them, so the overall discrimination ability is further degraded. To overcome the first limitation, the maximum likelihood

[1] Feature selection may be regarded as a special case of feature extraction in which each feature is either selected or not selected as a binary decision.

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part IV, LNCS 3954, pp. 308–320, 2006.

approach [1] and mixture discriminant analysis [2] have been proposed. More recently, Loog et al. [3] proposed a heteroscedastic extension to LDA based on the Chernoff criterion. Some methods have also been proposed to overcome the second limitation. For example, [4, 5, 6] proposed using a monotonically decreasing weighting function based on Euclidean distance to balance the contribution of different class pairs to the total optimization criterion. Loog et al. [7] proposed an approximate pairwise accuracy criterion which defines the weighting function based on Bayesian error information of the class pairs. More recently, Qin et al. [8] proposed the weighted pairwise Chernoff criterion which combines the strengths of the earlier works of Loog et al. [3, 7] while it overcomes the two limitations above simultaneously. In fact, the methods in [4, 5, 6, 7] may be regarded as special cases of [8].

On the other hand, those LDA-based algorithms generally suffer from the so-called small sample size problem which arises in many real-world applications when the number of examples is smaller than the input dimensionality, i.e., the data are undersampled. A traditional solution to this problem is to apply PCA in conjunction with LDA, as was done for example in Fisherfaces [9]. Recently, more effective solutions, sometimes referred to as direct LDA (DLDA) methods, have been proposed [10, 11, 12, 13, 14]. All DLDA methods focus on exploiting the discriminatory information in the null space of the within-class scatter matrix where most discriminatory information that is crucial for classification exists.

While LDA-based methods perform well for many classification applications, their performance is unsatisfactory for many other classification problems in which nonlinear decision boundaries are necessary. Motivated by kernel machines such as support vector machine (SVM) and kernel principal component analysis (KPCA) [15], nonlinear extension of LDA called kernel Fisher discriminant analysis (KFD) by applying the "kernel trick" has been shown to improve over LDA for many applications [16, 17, 18, 19, 20, 21, 22, 23, 24]. The basic idea of KFD is to map each input data point $\mathbf{x}$ via a nonlinear mapping ϕ implicitly to a feature space $\mathcal{F}$ and then perform LDA there. Mika et al. [16] first proposed a two-class KFD algorithm which was later generalized by Baudat and Anouar [17] to give the generalized discriminant analysis (GDA) algorithm for multi-class problems. Subsequently, a number of KFD algorithms [18, 19, 20, 21, 22, 23, 24] have been developed. However, these KFD-based algorithms suffer from the small sample size problem a lot more than the LDA-based ones since the kernel-induced feature space is typically of very high or even infinite dimensionality. Many methods have been proposed to address this problem. Mike et al. [16] proposed adding a small multiple of the identity matrix to make the inner product matrix invertible. Baudat and Anouar [17] and Xiong et al. [18] used QR decomposition to avoid the singularity of the inner product matrix. Park et al. [19] proposed the KFD/GSVD algorithm by employing generalized singular value decomposition (GSVD). Yang [20] adopted the technique introduced in Fisherfaces [9], i.e., kernel Fisherfaces. Lu et al. [21] proposed the kernel direct discriminant analysis (KDDA) algorithm based on generalization of the LDA algorithm in [11]. Recently, [22, 23] presented a further enhanced method called the kernel generalized

nonlinear discriminant analysis (KGNDA) algorithm which is based on the theoretical foundation established in [24]. More specifically, it attempts to exploit the crucial discriminatory information in the null space of the within-class scatter matrix in the feature space $\mathcal{F}$.

Similar to traditional LDA, however, most existing KFD-based algorithms, including KGNDA, are not optimal under the multi-class case as they tend to overemphasize the classes that are more separable and at the same time they are incapable of dealing with heteroscedastic data that are commonly found in real-world applications. In this paper, based on the idea of weighted pairwise Chernoff criterion proposed in [8], we further improve the overall discrimination ability of KGNDA by proposing a novel KFD algorithm called heteroscedastic kernel weighted discriminant analysis (HKWDA). We study the combination of the weighted pairwise Chernoff criterion and nonlinear techniques based on KFD directly, as the linear case can simply be seen as a special case when the mapping is linear, i.e., $\phi(\mathbf{x}) = \mathbf{x}$. Our method mainly focuses on improvement of the discriminatory information in the null space of the within-class scatter matrix, for two main reasons. First, this discriminatory information is crucial for improving the classification accuracy. Second, improving this discriminatory information is also the focus of other related works [10, 11, 12, 13, 14, 21, 22, 23, 24]. As a result, our proposed method has several appealing characteristics. First, like all kernel methods, it can handle nonlinearity efficiently in a disciplined manner. Second, by incorporating a weighting function that can capture heteroscedastic data distributions into the discriminant criterion, it can work under more realistic situations and hence can further enhance the classification accuracy in many real-world applications. Moreover, it can effectively deal with the small sample size problem. To demonstrate the efficacy of HKWDA, we compare it with several existing dimensionality reduction methods on face recognition where both the nonlinearity problem and the small sample size problem generally exist.

2 Existing Kernel Fisher Discriminant Analysis Algorithms

As discussed above, KFD algorithms essentially perform LDA in the feature space $\mathcal{F}$. Computation of the inner product of two vectors in $\mathcal{F}$ does not require applying the nonlinear mapping ϕ explicitly when the kernel trick is applied through using a kernel function $k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$. We regard a matrix as an operator in the feature space $\mathcal{F}$ which is a Hilbert space. Moreover, for any operator $\mathbf{A}$ in a Hilbert space $\mathcal{H}$ (which may be the feature space $\mathcal{F}$), we let $\mathbf{A}(0)$ denote the null space of $\mathbf{A}$, i.e., $\mathbf{A}(0) = \{\mathbf{x} | \mathbf{A}\mathbf{x} = 0\}$, and $\mathbf{A}^{\perp}(0)$ denote the orthogonal complement space of $\mathbf{A}(0)$, i.e., $\mathbf{A}(0) \bigoplus \mathbf{A}^{\perp}(0) = \mathcal{H}$.

Let $\mathbf{x}_i \, (i = 1, \ldots, N)$ denote N points in the training set $\mathcal{X}$. We partition $\mathcal{X}$ into c disjoint subsets $\mathcal{X}_i$, i.e., $\mathcal{X} = \bigcup_{i=1}^{c} \mathcal{X}_i$, where $\mathcal{X}_i$ consists of N_i points that belong to class i with $N = \sum_{i=1}^{c} N_i$. The between-class scatter operator $\mathbf{S}_b$, within-class scatter operator $\mathbf{S}_w$, and population scatter operator $\mathbf{S}_t$ can be expressed as follows [24]: $\mathbf{S}_b = \frac{1}{N} \sum_{i=1}^{c} N_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$, $\mathbf{S}_w =$

$\frac{1}{N}\sum_{i=1}^{c}\sum_{\mathbf{x}_j\in\mathcal{X}_i}(\phi(\mathbf{x}_j)-\mathbf{m}_i)(\phi(\mathbf{x}_j)-\mathbf{m}_i)^T$, and $\mathbf{S}_t=\mathbf{S}_b+\mathbf{S}_w=\frac{1}{N}\sum_{i=1}^{N}(\phi(\mathbf{x}_i)-\mathbf{m})(\phi(\mathbf{x}_i)-\mathbf{m})^T$, where $\mathbf{m}_i=\frac{1}{N_i}\sum_{\mathbf{x}_j\in\mathcal{X}_i}\phi(\mathbf{x}_j)$ denotes the sample mean of class i in $\mathcal{F}$ and $\mathbf{m}=\frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{x}_i)$ denotes the sample mean of all N points in $\mathcal{F}$. We maximize the following criterion function to find the optimal coefficients $\mathbf{w}$ for the discriminants:

$$J(\mathbf{w})=\frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}}. \tag{1}$$

However, many algorithms [16, 17, 18, 19, 20, 21] presented for KFD have not effectively solved the small sample size problem with respect to (5) and they generally discard the intersection space $\mathbf{S}_w(0)\bigcap\mathbf{S}_b^{\perp}(0)$ which potentially contains useful discriminatory information that can help to improve the classification accuracy. Recently, KGNDA was proposed to solve this problem [22, 23, 24]. To prevent the loss of crucial discriminatory information, the procedure of computing optimal discriminant coefficients in $\mathcal{F}$, which essentially can be considered as a nonlinear extension of DLDA [10, 12, 13, 14], is applied in KGNDA. KGNDA is based on the assumption that discriminatory information in F can be obtained from the intersection space $\mathbf{S}_w(0)\bigcap\mathbf{S}_t^{\perp}(0)$, since the intersection space $\mathbf{S}_w(0)\bigcap\mathbf{S}_t^{\perp}(0)$ is equivalent to the intersection space $\mathbf{S}_w(0)\bigcap\mathbf{S}_b^{\perp}(0)$ in practice. To obtain $\mathbf{S}_w(0)\bigcap\mathbf{S}_t^{\perp}(0)$, KGNDA first computes $\mathbf{S}_t^{\perp}(0)$ by the eigenanalysis of $\mathbf{S}_t$ in $\mathcal{F}$ (which essentially performs KPCA), and then obtains this intersection space by the eigenanalysis of the projection of $\mathbf{S}_w$ in $\mathbf{S}_t^{\perp}(0)$. Since $\mathbf{S}_w(0)\bigcap\mathbf{S}_t^{\perp}(0)$ can be obtained, KGNDA computes the discriminant coefficients in this intersection space without discarding the useful discriminatory information there. Besides this crucial discriminatory information in $\mathbf{S}_w(0)\bigcap\mathbf{S}_t^{\perp}(0)$, KGNDA also obtains some other discriminatory information in $\mathbf{S}_w^{\perp}(0)\bigcap\mathbf{S}_t^{\perp}(0)$ at the same time. More details can be found in [22, 23, 24]. Since it is generally believed that the subspace $\mathbf{S}_w(0)\bigcap\mathbf{S}_b^{\perp}(0)$ or $\mathbf{S}_w(0)\bigcap\mathbf{S}_t^{\perp}(0)$ contains most discriminatory information for classification, many recently developed discriminant analysis algorithms [10, 11, 12, 13, 14, 21, 22, 23, 24, 25] actually mainly focus on this subspace.

3 Our Heteroscedastic Kernel Weighted Discriminant Analysis Algorithm

Since KFD is essentially LDA in the feature space $\mathcal{F}$, the two limitations of traditional LDA, i.e., data homoscedasticity assumption and overemphasis on well-separated classes, as discussed in Section 1 are still applicable here. In this section, we present our HKWDA algorithm based on the weighted pairwise Chernoff criterion, by incorporating into the discriminant criterion in $\mathcal{F}$ a weighting function that does not rely on the restrictive homoscedasticity assumption. The theoretical results outlined in this section can be proved by applying tools from functional analysis in the Hilbert space, but their proofs are omitted here due to space limitation.

Based on the multi-class Chernoff criterion presented in [3], we replace the conventional between-class scatter operator $\mathbf{S}_b$ by a positive semi-definite between-class operator $\mathbf{S}_o$ as defined below:

$$\mathbf{S}_o = \frac{1}{N^2}\sum_{i=1}^{c}\sum_{j=i+1}^{c} N_i N_j \mathbf{S}_w^{1/2}\{(\mathbf{S}_w^{-1/2}\mathbf{S}_{i,j}\mathbf{S}_w^{-1/2})^{-1/2}\mathbf{S}_w^{-1/2}(\mathbf{m}_i - \mathbf{m}_j) \times (\mathbf{m}_i - \mathbf{m}_j)^T\mathbf{S}_w^{-1/2}(\mathbf{S}_w^{-1/2}\mathbf{S}_{i,j}\mathbf{S}_w^{-1/2})^{-1/2} + \frac{1}{\pi_i\pi_j}[\log(\mathbf{S}_w^{-1/2}\mathbf{S}_{i,j}\mathbf{S}_w^{-1/2}) - \pi_i \log(\mathbf{S}_w^{-1/2}\mathbf{S}_i\mathbf{S}_w^{-1/2}) - \pi_j \log(\mathbf{S}_w^{-1/2}\mathbf{S}_j\mathbf{S}_w^{-1/2})]\}\mathbf{S}_w^{1/2}, \quad (2)$$

where $\pi_i = N_i/(N_i + N_j)$ and $\pi_j = N_j/(N_i + N_j)$ are the prior probabilities of classes i and j, respectively, $\mathbf{S}_{i,j} = \pi_i\mathbf{S}_i + \pi_j\mathbf{S}_j$, and $\mathbf{S}_i$ and $\mathbf{S}_j$ the covariance operators of classes i and j, respectively. The detailed derivation is omitted here but can be found in [3].

Although the multi-class Chernoff criterion can effectively handle heteroscedastic data, it still cannot overcome the second limitation mentioned above. Moreover, direct computation of $\mathbf{S}_o$ in $\mathcal{F}$ is inconvenient or even computationally infeasible. To overcome the second limitation, we introduce a weighting function to the discriminant criterion as in [4, 5, 7], where a weighted between-class scatter operator is defined to replace the conventional between-class scatter operator. To overcome both limitations and make the computation in $\mathcal{F}$ tractable simultaneously, we define a weighted between-class scatter operator $\mathbf{S}_B$ on the Chernoff distance measure in $\mathcal{F}$ based on the previous work in [3, 4, 5, 7, 8]:

$$\mathbf{S}_B = \frac{1}{N^2}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c} N_i N_j w(d_{i,j})(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T, \quad (3)$$

with the weighting function defined as $w(d_{i,j}) = \frac{1}{2d_{i,j}^2}\mathrm{erf}(\frac{d_{i,j}}{2\sqrt{2}})$, where $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$ is the pairwise approximated Bayesian accuracy and $d_{i,j} = \frac{\pi_i\pi_j}{2}(\mathbf{m}_i - \mathbf{m}_j)\mathbf{S}_{i,j}^{-1}(\mathbf{m}_i - \mathbf{m}_j) + \frac{1}{2}(\log|\mathbf{S}_{i,j}| - \pi_i \log|\mathbf{S}_i| - \pi_j \log|\mathbf{S}_j|)$ is the pairwise Chernoff distance measure between the means of classes i and j in $\mathcal{F}$. From the definition of the weighting function $w(d_{i,j})$, it can be seen that classes that are closer together in the feature space and thus can potentially impair the classification performance should be more heavily weighted in the input space. In addition, by considering the pairwise Chernoff distance, the heteroscedastic characteristic can be explicitly taken into account. One method for computing the Chernoff distance between two classes in the feature space has been presented in [26], which is based on the kernel extension of the probabilistic principal component analysis [27].

Based on the weighted between-class scatter operator $\mathbf{S}_B$ defined in (3), we define a new population scatter operator $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_w$. Same as the traditional scatter operators, the new scatter operators satisfy the following properties.

Lemma 1. *Both the operators* $\mathbf{S}_B$ *and* $\mathbf{S}_T$ *are*

1. *bounded,*
2. *compact,*
3. *self-adjoint (symmetric), and*
4. *positive on the Hilbert space* $\mathcal{F}$.

From Lemma 1 and [24], we define our new kernel discriminant criterion as follows.

Definition 1. *The weighted pairwise Chernoff criterion in $\mathcal{F}$ is defined as*

$$J_1(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}} \qquad or \qquad J_2(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_T\mathbf{w}}. \tag{4}$$

From [21], both criteria are equivalent in that they should lead to the same solution. According to Lemma 1, Definition 1 and the recent work in [22, 23, 24], we assume the crucial discriminatory information with respect to $J_1(\mathbf{w})$ or $J_2(\mathbf{w})$ only exists in the intersection space $\mathbf{S}_w(0) \bigcap \mathbf{S}_B^{\perp}(0)$.

Lemma 2. *The space $\mathbf{S}_w(0) \bigcap \mathbf{S}_B^{\perp}(0)$ is equivalent to the space $\mathbf{S}_w(0) \bigcap \mathbf{S}_T^{\perp}(0)$.*

From Lemma 2, the crucial discriminatory information can also be obtained from the intersection space $\mathbf{S}_w(0) \bigcap \mathbf{S}_T^{\perp}(0)$.[2] However, it is intractable to compute this intersection space for two reasons. First, it is intractable to compute $\mathbf{S}_w(0)$ since the dimensionality of $\mathcal{F}$ may be arbitrarily large or even infinite. Second, it is intractable to compute $\mathbf{S}_T^{\perp}(0)$ by the eigenanalysis of $\mathbf{S}_T$, since $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_w$. Fortunately, we note the following two lemmas.

Lemma 3. *The discriminant vectors with respect to $J_1(\mathbf{w})$ and $J_2(\mathbf{w})$ can be computed in the space $\mathbf{S}_T^{\perp}(0)$ without any loss of the discriminatory information.*

Lemma 4. *The space $\mathbf{S}_T^{\perp}(0)$ is equivalent to the space $\mathbf{S}_t^{\perp}(0)$.*

According to Lemma 3, it is more reasonable to first compute $\mathbf{S}_T^{\perp}(0)$. Moreover, from Lemma 4, we can use $\mathbf{S}_w(0) \bigcap \mathbf{S}_t^{\perp}(0)$ in place of $\mathbf{S}_w(0) \bigcap \mathbf{S}_T^{\perp}(0)$.

From KGNDA [22, 23, 24], we can compute the intersection space $\mathbf{S}_w(0) \bigcap \mathbf{S}_t^{\perp}(0)$ by the eigenanalysis of $\mathbf{S}_t$ and $\mathbf{S}_w$ in $\mathcal{F}$, as follows:

- **Eigenanalysis of $\mathbf{S}_t$ in $\mathcal{F}$:**
  To obtain $\mathbf{S}_t^{\perp}(0)$, we need to compute the orthonormal basis of $\mathbf{S}_t^{\perp}(0)$ which can be obtained by applying KPCA. Then, $\mathbf{S}_t$ in (4) can be rewritten as:

$$\mathbf{S}_t = \sum_{i=1}^{N} \bar{\phi}(\mathbf{x}_i)\bar{\phi}(\mathbf{x}_i)^T = \boldsymbol{\Phi}_t\boldsymbol{\Phi}_t^T, \tag{5}$$

  where $\bar{\phi}(\mathbf{x}_i) = \sqrt{1/N}(\phi(\mathbf{x}_i) - \mathbf{m})$ and $\boldsymbol{\Phi}_t = [\bar{\phi}(\mathbf{x}_1), \ldots, \bar{\phi}(\mathbf{x}_N)]$. It is generally believed that direct computation of the orthonormal basis is intractable, since the order of the operator $\mathbf{S}_t$ is arbitrarily large or even infinite in $\mathcal{F}$. One solution is to compute the eigenvectors and eigenvalues of $N \times N$ matrix $\boldsymbol{\Phi}_t^T\boldsymbol{\Phi}_t$ [22, 23, 24, 26].

  For all training examples $\{\phi(\mathbf{x}_i)\}_{i=1}^N$ in $\mathcal{F}$, we can define an $N \times N$ kernel matrix $\mathbf{K}$ as $\mathbf{K} = [k_{ij}]_{N\times N}$, where $k_{ij} = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$. Hence, by the kernel trick, $\boldsymbol{\Phi}_t^T\boldsymbol{\Phi}_t$ can be expressed as

[2] In fact, direct computation of $\mathbf{S}_B^{\perp}(0)$ will lead to some loss of the crucial discriminatory information. See [23, 25] for analysis of KDDA [21].

$$\boldsymbol{\Phi}_t^T \boldsymbol{\Phi}_t = \frac{1}{N}\left[\mathbf{K} - \frac{1}{N}(\mathbf{K}\mathbf{1}_{N\times N} + \mathbf{1}_{N\times N}\mathbf{K}) + \frac{1}{N^2}\mathbf{1}_{N\times N}\mathbf{K}\mathbf{1}_{N\times N}\right], \quad (6)$$

where $\mathbf{1}_{N\times N}$ is an $N\times N$ matrix with all terms being one. Let λ_i and $\mathbf{e}_i\,(i=1,\ldots,m)$ be the ith positive eigenvalue and the corresponding eigenvector of $\boldsymbol{\Phi}_t^T\boldsymbol{\Phi}_t$, respectively. According to [22, 23, 24, 26], it is clear that $\mathbf{v}_i = \boldsymbol{\Phi}_t\mathbf{e}_i\lambda_i^{-1/2}\,(i=1,\ldots,m)$ constitute the orthonormal basis of $\mathbf{S}_t^{\perp}(0)$.

– **Eigenanalysis of $\mathbf{S}_w$ in $\mathcal{F}$:**
Projecting $\mathbf{S}_w$ onto the subspace spanned by $\mathbf{v}_i = \Phi_t\mathbf{e}_i\lambda_i^{-1/2}\,(i=1,\ldots,m)$, it is clear that the projection $\bar{\mathbf{S}}_w$ of $\mathbf{S}_w$ in this subspace can be expanded as

$$\bar{\mathbf{S}}_w = \mathbf{V}^T\mathbf{S}_w\mathbf{V} = \mathbf{E}^T\boldsymbol{\Xi}^T\boldsymbol{\Xi}\mathbf{E}. \quad (7)$$

Here, $\mathbf{V} = [\mathbf{v}_1,\ldots,\mathbf{v}_m]$, $\mathbf{E} = [\mathbf{e}_1\lambda_1^{-1/2},\ldots,\mathbf{e}_m\lambda_m^{-1/2}]$, and $\boldsymbol{\Xi} = \mathbf{K}/N - \mathbf{1}_{N\times N}\mathbf{K}/N^2 - \mathbf{A}_{N\times N}\mathbf{K}/N + \mathbf{1}_{N\times N}\mathbf{K}\mathbf{A}_{N\times N}/N^2$, where $\mathbf{A}_{N\times N} = \mathrm{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_m)$ is a block-diagonal matrix with $\mathbf{A}_i$ being an $N_i \times N_i$ matrix with all its terms equal to $1/N_i$.

Let $\mathbf{P} = [\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_l]$ be the corresponding eigenvectors of the zero eigenvalues of $\bar{\mathbf{S}}_w$. So it is clear that $\mathbf{S}_w(0)\bigcap\mathbf{S}_T^{\perp}(0)$ can be spanned by $\mathbf{VP}$. Then, the optimal discriminant vectors with respect to $J_1(\mathbf{w})$ or $J_2(\mathbf{w})$ can be computed in $\mathbf{S}_w(0)\bigcap\mathbf{S}_T^{\perp}(0)$ without the loss of crucial discriminatory information. From [22, 23, 24], since the between-class distance is equal to zero in $\mathbf{S}_w(0)\bigcap\mathbf{S}_T^{\perp}(0)$, the weighted pairwise Chernoff criterion in (9) can be replaced by $\hat{J}(\mathbf{w}) = \mathbf{P}^T\mathbf{V}^T\mathbf{S}_B\mathbf{P}\mathbf{V}$. By the kernel trick, it can be expanded as:

$$\hat{J}(\mathbf{w}) = \mathbf{P}^T\mathbf{V}^T\mathbf{S}_B\mathbf{V}\mathbf{P} = \mathbf{P}^T\mathbf{E}^T\left[\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}\left(\frac{\sqrt{N_iN_j}}{N^{3/2}}w(d_{i,j})\mathbf{Z}_{i,j}^T\mathbf{Z}_{i,j}\right)\right]\mathbf{E}\mathbf{P}, \quad (8)$$

where $\mathbf{P} = [\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_l]$, $\mathbf{V} = [\mathbf{v}_1,\ldots,\mathbf{v}_m]$, $\mathbf{E} = [\mathbf{e}_1\lambda_1^{-1/2},\ldots,\mathbf{e}_m\lambda_m^{-1/2}]$, $\mathbf{Z}_{i,j} = \mathbf{K}\mathbf{L}_i + \mathbf{H}\mathbf{K}\mathbf{L}_j - \mathbf{K}\mathbf{L}_j - \mathbf{H}\mathbf{K}\mathbf{L}_i$, $\mathbf{H}$ is an $N\times N$ matrix with all terms being $1/N$, $\mathbf{L}_i$ is an $N\times 1$ matrix where the terms corresponding to class i are $1/N_i$ and the remaining terms are zero. It is clear that the matrix $\mathbf{P}^T\mathbf{V}^T\mathbf{S}_B\mathbf{V}\mathbf{P}$ is a tractable $l\times l$ matrix. Let $\mathbf{z}_i\,(i=1,\ldots,l)$ be the eigenvectors of $\mathbf{P}^T\mathbf{V}^T\mathbf{S}_B\mathbf{V}\mathbf{P}$, sorted in descending order of the corresponding eigenvalues λ_i. According to [22, 23, 24], it is clear that $\mathbf{Y}_i = \mathbf{V}\mathbf{P}\mathbf{z}_i\,(i=1,\ldots,l)$ constitute the optimal discriminant vectors with respect to the weighted pairwise Chernoff criterion (4) in $\mathcal{F}$.

This gives the new HKWDA algorithm. For an input pattern $\mathbf{x}$, its projection onto the subspace spanned by $\boldsymbol{\Theta} = [\mathbf{Y}_1,\ldots,\mathbf{Y}_l]$ can be computed as $\mathbf{z} = \boldsymbol{\Theta}^T\phi(\mathbf{x})$. This expression can be rewritten via the kernel trick as follows: $\mathbf{z} = \sqrt{\frac{1}{N}}(\mathbf{z}_1,\ldots,\mathbf{z}_l)^T\mathbf{P}^T\mathbf{E}^T\mathbf{k}_x$, where $\mathbf{k}_x = (K(\mathbf{x},\mathbf{x}_1) - \frac{1}{N}\sum_{i=1}^{N}K(\mathbf{x},\mathbf{x}_i), \ldots, K(\mathbf{x},\mathbf{x}_N) - \frac{1}{N}\sum_{i=1}^{N}K(\mathbf{x},\mathbf{x}_i))^T$.

Thus, HKWDA can give a low-dimensional representation with enhanced discriminating power on the whole. Moreover, this method also effectively addresses the nonlinearity problem and the small sample size problem.

4 Experimental Results

To assess the performance of the HKWDA algorithm proposed in this paper, we conduct some face recognition experiments to compare HKWDA with other dimensionality reduction methods. Note that typical face recognition applications suffer from the small sample size problem and require nonlinear methods, which are particularly suitable for demonstrating the strengths of HKWDA. In addition, real-world face image databases seldom satisfy the restrictive homoscedasticity assumption.

Our experiments are performed on two different data sets:

1. Mixed data set of 1545 images from 117 subjects which are obtained from four different image sources:
   - 47 subjects from the FERET database, with each subject contributing 10 gray-scale images.
   - 40 subjects from the ORL database, with each subject contributing 10 gray-scale images.
   - 20 subjects from the UMIST database, with a total of 575 gray-scale images.
   - 10 subjects from the YaleB database, with each subject contributing 10 gray-scale images.
2. A subset of the FERET database: 200 subjects each with four different images.

The gray-level and spatial resolution of all images in both data sets are 256 and 92×112, respectively. Since there exist large variations in illumination, facial expression and pose in both data sets, the distribution of the face image patterns is highly nonlinear, complex, and heteroscedastic.

Both data sets are randomly partitioned into two disjoint sets for training and testing, respectively. For the mixed data set, five images per subject are randomly chosen for training while the rest for testing; for the subset of the FERET database, three images per subject are randomly chosen from the four images available for each subject for training while the rest for testing. For each feature representation obtained by a dimensionality reduction method, we use a simple minimum distance classifier [24] with Euclidean distance measure to assess the classification accuracy. Each experiment is repeated 10 times and the average classification rates are reported. For the kernel methods, we use the RBF kernel function $k(\mathbf{z}_1, \mathbf{z}_2) = \exp(\|\mathbf{z}_1 - \mathbf{z}_2\|^2/\sigma)$ and polynomial kernel function $k(\mathbf{z}_1, \mathbf{z}_2) = (\mathbf{z}_1^T \mathbf{z}_2/\sigma + 1)^2$ where $\sigma = 10^9$.

To reveal the fact that HKWDA can better utilize the crucial discriminatory information in the null space of the within-class scatter operator, our first experiment compares HKWDA with the corresponding part of KGNDA

[22, 23, 24] and a special case of HKWDA, referred to as Euclidean KWDA (EKWDA), which can be seen as HKWDA where the weighting function is defined based on the Euclidean distance instead of the Chernoff distance in the feature space. It is obvious that EKWDA is based on the homoscedasticity assumption. In addition, to show the effectiveness of the nonlinear extension, we also compare the corresponding part of the DLDA method [10, 12, 13, 14] which may be seen as the linear special case of KGNDA. The experimental results shown in Fig. 1 reveal that, as expected, HKWDA outperforms KGNDA, EKWDA and DLDA for both kernel functions on the two different data sets. From the results of paired t-test with significance level 0.05, we can conclude that the results of HKWDA are significantly better than those of the other three methods. Since DLDA is a linear method, it cannot effectively extract nonlinear features and hence the classification rate is very low. Comparing HKWDA and EKWDA, we can see that relaxing the homoscedasticity assumption of the face image data can result in significant improvement in classification performance.

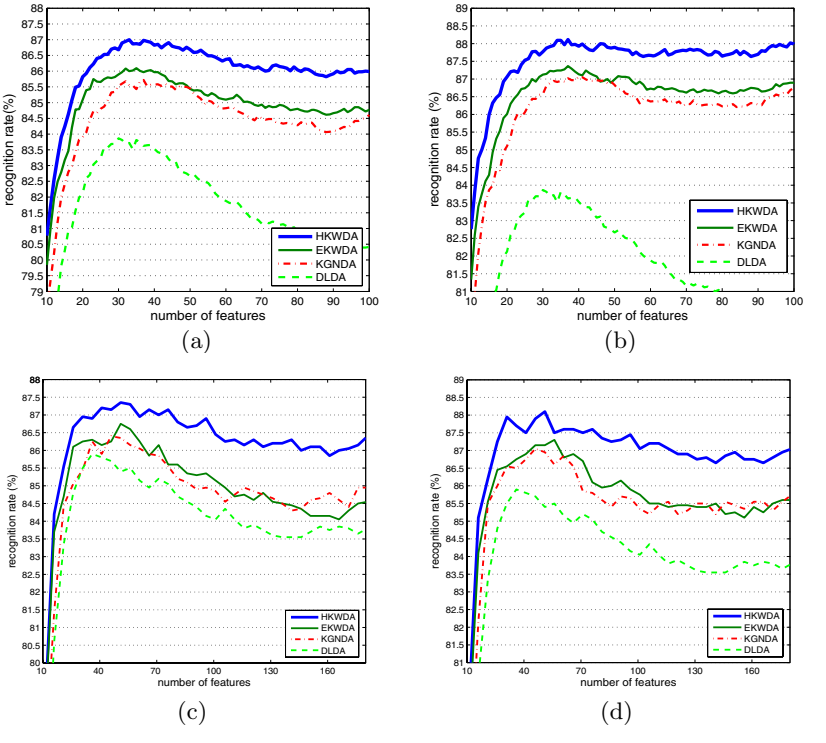

Fig. 1. Comparative performance of HKWDA, EKWDA, KGNDA and DLDA. (a) Polynomial kernel on the mixed data set; (b) RBF kernel on the mixed data set; (c) Polynomial kernel on the subset of FERET; (d) RBF kernel on the subset of FERET.

The second experiment compares HKWDA with several other kernel-based nonlinear dimensionality reduction methods, including KPCA [15], GDA [17], kernel Fisherfaces [20], KFD/QR [18], KFD/GSVD [19], and KDDA [21]. Previous works [22, 23, 24] also compare KGNDA with most of these methods in detail. Fig. 2 shows the classification rates for different methods based on the RBF kernel on both data sets. It can be seen that HKWDA is better than KPCA, GDA, kernel Fisherfaces, KFD/QR, KFD/GSVD and KDDA. In addition, we also compare different methods based on the average error percentage, which was originally proposed in [21] and can successfully evaluate the overall effectiveness of the proposed method compared with other methods. Specifically, in our experimental setting, the average percentage of the error rate of HKWDA over that of another method can be computed as the average of $(1-\alpha_i)/(1-\beta_i)$ $(i = 6, \ldots, J)$, where α_i and β_i are the recognition rates of HKWDA and another method, respectively, when i features are used. Using less than six features is not included in computing the average error percentages because the recognition rates are very low for all algorithms. Moreover, the value of J is set to 116 and 199,

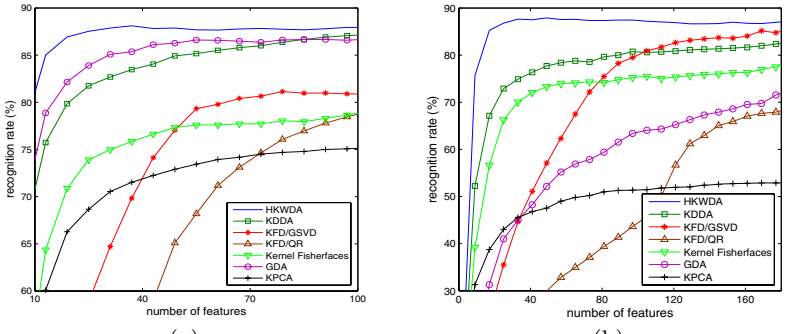

Fig. 2. Comparative performance of HKWDA and several other kernel methods based on the RBF kernel. (a) Mixed data set; (b) Subset of FERET.

Table 1. Average error percentages for different methods when compared with HKWDA

Algorithm	Mixed data set		Subset of FERET	
	Poly.	RBF	Poly.	RBF
DLDA [10, 12, 13, 14]	76.05%	67.85%	86.47%	82.31%
KPCA [15]	50.33%	45.19%	27.58%	26.38%
GDA [17]	71.17%	86.05%	33.13%	34.76%
Kernel Fisherfaces [20]	51.97%	53.01%	49.93%	27.80%
KFD/QR [18]	45.23%	39.98%	32.87%	58.79%
KFD/GSVD [19]	51.99%	51.87%	52.84%	58.79%
KDDA [21]	85.48%	81.63%	64.48%	63.35%
KGNDA [22, 23, 24]	90.57%	90.81%	91.11%	90.83%
EKWDA	93.39%	93.26%	91.42%	91.95%

respectively, for the mixed data set and the subset of FERET. The average error percentages for different methods are summarized in Table 1, showing that HKWDA is more effective than all other methods. We have performed more experiments but their results are not included in this paper due to space limitation. For example, we have performed similar experiments on another data set of 120 subjects selected from the AR database with each subject contributing 7 gray-scale images. All results consistently show that HKWDA outperforms other competing methods.

5 Conclusion

We have presented a new kernel Fisher discriminant analysis algorithm, called HKWDA, that performs nonlinear feature extraction for classification applications. By incorporating an appropriately chosen weighting function into the discriminant criterion, it can not only handle heteroscedastic data that are commonly found in real-world applications, but it can also put emphasis on classes that are close together for multi-class problems. Experimental results on face recognition are very encouraging, showing that HKWDA can consistently outperform other linear and nonlinear dimensionality reduction methods. Besides face recognition, we plan to apply HKWDA to other classification applications, including content-based image indexing and retrieval as well as video and audio classification.

Acknowledgment

This research has been supported by Competitive Earmarked Research Grant (CERG) HKUST621305 from the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China.

References

1. N. Kumar and A.G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMS for improved speech recognition. *Speech Communication*, 26:283–297, 1998.
2. T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixture. *Journal of the Royal Statistical Society, Series B*, 58:155–176, 1996.
3. M. Loog and R.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):32–739, June 2004.
4. Y.X. Li, Y.Q. Gao, and H. Erdogan. Weighted pairwise scatter to improve linear discriminant analysis. In *Proceedings of the 6th International Conference on Spoken Language Processing*, 2000.
5. R. Lotlikar and R. Kothari. Fractional-step dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):623–627, 2000.

6. J.W. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos. Face recognition using LDA-based algorithms. *IEEE Transactions on Neural Networks*, 14:195–200, 2003.
7. M. Loog, R.P.W. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(7):762–766, July 2001.
8. A.K. Qin, P.N. Suganthan, and M. Loog. Uncorrelated heteroscedastic LDA based on the weighted pairwise Chernoff criterion. *Pattern Recognition*, 2005.
9. P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:711–720, July 1997.
10. L.F. Chen, H.Y.M. Liao, M.T. Ko, J.C. Lin, and G.J. Yu. A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33:1713–1726, 2000.
11. H. Yu and J. Yang. A direct LDA algorithm for high-dimensional data with application to face recognition. *Pattern Recognition*, 34:2067–2070, 2001.
12. R. Huang, Q. Liu, H. Lu, and S. Ma. Solving the small size problem of LDA. In *Proceedings of the Sixteenth International Conference on Pattern Recognition*, volume 3, pages 29–32, August 2002.
13. J. Yang and J.Y. Yang. Why can LDA be performed in PCA transformed space? *Pattern Recognition*, 36:563–566, 2003.
14. H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana. Discriminative common vectors for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):4–13, 2005.
15. B. Schölkopf, A. Smola, and K.R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1999.
16. S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.R. Müller. Fisher discriminant analysis with kernels. In Y.H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Proceedings of the Neural Networks for Signal Processing IX*, pages 41–48, 1999.
17. G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12:2385–2404, 2000.
18. T. Xiong, J.P. Ye, Q. Li, V. Cherkassky, and R. Janardan. Efficient kernel discriminant analysis via QR decomposition. In *Advances in Neural Information Processing Systems 17*, 2005.
19. C.H. Park and H. Park. Nonlinear discriminant analysis using kernel functions and the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Application*, to appear (http://www-users.cs.umn.edu/ hpark/pub.html).
20. M.H. Yang. Kernel eigenfaces vs. kernel Fisherfaces: face recognition using kernel methods. In *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 215–220, May 2002.
21. J.W. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos. Face recognition using kernel direct discriminant analysis algorithms. *IEEE Transactions on Neural Networks*, 12:117–126, 2003.
22. G. Dai and Y.T. Qian. Modified kernel-based nonlinear feature extraction. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 17–21, 2004.
23. G. Dai and Y.T. Qian. Kernel generalized nonlinear discriminant analysis algorithm for pattern recognition. In *Proceedings of the IEEE International Conference on Image Processing*, pages 2697–2700, 2004.

24. J. Yang, A.F. Frangi, J.Y. Yang, D. Zhang, and Z. Jin. KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):230–244, 2005.
25. G. Dai and D.Y. Yeung. Nonlinear dimensionality reduction for classification using kernel weighted subspace method. In *Proceedings of the IEEE International Conference on Image Processing*, pages 838–841, 2005.
26. S.H. Zhou and R. Chellappa. From sample similarity to ensemble similarity. Technical Report SCR Technical Report (SCR-05-TR-774), Maryland University, 2005(http://www.umiacs.umd.edu/∼shaohua/).
27. M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999.