Pareto Discriminant Analysis

Karim T. Abou–Moustafa
Centre of Intelligent Machines
McGill University
karimt@cim.mcgill.ca

Fernando de la Torre
The Robotics Institute
Carnegie Mellon University
ftorre@cs.cmu.edu

Frank P. Ferrie
Centre of Intelligent Machines
McGill University
ferrie@cim.mcgill.ca

Abstract

Linear Discriminant Analysis (LDA) is a popular tool for multiclass discriminative dimensionality reduction. However, LDA suffers from two major problems: (1) It only optimizes the Bayes error for the case of unimodal Gaussian classes with equal covariances (assuming full rank matrices) and, (2) The multiclass extension maximizes the sum of pairwise distances between the classes, and does not "simultaneously" maximize each pairwise distance between the classes. This typically results in serious overlapping in the projected space between classes that are "close" in the input space. To solve these two problems, this paper proposes Pareto Discriminant Analysis (PARDA).

Firstly, PARDA explicitly models each of the classes as a multidimensional Gaussian with a sample covariance. Secondly, PARDA decomposes the multiclass problem to a set of pairwise objective functions representing the pairwise distance between different classes. Unlike existing extensions of Fisher discriminant analysis (FDA) to multiclass problems, that typically maximize the sum of pairwise distances between classes, PARDA simultaneously maximizes each pairwise distance, thus encouraging the case that all classes are equidistant from each other in the lower dimensional space. Solving PARDA is a multiobjective optimization problem – simultaneously optimizing more than one, possibly conflicting, objective functions – and the resulting solution is known to be "Pareto Optimal". Experimental results on synthetic data, several image data sets and data sets from the UCI repository show positive and encouraging results in favor of PARDA when compared with standard and state-of-the-art multiclass extensions of LDA.

1. Introduction

Fisher Discriminant Analysis (FDA) originally developed by Fisher in 1936 [6] is a technique for dimensionality reduction that is optimal for classification under two assumptions: (1) the number of classes c is exactly two, (2) the samples in each class are assumed to be generated from a multivariate Gaussian distribution with the same covariance (homoscedastic data) [7]. In this context, FDA is guaranteed to find a one dimensional subspace that will classify the samples with the optimal error rate, *Bayes error*, and the subspace is known to be *Bayes optimal* [7]. Rao [21] extended this approach to the multiclass homoscedastic case ($c > 2$), under the condition that the data features $d \geq c$ (assuming the number of samples $n > d$). The resultant $c - 1$ dimensional subspace is also guaranteed to be Bayes optimal, and the technique has become known as Linear Discriminant Analysis (LDA). Rao however, noted that if the lower dimensional subspace has dimensionality $k < c - 1$, the resultant subspace will not be Bayes optimal. Therefore, Rao proposed the well known formulation

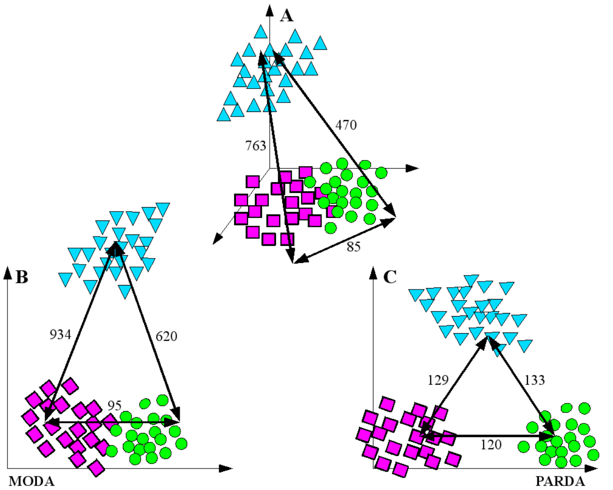

Figure 1. (A) A Synthetic example of a 3–class problem with three dimensional data. The numbers shown on arrows indicate the symmetric Kullback–Leibler divergence (KLD). (B) Projection using MODA on a two–dimensional space. Observe that the two classes that are close in the input space proportionally increase the KL divergence less than the classes that are further in the input space. (C) Projections obtained by Pareto Discriminant Analysis (PARDA) encourages the classes to be equally spread from each other in the lower dimensional space.

based on the Rayleigh quotient of the between–class scatter matrix $\mathbf{S}_b$ and the average of within–class scatters matrix $\mathbf{S}_w$ that provide an approximation to the optimal Bayes solution [7]. Recently, Hamsici and Martinez [9] pushed the homoscedastic case further and derived a Bayes optimal one dimensional subspace when $c > 2$.

Several researchers, backed by theoretical justifications, have scrutinized the limitations and non–optimality of LDA when its strong assumptions do not hold. The result was a plethora of extensions and modifications to the original LDA model which have been reported to perform well in a variety of application domains, most notably face recognition. A good review for these methods can be found in [2, 10, 13, 16, 18, 4, 25, 9].

To overcome the homoscedastic constraint and consider the hetereoscedastic case (classes have different means and different covariances), several researchers have proposed extensions to LDA derived from Gaussian assumptions [2, 13, 4, 9] and kernel extensions [19]. In particular, de la Torre and Kanade proposed Multimodal Oriented Discriminant Analysis (MODA) [4] which is a consistent generalization of FDA to multimodal Gaussian distributions with different means and covariances. As a measure of separation between two Gaussians, they use the symmetric Kullback–Leibler (KL) divergence and formulate an objective that maximizes this measure between the two classes. Note that the symmetric KL divergence (KLD) considers the difference in mean locations and covariance matrices. Unlike LDA which can define subspaces of maximum dimension $k; 1 \leq k \leq \min(c-1, d-1)$, MODA allows the extraction of more discriminative features than LDA does, pushing k to be $\leq min(n-1, d-1)$.

As noted by several researchers [16, 18, 23], even if all the homoscedastic assumptions are satisfied, LDA suffers from the serious problem of merging classes that are close to each other in the original input space, *a.k.a* the class separation problem [23]. This is due to the fact that both approaches, LDA and MODA, shift the 2–class formulation to the multiclass setting by maximizing the sum of pairwise KLDs[1] (one–versus–one) which is a good approximation when all classes are in proximity to each other in terms of KLD.

Figure (1A) depicts a synthetic example for a 3–class problem with three dimensional data. Traditional methods like LDA or MODA find projections that maximize the sum of pairwise Mahalanobis distance (LDA) or the symmetric KLD (MODA) between pairwise classes. Note that the Mahalanobis distance and the KLD are positive quadratic functions in the Euclidean distance between the means. From the optimization of minimax functions [3], it is known that the sum of positive powered functions, $\sum_{j=1}^{m}[f_j]^p$, where $p > 1$, is a smooth approximation for $\max_{1\leq j\leq m}[f_j]$, and hence $\sum_{j=1}^{m}[f_j]^p \approx [f_r]^p$ where $f_r > f_j \; \forall j \neq r$. Using this argument, it is possible to see that LDA and MODA are in fact maximizing a smooth approximation of the maximum of pairwise Mahalanobis distances and KLDs respectively. Hence, LDA and MODA intrinsically prefer solutions that encourage maximizing the largest distance in the input space to make it even larger in the lower dimensional subspace. In other words, LDA and MODA put needless effort to maximize already distant classes in the input space. This effect can be seen in Figure (1 B), where MODA's projection gives relatively better increase in terms of KL divergence to the classes that are farther away in the input space, while it only makes a slight effort to separate between classes that are very close in the input space.

In this research work, we consider a different approach for learning a discriminative subspace for the multiclass heteroscedastic case. We note that summing over all KLDs and maximizing that sum yields the class separation problem. In other words, maximizing the sum of KLDs does not consider each pairwise objective function independently, and does not search for a solution that best agrees with all independent and possibly conflicting objectives. Note that summing over all pairwise distances does not impose any constraint on the minimum distance between the classes. A desired projection would be that all classes are equally spread from each other in the lower dimensional space. We propose that each pair of different classes defines an independent objective function, using FDA or MODA, that needs to be optimized, and that the multiclass setting is *intuitively* a set of multiple pairwise objective (multiobjective) functions that need to be *simultaneously optimized*. Therefore, the scalarization for the problem of learning a discriminative subspace from the 2–class setting to the multiclass setting by summing over all pairwise KLDs is not the appropriate path to handle a multiobjective optimization problem [11], since by summing no maximal agreement is guaranteed between all pairwise objective functions.

Our contribution in this paper originates from the above observation. Specifically, we design a new objective function for learning a low dimensional discriminative subspace based on the multiobjective optimization framework. This objective function can easily adapt to the class topology of the classification problem. While LDA and MODA's objectives is to pull all the classes far apart from each other as much as possible, PARDA, or Pareto Discriminant Analysis, tries to equally spread all classes from each other. To this end, PARDA concentrates its effort on overlapping classes while safeguarding well separated classes from overlapping in the lower dimensional subspace. In other words, PARDA puts more effort in maximizing the distance between classes that are closer in the projected space, and will relax the constraint between classes that are

[1] A similar formulation was developed by Loog *et al.* [16] however as a sum of pairwise FDAs and not as a sum of pairwise KL divergences.

farther away. Figure (1C) shows the projection obtained by PARDA in a two dimensional space. Unlike MODA, Figure (1B), the two-dimensional projection obtained by PARDA encourages the case where classes are equally spread from each other in the lower dimensional space. To the best of our knowledge, this is the first paper to address multiclass dimensionality reduction as a multiobjective minimization problem.

2. Related work on LDA

Although there have been a lot of extensions to LDA, a detailed review is beyound the scope of the paper. However, we will review the previous literature related to heteroscedastic and multiclass extensions of LDA, and the small size sample problem.

In order to solve the heteroscedastic case, Campbell [2] derives a maximum likelihood approach to discriminant analysis. Assuming that all the classes have equal covariance matrices, Campbell shows that LDA is equivalent to imposing that the class means lie in a $(c-1)$–dimensional subspace. Following this approach, Kumar and Andreou [13] proposed heteroscedastic discriminant analysis, where they incorporate the estimation of the means and covariances in the low dimensional space. Saon *et al.* [22] define a new energy function to model the directionality of the data, $J(\mathbf{B}) = \prod_{i=1}^{c}(\frac{|\mathbf{B}^\top \mathbf{S}_b \mathbf{B}|}{|\mathbf{B}^\top \mathbf{\Sigma}_i \mathbf{B}|})^{n_i}$, where $\mathbf{\Sigma}_i$ is the class covariance matrix, $\mathbf{S}_b$ is the between–class scatter matrix and $\mathbf{B} \in \mathbb{R}^{d\times k}$ is the desired linear projection. De la Torre and Kanade [4] proposed a consistent generalization of LDA using the symmetric KL divergence and proposed extensions to the multimodal case. Recently, Zhu and Martinez [26] proposed techniques for subclass discriminant analysis in order to handle classes with multiple modalities.

To solve the class separation problem, Lu *et al.* [18] proposed a weighted variant of direct–LDA [24] and combine it with Fractional Step LDA [17]. For the between–class scatter matrix $\mathbf{S}_b$, they applied weights that are inversely proportional to the distance between class means. Alternatively, Loog *et al.* [16] suggested that the weights applied to $\mathbf{S}_b$ should link the distance between the class means to the amount of error they cause. Recently, Tao *et al.* [23] proposed a generalization of MODA under an information theoretic framework. GADA, or General Averaged Divergence Analysis, replaces the KL divergence in MODA with the general Bregman divergence, and replaces the sum of all pairwise divergences by a general mean divergence function. Similar to MODA, GADA does not consider each pair of classes separately, and hence it puts needless effort on already distant classes.

Another pathology of LDA for the case of images, is the small sample size problem. Hastie *et al.* [10] recast LDA as a linear regression problem. Under a least squares (LS) framework, linear regression can be generalized to more flexible nonlinear forms of regression, which in turn leads to a more flexible discriminant analysis (FDA). Moreover, LS allows for regularization to be applied on the variables (or features) and on the basis vectors yielding a penalized discriminant analysis (PDA) procedure within the FDA framework. Zhang and Sim [25] establishe a neat understanding for the four main subspaces that define LDA using the Fukunaga–Koontz transform (FKT). Based on their analysis they show that the FKT/LDA is equivalent to LDA/GSVD [12] and then provide a unification framework for other subspace methods such as: Fisherface, PCA+NULL, LDA/QR and LDA/GSVD.

3. Vector optimization

Multiobjective optimization (MOP), or vector optimization (VOP), is a branch of optimization science that is concerned with the optimization of more than one objective function simultaneously. In real world applications, it is often the case that the objectives are contradictory in a way that optimizing one of the objectives entails a poor performance of another. In such cases, one would require a good compromise solution which is suboptimal but acceptable as much as possible to the individual objective functions [11].

Let $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_\kappa(\mathbf{x})]^\top$ (*see notation* [2]) be the vector valued objective function to be optimized where $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^\kappa$, $\mathbf{x} \in \mathcal{R} \subseteq \mathbb{R}^d$ is the parameter vector for the set of objective functions, $f_j(\mathbf{x}) \in \mathbb{R}$ is the jth objective function, and $\mathcal{R}$ is the feasible set for the values of the parameter vector $\mathbf{x}$. The goal of VOP is to find $\mathbf{x}^*$ that simultaneously minimizes all $f_j(.)$'s. Since minimization presupposes in principle that various objective function values can be compared with each other, an appropriate ordering concept is needed on $\mathbb{R}^\kappa$. For reasons that will be shown later, it is difficult to have a total ordering that compares any two arbitrary elements in $\mathbb{R}^\kappa$, therefore a weaker, or a partial, ordering relation denoted by "$\leq$" will be used instead.

Definition (*Order relation "$\leq$" in $\mathbb{R}^\kappa$*) Let $\mathbf{y}_1$ and $\mathbf{y}_2$ be two points in $\mathbb{R}^\kappa$. The order relation "$\leq$" is defined as $\mathbf{y}_1 \leq \mathbf{y}_2 \iff \mathbf{y}_2 - \mathbf{y}_1 \in \mathbb{R}_+^\kappa$, where $\mathbb{R}_+^\kappa = \{\mathbf{y} \in \mathbb{R}^\kappa \mid y_i \geq 0, \text{ and } 1 \leq i \leq \kappa\}$ is the nonnegative orthant of $\mathbb{R}^\kappa$ and, $\forall i \in \{1, \ldots, \kappa\}$, $y_1^i \leq y_2^i$, $\exists j \in \{1, \ldots, \kappa\}$ s.t. $y_1^j < y_2^j$.

Since $\mathbb{R}_+^\kappa$ is a special case of the convex cone, then "$\leq$" is guaranteed to be compatible with the linear structure of $\mathbb{R}^\kappa$. In addition, properties such as reflexivity, transitivity, and antisymmetry are all satisfied [11] for the relation "$\leq$".

[2] Bold capital letters denote matrices: $\mathbf{A}$, bold lower-case letters denote a column vector: $\mathbf{x}$. Non-bold lower case letters represent scalar variables or indexes. $\mathbf{I}$ is the identity matrix of suitable dimension. $tr(\mathbf{A}) = \sum_i a_{ii}$ is the trace of the matrix $\mathbf{A}$. $\mathcal{N}(\,\cdot\,; \boldsymbol{\mu}, \mathbf{\Sigma})$ indicates a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$.

In optimization terms, if $\mathbf{y}_1 = \mathbf{f}(\mathbf{x}_1)$ and $\mathbf{y}_2 = \mathbf{f}(\mathbf{x}_2)$ represent two values of a vector valued objective function, then $\mathbf{y}_1 \leq \mathbf{y}_2$ implies that $\mathbf{y}_1$ is at least as small (as good) as $\mathbf{y}_2$ with regards to all objectives and it is strictly smaller (or better) with regards to at least one objective. In this case, $\mathbf{x}_1$ is said to *dominate* $\mathbf{x}_2$. There are vector pairs however for which neither $\mathbf{y}_1 \leq \mathbf{y}_2$ nor $\mathbf{y}_1 \geq \mathbf{y}_2$ are true; For instance $[2, 4]$ and $[4, 2]$. In such cases, the partial order relation reflects the fact that both objectives are of equal importance, and to select one solution, additional input is required from the domain expert, or *Decision Maker*. To this end, it is very important to emphasize the main difference between scalar valued optimization and vector valued optimization. While the former possesses a total ordering relation induced by the real numbers, the latter possesses only a partial ordering relation according to the definition above. In the following, we give two important definitions for the optimality of VOP.

Definition Let $\mathcal{Y} = \mathbf{f}(\mathcal{R}) \subseteq \mathbb{R}^\kappa$ be the image of the feasible set $\mathcal{R} \subseteq \mathbb{R}^d$. A point $\mathbf{y}^* \in \mathbf{f}(\mathcal{R})$ is called *Globally Efficient* with regards to the order relation "$\leq$" defined on $\mathbb{R}^\kappa$, if and only if there exists no other $\mathbf{y} \in \mathbf{f}(\mathcal{R})$ s.t. $\mathbf{y} \leq \mathbf{y}^*$ and $\mathbf{y} \neq \mathbf{y}^*$. A point $\mathbf{x}^* \in \mathcal{R}$ is called *Globally Pareto Optimal* if and only if $\mathbf{y}^* = \mathbf{f}(\mathbf{x}^*)$ is globally efficient.

Based on these definitions, VOP can be formally defined as finding efficient points $\mathbf{y}^* \in \mathbf{f}(\mathcal{R})$ with regards to the order relation "$\leq$" on $\mathbb{R}^\kappa$, and along with their Pareto optimal points $\mathbf{x}^*$ pertaining to them [5].

There are various approaches and frameworks to VOP and the interested reader can see [5] and [11] for a rigorous treatment of the subject. A class of these approaches are deterministic methods which "*scalarize*" the vector optimization problem through a parametric formulation. From the deterministic class, the weighted L_p-metric method [11], or the approximation of the ideal point – compromise method [5], is a well studied method for scalarizing VOPs with concrete theoretical results that guarantee Pareto optimal solutions.

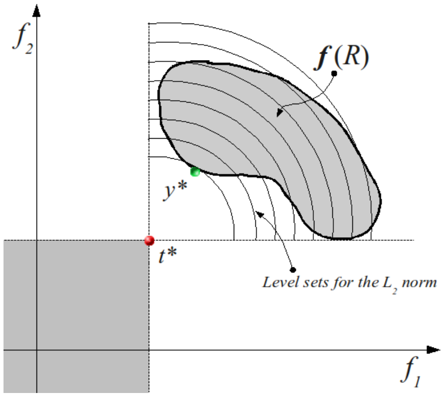

Figure 2. The intersection of level sets for the L_2-norm with $\mathcal{Y} = \mathbf{f}(\mathcal{R})$ in the objective space. Note that the ideal point $\mathbf{t}^* \notin \mathbf{f}(\mathcal{R})$ and the efficient (Pareto) point $\mathbf{y}^*$ is the closest to it.

In an ideal situation, the objective of VOP is to achieve the optimal solution for each individual objective function $f_j(\cdot)$. Let $\mathbf{t}^* \in \mathbb{R}^\kappa$ be such an ideal target point in the objective space. Then, $\forall\ \mathbf{y} \in \mathcal{Y}$, $\mathbf{t}^* \leq \mathbf{y}$ and $\mathbf{t}^*$ might or might not be in $\mathcal{Y}$. Since in real world problems, the individual objectives might conflict with each other, achieving $\mathbf{t}^*$ is impossible, however it can serve as a reference point with the goal of seeking a solution as close as possible to $\mathbf{t}^*$. Formally, given a distance function $dist : \mathbb{R}^\kappa \times \mathbb{R}^\kappa \to \mathbb{R}_+$, the L_p-metric method is given by: $\min_{\mathbf{x}\in\mathcal{R}} dist(\mathbf{f}(\mathbf{x}) - \mathbf{t}^*)$. Since the space $\mathbb{R}^\kappa$ is endowed with a vector norm $\|\cdot\|$ then the induced weighted distance, or the L_p-metric method can be defined as follows:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{R}} \ell(\mathbf{x}), \qquad (1)$$

$$\text{where} \quad \ell(\mathbf{x}) = \left(\sum_{j=1}^{\kappa} w_j |f_j(\mathbf{x}) - t_j^*|^p \right)^{\frac{1}{p}},$$

$p \in [1, \infty]$, $w_j > 0$ is the weight for the j-th objective function, and $\sum_{j=1}^{k} w_j = 1$. The weight w_j reflects the significance of the objective function $f_j(\cdot)$. In turn, w_j can reflect some *a priori* knowledge from the problem domain, or impose some bias on the final solution $\mathbf{x}^*$.

Finally, we can state Theorem 4.20 from [5] (see proof in pp. 112), that links the monotonicity of a norm to the solution obtained by Equation (1), in order to introduce our main result of this section in Corollary 3.2.

Theorem 3.1 *If $\|\cdot\|$ is a strictly monotonic norm and $\mathbf{x}^*$ is an optimal solution of Equation (1), then $\mathbf{x}^*$ is Pareto Optimal.*

Corollary 3.2 *For the L_p-norm $\|\cdot\|_p$, if $1 \leq p < \infty$ and $\mathbf{x}^*$ is the optimal solution for Equation (1), then $\|\cdot\|_p$ is strictly monotonic, and $\mathbf{x}^*$ is Pareto Optimal.*

The L_p-metric method has a nice interpretation in terms of level sets [5] $\{\mathbf{y} \in \mathbb{R}^\kappa \mid \|\mathbf{y} - \mathbf{t}^*\|_p \leq u\}$ where such sets contain all points of distance u or less to t^*. From that perspective, the goal of the L_p-metric method is to search for the smallest u such the intersection of the corresponding level set with $\mathcal{Y} = \mathbf{f}(\mathcal{R})$ is nonempty. Figure (2) illustrates this concept for the L_2-norm.

4. Pareto discriminant analysis

It is possible now to formulate the proposed method for learning a discriminant subspace based on the Pareto optimality introduced in the previous section. From FDA and

MODA [4], it is known that both are optimal for the 2-class homoscedastic and heteroscedastic cases respectively. To account for the multiclass setting, both approaches use the same scalarization by summing all pairwise KLDs and maximize that sum, which intrinsically yields the class separation problem. Here we propose a different approach for the scalarization of the multiclass heteroscedastic setting. Each pair of classes, C_i and C_j, $1 \leq i \neq j \leq c$, is considered as an individual objective function, and all $\kappa = c(c-1)/2$ pairs of classes result in κ objective functions that need be optimized simultaneously. Since it is expected that the objective functions can conflict with each other, using the VOP framework can guarantee that the obtained subspace will be in maximal agreement with all pairwise objectives but might be suboptimal for each individual objective. Using the L_p-metric method and setting an appropriate target vector $\mathbf{t}^*$, the optimization effort will be distributed according to the class topology of the classification problem. The simultaneous optimization of the objective functions with regard to the target vector will put more effort on overlapping classes while safeguarding distant classes from overlapping in the lower dimensional subspace.

For maximal separation between two heteroscedastic classes, MODA's formulation based on maximizing the symmetric KLD can define an appropriate objective function. For two classes C_i and C_j with different Gaussian distributions, $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, the symmetric KLD is given by [7]:

$$\begin{aligned} KL(i||j) \;=\; & \operatorname{tr}\left(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j^{-1} + \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j - 2\mathbf{I}\right) \\ & + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \end{aligned} \tag{2}$$

MODA seeks a linear transformation $\mathbf{B} \in \mathbb{R}^{d \times k}$ with $k < d$, for both classes – $\mathcal{N}(\mathbf{B}^\top\boldsymbol{\mu}_i, \mathbf{B}^\top\boldsymbol{\Sigma}_i\mathbf{B})$ and $\mathcal{N}(\mathbf{B}^\top\boldsymbol{\mu}_j, \mathbf{B}^\top\boldsymbol{\Sigma}_j\mathbf{B})$ – such that it maximizes the KLD between the two classes in the lower dimensional subspace. Note that the linear transformation $\mathbf{B}$ can have any number of bases k such that $1 \leq k \leq d-1$. This is another advantage of PARDA over FDA/LDA that can only define subspaces of dimensionality $k \leq \min(c-1, d-1)$. Let the KLD in the lower dimensional subspace be defined as follows:

$$\begin{aligned} g_{ij}(\mathbf{B}) \;=\; & \operatorname{tr}\{(\mathbf{B}^\top\boldsymbol{\Sigma}_i\mathbf{B})(\mathbf{B}^\top\boldsymbol{\Sigma}_j\mathbf{B})^{-1} + \\ & (\mathbf{B}^\top\boldsymbol{\Sigma}_i\mathbf{B})^{-1}(\mathbf{B}^\top\boldsymbol{\Sigma}_j\mathbf{B}) - 2\mathbf{B}^\top\mathbf{B}\} + \\ & \mathbf{u}_{ij}^\top\mathbf{B}\left[(\mathbf{B}^\top\boldsymbol{\Sigma}_i\mathbf{B})^{-1} + (\mathbf{B}^\top\boldsymbol{\Sigma}_j\mathbf{B})^{-1}\right]\mathbf{B}^\top\mathbf{u}_{ij}, \end{aligned} \tag{3}$$

where $\mathbf{u}_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$. After some algebraic manipulation, Equation (3) can be rewritten as follows:

$$\begin{aligned} g_{ij}(\mathbf{B}) \;=\; & \underbrace{\operatorname{tr}\left[(\mathbf{B}^\top\boldsymbol{\Sigma}_i\mathbf{B})^{-1}(\mathbf{B}^\top\mathbf{A}_{ij}\mathbf{B})\right]}_{\mathrm{I}(\mathbf{B})} \;+ \\ & \underbrace{\operatorname{tr}\left[(\mathbf{B}^\top\boldsymbol{\Sigma}_j\mathbf{B})^{-1}(\mathbf{B}^\top\mathbf{A}_{ji}\mathbf{B})\right]}_{\mathrm{II}(\mathbf{B})}, \end{aligned} \tag{4}$$

where $\mathbf{A}_{ij} = \mathbf{u}_{ij}\mathbf{u}_{ij}^\top + \boldsymbol{\Sigma}_j$ and $\mathbf{A}_{ji} = \mathbf{u}_{ij}\mathbf{u}_{ij}^\top + \boldsymbol{\Sigma}_i$. For the 2–class heteroscedastic setting, maximizing Equation (4) with respect to $\mathbf{B}$ yields a basis $\mathbf{B}_{ij}^*$ that is optimal for classes C_i and C_j only. In the c–class heteroscedastic setting however, there are κ pairwise objective functions that need to be maximized simultaneously. This is the major difference between MODA and LDA on one side and Pareto discriminant analysis (PARDA) on the other side. While MODA (and LDA in the homoscedastic case) sum over all $g_{ij}(\mathbf{B})$'s, and search for a basis that maximizes that sum, PARDA plugs all the pairwise objective functions in a multiobjective optimization framework and searches for a basis that is in maximal agreement with all pairwise objective functions and simultaneously maximizes them.

Formally, using the scalarization of the L_p-metric method in Equation (1) and Corollary (3.2), Pareto discriminant analysis is defined as follows:

$$\mathbf{B}^* \;=\; \underset{\mathbf{B} \in \mathcal{R}}{\arg\min}\;\; \ell(\mathbf{B}) = \sum_{i=1}^{c}\sum_{j=i+1}^{c} w_{ij}(g_{ij}(\mathbf{B}) - t_{ij}^*)^2, \tag{5}$$

where $\mathcal{R} \subseteq \mathbb{R}^{d \times k}$, and we have set $p = 2$ to guarantee the strict monotonicity of the norm, and decomposed the sum over all pairwise classes. According to Corollary (3.2) and given a proper target vector $\mathbf{t}^*$, the obtained solution $\mathbf{B}^*$ from Equation (5) will be Pareto Optimal.

4.1. The target vector $\mathbf{t}^*$ and the weight w_{ij}

The target vector $\mathbf{t}^*$ plays an important role for PARDA, and different target vectors lead to different Pareto optimal solutions. Although intuitively appealing heuristics or some a priori knowledge might be used to select the target vector, here we propose an approach inspired by the *"ideal point"* introduced in the previous section. That is, each pairwise objective function $g_{ij}(\mathbf{B})$ in Equation (4) is maximized using MODA, and the maximum obtained KLD is set as the target value t_{ij}^* in Equation (5). In all our experiments presented in Section 5, this approach was used to set all the target vectors for PARDA.

The weight w_j can drive the optimization procedure to favor some objective functions over others. For discriminant analysis, one would desire to bias the solution towards classes that will overlap in the lower dimensional space. Using the ideal target vector, we derive a simple notion for setting the weight w_{ij}. If $g_{ij}(\mathbf{B})$ is very close from its ideal value t_{ij}^*, then w_{ij} should be relatively small, while if $g_{ij}(\mathbf{B})$ is faraway from its ideal target value, then its corresponding weight should be large. Such a setting of weights will encourage the optimization procedure to concentrate its effort on faraway targets. To formulate this notion, we used a simple rule that provided good results in practice. That is, $w_{ij} = \tilde{w}_{ij} / \sum_{ij} \tilde{w}_{ij}$, where $\tilde{w}_{ij} = 1/|g_{ij}(\mathbf{B}_0) - t_{ij}^*|^2$ and $\mathbf{B}_0$ is the initial basis for PARDA.

4.2. Minimization of PARDA

Computationally, since there is no closed form solution for the maximization of the objective function in Equation (4), it follows that minimizing PARDA will not have a closed form solution as well, and an iterative algorithm based on gradient descent is used instead:

$$\begin{aligned}
\mathbf{B}^{t+1} &= \mathbf{B}^{t} - \eta \frac{\partial\, \ell(\mathbf{B})}{\partial \mathbf{B}}\,, \quad \text{where} \qquad (6)\\
\frac{\partial\, \ell(\mathbf{B})}{\partial \mathbf{B}} &= \sum_{i=1}^{c} \sum_{j=i+1}^{c} 2 w_{ij} \left(g_{ij}\left(\mathbf{B}\right) - t_{ij}^{*}\right) \frac{\partial\, g_{ij}(\mathbf{B})}{\partial \mathbf{B}},\\
\frac{\partial\, g_{ij}(\mathbf{B})}{\partial \mathbf{B}} &= \frac{\partial\, \mathrm{I}(\mathbf{B})}{\partial \mathbf{B}} + \frac{\partial\, \mathrm{II}(\mathbf{B})}{\partial \mathbf{B}}\ ,\\
\frac{\partial\, \mathrm{I}(\mathbf{B})}{\partial \mathbf{B}} &= 2\mathbf{A}_{ij}\mathbf{B}\boldsymbol{\Phi}_i - 2\boldsymbol{\Sigma}_i\mathbf{B}\boldsymbol{\Phi}_i(\mathbf{B}^{\top}\mathbf{A}_{ij}\mathbf{B})\boldsymbol{\Phi}_i\ ,\\
\frac{\partial\, \mathrm{II}(\mathbf{B})}{\partial \mathbf{B}} &= 2\mathbf{A}_{ji}\mathbf{B}\boldsymbol{\Phi}_j - 2\boldsymbol{\Sigma}_j\mathbf{B}\boldsymbol{\Phi}_j(\mathbf{B}^{\top}\mathbf{A}_{ji}\mathbf{B})\boldsymbol{\Phi}_j\ ,
\end{aligned}$$

$\boldsymbol{\Phi}_i = (\mathbf{B}^{\top}\boldsymbol{\Sigma}_i\mathbf{B})^{-1}$ and $\boldsymbol{\Phi}_j = (\mathbf{B}^{\top}\boldsymbol{\Sigma}_j\mathbf{B})^{-1}$.

The gradient descent procedure starts with a reasonable step length and it is decreased by 50% if it decreases the value of $\ell(\mathbf{B})$. Other strategies such as line search are possible but this simple method has provided good results. The loss function in problem (5) is a non–convex problem and any gradient descent method can be trapped into local minima. Therefore, we typically start the algorithm with multiple initializations and select the solution with lowest error. Alternatively, the best solution can be chosen with cross-validation.

5. Experimental results

Two sets of experiments were conducted to evaluate the performance of PARDA. In the first set, PARDA was tested on a synthetic data set to verify its behavior under ideal conditions. In the second set, PARDA was tested on a diversity of image databases and data sets from the UCI machine learning repository. The base line algorithms to compare with were: PCA, the weighted mean variant of direct LDA (dLDA) by Lu *et al.* [18] and modern approaches such as RCA [1] and MODA [4].

5.1. Synthetic data

To verify the behavior of PARDA in ideal situations, we tested PARDA on the same synthetic data set described in [4, 23]. The data set consists of 100 groups where each group consists of 400 samples, 200 for training and 200 for test, that are generated from five 20–dimensional Gaussian distributions. Each sample from class C_i is generated as $\mathbf{y}_i = \mathbf{T}_i\mathbf{b} + \boldsymbol{\mu} + \mathbf{n}$, where $\mathbf{y} \in \mathbb{R}^{20}$, $\mathbf{T}_i \in \mathbb{R}^{20\times 7}$, $\mathbf{b} \sim \mathcal{N}_7(0, \mathbf{I})$, and $\mathbf{n} \sim \mathcal{N}_{20}(0, \mathbf{I})$. The basis $\mathbf{T}_i$ are random matrices where each element is generated from $\mathcal{N}(0, 5)$. The means of each class were as follows: $\boldsymbol{\mu}_1 = (2\mathcal{N}(0,1) + 4)\mathbf{1}$, $\boldsymbol{\mu}_2 = \mathbf{0}_{20}$, $\boldsymbol{\mu}_3 = (2\mathcal{N}(0,1) - 4)[\mathbf{0}_{10}, \mathbf{1}_{10}]^{\top}$, $\boldsymbol{\mu}_4 = (2\mathcal{N}(0,1) + 4)[\mathbf{1}_{10}, \mathbf{0}_{10}]^{\top}$, and $\boldsymbol{\mu}_2 = (2\mathcal{N}(0,1)+4)[\mathbf{1}_5, \mathbf{0}_5, \mathbf{1}_5, \mathbf{0}_5]^{\top}$. Once the data are projected onto the lower dimensional space, we used the one nearest neighbor (1-NN) classifier with the Euclidean distance. Note that subspaces obtained from MODA and PARDA were orthogonalized with a Gram-Schmidt procedure. We run the algorithm 10 times from different initializations and select the solution with lowest error on the training set. The generalization error with its standard deviation was averaged over the test sets of the 100 groups. Table (3) shows the error rate for PCA, dLDA, MODA and PARDA as the number of projection dimensions increases. In this dataset PARDA consistently has the lowest error among other techniques.

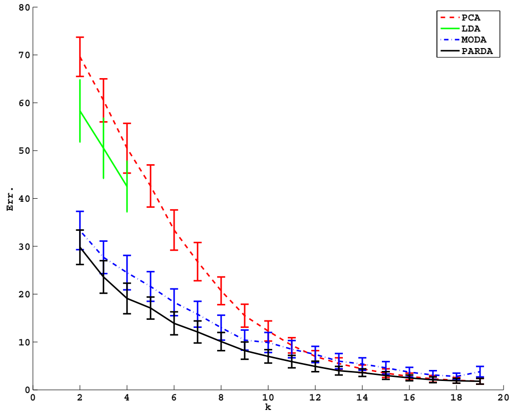

Figure 3. Generalization error (with standard deviation) for the synthetic data set described in [4]. Note that dLDA can find discriminative subspaces of dimension $k \leq \min(c-1, d-1)$, where c is the number of classes, and d is the number of features.

5.2. Image data sets and the UCI ML Repository

We tested PARDA on real data sets from standard benchmark databases. For this purpose, we used a large variety of data sets from different domains with various size and dimensionality. Such diversity of data sets is crucial to verify whether the concept of multiobjective optimization can be generalized on a large variety of problems. Please see Table (1) for a detailed description of the data sets used in our experiments. Similar to the previous experiments we used the 1–NN classifier with the Euclidean distance in the lower dimensional subspace.

The experimental setting proceeded as follows. For data sets with explicit training and test sets such as MNIST, USPS, UCI Isolet, the generalization error was measured directly on the test set. For other data sets, we used m–folds cross validation ($m = 50$) where the data sets were ran-

Table 1. Specifications of the data sets used in our experiments where number of classes, size and dimensionality are denoted by c, n and d respectively. Dashed numbers in the third column indicate the size of training and test sets respectively.

Data set	c	n	d	Ref.
Digits MNIST	10	20000 – 10000	24×24	[14]
Digits USPS	10	7291 – 2007	256	[15]
Faces Yale–B	38	2414	1024	[8]
Sitting Postures	10	2500	1080	[27]
UCI Bupa	2	345	6	[20]
UCI German	2	1000	24	”
UCI HouseVotes	2	435	16	”
UCI Iris	3	150	4	”
UCI Isolet	26	6238 – 1559	617	”
UCI Letter	26	20000	16	”
UCI NewThyroid	3	215	5	”
UCI Pima	2	768	8	”
UCI Segment	7	2310	18	”
UCI Spam	2	4601	57	”
UCI TicTacToe	2	958	9	”
UCI Vowel	11	990	11	”

domly partitioned to 80% – 20% for training and test sets respectively. The generalization error and standard deviation were measured on the test sets only. For data sets with very high dimensionality such as MNIST, Yale–B, Sitting Postures and UCI Isolet, PCA was used to reduce the dimensionality of these data sets keeping 97% – 98% of their total variance. Since the original size of the MNIST training set is so large (60000), we only used the first third of the samples in each class resulting in a training set of 20000 samples. The test set of MNIST was kept untouched.

The generalization error was measured at two different projection dimensions: 1) $k = \min(c-1, d-1)$, and 2) $c - 1 \leq k \leq d$. Table (2) shows the generalization error for the four techniques PCA, dLDA, MODA and PARDA when $k = min(c-1, d-1)$. This is where FDA/LDA should perform optimally. Except for the Isolet data set, PARDA significantly outperforms the standard FDA/LDA approach even for problems with two classes. Similar behavior is observed when comparing PARDA with MODA for problems with more than two classes. For 2–class problems, MODA and PARDA have almost an identical performance with slight improvement for PARDA in some cases. Table (3) shows the generalization error for some data sets when $c-1 \leq k \leq d-1$. This can be only obtained by PCA, RCA, MODA and PARDA but not by FDA/LDA. The optimal k for each technique was set according to the lowest generalization error achieved by the algorithm. We note that in most of the multiclass problem, PARDA maintains a low generalization error with respect to other approaches, while competitive with other approaches in 2–class problems.

Table 2. Generalization error (with standard deviation) for image data sets and some standard benchmark data sets from the UCI Machine Learning Repository. Here $k = \min(c-1, d-1)$.

Data set	PCA	dLDA	RCA	MODA	PARDA
MNIST	11.4	10.6	10.3	15.3	**9.5**
USPS	7.5	6.5	**4.8**	12.2	**5.7**
Yale–B	49.0(1.3)	40.2(1.8)	**3.9**(1.1)	49.6(5.6)	**33.6**(2.3)
Sit–Post	**21.7**(1.7)	23.0(1.4)	29.7(1.4)	29.8(1.9)	25.7(1.3)
Bupa	46.2(5.0)	44.9(5.8)	**40.6**(6.6)	**42.3**(6.4)	**42.3**(6.6)
German	38.4(2.8)	41.2(2.9)	**30.9**(2.8)	**37.7**(3.9)	**37.8**(3.4)
House	15.2(3.5)	13.3(3.5)	**4.6**(2.1)	6.8(2.8)	**6.5**(2.3)
Iris	4.2(3.5)	7.0(4.2)	4.4(3.4)	4.4(3.4)	**3.6**(3.6)
Isolet	15.3	**9.3**	**6.6**	25.4	12.7
Letter	4.5(0.3)	4.4(0.3)	4.2(0.2)	5.4(2.6)	**3.9**(0.3)
NewThy.	8.0(3.0)	11.8(4.2)	**4.4**(5.7)	8.4(4.4)	**6.3**(3.4)
Pima	37.5(3.5)	39.6(4.0)	**31.4**(4.1)	**33.8**(3.5)	**33.8**(3.9)
Segment	6.1(1.1)	5.0(0.9)	26.1(18.6)	5.2(1.1)	**4.7**(0.8)
Spam	36.4(1.1)	34.7(1.3)	**14.4**(1.0)	**29.1**(1.7)	**29.1**(1.4)
TicTac	44.5(8.2)	20.7(11.2)	3.1(1.2)	3.9(4.0)	**2.5**(1.1)
Vowel	1.5(0.8)	**1.4**(0.8)	1.6(0.8)	1.8(0.9)	**1.4**(0.9)

Table 3. Generalization error (with standard deviation) for some data sets from Table (2) when the error is significantly decreased by adding more projection dimensions ($c - 1 \leq k \leq d - 1$).

Data set	PCA	RCA	MODA	PARDA
USPS	2.3	4.8	13.5	**1.9**
Bupa	40.2(5.3)	39.5(4.9)	40.7(5.5)	**38.8**(4.7)
German	37.0(2.4)	**30.7**(2)	34.9(3.0)	**32.9**(3.1)
House	15.2(3.5)	**4.9**(5.1)	6.8(2.8)	**5.1**(2.1)
Isolet	13.8	**6.4**	20.7	**10.1**
NewThy.	8.2(3.2)	4.4(5.7)	4.7(2.8)	**4.0**(2.6)
Pima	36.0(3.8)	**30.7**(3.7)	33.5(3.4)	**31.9**(3.3)
Spam	24.3(1.1)	**11.4**(1.8)	21.2(1.3)	**18.3**(1.4)

As an illustration for the projection obtained by PARDA with a real data set, we used a data set from the UCI repository, New-Thyroid, and compared it to projections obtained by PCA, dLDA and MODA. Figure (4) shows the four projections obtained by the four different algorithms together with the classification error. The symmetric KLD for $\{(C_1, C_2), (C_1, C_3), (C_2, C_3)\}$ for the four algorithms are: PCA $\{18.9, 91, 3, 560.9\}$, dLDA $\{22.8, 122.7, 1991\}$, MODA $\{16.3, 236.2, 2217.2\}$, and PARDA $\{29.1, 50, 23.1\}$. Note that, although dLDA and MODA has maximized the large KLDs, they did not yield interesting projections nor the lowest error.

6. Concluding remarks

We close by remarking that the literature has proposed various extensions of FDA to the multiclass case by summing over all pairwise KLDs. However, as our study might suggest, this might not be the most appropriate

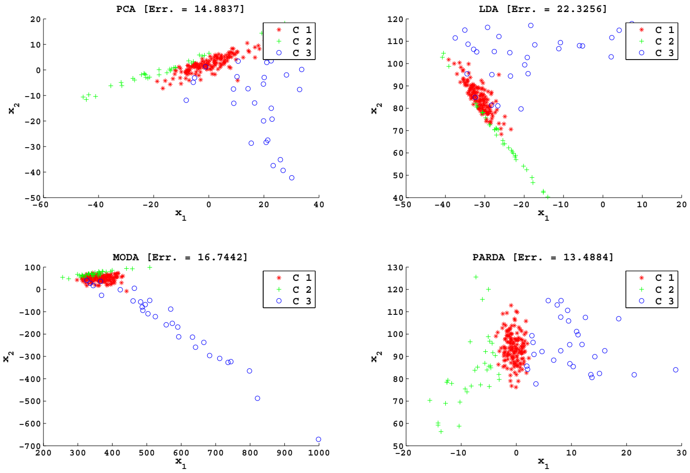

Figure 4. Projection of the New-Thyroid data set using PCA, dLADA, MODA and PARDA. See text for the Symmetric KLD values between all classes for the four algorithms.

approach for handling the multiclass case. To this end, our paper proposes a fundamentally different approach to deal with multiclass dimensionality reduction based on the machinery of multiobjective optimization and the concept of Pareto optimality. Experimental results on synthetic data, images and several data sets from the UCI Machine Learning Repository show how PARDA outperforms current approaches. Such encouraging results suggest that PARDA and Pareto optimality are promising research directions that are worth further investigation with other problems in pattern recognition and machine learning.

Acknowledgment K. Abou-Moustafa acknowledges the financial support of REPARTI and FQRNT. This research was mainly supported by REPARTI–FQRNT Award for International Training, and NSERC Discovery Grant Rgpin 36560-06.

References

[1] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a Mahalanobis metric from equivalence constraints. *J. of Machine Learning Research*, 6:937–965, 2005. 6

[2] N. A. Campbell. Canonical variate analysis – a general formulation. *Australian Journal of Statistics*, 26:86–96, 1984. 2, 3

[3] R. Chen. Solution of MINIMAX problems using equivalent differentaible functions. *Computers & Mathematics with Applications*, 12:1165–1169, 1985. 2

[4] F. De La Torre and T. Kanade. Multimodal oriented discriminant analysis. In *ACM Proc. of ICML*, pages 177–184, 2005. 2, 3, 5, 6

[5] M. Ehrgott. *Multicriteria Optimization*. Springer, 2005. 4

[6] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936. 1

[7] K. Fukunaga, editor. *Introduction to Statistical Pattern Recognition*. Academic Press, 1972. 1, 2, 5

[8] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. PAMI*, 23(6):643–660, 2001. 7

[9] O. Hamsici and A. Martinez. Bayes optimality in linear discriminant analysis. *IEEE Trans. PAMI*, 30(4):647–657, Apr 2008. 2

[10] T. Hastie, R. Tibshirani, and B. Andreas. Flexible discriminant and mixture models. In *Statistics and neural networks: advances at the interface*, pages 1–23, 1999. 2, 3

[11] C. Hillermeier. *Nonlinear multiobjective optimization*. Birkhäuser Verlag, 2001. 2, 3, 4

[12] P. Howland and H. Park. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Trans. PAMI*, 26(8):995–1006, Aug 2004. 3

[13] N. Kumar and A. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication*, 26(4):283 – 297, 1998. 2, 3

[14] Y. LeCun. The MNIST database of handwritten digits, 1998. `http://yann.lecun.com/exdb/mnist/`. 7

[15] Y. LeCun. United States Postal Service (USPS) data set, 1998. `www.gaussianprocess.org/gpml/data/`. 7

[16] M. Loog, R. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Trans. PAMI*, 23:762–766, Jul 2001. 2, 3

[17] R. Lotlikar and R. Kothari. Fractional–step dimensionality reduction. *IEEE Trans. PAMI*, 22(6):623–627, Jun 2000. 3

[18] J. Lu, K. Plataniotis, and A. Venetsanopoulos. Face recognition using LDA–based algorithms. *IEEE Trans. Neural Networks*, 14(1):195–200, Jan 2003. 2, 3, 6

[19] S. Mika. Kernel Fisher Discriminants. In *PhD Thesis, University of Technology, Berlin*, 2002. 2

[20] D. Newman, S. Hettich, C. Blake, and C. Merz. UCI Repository of Machine Learning Databases, 1998. `www.ics.uci.edu/∼mlearn/MLRepository.html`. 7

[21] C. R. Rao. *Linear statistical inference and its applications*. John Wiley & Sons, New York, 1965. 1

[22] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen. Maximum likelihood discriminant feature spaces. In *ICASSP*, 2000. 3

[23] D. Tao, X. Li, X. Wu, and S. Maybank. General averaged divergence analysis. In *IEEE Proc. of Seventh ICDM*, pages 302–311, 2007. 2, 3, 6

[24] H. Yu and J. Yang. A direct LDA algorithm for high–dimensional data – with application to face recognition. *Pattern Recognition*, 34:2067–2070, 2001. 3

[25] S. Zhang and T. Sim. Discriminant subspace analysis: A Fukunaga–Koontz approach. *IEEE Trans. PAMI*, 29(10):1732–1745, Oct 2007. 2, 3

[26] M. Zhu and A. Martinez. Subclass discriminant analysis. *IEEE Trans. PAMI*, 28(8):1274–1286, Aug 2006. 3

[27] M. Zhu and A. Martinez. Pruning noisy bases in discriminant analysis. *IEEE Trans. Neural Networks*, 19(1):148–157, 2008. 7