A Hybrid PCA-LDA Model for Dimension Reduction

Nan Zhao, Washington Mio and Xiuwen Liu

Abstract— **Several variants of Linear Discriminant Analysis (LDA) have been investigated to address the vanishing of the within-class scatter under projection to a low-dimensional subspace in LDA. However, some of these proposals are ad hoc and some others do not address the problem of generalization to new data. Meanwhile, even though LDA is preferred in many application of dimension reduction, it does not always outperform Principal Component Analysis (PCA). In order to optimize discrimination performance in a more generative way, a hybrid dimension reduction model combining PCA and LDA is proposed in this paper. We also present a dimension reduction algorithm correspondingly and illustrate the method with several experiments. Our results have shown that the hybrid model outperform PCA, LDA and the combination of them in two separate stages.**

I. Introduction

INFOMAX principle [1][2] tries to explain neural networks as mapping the input (sensory information) to a more efficient output (internal representation). In this mapping, the input and the output are considered as two sets of variables (or dimensions). The redundancy in the input variable set is reduced due to a constrained optimization on mutual information between the inputs and outputs. Clearly, neural networks for recognition and other higher level tasks require not only expressive information for structure representation but also most discriminative information.

Dimension reduction is widely used for applications such as face recognition[3][4][5] and text classification[6]. It is similar to how neural networks work. The resultant subspace gives an effective discriminative representation of the original space in a more efficient way. Linear Discriminant Analysis (LDA) is a classical supervised dimension reduction technique. It is designed to optimally cluster different classes of objects under a projection to a low dimensional subspace [7][8]. More precisely, given m-dimensional feature vectors representing different classes of objects, one uses labeled training data to learn a p-dimensional subspace, $p < m$ fixed, for which the ratio R of the total between-class and within-class scatter of the projected data is maximized. To simplify the discussion, we consider the case $p = 1$, that is, reduction to a 1-dimensional subspace.

In many applications, the dimension m of the original feature vectors is rather large as compared to the number T of training samples. This is often the case, for example, when the feature vectors are images and m is the number of pixels. In such cases, it is almost always possible to find a 1-dimensional projection that collapses each cluster in the training set to a single point, making the total within-class scatter vanish, or equivalently, making the ratio R become infinite. Although the objects in the training set get clustered perfectly under such a projection, the discrimination of new test data is generally very poor. This shortcoming can be traced to small sample size, that is, $T \ll m$. To prevent the vanishing of the within-class scatter, the LDA cost function has been regularized in [9] by adding a small number $\epsilon > 0$ to the denominator. However, this does not address the key issue of poor generalization. Two-stage approaches to LDA improve the situation somewhat [10][11]. On a first step, one performs a preliminary dimension reduction to a k-dimensional subspace using Principal Component Analysis (PCA) and then applies LDA to the reduced k-dimensional data. Experimental results show noticeable improvement on generalization to new test data with this strategy. Note, however, that the choice of k is *ad hoc* and there is no clear learning model underlying this approach. Other variants of LDA have been proposed in [3][6] to address the vanishing of the within-class scatter.

The goal of this paper is to develop a discriminative dimension reduction model for the choice of a 1-dimensional subspace that yields an optimal balance between generalization and class discrimination to new data. We take the viewpoint that what enables a two-stage type approach to improve the generalization to new data is that PCA is designed to preserve, as much as possible, the geometry and clustering patterns observed in the original set of feature vectors. As such, discrimination learned with LDA on the reduced representation, which is not subject to the small-sample-size problem, better extrapolates to test data. In fact, a combination between PCA and LDA is reasonable for classification rather than pure LDA in that LDA is not always superior to PCA for classification [12]. However, since PCA may lose some potential discriminative information in the ignored principle components while involving useless information for classification in the first few principle components (like variations due to illumination and viewing direction in face recognition), PCA and LDA should be combined in a more intrinsic way rather than in two separate stages. Following this philosophy, we propose a hybrid model guided by a cost function that is a linear interpolation of the PCA and regularized LDA cost functions. This gives a family of models indexed by an interpolation parameter t, which takes values on the interval $[0, 1]$, with $t = 0$ corresponding to PCA and $t = 1$ to regularized LDA. We then use cross-validation data to choose a value of t that maximizes classification performance. Therefore, the maximization of

Nan Zhao and Xiuwen Liu are with the Department of Computer Science, Florida State University, Tallahassee, Florida 32306, USA (email: {nzhao, liux}@cs.fsu.edu).

Washington Mio is with the Department of Mathematics, Florida State University, Tallahassee, Florida 32306, USA (email: mio@math.fsu.edu).

This research was supported in part by NSF grant DMS-0713012 and NIH Roadmap for Medical Research grant U54 RR021813.

the cost function is modeled as an optimization problem concerning t in this linear interpolation.

The paper is organized as follows. In Section II, we present the hybrid PCA-LDA model for dimension reduction. The optimization problem that arises in the estimation of the optimal subspace is discussed in Section III, including an algorithm to solve the problem. Section IV illustrates the methodology with several experiments. Section V gives a brief summary and discussion on the text.

II. The Hybrid PCA-LDA Model

The problem using this model can be simply described as following: given a set of labeled training data from different classes and another set of unlabeled testing data from the same group of classes, identify each testing data relying the new model. Both sets consist of feature vectors in some high-dimensional Euclidean space $\mathbb{R}^m$ representing K different classes of objects. In the training set, the feature vectors representing the ith class are denoted $x_{ij} \in \mathbb{R}^m$, with $1 \leqslant i \leqslant K$ and $1 \leqslant j \leqslant n_i$, where n_i is the number of samples in the ith class. Thus, the total number of objects in the training set is $N = n_1 + \ldots + n_K$.

The most straightforward method for such problem is applying a nearest neighbor rule on the data space, trying to identify each testing data through assigning to it the label of the training data with the closest distance in the data space. However, the drawbacks are that the computational cost in time and amount of storage required are both very expensive, especially when the data dimension m is high. Therefore, it is natural to apply a dimension reduction model on both data sets and a nearest neighbor classifier is then used in the resultant lower-dimensional feature space. As widely used in the field of dimension reduction, PCA was involved in many applications [4] for such purpose. PCA is a linear projection from an original m-dimensional space to a lower p-dimensional space ($m > p$) relying on maximization of the total scatter matrix of projected samples. Specifically, given a training data set as defined above, the new feature vectors $y_{ij} \in \mathbb{R}^p$ after such projection can be defined as following:

$$\mathbf{y_{ij}} = M_{pca}^T \mathbf{x_{ij}} \qquad 1 \leqslant i \leqslant K \; and \; 1 \leqslant j \leqslant n_i \tag{1}$$

where $M_{pca} \in \mathbb{R}^{m \times p}$ is an orthonormal matrix.

Let

$$\mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \tag{2}$$

be the sample mean of the ith class and

$$\mu = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ij} \tag{3}$$

the mean of the entire training set. Then, the scatter matrix S of all the training data is defined as (cf. [7]):

$$S = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \mu)(x_{ij} - \mu)^T. \tag{4}$$

Thus, after applying the linear projection M_{pca}, the scatter matrix of the reduced feature vectors is $M_{pca}^T S M_{pca}$. The determinant of this projected total scatter matrix is maximized in PCA so as to optimize the transformation matrix M_{pca}:

$$\begin{aligned} M_{pca} &= \underset{M_{pca}}{\mathrm{a}rg\; max} |M_{pca}^T S M_{pca}| \\ &= [\mathbf{m_1 m_2 ... m_p}] \end{aligned} \tag{5}$$

where $\{\mathbf{m_i} | i = 1, 2, ..., p\}$ correspond to the p largest eigenvalues ($p \leqslant K$). However, PCA does not only maximize the between-class scatter but also the within-class scatter. Thus, in the first few principle components, some unwanted information for classification may be preserved while useful information for discrimination may be lost [5].

As labels are known for the training data set, it is reasonable to build class specific model for discriminative dimension reduction on the feature space. A well defined class specific model is LDA, such as Fisher's Linear Discriminant (FLD) [18]. It tries to perform more reliable dimension reduction via linear mapping and still maintains the linear separability among different classes. After applying this linear mapping M_{lda}, the between-class scatter and the within-class scatter of the transformed feature vectors are $M_{lda}^T S_B M_{lda}$ and $M_{lda}^T S_W M_{lda}$ accordingly. Instead of maximizing the determinant of the total scatter matrix in PCA, a ratio of the between-class scatter matrix and the within-class scatter matrix is maximized. The between-class scatter matrix S_B and the within-class scatter matrix S_W are formally defined as following:

(i) $S_i = \sum_{j=1}^{n_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T$, the scatter of the ith class.

(ii) $S_W = \sum_{i=1}^{K} S_i$, the within-class scatter.

(iii) $S_B = \sum_{i=1}^{K} n_i (\mu_i - \mu)(\mu_i - \mu)^T$, the total between-class scatter.

If S_W is non-singular, M_{lda} is the orthonormal matrix maximizing the ratio between the determinant of the between-class scatter and the determinant of the within-class scatter:

$$\begin{aligned} M_{lda} &= \underset{M_{lda}}{\mathrm{a}rg\; max} \frac{|M_{lda}^T S_B M_{lda}|}{|M_{lda}^T S_W M_{lda}|} \\ &= [\mathbf{m_1 m_2 ... m_p}] \end{aligned} \tag{6}$$

where $\{\mathbf{m_i} | i = 1, 2, ..., p\}$ correspond to the p largest eigenvalues ($p \leqslant K$).

In order to lower the computational cost via PCA while preserve the linear separability among different classes by LDA, a 2-stage dimension reduction model was proposed [10]: the original image space is projected to a lower dimensional feature space (N-K) via PCA in the first stage and then apply LDA to reduce the dimension to even lower dimensional space (K-1) in the second stage. In this case, the transformation matrix M for dimension reduction is defined as:

$$M^T = M_{lda}^T M_{pca}^T \tag{7}$$

where

$$\begin{aligned} M_{pca} &= \underset{M}{\mathrm{a}rg\ max}|M^T S M| \\ M_{lda} &= \underset{M}{\mathrm{a}rg\ max}\frac{|M^T M_{pca}^T S_B M_{pca} M|}{|M^T M_{pca}^T S_W M_{pca} M|} \end{aligned} \tag{8}$$

The problem of the two-stage model is that it cannot avoid loss of discriminating information while involving unwanted information for classification in the first stage (PCA). Therefore, the first few principle components may not be suitable for building a discriminative model in the second stage (LDA). To overcome this problem, we propose an alternative to the dimension reduction model preserving both low computational cost and linear separability. This model for dimension reduction is called hybrid PCA-LDA model. Instead of separating PCA and LDA as two stages, we combine them in a linear combination. To clarify our idea in this paper, here we only discuss the case of reducing the feature space to 1-dimension subspace. A 1-dimensional subspace will be represented by a unit vector e and the usual dot product of vectors $u, v \in \mathbb{R}^m$ will be written as $\langle u, v\rangle$. The cost functions associated with PCA [14] and regularized LDA [9] are

$$\begin{aligned} &H_1(e) = \langle e, Se\rangle \qquad \text{and} \\ &H_2(e) = \frac{\langle e, S_B e\rangle}{\langle e, (S_W + \epsilon I)e\rangle} = \frac{\langle e, S_B e\rangle}{\langle e, S_W e\rangle + \epsilon}\,, \end{aligned} \tag{9}$$

respectively, where $\epsilon > 0$ is a small number that prevents the occurrence of vanishing denominators. For each t, $0 \leqslant t \leqslant 1$, we propose the linear interpolation

$$\begin{aligned} F_t(e) &= \frac{1-t}{2} H_1(e) + \frac{t}{2} H_2(e) \\ &= \frac{1-t}{2}\langle e, Se\rangle + \frac{t}{2}\,\frac{\langle e, S_B e\rangle}{\langle e, S_W e\rangle + \epsilon} \end{aligned} \tag{10}$$

of the cost functions for PCA and regularized LDA. The goal is to find a unit vector $e_t \in \mathbb{R}^n$ that maximize the proposed cost function $F_t(e)$:

$$e_t = \underset{e_t}{\arg\min}\, F_t(e) \tag{11}$$

The main computational task is to calculate e_t. Once e_t is known, we choose t so that the classification performance, with the nearest-neighbor classifier applied to cross-validation data, is optimized under projection to the subspace spanned by e_t.

III. Maximization of the Cost Function

As the objective is to maximize F_t under the constraint $\|e\|^2 = 1$, the computational task on e_t can be modeled as a Lagrange optimization problem. First we construct the Lagrange function

$$L(e\lambda) = F_t(e) + \lambda(\|e\|^2 - 1) \tag{12}$$

where λ is a scalar called Lagrange multiplier. This constrained optimization problem can be converted into an unconstrained problem via taking the derivative on both side:

$$\frac{\partial L(e,\lambda)}{\partial e} = \frac{\partial F_t(e)}{\partial e} + \lambda\frac{\partial(\|e\|^2 - 1)}{\partial e} = 0 \tag{13}$$

A calculation shows that the gradient of F_t is

$$\begin{aligned} \nabla F_t(e) = (1-t)Se &+ \frac{t}{\langle e, S_W e\rangle + \epsilon} S_B e \\ &- t\frac{\langle e, S_B e\rangle}{(\langle e, S_W e\rangle + \epsilon)^2} S_W e\,. \end{aligned} \tag{14}$$

Therefore, our goal is to find e such that

$$\nabla F_t(e) = \lambda e. \tag{15}$$

In other words, to make $\nabla F_t(e)$ parallel to e. Using the fact that S is positive semi-definite, it follows from (14) that $\langle \nabla F_t(e), e\rangle \geqslant 0$. By (15), $\lambda = \langle \nabla F_t(e), e\rangle$, so that λ must be non-negative. If $\lambda > 0$, we can normalize both sides of (15), which then becomes

$$\frac{\nabla F_t(e)}{\|\nabla F_t(e)\|} = e\,. \tag{16}$$

If we let $T(e) = \nabla F_t(e)/\|\nabla F_t(e)\|$, Equation 16 translates to finding a fixed point of the mapping T, which maps unit vectors to unit vectors. The problem with this argument is that λ may vanish. This can be corrected as follows. Add a positive multiple of e to both sides of (15) and then normalize. That is, change the mapping T to

$$T(e) = \frac{\alpha e + \nabla F_t(e)}{\|\alpha e + \nabla F_t(e)\|}\,, \tag{17}$$

with $\alpha > 0$. A fixed point of this modified version of T gives a unit vector e where the constrained gradient of F_t vanishes.

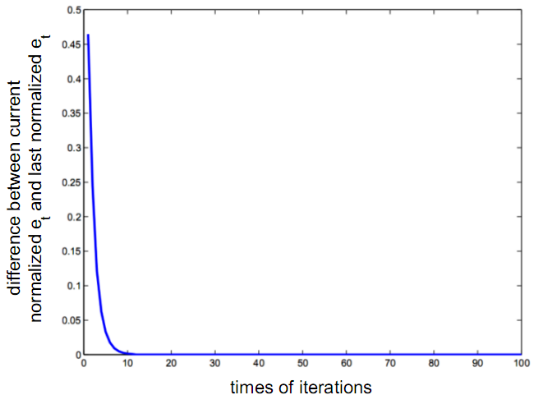


Fig. 1. A sample convergence of e_t. The vertical axis shows the difference between current normalized e_t and last normalized e_t in each iteration.

We adopt an iterative scheme to search for a fixed point of T, a procedure that closely resembles constrained gradient descent. Initialize the search arbitrarily. Then, use the update $e_{n+1} = T(e_n)$. A sample process of obtaining the convergence is shown in Fig. 1. Note that, for $t = 0$, this iterative

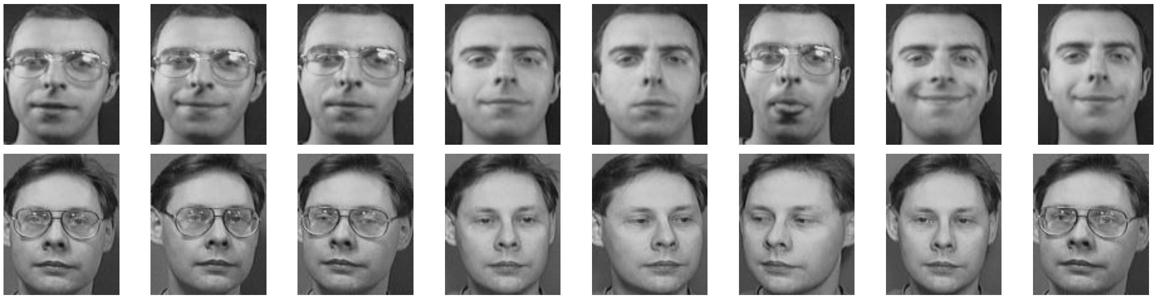


Fig. 2. Sample images of 2 individuals from the AT&T Database of Faces. Images in each row was taken from an individual.

procedure is just the power method to find the first principal direction of S. For $t = 1$, this gives an alternative approach to regularized LDA. A pseudo-code for this algorithm is give in **Algorithm 1**.

Algorithm 1: Hybrid PCA-LDA algorithm

Input: Date matrix A

Output: Reduced data matrix A^L

```
begin construct the scatter matrices $S_W, S_B$ and $S$ in (ii), (iii) and (4).
  Initialize a random unit vector $e_t \in \mathbb{R}^n$.
  do apply normalization on $\alpha e_t + \nabla F_t(e_t)$ as $T(e)$,
     where $\nabla F_t(e)$ is given by (14).
    $e_t \leftarrow T(e)$.
  until no significant change in $e_t$.
  return $A^L \leftarrow Ae_t$.
end
```

IV. Experimental results

A. Data Sets

In this section, we present and discuss the properties of proposed hybrid dimension reduction model using two different type of data sets: the facial recognition data set from AT&T Laboratories Cambridge [16] and the UCI wine data set [13].

In the first data set, also formally called ORL data set of faces, it consists of 400 facial images, 10 each from 40 individuals. The original image size is 92×112, with 256 grey-scale value in pixel. We sub-sampled the images as 28×23 and the dimension of each image, as an instance, is therefore reduced to $28 \times 23 = 644$. All the images were well centralized. For most of the individuals the facial images were taken at different times and under various lighting conditions but all with a homogeneous background in darkness. The major challenge of this data set is the variation of poses, expressions and facial details. Some individuals were both captured with and without glasses. However, there is minor occlusion due to the presence of glasses.

The second data set is based on a chemical analysis of wine from the same region in Italy, but from three different cultivars. We refer to them as types 1, 2 and 3. In each of these 3 types of wines, 13 constituents of wine are taken into account and each instance is thus a 13-dimensional feature vector. The 13 variables quantify the following constituents or properties of wine: alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavonoids, nonflavonoid phenols and proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines, proline. The full data set comprise 178 feature vectors: 59 from class 1, 71 from class 2 and 48 from class 3.

B. Choice of ϵ

As the proposed subspace learning model involves a parameter $\epsilon > 0$, our first experiment was designed to provide evidence that the optimal choice of t, based on performance on cross-validation data, is not very sensitive to the choice of ϵ, so long as ϵ be small. In this preliminary experiment, we used a small subset of the ORL data set, 2 individuals with 20 facial images. Each image is rescaled into a vector with 645 dimensions (with one more dimension indicating the label). Figure 2 shows a few samples. The interpolation parameter t was discretized into 21 values obtained by dividing the interval $[0, 1]$ into 20 equal parts. We carried out 1,000 experiments with randomized 5-fold cross-validation for a total of 5,000 runs of the algorithm. In that sense, the training set is randomly split into 5 disjoint sets of equal size and the classifier is then trained 5 times. Each time a different set is considered as a validation set for estimating the generalization error and the other sets combined as a training set used to adjust parameter e_t in the hybrid model. The estimated performance is generated by the mean of these 5 generalization errors. Figure 3 shows the average classification performance as a function of t for four different choices of the parameter ϵ. The graphs are nearly the same for small values of ϵ, as are the values of t where the validation performance peaks. Therefore, we fix ϵ as 0.01 for the experiments below. Note that, in this case, LDA ($t = 1$) does not perform well because the number of pixels is much larger than the number of elements in the training set, as explained above.

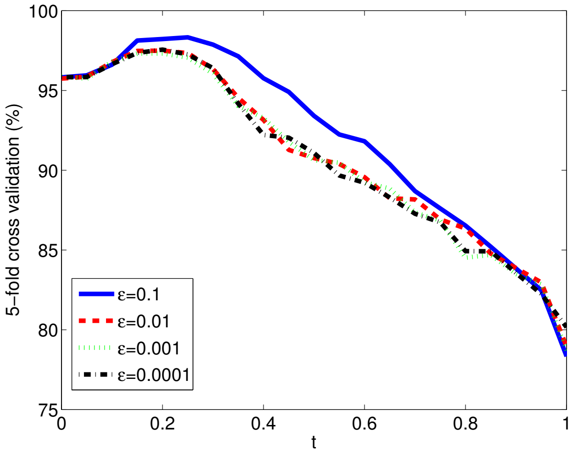


Fig. 3. Classification performance for different values of the parameter $\boldsymbol{\epsilon}$. The horizontal axis represents parameter t in the hybrid model (10) and the vertical axis shows the classification performance under 5-fold cross validation.

We carried out another face classification experiment with two different individuals. The parameter was set to $\epsilon = 0.01$ and 10,000 experiments were carried out with 5-fold cross validation. Figure 4 shows the average performance over the 50,000 runs. The optimal value occurs near $t = 0.2$, showing that optimal classification is achieved by a hybrid PCA-LDA model.

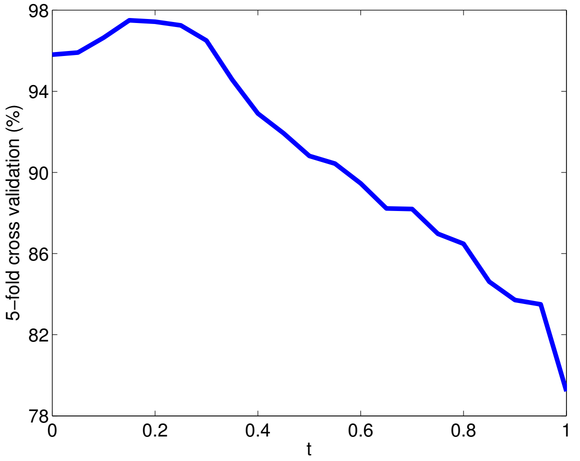
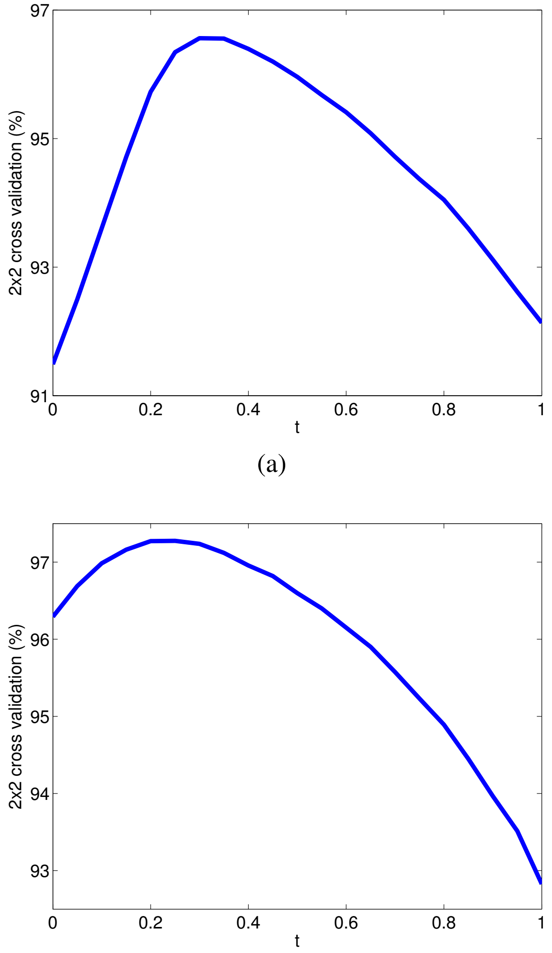


Fig. 4. Average performance on a subset of the AT&T Database of Faces. The horizontal axis represents parameter t in the hybrid model (10) and the vertical axis shows the classification performance under 5-fold cross validation.

C. A Comparison Between the Hybrid Model and the two-stage Model

To demonstrate the superiority of proposed hybrid PCA-LDA model over PCA, LDA and the two-stage model, we designed another set of experiments to compare the performance of these models when reducing original data space to a final 1-dimensional space. The second set of experiments, with the UCI wine database, illustrate the fact that even in cases where the dimension m of the feature space is relatively small, the hybrid model can boost classification performance obtained with either PCA or LDA alone. We first used 59 samples of type 1 and 71 samples of type 2. We performed 10,000 experiments with randomized 2×2 cross validation for a total of 40,000 runs. The parameter was set to $\epsilon = 0.01$. Figure 5(a) shows the average performance over all runs using the hybrid model. In addition, figure 5(b) shows the results of a similar experiment with 71 samples of type 2 and 48 samples of type 3. Obviously, neither pure LDA ($t = 1$ in Fig. 5) nor pure PCA ($t = 0$ in Fig. 5) outperform the hybrid model (the peaks in Fig. 5).



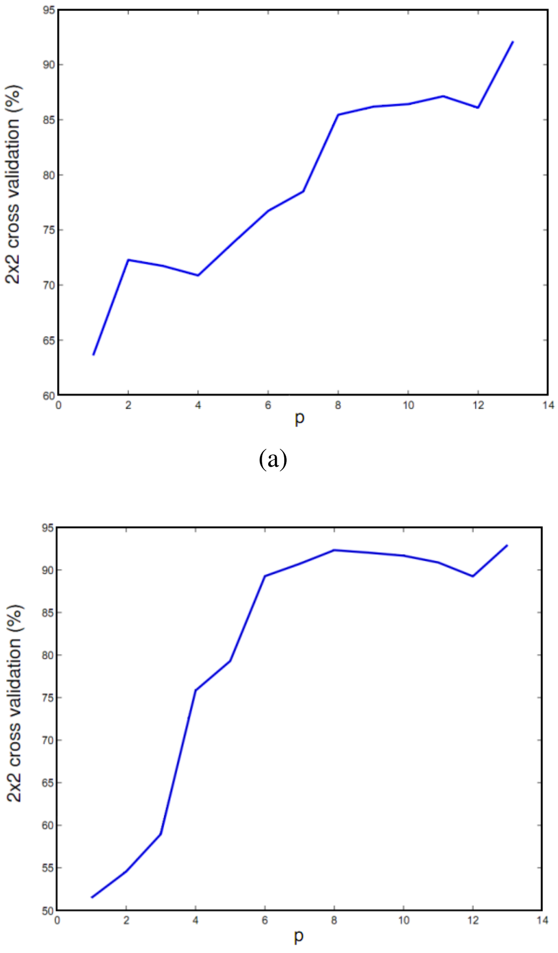
(a)

(b)

Fig. 5. Average classification performance for two different subsets of the UCI wine database ((a) and (b) respectively) using the hybrid model. The horizontal axis represents parameter t in the hybrid model (10) and the vertical axis shows the classification performance under 2×2 cross validation.

In contrast, figure 6 shows the performance of the two-stage model [10]. In such model, training set is reduced to a p-dimensional subspace in the PCA stage, where p ranges from 1 to the number of dimensions in original data. LDA is then applied on the p-dimensional subspace to obtain a new subspace reserving discriminating information. As in the experiment for our hybrid model, 59 samples of type 1 and 71 samples of type 2 are used at first and figure

6(a) shows the average performance over all runs using the two-stage model. 10,000 experiments with $\epsilon = 0.01$ and randomized 2×2 cross validation for a total of 40,000 runs are performed as well. Figure 6(b) shows the results with 71 samples of type 2 and 48 samples of type 3 using the two-stage model correspondingly. As shown in figure 5 and 6, in both cases, our hybrid model is superior to the two-stage model concerning the classification performance. There is an optimal solution between PCA and LDA such that both most expressive information and most discriminating information can be preserved. In contrast, the performance of the 2-stage model may suffer from a large amount of undiscriminating information inside the first few principle components. Therefore, a hybrid model combining PCA and LDA may be better for discriminative dimension reduction rather than incorporating them via a two-stage system.



(a)

(b)

Fig. 6. Average classification performance for two different subsets of the UCI wine database ((a) and (b) respectively) using the two-stage model. The horizontal axis represents parameter p, which is the dimension reserved in PCA stage, and the vertical axis shows the classification performance under 2×2 cross validation.

V. Summary and Discussion

We proposed a dimension reduction method that may be interpreted as a hybrid of principal component analysis and linear discriminant analysis. The main goal is to enhance data discrimination that can be achieved with subspaces learned with either PCA or LDA alone. The learning mechanism differs from existing proposals in that it is guided by a hybrid model and thus addresses the problem of generalization to new data in a more direct way. In addition to the model, we developed computational strategies to estimate optimal subspaces. The method was illustrated with applications to facial classification experiments and the discrimination of different types of wine based on results of chemical analysis of their constituents and properties.

In this paper, we only considered reduction to a 1-dimensional subspace. This certainly limits the discrimination performance on data that exhibit more intricate clustering patterns. In future work, the model and the methodology will be extended to subspaces of higher dimension, including experiments to illustrate the gains that can be achieved with subspaces of dimension greater than 1. Note that after the low dimensional representation is learned, we can use any classifier, including SVM and neural networks for classification and comparison. In addition, the proposed hybrid PCA-LDA model can be considered as a special case of generative-discriminative models for classification. Therefore PCA can be replaced by other generative models and LDA by other discriminative algorithms (such as optimal component analysis [15] and optimal factor analysis [17]), leading to a family of new models.

References

[1] E. Oja, “Neural networks, principle components, and subspaces,” *International Journal of Neural Systems*, vol. 1, pp. 61-68, 1989.

[2] K. Friston, “The free-energy principle: a unified brain theory?,” *Nature Reviews Neuroscience*, vol. 11, pp. 127-138, 2010.

[3] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu “A new LDA-based face recognition system which can solve the small sample size problem,” *Pattern Recognition*, vol. 33, pp. 1713-1726, 2000.

[4] M. Turk and A. Pentland, “Eigenfaces for Recognition,” *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.

[5] P.-N. Belhumeour, J.-P. Hespanha, and D.-J. Kriegman, “Eigenfaces vs. Fisherfaces: recognition using class specific linear projection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[6] P. Howland, M. Jeon, and H. Park, “Structure preserving dimension reduction for clustered text data based on the generalized SVD,” *SIAM Journal on Matrix Analysis and Applications*, vol. 25, no. 1, pp. 165-179, 2003.

[7] R. Duda, P. Hart, and D. Stork, *Pattern Classification* John Wiley and Sons, 2000.

[8] R. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, pp. 179-188, 1936.

[9] J. Friedman, “Regularized discriminant analysis,” *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165-175, 1989.

[10] J. Yang and J.-Y. Yang, “Why can LDA be performed in PCA transformed space?” *Pattern Recognition*, vol. 36, pp. 563-566, 2003.

[11] D.-L. Swets and J. Weng, “Using discriminant eigenfeatures for image retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831-836, 1996.

[12] A.-M. Martinez and A.-C. Kak, “PCA versus LDA,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228-233, 2001.

[13] A. Asuncion and D. Newman, UCI machine learning repository, 2007.

[14] I.T. Jolliffe, *Principal component analysis*, Springer series in statistics. Springer-Verlag, 1986.

[15] X. Liu, A. Srivastava, and K. Gallivan, “Optimal linear representations of images for object recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 662-666, 2004.

[16] F. Samaria and A. Harter, “Parameterisation of a stochastic model for human face identification,” *Proc. of the 2nd IEEE Workshop on Applications of Computer Vision*, pp. 138-142, 1994.

[17] Y. Zhu, W. Mio, and X. Liu, “Optimal factor analysis and applications to content-based image retrieval,” In J. Braz, A. Ranchordas, H. Araujo, and J. Pereira, editors, *Computer Vision and Computer Graphics: Theory and Applications*, vol. 21 of *Communications in Computer and Information Sciences*, pp. 164-176, Springer, 2009.

[18] R.-A. Fisher, “The use of multiple measures in taxonomic problems,” *Annals of Eugenics*, vol. 7, pp. 179-188, 1936.