Robust Learning of Discriminative Projection for Multicategory Classification on the Stiefel Manifold

Duc-Son Pham and Svetha Venkatesh
Dept. of Computing, Curtin University of Technology
GPO Box U1987, Perth, WA 6845, Australia
dspham@ieee.org, svetha@cs.curtin.edu.au

Abstract

Learning a robust projection with a small number of training samples is still a challenging problem in face recognition, especially when the unseen faces have extreme variation in pose, illumination, and facial expression. To address this problem, we propose a framework formulated under statistical learning theory that facilitates robust learning of a discriminative projection. Dimensionality reduction using the projection matrix is combined with a linear classifier in the regularized framework of lasso regression. The projection matrix in conjunction with the classifier parameters are then found by solving an optimization problem over the Stiefel manifold. The experimental results on standard face databases suggest that the proposed method outperforms some recent regularized techniques when the number of training samples is small.

1. Introduction

Face recognition is an active area of research in pattern recognition. Although the dimensionality of the problem is high, the useful information for classification resides in a much lower dimensional manifold. In the face recognition context, techniques for dimensionality reduction, either unsupervised like principal component analysis (PCA) [27] or supervised like linear discriminant analysis (LDA) [2], have been developed. The traditional approach is based on the power of linear algebra to find a linear subspace for dimensionality reduction and the majority of techniques find a set of eigenvectors derived from a certain formulation. For example, it is the maximum variance under PCA or maximum discrimination ratio of the between-class and within-class variances in the reduced subspace under LDA [2]. Recent approaches to dimensionality reduction use a nonlinear manifold [9, 25]. Despite its simplicity in formulation, linear subspace techniques perform very well in practice over a wide range of face databases, with the advantage of relatively low computational cost. This implies that in the commonly used face databases, the assumption of a linear face manifold is still a reasonable one.

With dimensionality reduction, typically the system is divided into two stages wherein dimensionality reduction takes place first and the actual classification follows. Under unsupervised approaches such as PCA, the two stages are completely independent and the projection is only optimal in a sense of, for example, preserving the data variance (i.e., keeping the most interesting directions). In contrast, supervised techniques like LDA find a projection that is more suitable for the classification stage. By projecting the original data into a much smaller dimensional space, one implicit advantage is that the effective noise in the original data can be reduced. The formulation in the second stage leads to a different supervised linear projection. For example, a family of linear projection techniques such as locality preserving projection (LPP) [13], orthogonal Laplacianfaces (OLPP) [4] and Fisherfaces [2, 5] when viewed under a generalized graph embedding framework [29] resort to different choices of the graph configurations. The choices imply either the local structure is preserved when projected or the ratio of between-class and within-class variances is maximized.

However, it is noted that the majority of linear projection techniques are based on an implicit assumption of nearest neighbor classification. This assumption leads to the need to preserve local structure in the low dimensional subspace. The nearest neighbor approach is a simple, yet effective multiclass classifier. However, it might not be the optimal choice when training images do not have clear nearest neighbor characteristics.

In this work, we propose to make the classification stage explicit in the formulation to find the linear projection under a statistical learning theory framework [28]. To do so, we select a family of linear classifiers as the hypothesis space of learning functions and select ℓ_1-norm for regularization, which turns into a multivariate, lasso-based classifier. The objective function is then the ℓ_1-regularized em-

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

pirical risk. As the classifier is designed for the projected subspace, the projection matrix to be learned automatically enters the above regularized formulation. We then apply optimization theory to derive an algorithm that finds this projection matrix and its matching linear classifier over the Stiefel manifold. Doing so gives us two advantages. Firstly, the formulation solves the problem in a rigorous statistical framework. Secondly, the projection matrix found is robust to small numbers of training samples as it inherits good generalization ability when paired with the classifier found under this framework.

When viewed under the information-theoretic perspective, our proposed framework can be regarded as *joint source and channel coding* [8]. We note that the recent work of [16] also employs this concept and the projection matrix is found by maximizing the effective information over the Grassman manifold. However, we believe that the effective information, just as in the case of pure empirical risk, can face the problem of overfitting with small training size, as no regularization is involved. Recently published work [5] clearly indicates that regularization is the key to success when dealing with small training size.

The paper is organized as follows. To facilitate readability, the background in embedded in relevant sections and not separated out. In Section II, we describe our proposed general framework for robust learning of the linear projection. In Section III, we apply the multivariate lasso method and derive an algorithm for learning this linear projection over the Stiefel manifold. Section IV contains experimental results on standard face databases that show the advantages of our proposed method over some recently proposed techniques. Finally, Section V contains concluding remarks.

2. Proposed Learning Framework

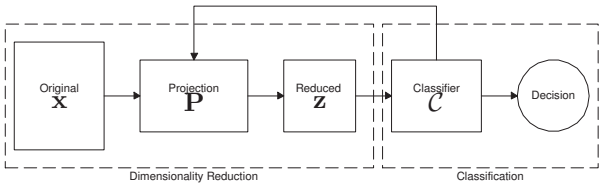

Figure 1. Supervised learning of projection matrix.

Consider the original input data, presented in vector format $\mathbf{x} \in \mathbb{R}^n$. A linear projection $\mathbf{P} \in \mathbb{R}^{n \times k}$ projects the original input data $\mathbf{x}$ to $\mathbf{z} = \mathbf{P}^T\mathbf{x}$. This projection not only reduces the dimension to $k \ll n$ but also suppresses effective noise in the original data before passing $\mathbf{z}$ to the classifier $\mathcal{C}$ (see Fig. 1). In an unsupervised manner, the projection $\mathbf{P}$ can be chosen independently of the classifier $\mathcal{C}$. In a supervised manner, the projection $\mathbf{P}$ is chosen so as to ensure, for example, in the case of nearest neighbor classification, local or global similarity in the original domain $\mathbf{x}$ is preserved in the reduced space $\mathbf{z}$. This requirement often intuitively translates to generalized eigenvalue problems and the projection matrix $\mathbf{P}$ can be solved in a regularized or unregularized manner [4].

However, nearest neighbor is not the only choice for the structure of the classifier $\mathcal{C}$. Following statistical learning theory, in principle we can associate the structure of the classifier with a family of learning functions $f(\mathbf{z})$ in the hypothesis space $\mathcal{H}$. Under the empirical risk minimization principle, one seeks to find the suitable learning function by minimizing the empirical risk:

$$\arg\min_{f \in \mathcal{H}} \; R_{\text{emp}}[f]. \tag{1}$$

When the number of training samples is small, one may wish to avoid overfitting and improve generalization ability by using a regularized version:

$$\arg\min_{f \in \mathcal{H}} \; R_{\text{emp}}[f] + R_{\text{reg}}[f]. \tag{2}$$

Such regularization would favor certain smoothing characteristics of the learning functions in the given hypothesis space [21]. Some form of regularization would have direct meaning, such as maximizing the margin in the case of support vector machines (SVM).

Recall that the reduced space $\mathbf{z}$ is obtained as the projection from $\mathbf{x}$. Hence, in order to find a discriminative projection for the specified structure of the classifier, we propose to solve the following regularized empirical risk problem

$$\arg\min_{\mathbf{P}, f \in \mathcal{H}} \; R_{\text{emp}}[f(\mathbf{P}^T\mathbf{x})] + R_{\text{reg}}[f(\mathbf{P}^T\mathbf{x})] \tag{3}$$

subject to the condition $\mathbf{P}^T\mathbf{P} = \mathbf{I}$ and $\text{rank}(\mathbf{P}) = k$. The set of such orthogonal projection matrices $\mathbf{P}$ is either the Grassmann or Stiefel manifold depending on the objective function [10].

The difficulty of the optimization problem (3) depends on the hypothesis space and the regularization form. Even if the class of functions is easy enough, the global solution might not be available, as it may not be convex in both $\mathbf{P}$ and f. We therefore resort to a sub-optimal solution using a sequential minimization technique, that seeks to alternate between $\mathbf{P}$ and f, and solve for each one at a time, so that at least a local solution can be found. In the next section, we describe our special choice of the loss function and the linear classifier structure.

3. Multivariate Lasso Regression

3.1. Multivariate embedding

In order to cater for multi-category classification using a regression technique, the simplest way is to do is to build a binary classifier for each category against the rest. The other

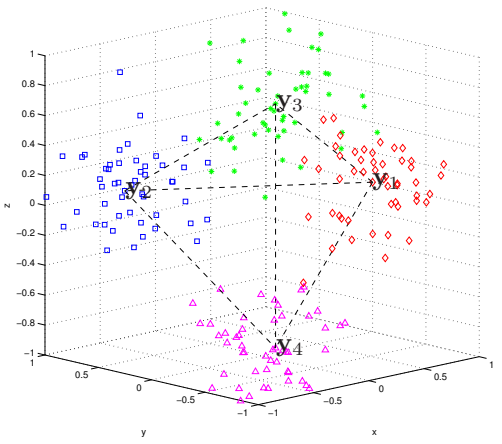

Figure 2. Illustration of the embedding symmetric simplex for the case $k = 4$. The linear function will enforce the mapping of each datapoint from each class to its target point in this embedded space.

approach is to generalize to multivariate regression [24]. Effectively, the class labels are embedded in this multivariate regression space, where each vertex $\mathbf{y}_i = [y_i^1, \ldots, y_i^{k-1}]^T$ is the central or target point of a class. An embedded point closest to $\mathbf{y}_i$ will be classified as belonging to class i (see Fig. 2 for an example with $k = 4$). Suppose the linear regression model is used. If the input data is centralized and the intercept is set to zero, then it is relevant to consider the symmetric simplex as described in [1, 15]. It is also straightforward to see subsequently, the performance is invariant to any permutation of class assignment to the vertices and any rotation of the simplex.

3.2. Lasso

The lasso regression technique was introduced by Tibshirani [26] as a robust alternative to many existing methods such as subset selection and ridge regression. Consider a linear regression model

$$y_i = \mathbf{w}^T \mathbf{z}_i + \beta_0, \;\; i = 1, \ldots, L \tag{4}$$

the lasso finds the regression parameters $\mathbf{w}$ via

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{L} (y_i - \mathbf{w}^T \mathbf{z}_i - \beta_0)^2, \quad \text{subject to } \| \mathbf{w} \|_1 \leq s. \tag{5}$$

It can be shown that this problem is equivalent to

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{L} (y_i - \mathbf{w}^T \mathbf{z}_i - \beta_0)^2 + \lambda \| \mathbf{w} \|_1, \tag{6}$$

for a suitable value of λ, which is often found by cross-validation. As can be seen, the difference to ridge regression is in the choice of ℓ_1 norm regularization instead of ℓ_2. This is known to generate *sparse* coefficients. For related discussion, please see [12, 26]. We just note the important point that the lasso is considerably favorable over other choices such as subset selection, ridge regression, and Breiman's garrote, for a number of statistical problems with small number of observations [26]. As we specifically address the issue of small training size, the choice of lasso is justified.

When extending the lasso to the multivariate case with $\mathbf{y} \in \mathbb{R}^{k-1}$, one considers the linear regression model

$$\mathbf{y}_i = \mathbf{W}^T \mathbf{z}_i + \mathbf{b}, \;\; i = 1, \ldots, L, \tag{7}$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{k-1}]$. When the embedding of $\mathbf{y}_i$ is symmetric and the input is centralized, we can set the intercept $\mathbf{b} = \mathbf{0}$. The optimization (6) can be written in matrix form as follows

$$\arg\min_{\mathbf{W}} \| \mathbf{Y} - \mathbf{W}^T \mathbf{Z} \|_F^2 + \lambda \| \mathbf{W} \|_1, \tag{8}$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_L]$ and $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_L]$. When both $\mathbf{Z}$ and $\mathbf{Y}$ are known, the lasso regression problem can be solved numerically though an explicit solution is not available. This is also known as ℓ_1-norm regularized least-squares problem. Techniques for solving ℓ_1-norm regularized least-squares such as basis pursuit [17], matching pursuit [7] etc. are readily available. We found that the latest `l1_ls` package from Stanford [14] performs well and is scalable to large problems with millions of variables. Other ℓ_1 solvers include [6, 7, 17, 23].

4. Optimization on the Stiefel Manifold

When the loss function is quadratic and the ℓ_1-norm is used for regularization, our proposed method is essentially the lasso. We are motivated from the desired statistical properties of the lasso as a regularized least squares method, particularly suitable for small training sizes. In the view that $\mathbf{Z} = \mathbf{P}^T \mathbf{X}$ we propose to jointly find the linear projection $\mathbf{P}$ and select the suitable linear classifier from the following optimization problem

$$\arg \min_{\mathbf{P} \in \mathcal{S}_{n,k}, \mathbf{W} \in \mathbb{R}^{k \times m}} \| \mathbf{Y} - \mathbf{W}^T \underbrace{\mathbf{P}^T \mathbf{X}}_{\mathbf{Z}} \|_F^2 + \lambda \| \mathbf{W} \|_1 . \tag{9}$$

where the set

$$\mathcal{S}_{n,k} = \{ \mathbf{P} \in \mathbb{R}^{n \times k} : \mathbf{P}^T \mathbf{P} = \mathbf{I} \} \tag{10}$$

is the real-valued Stiefel manifold. The optimization is solved over the Stiefel manifold $\mathcal{S}_{n,k}$ for $\mathbf{P}$ because the objective function is not invariant to a right orthogonal transformation on the rank-k projection matrix $\mathbf{P}$. Finding the

suitable $f \in \mathcal{H}$ is equivalent to finding $\mathbf{W}$. As mentioned earlier, a suboptimal solution is to alternate between solving for $\mathbf{P}$ and for $\mathbf{W}$. The objective function in (9) is convex with respect to either $\mathbf{W}$ or $\mathbf{P}$, but not for both. A viable suboptimal solution is then to alternate between solving for each variable until convergence is found. When $\mathbf{P}$ is held fixed, the numerical solution for $\mathbf{W}$ can be readily obtained using a ℓ_1-solver of the following problem:

$$\arg \min_{\mathbf{W} \in \mathbb{R}^{k \times m}} \| \mathbf{Y} - \mathbf{W}^T \mathbf{Z} \|_F^2 + \lambda \| \mathbf{W} \|_1 . \qquad (11)$$

When $\mathbf{W}$ is kept fixed, we need to solve the following convex optimization problem

$$\arg \min_{\mathbf{P} \in \mathcal{S}_{n,k}} g(\mathbf{P}) = \| \mathbf{Y} - \mathbf{W}^T \mathbf{P}^T \mathbf{X} \|_F^2 . \qquad (12)$$

Techniques for solving this type of optimization problem often endow the manifold with a Riemannian structure and extend classical optimization techniques. The well-known work of Edelman, Arias and Smith [10] chooses to move along the geodesics. However, this often results in higher computational cost to compute the path of a geodesic. In [18], a simpler strategy is proposed which leads to a moderate reduction in computational cost. This strategy can be used for both steepest descent and Newton methods. In what follows, we show how to apply this strategy to solve our optimization problem. In particular, we choose the steepest descent method for simplicity. For notational clarity, we start with the preliminary result in [18].

Proposition 1 *(Projection on the Stiefel manifold.) Let $\mathbf{X} \in \mathbb{R}^{n \times k}$ be a rank-k matrix. Define the projection operator $\pi : \mathbb{R}^{n \times k} \mapsto \mathcal{S}_{n,k}$ as*

$$\pi(\mathbf{X}) = \arg \min_{\mathbf{Q} \in \mathcal{S}_{n,k}} \| \mathbf{X} - \mathbf{Q} \|_F^2 . \qquad (13)$$

If the singular value decomposition of $\mathbf{X}$ is $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ then

$$\pi(\mathbf{X}) = \mathbf{U}\mathbf{I}_{n,k}\mathbf{V}^T \qquad (14)$$

where $\mathbf{I}_{n,k}$ is the first k columns of the identity matrix $\mathbf{I} \in \mathbb{R}^{n \times n}$.

Using this notation, the application of the steepest descent method is detailed as follows.

- **Initialization:** Initialize $\mathbf{P} = \mathbf{P}_0$ (from PCA, for example).
- **Iteration** i.

$$\mathbf{F} = \frac{\partial g(\mathbf{P})}{\partial \mathbf{P}} = 2(\mathbf{X}\mathbf{X}^T\mathbf{P}\mathbf{W}\mathbf{W}^T - \mathbf{X}\mathbf{Y}^T\mathbf{W}^T)$$

  - Step 1: Set $\mathbf{G} = \mathbf{F} - \mathbf{P}\mathbf{F}^T\mathbf{P}$, $\mathbf{H} = -\mathbf{G}$.
  - Step 2: Compute

$$\langle \mathbf{H}, \mathbf{H} \rangle = \mathrm{tr}[\mathbf{H}^T(\mathbf{I} - (1/2)\mathbf{P}\mathbf{P}^T)\mathbf{H}]$$

  - Step 3: Check if $\langle \mathbf{H}, \mathbf{H} \rangle \leq \varepsilon$ then stop
  - Step 4: If $g(\mathbf{P}) - g(\pi(\mathbf{P} + 2\gamma\mathbf{H})) \geq \gamma\langle \mathbf{H}, \mathbf{H} \rangle$ set $\gamma \Rightarrow 2\gamma$ then repeat Step 4
  - Step 5: If $g(\mathbf{P}) - g(\pi(\mathbf{P} + \gamma\mathbf{H})) \leq (1/2)\gamma\langle \mathbf{H}, \mathbf{H} \rangle$ set $\gamma \Rightarrow (1/2)\gamma$, then repeat Step 5.
  - Step 6: Set $\mathbf{P} \Rightarrow \pi(\mathbf{P} + \gamma\mathbf{H})$, where π is the projection onto the Stiefel manifold. If the objective function in (12) has not converged, go back to Step 1.

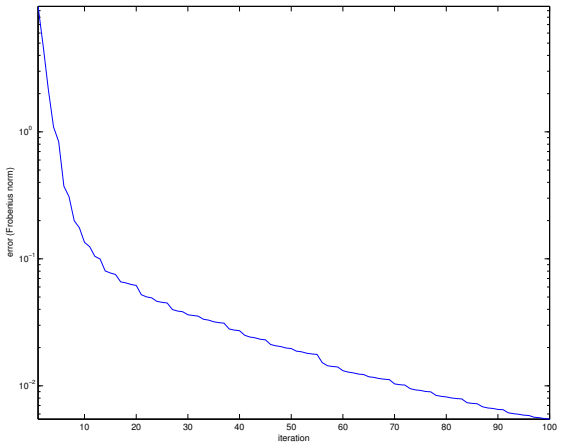

Figure 3. Typical convergence of the steepest descent algorithm on the Stiefel manifold

It is noted that though it is possible to use modified Newton method on the Stiefel manifold, we found in practice that the steepest descent is sufficient for a moderate accuracy, with the advantage of being simple and fast. A typical convergence of the quadratic objective function is shown in Fig. 3. When being used with the lasso to solve the optimization (9), we found that the solution is found with a reasonable accuracy within a few iterations (see Fig. 4).

5. Experimental Results

In this section, we compare the performance of our proposed method to other robust counterparts, including the regularized and smooth linear discriminant analysis (LDA) and local preserving projection (LPP). These methods have been demonstrated to significantly outperform other techniques when the number of training samples is small [5]. We also include the performance of the baseline PCA method and the orthogonal Laplacianfaces [4]. We note

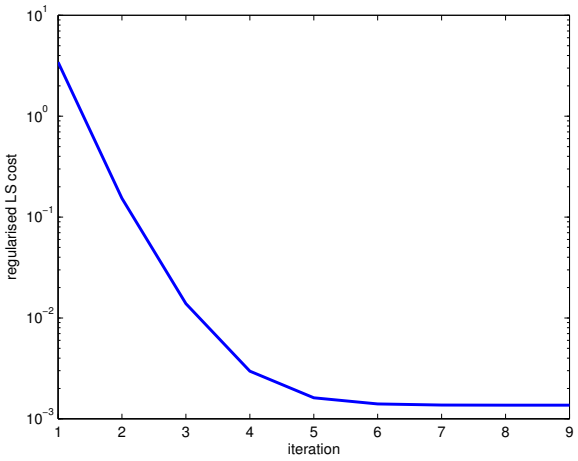

Figure 4. Convergence of the alternating optimization algorithm

that all these methods use nearest neighbor for classification. To facilitate a fair comparison and to illustrate the advantage of the joint dimensionality reduction and classification framework, we also include the method that use PCA as dimensionality reduction and lasso as classification, but in a disjoint manner. We shall denote this method as PCA+LASSO. Our proposed method will be denoted as MLASSO. We are only interested in the classification problem in this experiment.

The datasets used in this experiment are well-known datasets in the face recognition community. The original PIE database from Carnegie Mellon University [22] consists of 68 individuals with 41,368 images. This experiment only uses a near-frontal subset of the PIE database that we downloaded from [3] [1]. In this subset, there are approximately 170 images per individual over about five different poses. We also select the Yale B face database [11] with 38 individuals, each having a total of 64 near-frontal images, which is downloaded from [3]. They are known to be difficult datasets in face recognition. With these two datasets, we generate 20 random splits. In each split, the images are randomly selected from each class for training, and the rest is used for testing. Then, we report the average and standard deviation of the measured error rate.

In all cases, the pre-processing step involves cropping and resizing the faces to 32×32 gray-scale images, then centralizing about the mean, and finally normalizing each vector to unit norm[2]. As we are only interested in small training sizes, we select 2, 3, and 4 images for training. The parameters for the regularized S-LDA and S-LPP are the suggested values from the authors (in particular the choice of weighting matrix and the regularization $\alpha = 0.1$). To make it comparable to S-LDA, we set the projection onto a subspace with the dimension being the number of classes minus one. The reported PCA method is also based on the assumption of the same dimension.

The results on these two databases are reported in Tables 1 and 2. Among the compared methods, we note that S-LDA performs well and especially when the training sizes get larger. For example, on the PIE database its performance gap with OLPP increases from 1% to 3% when the training size increases from 2 to 4. They also outperform the baseline PCA method in both cases. Our proposed method (MLASSO) outperforms the best method by 7%. This increase in performance gain illustrates the clear advantage in terms of robustness. Finally, when we compare our proposed method MLASSO to PCA+LASSO, significant performance gain is observed. For example, MLASSO has 4% to 9% lower error rates compared with PCA+LASSO when the training samples increase from 2 to 4 on the PIE database, whilst that figure is 7% to 13% on the extended Yale B database. This is as expected because separating PCA and LASSO implies discriminative information being lost during projection. The observation that the performance gap with PCA+LASSO increases with the number of training samples suggests that the angle between the optimal subspace of PCA and the discriminative subspace gets larger, hence separating dimensionality reduction from classification just makes the classification worse. We note that whilst linear techniques like PCA and S-LDA are generally very fast, our proposed method takes longer time to run as it needs to solve the optimization problem on the Stiefel manifold and lasso regression. However, the increase in performance justifies the choice.

Table 1. Error rate on PIE database

Train	2	3	4
PCA	0.86±0.01	0.83±0.01	0.80±0.01
S-LDA	0.60±0.02	0.48±0.02	0.40±0.02
S-LPP	0.66±0.01	0.59±0.01	0.52±0.02
OLPP	0.61±0.03	0.50±0.03	0.43±0.02
PCA+LASSO	0.58±0.01	0.49±0.01	0.43±0.01
MLASSO	0.54±0.01	0.43±0.01	0.34±0.01

The results above are measured over random splits of the images for small-size datasets. We now show that when extended to *a much larger set* and standard benchmarks for performance measure, our method still has a competitive advantage over compared methods. We select the Feret

[1] It appears that there are different ways of pre-processing the images from the original databases. As a consequence, the absolute accuracy rate might be reported higher if different pre-processing is used. We however found that the pre-processing of [3] is very challenging for recognition techniques and use in this experiment.

[2] We note that some performance on some datasets might be slightly improved with different pre-processing techniques but the relative performance between different methods should stay approximately the same.

Table 2. Error rate on Yale B database

Train	2	3	4
PCA	0.84±0.02	0.80±0.01	0.77±0.01
S-LDA	0.60±0.03	0.48±0.02	0.40±0.02
S-LPP	0.64±0.03	0.56±0.03	0.50±0.03
OLPP	0.57±0.03	0.45±0.02	0.39±0.03
PCA+LASSO	0.58±0.02	0.54±0.02	0.50±0.02
MLASSO	0.51±0.03	0.40±0.02	0.33±0.02

database [20, 19] for this purpose[3]. The original Feret database has a total of 1,199 individuals, taken at different times and wide range of variations [19]. The galleries fa, fb consist of near-frontal images of more than 200 individuals and they are used for training. The database also contains standard testing sets for evaluation purpose namely dup1 and dup2 (duplicate probes) which were typically taken on different days. These sets can be used for classification and verification. We select the majority of dup1 for this classification. This results in a set of 250 individuals with 499 images for training and 736 images for testing. The preprocessing is the same as the previous experiment. The result is reported in Table 3. Once again, it can be seen that over this standard test set, our proposed method clearly outperforms the best of the compared methods, which in this case is S-LDA, by a margin of as much as 7%[4]. When compared with the disjoint framework PCA+LASSO, our method also yields a better error rate by 2%. This clearly demonstrates that our method retains a strong advantage when there is a large number of classes.

Table 3. Error rate on Feret database

Train	2
PCA	0.58
S-LDA	0.50
S-LPP	0.51
OLPP	0.65
PCA+LASSO	0.45
MLASSO	0.43

6. Discussion

In this work, we have chosen the quadratic loss over the class of linear learning functions and with ℓ_1-norm regularization in a statistical learning framework, which results in a lasso-based classifier. This loss function puts a high penalty for embedded points far from the embedded vertices of the classes. This is particularly satisfactory for a regression problem. But if classification is the ultimate result, improvement could be made by making the risk more closely related to the classification error. In other words, it is possible to use our framework with other type of loss functions and regularization, rather than the lasso. For example, one can incorporate the idea of learning the projection matrix in the context of SVM formulation so that the risk is directly related to the empirical risk. Of course, the choice should be sensible so that such an optimization on the Stiefel or Grassman manifold is tractable. The classification performance gain of our particular choice of the lasso over other robust alternatives suggests that further improvement is feasible.

In terms of implementation, our method requires several specifications. The initialization for the projection matrix $\mathbf{P}$ can be obtained from PCA or LDA and takes the better of the two in terms of the final objective function. For the regularization parameter λ, it is possible to use cross-validation to optimize it within the range between 0.01 and 0.2. Within this range, the variation of performance is only few percent which implies that model selection is not a critical problem with our method.

7. Conclusion

We have presented a new approach for learning the linear projection onto the face manifold. The idea of this approach is to form an optimization problem in a regularized learning framework that involves both dimensionality reduction and classification. For the choice of multivariate lasso regression, the projection matrix is found over the Stiefel manifold jointly with the linear classifier's parameters. We also derive an algorithm to achieve a local solution of the formulated optimization problem. This results in a method that is robust to few training samples and this has been demonstrated through experiments on several well-known and publicly available face datasets. The marked performance gain over recently published methods clearly supports the approach outlined in this work. ~

References

[1] S. An, W. Liu, and S. Venkatesh. Face recognition using kernel ridge regression. In *Proceedings CVPR*, 2007.

[2] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[3] D. Cai. Codes and datasets for subspace learning. Available online at www.cs.uiuc.edu/homes/dengcai2/Data/data.html, As retrieved on October 2007.

[4] D. Cai, X. He, J. Han, and H. J. Zhang. Orthogonal Laplacianfaces for Face Recognition. *IEEE Transactions on Image Processing*, 5(11):3608–3614, 2006.

[3] The Feret database is publicly available free of charge and is the standard of commercial face recognition testing platform.

[4] Please note that with this standard testing set, there is no random split hence the standard deviation is 0

[5] D. Cai, X. He, Y. Hu, J. Han, and T. Huang. Learning a spatially smooth subspace for face recognition. In *Proceedings CVPR*, 2007.

[6] E. Candes. L1-magic: Recovery of sparse signals. *www.acm.caltech.edu/l1magic/*.

[7] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

[8] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley, 2nd edition, 2006.

[9] D. L. Donoho and C. Grimes. Image manifolds which are isometric to Euclidean space. *Journal of Mathematical Imaging and Vision*, 23(1):5–24, 2005.

[10] A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal of Matrix Analysis and Applications*, 20(2):303–353, Oct 1998.

[11] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2001.

[13] X. He and P. Niyogi. Locality preserving projection. In *Proceedings NIPS*, 2003.

[14] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. A method for large-scale ℓ_1-regularized least squares. *IEEE Journal on Selected Topics in Signal Processing*, to appear. (Software package available at www.stanford.edu/~boyd/l1_ls/).

[15] F. Lazebnik. On regular simplex in $\mathbb{R}^n$. Available online at www.math.udel.edu/ lazebnik/Info/teaching.html, as retrieved in October 2007.

[16] D. Lin, S. Yan, and X. Tang. Pursuing informative projection on Grassman manifold. In *Proceedings CVPR*, 2006.

[17] S. G. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, Dec 1993.

[18] J. H. Manton. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50(3):4311–4322, Mar 2002.

[19] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[20] P. Phillips, H. Wechsler, J. Huang, and P. J.Rauss. FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[21] B. Schölkopf and Smola. *Learning with kernels*. MIT Press, 2002.

[22] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) databse. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003.

[23] SparseLab. Seeking sparse solution to linear systems of equations. Available online at sparselab.stanford.edu, as retrieved in October 2007.

[24] M. V. Srivastava. *Methods of multivariate statistics*. Wiley-Interscience, 2002.

[25] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, Dec 2000.

[26] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.

[27] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proceedings CVPR*, pages 586–591, 1991.

[28] V. Vapnik. *Statistical learning theory*. Wiley, 1998.

[29] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: a general framework for dimensionalityreduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, Jan 2007.