ENERGY-CONSTRAINED DISCRIMINANT ANALYSIS

Scott Philips, Visar Berisha

MIT Lincoln Laboratory
244 Wood Street Lexington, MA 02420, USA

Andreas Spanias

Arizona State University
Department of Electrical Engineering
Box 5706, Tempe, AZ 85287 USA

ABSTRACT

Dimensionality reduction algorithms have become an indispensable tool for working with high-dimensional data in classification. Linear discriminant analysis (LDA) is a popular analysis technique used to project high-dimensional data into a lower-dimensional space while maximizing class separability. Although this technique is widely used in many applications, it suffers from overfitting when the number of training examples is on the same order as the dimension of the original data space. When overfitting occurs, the direction of the LDA solution can be dominated by low-energy noise and therefore the solution becomes non-robust to unseen data. In this paper, we propose a novel algorithm, energy-constrained discriminant analysis (ECDA), that overcomes the limitations of LDA by finding lower dimensional projections that maximize inter-class separability, while also preserving signal energy. Our results show that the proposed technique results in higher classification rates when compared to comparable methods. The results are given in terms of SAR image classification, however the algorithm is broadly applicable and can be generalized to any classification problem.

Index Terms— Dimensionality reduction, discriminant analysis, machine learning, pattern recognition, principal components analysis

1. INTRODUCTION

With continuing growth in sensor capabilities and database complexity, pattern recognition applications will have to contend with ever-increasing data dimensionality. Extracting appropriate information from such data sets has been a long-standing problem, commonly referred to as the curse of dimensionality. Commonly used approaches to mitigating this problem include linear discriminant analysis (LDA), principal components analysis (PCA), and derivatives of them [1, 2, 3, 4, 5].

Linear discriminant analysis identifies a linear transform that projects high dimensional data into a low dimensional space subject to a classification constraint. More specifically, LDA determines the optimal transform that maximizes the distance between the mean of each class while minimizing the variability within each class [2]. Consider the two-class, two-dimensional data set in Fig. 1. For this data set, LDA identifies the optimal dimension for classification as the vector that lies along dimension 2. Intuitively this makes sense because the data projected along this dimension is most separable. Unfortunately, this is a dimension of very low energy and therefore does not capture significant signal information. In addition, in problems where the number of data samples is on the order of the dimensionality of data space, LDA tends to suffer from overfitting [4]. When overfitting occurs, the direction of the LDA solution can also be dominated by low energy noise. To avoid such problems, principal component analysis is often used.

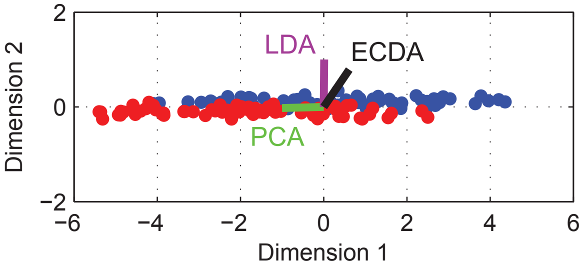

Fig. 1. Scatter plot of a two-dimensional, two class dataset with PCA, LDA and ECDA directional vectors.

As opposed to LDA, PCA determines a transform that maps the high dimensional data onto the low dimensional space that preserves most signal energy [3]. In Fig. 1, PCA identifies dimension 1 as the dimension containing the most energy. This makes sense as the variance of the data set is highest along this dimension. Although this dimension captures much of the signal information, there is no explicit classification criterion used.

Attempts to overcome the limitations of these techniques include regularized LDA techniques [5] and PCA plus LDA [6]. In regularized LDA, covariance matrices are typically diagonally loaded when the number of training samples is scarce [5]. Although these techniques have been shown to provide an improvement when the covariance matrices are of low rank, the diagonal loading is often ad-hoc making the algorithm very sensitive to the tuning parameter. PCA plus LDA techniques use PCA as a preprocessing step before applying LDA. Although this approach can be useful when the number of training samples is low, using PCA to completely discard lower-energy dimensions can reduce classification performance, as we will show later [6].

In this paper, we attempt to overcome the limitations of the existing techniques by finding optimal projections for discrimination, subject to minimally-acceptable energy criteria. In other words, we combine both PCA and LDA in a single algorithm, energy-constrained discrimination analysis (ECDA), that allows the user to take advantage of the benefits of each of the two approaches. Consider the PCA, LDA, and ECDA directions in Fig. 1. The optimal PCA projection contains very little discrimination power, whereas the optimal LDA projection contains very little energy. ECDA is

This work was sponsored by the Defense Advanced Research Projects Agency/IPTO under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are not necessarily endorsed by the United States Government.

a reasonable compromise that maintains discrimination power with improve robustness. We show that the proposed technique results in improved classification rates when compared to four other existing algorithms. Although the results are given in terms of SAR image classification, the algorithm is broadly applicable and can be generalized to any classification problem.

This paper is organized as follows. Section 2 provides a description of the proposed algorithm. In Section 3 we analyze and compare the performance of ECDA to other techniques. Section 4 provides concluding remarks.

2. PROPOSED ALGORITHM

Consider a data set $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with n points residing in some d-dimensional space and belonging to one of r classes. We define the total data matrix, $\mathbf{X}$, and class data matrices, $\mathbf{X}_i$ $(1 \leq i \leq r)$, as follows:

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] = [\mathbf{X}_1, ..., \mathbf{X}_r], \tag{1}$$

where the subset of vectors in class i $(1 \leq i \leq r)$ is represented as block matrix $\mathbf{X}_i \in \mathbf{R}^{d \times n_i}$. Each class i contains n_i observations with $n = \sum_i n_i$. Define the between-class scatter matrix $\mathbf{S}_\mathrm{B}$ as

$$\mathbf{S}_\mathrm{B} = \sum_{i=1}^{r} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^\mathrm{T}, \tag{2}$$

where $\mathbf{m}$ is mean of the data and $\mathbf{m}_i$ is the mean of class i. Define the within-class scatter matrix $\mathbf{S}_\mathrm{W}$ as

$$\mathbf{S}_\mathrm{W} = \sum_{i=1}^{r} \sum_{j \in \mathcal{C}_i} (\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^\mathrm{T}, \tag{3}$$

where $\mathcal{C}_i$ is the set of cardinality n_i containing all data vectors in class i. It can be shown that the total scatter matrix of the data set $\mathbf{X}$ is $\mathbf{S}_\mathrm{T} = \mathbf{S}_\mathrm{W} + \mathbf{S}_\mathrm{B}$.

The goal of LDA is to find a directional vector $\mathbf{v}_\mathrm{LDA} \in \mathbf{R}^{d \times 1}$ such that the projection $\mathbf{v}_\mathrm{LDA}^\mathrm{T}\mathbf{X}$ maximizes the distance between vectors of different classes ($\mathbf{S}_\mathrm{B}$) while also minimizing the distance between vectors within the same class ($\mathbf{S}_\mathrm{W}$) [2]. Written explicitly, the optimal LDA vector is defined as

$$\mathbf{v}_\mathrm{LDA} = \operatorname{argmax} \frac{\mathbf{v}^\mathrm{T}\mathbf{S}_\mathrm{B}\mathbf{v}}{\mathbf{v}^\mathrm{T}\mathbf{S}_\mathrm{W}\mathbf{v}}. \tag{4}$$

PCA, on the other hand, preforms dimensionality reduction, without making use of class information, by finding the vector $\mathbf{v}_\mathrm{PCA}$ such that the projection $\mathbf{v}_\mathrm{PCA}^\mathrm{T}\mathbf{X}$ maximizes the total scatter [3]. Mathematically, this is defined as:

$$\mathbf{v}_\mathrm{PCA} = \operatorname{argmax} \mathbf{v}^\mathrm{T}\mathbf{S}_\mathrm{T}\mathbf{v}. \tag{5}$$

Energy-constrained discriminant analysis (ECDA) finds the optimal direction for discrimination, subject to energy criteria. In other words, we seek the vector $\mathbf{v}_\mathrm{ECDA}$ such that the projection $\mathbf{v}_\mathrm{ECDA}^\mathrm{T}\mathbf{X}$ maximizes the separability of the data while also maintaining a minimally-acceptable amount of energy. Written explicitly, the optimal ECDA vector is the maximizer of

$$\begin{array}{ll} \underset{\mathbf{v}}{\text{maximize}} & \dfrac{\mathbf{v}^\mathrm{T}\mathbf{S}_\mathrm{B}\mathbf{v}}{\mathbf{v}^\mathrm{T}\mathbf{S}_\mathrm{W}\mathbf{v}} \\ \text{subject to} & \dfrac{\mathbf{v}^\mathrm{T}}{||\mathbf{v}||}\mathbf{S}_\mathrm{T}\dfrac{\mathbf{v}}{||\mathbf{v}||} \geq \alpha\lambda_\mathrm{PCA}, \end{array} \tag{6}$$

where λ_PCA is the principal eigenvalue of PCA, and the user-defined parameter $\alpha \in [0, 1]$ specifies the energy constraint. It is important to note that for $\alpha = 0$ this formulation results in LDA and for $\alpha = 1$ the result is PCA. For $0 < \alpha < 1$, the proposed approach finds robust discrimination directions that generalize better to testing data, as will be shown later. In this analysis, we only show the procedure to find the optimal ECDA direction, however the approach can be easily generalized to multiple ECDA directions by iteratively removing existing ECDA directions from the data and finding new ECDA directions from the remaining subspace.

Since the objective function in (6) is invariant with respect to rescaling of the vector $\mathbf{v}$, we can fix the denominator to an arbitrary constant and optimize only the numerator. This allows us to transform the problem of maximizing (6) into the constrained problem shown below, in which we have added an additional constraint and rewritten the existing constraint:

$$\begin{array}{ll} \underset{\mathbf{v}}{\text{minimize}} & -\mathbf{v}^\mathrm{T}\mathbf{S}_\mathrm{B}\mathbf{v} \\ \text{subject to} & \mathbf{v}^\mathrm{T}\mathbf{S}_\mathrm{W}\mathbf{v} = 1, \\ & \mathbf{v}^\mathrm{T}\mathbf{S}_\mathrm{IT}\mathbf{v} \leq 0, \end{array} \tag{7}$$

where $\mathbf{S}_\mathrm{IT} = [\alpha\lambda_\mathrm{PCA}\mathbf{I} - \mathbf{S}_\mathrm{T}]$. The resulting formulation is a quadratic-constrained quadratic programming (QCQP) problem.

2.1. Relaxing the Constraints

It is easy to see that the matrix $\mathbf{S}_\mathrm{IT}$ in (7) is generally not positive semidefinite, thereby resulting in a nonconvex optimization problem. Although a number of methods have been used in the literature to solve such problems [7, 8, 9], these are generally not robust and often settle in non-optimal local minima when the dimensionality of the problem is large. In this paper, we make use of a semidefinite relaxation of the QCQP problem in (7) that results in a convex optimization problem [10].

We use $\mathbf{v}^\mathrm{T}\mathbf{S}\mathbf{v} = \mathrm{tr}(\mathbf{v}\mathbf{v}^\mathrm{T}\mathbf{S}) = \mathrm{tr}(\mathbf{V}\mathbf{S})$ to rewrite (7) as follows:

$$\begin{array}{ll} \underset{\mathbf{V},\mathbf{v}}{\text{minimize}} & -\mathrm{tr}(\mathbf{V}\mathbf{S}_\mathrm{B}) \\ \text{subject to} & \mathrm{tr}(\mathbf{V}\mathbf{S}_\mathrm{W}) \leq 1, \\ & \mathrm{tr}\left(\mathbf{V}\left[\alpha\lambda_\mathrm{PCA}\mathbf{I} - \mathbf{S}_\mathrm{T}\right]\right) \leq 0, \\ & \mathbf{V} = \mathbf{v}\mathbf{v}^\mathrm{T}. \end{array} \tag{8}$$

The formulation of the optimization problem in (8) does not change the underlying solution, however, it allows us to better understand the relaxation. By relaxing the final constraint from $\mathbf{V} = \mathbf{v}\mathbf{v}^\mathrm{T}$ (a nonconvex constraint) to $\mathbf{V} - \mathbf{v}\mathbf{v}^\mathrm{T} \succeq 0$ (the convex constraint that the difference matrix is positive semidefinite), the optimization criteria is converted to a semidefinite programming problem. The final formulation is:

$$\begin{array}{ll} \underset{\mathbf{V},\mathbf{v}}{\text{minimize}} & -\mathrm{tr}(\mathbf{V}\mathbf{S}_\mathrm{B}) \\ \text{subject to} & \mathrm{tr}(\mathbf{V}\mathbf{S}_\mathrm{W}) \leq 1, \\ & \mathrm{tr}\left(\mathbf{V}\left[\alpha\lambda_\mathrm{PCA}\mathbf{I} - \mathbf{S}_\mathrm{T}\right]\right) \leq 0, \\ & \mathbf{V} - \mathbf{v}\mathbf{v}^\mathrm{T} \succeq 0. \end{array} \tag{9}$$

It can be shown that this approach provides a lower bound on the optimal solution, rather than the optimal solution [10]; however, due to the convexity of the problem, it is guaranteed to result in a single solution every time. Furthermore, semidefinite programming problems are well-known in the literature and there exist numerous robust methods for obtaining solutions to such problems [10] [12]. In this paper we make use of the well-known Matlab packages Yalmip [11] and Sedumi [12] to solve all semidefinite programming problems.

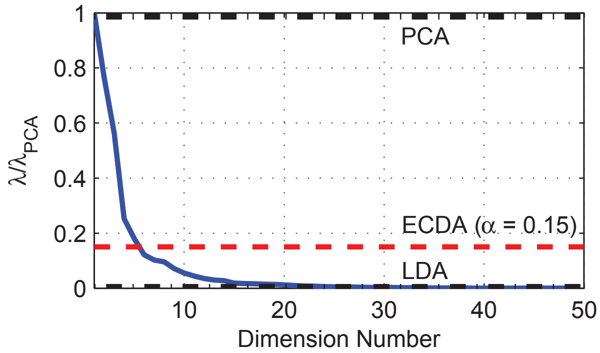

Fig. 2. Normalized and sorted PCA eigenvalues for a dataset containing 50 SAR images. Dotted lines shows normalized signal energy for the PCA, LDA, and ECDA ($\alpha = 0.15$) solutions.

2.2. Setting α

The α parameter in the proposed technique restricts the energy of the principal ECDA vector to a fraction of the primary PCA vector. More specifically, ECDA requires that the optimal discrimination vector is found subject to the constraint that the selected direction captures $\alpha\lambda_{\text{PCA}}$ of the data energy. For small values of α, the algorithm may select low-energy dimensions for discrimination that do not generalize very well to unseen test data. For large values of α, the algorithm is restricted to only high-energy portions of the signal that may be similar among different classes. As a result, we seek an appropriate value for α that finds directions of separability and generalizes to unseen test data.

We propose two methods for determining α. The first method involves a cross-validation procedure over which a set of α's from 0 to 1 are used to generate ECDA features and train a classifier. A testing set is then used to analyze the performance of the classifier and the α maximizing classifier performance is selected. A second, computationally less complicated method of determining α involves analyzing the normalized PCA eigenvalue vs. dimension plot. In Fig. 2 we show the normalized and sorted PCA eigenvalues, $\frac{\lambda}{\lambda_{\text{PCA}}}$, for a dataset containing 50 SAR images. We determine a value for α by seeking a knee in the curve of this plot. In Fig. 2, $\alpha = 0.15$ seems to capture the pertinent signal energy, while excluding low-energy noise. Notice that the PCA solution corresponds to $\alpha = 1$, whereas the LDA solution corresponds to $\alpha = 0$.

3. RESULTS

In this section, we evaluate the discrimination performance of ECDA using the public release portion of the MSTAR database of SAR images [13]. The SAR images are first pre-processed using a constant false alarm rate (CFAR) algorithm that centers a 50 pixel x 50 pixel frame around each target. These target images comprise a two class scenario, namely the T-72 tank and the BMP-2 infantry fighting vehicle. The data set is divided into two groups, the training set (depression angle 17°) and the testing set (depression angle 15°). The training set consists of relatively few examples (50) while the testing set consists of many more (1169).

Example SAR images from this dataset are shown in Fig. 3. The figure shows 50 pixel x 50 pixel images of two targets, one from the class of T-72 tanks and another from the class of BMP-2 infantry fighting vehicle. The classification of these images requires performing analysis in a $\mathbb{R}^{2500}$ space. From the well-known curse of dimensionality, it is clear that the density of the data set in this high-dimensional space is sparse and, although the data set is represented in $\mathbb{R}^{2500}$, information relevant for the classification of these images is likely to reside in a much lower-dimensional subspace. This motivates the need for dimensionality reduction.

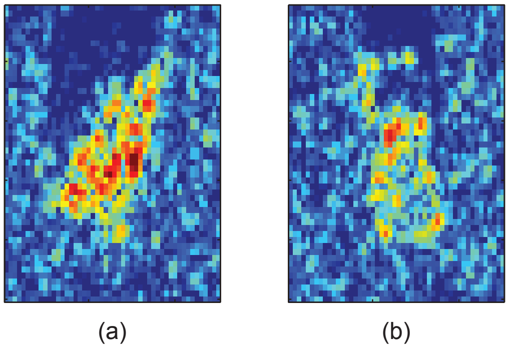

Fig. 3. Sample SAR images (in dB scale) of (a) a T-72 tank and (b) a BMP-2 infantry fighting vehicle.

As a first step, we do an initial dimensionality reduction using a 2-D, level-3 'Haar' wavelet decomposition. This effectively downsamples the image in order to reduce the dimensionality from very large $\mathbb{R}^{2500}$ to a more reasonable $\mathbb{R}^{49}$. From these 7 pixel x 7 pixel images we compared the discrimination performance of ECDA against the standard techniques of PCA [3], LDA [2], regularized LDA [5], and PCA plus LDA [6].

As an alternative to evaluating ECDA using a specific classifier, we use the Henze-Penrose (HP) divergence for measuring relative interclass separability [14, 15]. The underlying assumption is that independent of any particular classifier, feature sets that exhibit more divergence (or separability) should, in general, be of greater utility than feature sets that exhibit less divergence (or separability). For two-class data sets with an equal number of samples in each class, HP divergence values range from 0.5 to 1, with 0.5 implying the classes cannot be separated and 1 implying that the two classes are completely separable.

Using the 50 training examples we identify the 1-D optimal vectors $\mathbf{v}_{\text{LDA}}$, $\mathbf{v}_{\text{PCA}}$, and $\mathbf{v}_{\text{ECDA}}$ according to (4), (5), and (9), respectively. These vectors are then used to project the 1169 test examples down to a single dimension. The distribution of values along this single dimension is shown in Fig. 4. As expected, the total variance (or spread) of the PCA projection is the largest, although the separation between the two classes is negligible. This is due to the fact that PCA explicitly maximizes the total scatter, while ignoring class distinction. The LDA projection, on the other hand, captures very little variance while also providing very little class separation. This is because LDA overfits to low-energy noise in the training data and therefore does not generalize to unseen test data. Finally, the ECDA projection is a balance between the other two projections. ECDA contains a relatively moderate amount of total variance while also capturing class separation.

Other methods that attempt to address the above problems with PCA and LDA are regularized LDA and PCA plus LDA. In regularized LDA the within-class scatter matrix $\mathbf{S}_{\text{W}}$ is modified as follows

$$\hat{\mathbf{S}}_{\text{W}} = \mathbf{S}_{\text{W}} + \beta\mathbf{I}, \tag{10}$$

where I is the identity matrix and β is a tuning parameter. This ap-

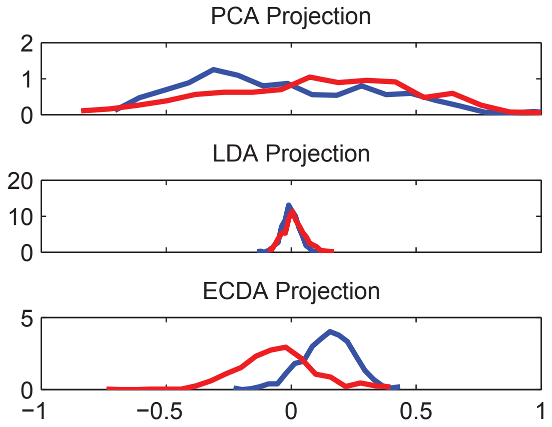

Fig. 4. Distributions of the test data projected along the vectors $\mathbf{v}_{\mathrm{PCA}}$, $\mathbf{v}_{\mathrm{LDA}}$, and $\mathbf{v}_{\mathrm{ECDA}}$.

proach is motivated by the fact that in undersampled problems, $\mathbf{S}_{\mathrm{W}}$ is often close to singular or ill-conditioned; diagonal loading insures that $\mathbf{S}_{\mathrm{W}}$ has full rank. In PCA plus LDA, PCA data's dimensionality is first reduced by PCA to capture a minimum acceptable amount of energy; LDA is then performed on the reduced data.

To rigorously compare the discrimination power of each of these methods we extend the ECDA and PCA projections into multiple dimensions; that is we find projections of the test data into an N-dimensional subspace, where $N = [1, ..., 5]$. This allows us to test the divergence of the test set for multiple dimensions. For standard LDA and its derivatives (regularized LDA and PCA plus LDA) this extension is not possible because the rank of the matrix $\mathbf{S}_{\mathrm{W}}^{-1}\mathbf{S}_{\mathrm{B}}$ is $r-1$, which limits the dimensionality of the projection to the number of classes minus one [4]. In our two-class case, that dimensionality is one.

Figure 5 shows the HP divergence as a function of dimensionality. Methods with only a one-dimensional solution are shown as horizontal dashed lines. LDA and PCA plus LDA result in the lowest divergence values (0.5 and 0.62, respectively). PCA, while having poor discrimination power in low dimensions, increases in performance as the dimensionality increases. This can be expected because as more signal energy is captured, the additional information improves class separability. Regularized LDA, with a $\beta = 0.1$ determined through cross-validation, is able to overcome some of the inherent problems of LDA, thereby resulting in improved performance. It results in a divergence of 0.7 and is on par with ECDA in one dimension. As ECDA is not limited to one dimension, it increases in divergence as dimensionality is increased, with a maximum near 0.8.

4. CONCLUSION

In this paper, we proposed an algorithm for performing dimensionality reduction based on energy and classification criteria. The ECDA algorithm combines the benefits of both LDA and PCA to generate low-dimensional projections that help separate the data and also preserve its energy. Our results show that the technique significantly outperforms LDA, PCA, regularized LDA, and PCA plus LDA when the number of available training samples is limited. Future work will focus on more explicit methods for determining α that do not require cross-validation. Furthermore, techniques of lower-computational complexity for determining the optimal ECDA vectors can help in speeding up the solution to the optimization problem. In addition, methods for extending such a technique to non-linear manifolds through the use of kernel functions will also be analyzed.

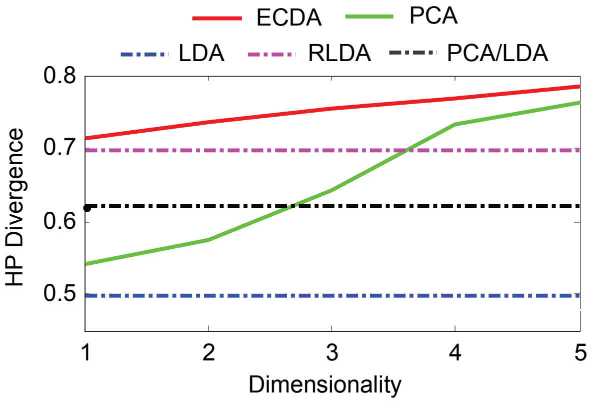

Fig. 5. HP Divergence vs. dimensionality for ECDA, PCA, LDA, regularized LDA, and PCA plus LDA. Methods with only a one-dimensional solution are shown as horizontal dashed lines (LDA, regularized LDA, and PCA plus LDA).

5. REFERENCES

[1] R.A. Fisher "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics*, 7: 179-188, 1936.

[2] G.J. McLachlan, "Discriminant Analysis and Statistical Pattern Recognition." *Wiley-Interscience*, 2004.

[3] I.T. Jolliffe "Principal Component Analysis," *Series: Springer Series in Statistics,* 2nd ed., Springer, NY, 2002. ISBN 978-0-387-95442-4.

[4] A.M. Martinez and A.C. Kak, "PCA versus LDA," in *IEEE Trans. Pattern Analysis and Mach. Intel.*, vol.23, no.2, pp.228-233, Feb 2001.

[5] C.H. Park and H. Park, "A comparison of generalized linear discriminant analysis algorithms," in *Pattern Recog.*, Elsevier Science, vol.41, no.3, pp 1083 – 1097, 2008.

[6] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection" in *IEEE Trans. on Patt. Anal. and Mach. Learn.*, vol. 19, pp. 711-720. July 1997.

[7] D. Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning," Kluwer Academic Publishers, 1989, Boston, MA.

[8] M. Mitchell, "An Introduction to Genetic Algorithms," MIT Press, Cambridge, MA.

[9] S. Kirkpatrick and C. D. Gelatt and M. P. Vecchi, "Optimization by Simulated Annealing," *Science,* vol. 220, no. 4598, pp. 671-680, 1983.

[10] *Handbook of Semidefinite Programming,* Editors: H. Wolkowicz, R. Saigal, and L. Vandenberghe, Kulwer Academic Publishers, 2000.

[11] J. Lfberg "YALMIP : A Toolbox for Modeling and Optimization in MATLAB." In *Proc. of the CACSD Conference*, Taipei, Taiwan, 2004.

[12] J.F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optim. Meth. and Soft.*, pp. 625-653, 1999.

[13] E.R. Keydel, "MSTAR Extended Operating Conditions," *SPIE Proceedings*, vol. 2757, pp. 228-242, 1996.

[14] N. Henze and M.D. Penrose, "On the Multivariate Runs Test," *The Annals of Statistics*, vol. 27, pp. 290-298, 1999.

[15] V. Berisha et al. "Sparse Manifold Learning With Applications to SAR Image Classification", in *IEEE ICASSP*, vol. 3, pp 1089 - 1092, April 2007.