12-1-2007

Weighted Additive Criterion for Linear Dimension Reduction

Jing Peng
Montclair State University, pengj@mail.montclair.edu

Stefan Robila
Montclair State University, robilas@mail.montclair.edu

Weighted Additive Criterion for Linear Dimension Reduction

Jing Peng & Stefan Robila
Computer Science Department, Montclair State University
Montclair, NJ 07043
{peng,robila}@pegasus.montclair.edu

Abstract

Linear discriminant analysis (LDA) for dimension reduction has been applied to a wide variety of face recognition tasks. However, it has two major problems. First, it suffers from the small sample size problem when dimensionality is greater than the sample size. Second, it creates subspaces that favor well separated classes over those that are not. In this paper, we propose a simple weighted criterion for linear dimension reduction that addresses the above two problems associated with LDA. In addition, there are well established numerical procedures such as semi-definite programming for efficiently computing the proposed criterion. We demonstrate the efficacy of our proposal and compare it against other competing techniques using a number of examples.

1 Introduction

In pattern classification, a large number of features or attributes often makes the design of a classifier difficult and degrades its performance. This is particularly pronounced when the number of examples is small relative to the number of features [10, 16, 17]. This fact is due to the curse of dimensionality [2]. It states in simple terms that the number of examples required to properly compute a classifier grows exponentially with the number of features. For example, assuming features are correlated, approximating a binary distribution in a n dimensional feature space requires estimating $O(2^n)$ unknown variables [3]. In such situations, computational complexity often becomes intractable. This calls for reducing the number of features in constructing classifiers.

There are many dimensionality reduction techniques in the literature. The two most popular ones are principal components analysis (PCA) [11] and linear discriminant analysis (LDA) [6, 8]. Both techniques have been successfully applied to a wide variety of practical problems. By projecting data onto a linear subspace spanned by principal components, PCA achieves dimension reduction with the minimal data reconstruction error. On the other hand, without taking into account class information PCA cannot compute discriminant information required by classifiers. In this pape, we are concerned with LDA.

In LDA, we are given a set of l examples: $z = \{(x_i, y_i)\}_{i=1}^{l}$. These examples are independently and identically distributed (i.i.d.) from the probability space $Z = X \times Y$. Here probability measure ρ is defined but unknown, $x_i \in \Re^n$ are the n-dimensional inputs, and y_i are scalar labels. According to Fisher's criterion, one has to find a projection matrix W that maximizes:

$$J(W) = \frac{|W^T S_b W|}{|W^T S_w W|} \tag{1}$$

where S_b and S_w are so-called between-class and within-class matrices, respectively. In practice, the "small sample size" (SSS) problem is often encountered, where S_w is singular. Therefore, the maximization problem can be difficult to solve.

To address this issue, the term εI is added, where ε is a small positive number and I the identity matrix of proper size. This results in maximizing $J(W) = |W^T S_b W|/|W^T(S_w + \varepsilon I)W|$. It can then be solved without any numerical problems. This is a special case of Friedman's regularized discriminant analysis with regard to the small sample size problem [7].

Another problem with Fisher's criterion is that in multiclass problems, it creates subspaces that favor well separated classes over those that are not [14]. This is because the solution to (1) is a linear transform that maximizes the mean squared distance between the classes in the transformed space. As a result, a outlier (far away) class can be further separated from the remaining classes that really need a clear separation.

The purpose of this paper is to present a weighted additive criterion for dimension reduction that potentially provides a solution to the problems implied by the above discussions. In particular, we show that (1) our weighted additive criterion for dimension deduction is closely related to the maximum margin criterion [12]; (2) our criterion

1550-4786/07 $25.00 © 2007 IEEE
DOI 10.1109/ICDM.2007.81

has the potential to help alleviate Fisher's bias toward outlier classes in multi-class problems [14]; and (3) our objective can be optimized using efficient algorithms such as semi-definite programming, thereby avoiding the inverse of S_w and thus the potential small sample size problem. We demonstrate the efficacy of our proposed technique using a variety of examples.

2 Related Work

A number of proposals has been tabled to address the computational difficulty associated with LDA when the small sample size problem occurs (S_w becomes singular). A straightforward method (PCA+LDA) is to use the pseudo-inverse of S_w^+ in place of S_w^{-1} [22]. While simple, the method does not guarantee that Fisher's objective will be optimized by the eigenvector matrix of $S_w^+ S_b$. Another simple method is to first use PCA to remove the null space of S_w, and then apply LDA to the reduced representation. Fisherface is one such example [1]. However, this method remains sub-optimal because the null space of S_w potentially contains discriminant information [4].

Another technique, newLDA [4], first transforms the data into the null space of S_w. It then applies PCA to maximize the between-class scatter matrix in the transformed space. While newLDA mitigates the small sample size problem to the extent possible, its performance degrades with decreasing dimensions of the null space. A variant of LDA+PCA is proposed in [9]. The method first discards the null space of $S_w + S_b$ that is the common null space of both S_w and S_b. And as such, discarding this null space does not lose any discriminant information. The method then applies LDA+PCA to the reduced representation in the transformed space. A direct LDA (DLDA) is a method that throws away the the null space of S_b [25]. If $S_w + S_b$ replaces S_w, DLDA reduces to PCA+LDA [25]. We will have more to say about these null space methods later in the paper.

More recently, discriminant analysis based on maximum margin criterion is proposed [12]. The technique is closely related to LDA but does not involve inverting matrices. Since the criterion ($tr(W^t(S_b - S_w)W)$ is additive, the technique does not suffer from the small sample size problem. Classic multiclass LDA creates subspaces that favor well separated classes over those that are not. A technique based on weighted pairwise Fisher criteria is proposed that works well even when outlier classes exist [14].

3 Weighted Additive Criterion for Discriminant Analysis

In this section, we first review LDA using Fisher's criterion, and then go on to investigate discriminant analysis using a weighted additive criterion and related optimization techniques.

3.1 Linear Discriminant Analysis

In LDA, within-class, between-class, and mixture scatter matrices are used to formulate the criteria of class separability. Consider a J-class problem, where m_0 is the mean vector of all data, and m_j is the mean vector of j-th class data. A within-class scatter matrix characterizes the scatter of samples around their respective class mean vectors, and it is expressed by $S_w = \sum_{j=1}^{J} \sum_{i=1}^{l_j} (x_i^j - m_j)(x_i^j - m_j)^T$, where l_j is the size of the data in the jth class. A between-class scatter matrix characterizes the scatter of the class means around the mixture mean m_0. It is expressed by $S_b = \sum_{j=1}^{J} l_j (m_j - m_0)(m_j - m_0)^T$. The mixture scatter matrix is the covariance matrix of all samples, regardless of their class assignment, and it is given by $S_m = \sum_{i=1}^{l} (x_i - m_0)(x_i - m_0)^T = S_w + S_b$. Fisher's criterion is used to find the projection matrix that maximizes the objective (1). In order to determine the matrix W that maximizes $J(W)$, one can solve the generalized eigenvalue problem: $S_b w_i = \lambda_i S_w w_i$. The eigenvectors corresponding to the largest eigenvalues form the columns of W. For a two class problem, it can be written in a simpler form: $S_w w = m = m_1 - m_2$, where m_1 and m_2 are the means of the two classes.

3.2 Weighted Additive Criterion for Linear Dimension Reduction

Here we first focus on two class problems. The multiclass case will be discussed later. The goal of LDA is to find a direction w that simultaneously places two classes afar and minimizes within class variations. Fisher's criterion 1 achieves this goal. Alternatively, we can achieve this by

$$\max_w : w^t(\lambda S_b - S_w)w, \quad \text{subject to:} \|w\| = 1 \qquad (2)$$

where $\lambda > 0$ is a constant that weighs relative importance of the two terms S_b and S_w in determining the outcome of w. Notice that λ must be sufficiently large to ensure $(\lambda S_b - S_w)$ to be positive semi-definite. Large λ values prefer directions along which the two classes can be well separated over those that are not. On the other hand, small λ values penalize discriminants that result in large within class spread.

3.3 Relations to Maximum Margin Criterion

We show here that the objective to be maximized $J(w) = w^t(\lambda S_b - S_w)w$ is closely related to the maximum margin criterion for feature extraction proposed in

[12]. Let $\lambda = 1$, without loss of generality. We can rewrite $J(w)$ as $J(w) = tr(w^t(\lambda S_b - S_w)w)$, where tr denotes the trace operator. Then it can be shown that maximizing $tr(w^t(S_b - S_w)w)$ is equivalent to maximizing

$$J = \frac{1}{2}\sum_{i}^{2}\sum_{j}^{2} p_i p_j d(C_i, C_j) \tag{3}$$

where p_i denotes the probability of class C_i. Here the interclass distance d is defined as $d(C_i, C_j) = d(m_i, m_j) - tr(S_i) - tr(S_j)$, where S_i represents the covariance of class C_i. $tr(S_i)$ measures the overall variance of class C_i. That is, $d(C_i, C_j)$ measures the average margin between two classes. Thus, (3) is an average margin criterion. In contrast, the minimum margin criterion is used in SVMs [5, 24].

4 Computing Linear Discriminants with Semi-Definite Programming

We notice that (2) is a constraint optimization problem. It can be reformulated as follows:

$$\begin{aligned} w^t(\lambda S_b - S_w)w &= tr((\lambda S_b - S_w)ww^t) \\ &= tr((\lambda S_b - S_w)X). \end{aligned} \tag{4}$$

The notation here $X = ww^t \succeq 0$ means that the symmetric matrix X is positive semi-definite. This problem is equivalent to

$$\begin{aligned} \max_X \quad & tr((\lambda S_b - S_w)X) \\ & X \succeq 0 \end{aligned} \tag{5}$$

This is a semi-definite program (SDP), where the objective is linear with linear matrix inequality and affine equality constraints. Because linear matrix inequality constraints are convex, SDPs are convex optimization problems. SDPs arise in many applications. There are algorithms that have a good theoretical foundation to solve SDPs efficiently [23].

Assume rank(X) = 1. Since X is symmetric, one can show that rank(X) = 1 iff $X = ww^t$ for some vector w [19]. Therefore, we can recover w from X as follows. Select any column (say the ith column) of X such that $X(1, i) \neq 0$, and let

$$w = X(:, i)/X(1, i), \tag{6}$$

where $X(:, i)$ denotes the ith column of the matrix X. Thus, our goal here is to ensure the solution X to (6) has rank at most 1.

One way to guarantee rank(X) = 1 is to reformulate (6) as follows:

$$\begin{aligned} \max_X \quad & tr((\lambda S_b - S_w)X) \\ & X \succeq 0 \\ & \text{rank}(X) = 1 \end{aligned} \tag{7}$$

However, the last constraint rank(X) = 1 is not convex. Alternatively, we can replace the constraint rank(X) = 1 by the constraint $\sum_i^n X_i i = 1$. That is,

$$\begin{aligned} \max_X \quad & tr((\lambda S_b - S_w)X) \\ & X \succeq 0 \\ & I \bullet X = 1 \end{aligned} \tag{8}$$

where I is the identity matrix and the inner product of symmetric matrices is $A \bullet B = \sum_{i,j}^n a_{ij} b_{ij}$.

The constraint $I \bullet X = 1$ ensures rank(X) = 1. Here we appeal to the following theorem by Pataki [15]. First, let S be a closed convex set. A convex subset F of S is called a *face* of S if $xinF$, $y, x \in S$, $x = \frac{1}{2}(y + z)$ implies that y and z must both in F. Also, a *vertex* or an *extreme point* of S is a face consisting of a single element.

Theorem 1 *Suppose $X \in F$, where F be a face of the feasible set of (8). Let $d = dim(F)$, $r = rank(X)$. Also, let m be the number of linear constraints in (8). Then*

$$r(r + 1)/2 \leq m + d.$$

Here the *dimension* of a convex set S is $dim(S) = \max\{p | v_1, \cdots, v_p \in S\} - 1$, where the vectors $v_1, \cdots, v_p \in \Re^n$ are affinely independent. Thus if $\sum_i^p \mu_i v_i = 0$, $\sum_i^p \mu_i = 0$ implies that $\mu_i = \cdots = \mu_p = 0$.

The existence of the optimal solution X^* to the optimization problem (8) implies that F is an extreme point, i.e., a face having a single element. Therefore, at the optimum, $d = dim(F) = 0$. Also, m represents the number of linear constraints in (8). In our case, $m = 1$. Together with $r > 0$, we obtain $r = 1$. Therefore, our procedure for computing w from the matrix X (Eq. 6) is guaranteed to produce the correct answer. We call our algorithm SDP-LDA.

5 Multi-class DLA

We have presented a weighted additive criterion as an alternative to Fisher's criterion. We have shown how to optimize our criterion with semi-definite programming to obtain the optimal linear transform in two class problems, where one dimensional projection is adequate. However, LDA is generally used to find a subspace with d dimensions for multiple class problems. In this section we extend our SDP approach to LDA to the multi-class case.

Notice that LDA in a multiple class problem can be decomposed into l two class problems. In the ith two class problem, it treats the ith class as one class and all remaining classes as the second class. Each binary class problem is solved first, and after finding all subspaces, PCA is applied to find eigenvectors having the largest eigenvalues. These new eigenvectors are the solution of the original multi-class LDA problem.

To be more precise, let us look at (2) again. We can use other constraints instead of $\|w\| = 1$. For example, we can maximize $w^t \lambda S_b w$, subject to $w^t S_w w = 1$. One can show that maximizing $w^t(\lambda S_b - S_w)w$ under such a constraint results in LDA. Thus, our SDP based LDA can be naturally extended to the multi-class case. Simply, in the decomposition step, we replace S_b in two class LDA by λS_b. In the linear case, it turns out that the SDP based LDA algorithm in the multi-class case simply solves $\lambda S_b W = S_w W \Lambda$.

6 Nonlinear Extension

Our weighted criterion (2) can be extended to the nonlinear case as well. We can follow the approach (kernel trick) of nonlinear SVMs to kernelize the linear feature extractor [5, 24]. In general, we use kernel functions to implicitly perform a nonlinear mapping ϕ to embed the data into a feature space F, where linear feature extraction is carried out. Common kernels are Gaussian $k(x, y) = e^{-\gamma\|x-y\|^2}$ and degree d polynomials $k(x, y) = (1 + x \cdot y)^d$.

We skip the detail of the derivations. Briefly, write $w = \sum_i^l \alpha_i \phi(x_i)$, and

$$\hat{m}_i = (\frac{1}{n_i}\sum_{j=1}^{n_i} k(x_1, x_j^{(i)}), \cdots, \frac{1}{n_i}\sum_{j=1}^{n_i} k(x_l, x_j^{(i)}))^t,$$

where $x_j^{(i)}$ is in the ith class, and n_i the number of examples in the ith class. Also, $\hat{m} = \sum_{i=1}^2 p_i \hat{m}_i$. Then, using the kernel trick we can show that we can rewrite (2) in feature space as $\max_w : \alpha^t(\lambda \hat{S}_b - \hat{S}_w)\alpha$, subject to: $\|\alpha\| = 1$, where $\hat{S}_b = \sum_{i=1}^2 p_(\hat{m}_i - \hat{m})(\hat{m}_i - \hat{m})^t$, and

$$\hat{S}_w = \sum_{i=1}^{2} p_i \frac{1}{n_i}\mathbf{K}_i(I_{n_i} - \frac{1}{n_i}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t)\mathbf{K}_i^t.$$

After solving $\alpha = [\alpha_1, \cdots, \alpha_l]^t$, any given new data x will be projected onto the subspace by $s = \sum_{i=1}^l \alpha_i k(x_i, x)$.

7 Discussions

We are interested in comparing the proposed SDP-LDA algorithm to various LDA algorithms, such as PCA+LDA [1, 21], scatter-LDA [13, 8], and newLDA [4]. These techniques are mostly proposed for solving face recognition problems where the SSS problem will always occur. We summarize them briefly:

PCA+LDA: Apply PCA to remove the null space of S_w first, then maximize $J(W) = |W^T S_b W|/|W^T S_w W|$.

Scatter-LDA: Same as PCA+LDA but maximizing $J(W) = |W^T S_b W|/|W^T S_m W|$.

newLDA: If S_w is full rank then solve regular LDA; else in the null space of S_w, find the eigenvectors of S_b with largest eigenvalues.

It should be noted that PCA+LDA and Scatter-LDA can be equivalent when S_w and S_m span the same subspace. However, they are different when S_b totally or partially spans the null space of S_w, thus S_w and S_m span different subspaces. For face recognition the latter case turns out to be more common. In [4], Chen et al. prove that the null space of S_w contains discriminant information. They also show that Scatter-LDA is not "optimal" in that it fails to distinguish the most discriminant information in the null space of S_w. Thus they propose the newLDA method. However, newLDA fell short of making use of any information outside of that null space.

All these techniques make "hard" decisions, either discarding a null space, or only working in a null space. On the other hand, our weighted additive criterion does not make any "hard" decisions. Instead, it explores λ (2) to judicially extract information from both subspaces. In addition, cross-validated λ values can be leveraged to alleviate Fisher's bias toward outlier classes in multi-class problems [14].

8 Experiments

In this section we compare the SDP-LDA algorithm against several competing methods: PCA+LDA, newLDA, and Scatter-LDA on several multi-class (facial images) and binary data sets. To solve the semi-definite program (2), we used the general purpose optimization software SeDuMi [20].

8.1 Facial Images

Here we used the ORL data set [18]. The size of each image is 92×112. We extracted 120 images, where there are 40 subjects with three images from each. As a result, we are facing the challenge of the small sample size problem.

We randomly choose two images per person for training, and the remaining one for testing. We have 80 training and 40 test images. Subspaces are calculated from the training data, and the 1 nearest neighbor (NN) classifier is used to obtain accuracy rates after projecting the data onto the subspace. To obtain average performance, each method is repeated 10 times. The term λ in (2) is chosen by 10-fold cross-validation.

The average accuracy as a function of dimensionality is shown 1. The X-axis represents the dimensionality of the subspace. For each technique, the higher the dimension, the less discriminant the dimension. For most techniques, the accuracy rate increases quickly around the first 10 dimensions, and then increase slowly with additional dimensions.

SDP-LDA is uniformly better than any other algorithms on both problems, demonstrating its efficacy. It achieves the highest accuracy rate of 0.8875 on ORL. newLDA performs quite well in these experiments, again demonstrating that most discriminant information is in the null space of S_w, for the facial recognition tasks. On the other hand, Scatter-LDA does not perform well at lower dimensional subspaces. But it eventually performs better than PCA+LDA, when the number of dimensions is large enough. All methods achieve their highest accuracy rate with a 39 dimensional subspace, which is not surprising, for it is a 40 class problem. It should be noted that the performance of newLDA and Scatter-LDA (its tail is not shown in the plot) drops quickly with unnecessary dimensions.

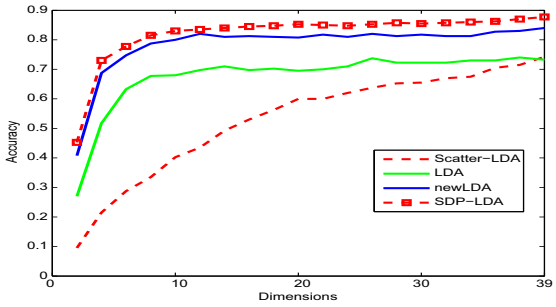

Figure 1. Comparison of SDP-LDA, LDA, newLDA, and Scatter-LDA on the ORL image data.

8.2 Binary Data Sets

In these experiments, we compare the four competing methods on a number of two class classification problems. We use 11 data sets from the UCI database and the cat and dog (CatDog) data. They are all two class classification problems.

For each data set, we randomly choose 60% as training and the remaining 40% as testing. We train the four methods on the training data and obtain projections. We then project both training and test data on the chosen subspace and use the 1NN classifier to obtain error rates. Note that for the two class case, one dimensional subspace is sufficient. Again, λ in Eq. 2 is chosen through ten-fold cross-validation. We repeat the experiments 10 times on each data set to obtain the average accuracy rates.

The results are shown in Table 1. On 6 data sets out of 12, SDP-LDA performed the best. It came second on five data sets. It has the overall best average. Another way to look at these methods is to measure robustness. For each method m we compute the ratio b_m between its error rate e_m and the smallest error rate over all methods being compared in a particular example: $b_m = e_m / \min_{1 \le k \le 12} e_k$.

Figure 2 plots the distribution of b_m for each method over the 12 data sets. The box area represents the lower and upper quartiles of the distribution that are separated by the median. The outer vertical lines show the entire range of values for the distribution. As shown in Figure 2, the spread of the error distribution for SDP-LDA is narrow and close to 1, followed by Scatter-LDA. The results clearly demonstrate that SDP-LDA obtained the most robust performance over these data sets.

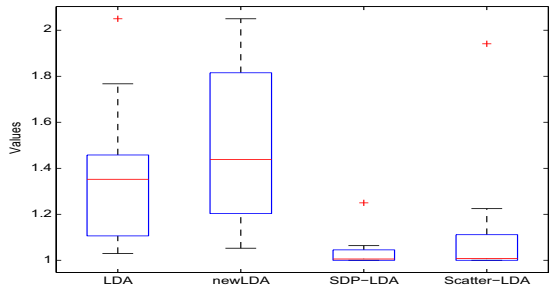

Figure 2. Error distributions of SDP-LDA, LDA, newLDA, and Scatter-LDA on the 12 data sets.

9 Summary

This paper presents a weighted additive criterion for dimension reduction that potentially provide a solution to the small sample size problem, often associated with Fisher's criterion. In particular, the paper has shown that (1) the proposed weighted additive criterion (2) for dimension deduction is closely related to the maximum margin criterion; (2) the criterion has the potential to help alleviate Fisher's bias toward outlier classes in multi-class problems; and (3) the criterion can be optimized using efficient algorithms such as semi-definite programming, thereby avoiding the inverse of Sw and thus the potential small sample size problem. The paper demonstrates the efficacy of the proposed technique using a number of real examples, and the results show that the proposed technique registered superior performance over several competing methods in several examples.

References

[1] V. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

Table 1. Classification accuracy rates in subspaces computed by the four competing methods using 1NN classifier, on 12 data sets.

Data Set	PCA+LDA	newLDA	SDP-LDA	Scatter-LDA
Breast Cancer	0.6400	0.6400	0.6373	0.6591
Cancer Wisconsin	0.9502	0.9502	0.9626	0.9542
Credit	0.7284	0.6364	0.8050	0.8019
Heart Cleveland	0.7432	0.7432	0.7644	0.7780
Heart Hungary	0.7607	0.6504	0.7675	0.7547
Ionosphere	0.7557	0.7557	0.8200	0.8221
New Thyroid	0.6400	0.6400	0.8244	0.6591
Pima	0.6749	0.6749	0.6821	0.6912
Glass	0.8565	0.8565	0.9165	0.9188
Sonar	0.6084	0.5205	0.7229	0.7181
Iris	0.9375	0.9375	0.9375	0.9500
CatDog	0.6937	0.6203	0.7962	0.7620
Average	0.7491	0.7188	0.8030	0.7891

[2] R. E. Bellman. *Adaptive Control Precesses: A Guided Tour*. Princeton University Press, 1961.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[4] L. Chen, H. M. Liao, M.T.Ko, J. Lin, and G. Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33:1713–1726, 2001.

[5] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, UK, 2000.

[6] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley-Sons, New York, 1 edition, 1973.

[7] J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[8] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, 1990.

[9] R. Huang, Q. Liu, H. Lu, and S. Ma. Solving the small sample size problem of lda. In *Proceedings of 16th International Conference on Pattern Recognition*, volume 3, pages 29–32, 2002.

[10] A. K. Jain and B. Chandrasekaran. Dimensionality and sample size considerations in pattern recognition practice. *Handbook of Statistics*, 2:835–855, 1982.

[11] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag:New York, 1986.

[12] H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximum margin criterion. *IEEE Transactions on Neural Networks*, 17(1):157–165, January 2006.

[13] K. Liu, Y. Cheng, and J. Yang. A generalized optimal set of discriminant vectors. *Pattern Recognition*, 25(7):731–739, 1992.

[14] M. Loog, R. Duin, and R. Haeb-Umbach. Muliclass linear dimension reduction by weighted pairwise fisher criteria. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(7):762–766, 2001.

[15] G. Pataki. On the rank of extrmem matrices in semi-definite programs and the multiplicity of optimal eigenvalues. *Mathematics of Operations Research*, 23:339–358, 1998.

[16] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(2):252–264, 1991.

[17] S. J. Raudys and V. Pikelis. On dimensionality, sample size, classification error, and complexity of classification algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2:243–251, 1980.

[18] F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, 1994.

[19] G. Stewart. *Introduction to Matrix Computations*. Academic Press, INC, 1973.

[20] J. F. Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11(12):625–653, 1999.

[21] D. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(8):831–836, 1996.

[22] Q. Tian, M. Barbero, Z. H. Gu, and S. H. Lee. Image classification by the foley-sammon transform. *Optical Engineering*, 25(7):834–840, 1986.

[23] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.

[24] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[25] H. Yu and J. Yang. A direct lda algorithm for high-dimension data with application to face recognition. *Pattern Recognition*, 34:2067–2070, 2001.