Chapter 3
Face Subspace Learning

Wei Bian and Dacheng Tao

3.1 Introduction

The last few decades have witnessed a great success of subspace learning for face recognition. From principal component analysis (PCA) [43] and Fisher's linear discriminant analysis [1], a dozen of dimension reduction algorithms have been developed to select effective subspaces for the representation and discrimination of face images [17, 21, 45, 46, 51]. It has demonstrated that human faces, although usually represented by thousands of pixels encoded in high-dimensional arrays, they are intrinsically embedded in a vary low dimensional subspace [37]. The using of subspace for face representation helps to reduce "the curse of dimensionality" in subsequent classification, and suppress variations of lighting conditions and facial expressions. In this chapter, we first briefly review conventional dimension reduction algorithms and then present the trend of recent dimension reduction algorithms for face recognition.

The earliest subspace method for face recognition is Eigenface [43], which uses PCA [23] to select the most representative subspace for representing a set of face images. It extracts the principal eigenspace associated with a set of training face images. Mathematically, PCA maximizes the variance in the projected subspace for a given dimensionality, decorrelates the training face images in the projected subspace, and maximizes the mutual information between appearance (training face images) and identity (the corresponding labels) by assuming that face images are Gaussian distributed. Thus, it has been successfully applied for face recognition. By projecting face images onto the subspace spanned by Eigenface, classifiers can be used in the subspace for recognition. One main limitation of Eigenface is that the

W. Bian (✉) · D. Tao
Centre for Quantum Computation & Intelligence Systems, FEIT, University of Technology, Sydney, NSW 2007, Australia
e-mail: wei.bian@student.uts.edu.au

D. Tao
e-mail: dacheng.tao@uts.edu.au

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_3, © Springer-Verlag London Limited 2011

class labels of face images cannot be explored in the process of learning the projection matrix for dimension reduction. Another representative subspace method for face recognition is Fisherface [1]. In contrast to Eigenface, Fisherface finds class specific linear subspace. The dimension reduction algorithm used in Fisherface is Fisher's linear discriminant analysis (FLDA), which simultaneously maximizes the between-class scatter and minimizes the within-class scatter of the face data. FLDA finds in the feature space a low dimensional subspace where the different classes of samples remain well separated after projection to this subspace. If classes are sampled from Gaussian distributions, all with identical covariance matrices, then FLDA maximizes the mean value of the KL divergences between different classes. In general, Fisherface outperforms Eigenface due to the utilized discriminative information.

Although FLDA shows promising performance on face recognition, it has the following major limitations. FLDA discards the discriminative information preserved in covariance matrices of different classes. FLDA models each class by a single Gaussian distribution, so it cannot find a proper projection for subsequent classification when samples are sampled from complex distributions, for example, mixtures of Gaussians. In face recognition, face images are generally captured with different expressions or poses, under different lighting conditions and at different resolution, so it is more proper to assume face images from one person are mixtures of Gaussians. FLDA tends to merge classes which are close together in the original feature space. Furthermore, when the size of the training set is smaller than the dimension of the feature space, FLDA has the undersampled problem.

To solve the aforementioned problems in FLDA, a dozen of variants have been developed in recent years. Especially, the well-known undersample problem of FLDA has received intensive attention. Representative algorithms include the optimization criterion for generalized discriminant analysis [44], the unified subspace selection framework [44] and the two stage approach via QR decomposition [52]. Another important issue is that FLDA meets the class separation problem [39]. That is because FLDA puts equal weights on all class pairs, although intuitively close class pairs should contribute more to the recognition error [39]. To reduce this problem, Lotlikar and Kothari [30] developed the fractional-step FLDA (FS-FLDA) by introducing a weighting function. Loog et al. [28] developed another weighting method for FLDA, namely the approximate pairwise accuracy criterion (aPAC). The advantage of aPAC is that the projection matrix can be obtained by the eigenvalue decomposition. Both methods use weighting schemes to select a subspace that better separates close class pairs. Recently, the general mean [39] (including geometric mean [39] and harmonic mean [3]) base subspace selection and the max-min distance analysis (MMDA) [5] have been proposed to adaptively choose the weights.

Manifold learning is a new technique for reducing the dimensionality in face recognition and has received considerable attentions in recent years. That is because face images lie in a low-dimensional manifold. A large number of algorithms have been proposed to approximate the intrinsic manifold structure of a set of face images, such as locally linear embedding (LLE) [34], ISOMAP [40], Laplacian eigenmaps (LE) [2], Hessian eigenmaps (HLLE) [11], Generative Topographic Mapping

(GTM) [6] and local tangent space alignment (LTSA) [53]. LLE uses linear coefficients, which reconstruct a given measurement by its neighbors, to represent the local geometry, and then seeks a low-dimensional embedding, in which these coefficients are still suitable for reconstruction. ISOMAP preserves global geodesic distances of all pairs of measurements. LE preserves proximity relationships by manipulations on an undirected weighted graph, which indicates neighbor relations of pairwise measurements. LTSA exploits the local tangent information as a representation of the local geometry and this local tangent information is then aligned to provide a global coordinate. Hessian Eigenmaps (HLLE) obtains the final low-dimensional representations by applying eigen-analysis to a matrix which is built by estimating the Hessian over neighborhood. All these algorithms have the out of sample problem and thus a dozen of linearizations have been proposed, for example, locality preserving projections (LPP) [20] and discriminative locality alignment (DLA) [55]. Recently, we provide a systematic framework, that is, patch alignment [55], for understanding the common properties and intrinsic difference in different algorithms including their linearizations. In particular, this framework reveals that: i) algorithms are intrinsically different in the patch optimization stage; and ii) all algorithms share an almost-identical whole alignment stage. Another unified view of popular manifold learning algorithms is the graph embedding framework [48]. It is shown that manifold learning algorithms are more effective than conventional dimension reduction algorithms, for example, PCA and FLDA, in exploiting local geometry information.

In contrast to conventional dimension reduction algorithms that obtain a low dimensional subspace with each basis being a linear combination of all the original high dimensional features, sparse dimension reduction algorithms [9, 24, 59] select bases composed by only a small number of features of the high dimensional space. The sparse subspace is more interpretable both psychologically and physiologically. One popular sparse dimension reduction algorithm is sparse PCA, which generalizes the standard PCA by imposing sparsity constraint on the basis of the low dimensional subspace. The Manifold elastic net (MEN) [56] proposed recently is another sparse dimension reduction algorithm. It obtains a sparse projection matrix by imposing the elastic net penalty (i.e., the combination of the lasso penalty and the L_2-norm penalty) over the loss (i.e., the criterion) of a discriminative manifold learning, and formulates the problem as lasso which can be efficiently solved. In sum, sparse learning has many advantages, because (1) sparsity can make the data more succinct and simpler, so the calculation of the low dimensional representation and the subsequent recognition becomes more efficient. Parsimony is especially important for large scale face recognition systems; (2) sparsity can control the weights of original variables and decrease the variance brought by possible over-fitting with the least increment of the bias. Therefore, the learn model can generalize better and obtain high recognition rate for distorted face images; and (3) sparsity provides a good interpretation of a model, thus reveals an explicit relationship between the objective of the model and the given variables. This is important for understanding face recognition.

One fundamental assumption in face recognition, including dimension reduction, is that the training and test samples are independent and identically distributed (i.i.d.) [22, 31, 38]. It is, however, very possible that this assumption does not hold, for example, the training and test face images are captured under different expressions, postures or lighting conditions, letting alone test subjects do not even appear in the training set [38]. Transfer learning has emerged as a new learning scheme to deal with such problem. By properly utilizing the knowledge obtained from the auxiliary domain task (training samples), it is possible to boost the performance on the target domain task (test samples). The idea of cross domain knowledge transfer was also introduced to subspace learning [31, 38]. It has shown that by using transfer subspace learning, the recognition performance on the cases where the face images in training and test sets are not identically distributed can be significantly improved compared with comparison against conventional subspace learning algorithms.

The rest of this chapter presents three groups of dimension reduction algorithms for face recognition. Specifically, Sect. 3.2 presents the general mean criterion and the max-min distance analysis (MMDA). Section 3.3 is dedicated to manifold learning algorithms, including the discriminative locality alignment (DLA) and manifold elastic net (MEN). The transfer subspace learning framework is presented in Sect. 3.4. In all of these sections, we first present principles of algorithms and then show thorough empirical studies.

3.2 Subspace Learning—A Global Perspective

Fisher's linear discriminant analysis (FLDA) is one of the most well-known methods for linear subspace selection, and has shown great value in subspace based face recognition. Being developed by Fisher [14] for binary-class classification and then generalized by Rao [33] for multiple-class tasks, FLDA utilizes the ratio of the between-class to within-class scatter as a definition of discrimination. It can be verified that under the homoscedastic Gaussian assumption, FLDA is Bayes optimal [18] in selecting a $c-1$ dimensional subspace, wherein c is the class number. Suppose there are c classes, represented by homoscedastic Gaussians $N(\mu_i, \Sigma \mid \omega_i)$ with the prior probability p_i, $1 \le i \le c$, where μ_i is the mean of class ω_i and Σ is the common covariance. The Fisher's criterion is given by [15]

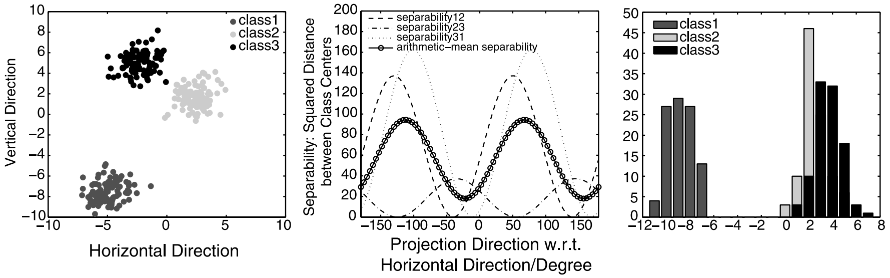
$$\max_W \operatorname{tr}\bigl(\bigl(W^{\mathrm{T}} \Sigma W\bigr)^{-1} W^{\mathrm{T}} S_b W\bigr) \tag{3.1}$$

where

$$S_b = \sum_{i=1}^{c} p_i (\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}}, \quad \text{with } \mu = \sum_{i=1}^{c} p_i \mu_i. \tag{3.2}$$

It has been pointed out that the Fisher's criterion implies the maximization of the arithmetic mean of the pairwise distances between classes in the subspace. To see this, let us first define the distance between classes ω_i and ω_j in the subspace W as

$$\Delta(\omega_i, \omega_i \mid W) = \operatorname{tr}\bigl(\bigl(W^{\mathrm{T}} \Sigma W\bigr)^{-1} W^{\mathrm{T}} D_{ij} W\bigr), \quad \text{with } D_{ij} = (\mu_i - \mu_j)(\mu_i - \mu_j)^{\mathrm{T}}. \tag{3.3}$$



(a) (b) (c)

Fig. 3.1 An illustrative example on the class separation problem of FLDA. **a** 2-dimensional scatter plot of three classes, **b** plots of pairwise separabilities and the arithmetic mean (FLDA) separability verse projection directions, from −180 degree to 180 degree with respect to horizontal direction in (**a**), and **c** shows the histogram of three classes projected onto the FLDA direction, which is around 66 degree

Then, simple algebra shows that (3.1) is equivalent to the arithmetic mean criterion below

$$\max_W A(W) = \sum_{1 \le i < j \le c} p_i p_j \Delta(\omega_i, \omega_j \mid W). \tag{3.4}$$

We call it arithmetic mean based subspace selection (AMSS). Since the arithmetic mean of all pairwise distance is used as the criterion, one apparent disadvantage of (3.4) is that it ignores the major contributions of close class pairs to classification error and may cause the merge of those class pairs in the selected subspace. Such phenomenon of FLDA or AMSS is called the class separation problem [39].

Figure 3.1 illustrates the class separation problem of FLDA [5]. In the toy example, three class are represented by homoscedastic Gaussian distributions on the two dimensional space. And we want to find a one dimensional subspace (or projection direction) such that the three classes can be well separated. Varying the one dimensional subspace, that is, changing the angle of projection direction with respect to the horizontal direction, the three pairwise distances change. FLDA finds the subspace that maximizes the average of the three pairwise distances. However, as illustrated, the obtained one dimensional subspace by FLDA severely merges the blue and green classes.

3.2.1 General Mean Criteria

To improve the separation between close class pairs, the general mean criteria has been proposed by Tao et al., of which two examples are the geometric mean based subspace selection (GMSS) [39] and the harmonic mean based subspace selection

(HMSS) [3]

$$\max_W G(W) = \prod_{1\le i<j\le c} \Delta(\omega_i, \omega_j \mid W)^{(p_i p_j)} \quad \text{(GMSS)} \tag{3.5}$$

and

$$\max_W H(W) = \left[\sum_{1\le i<j\le c} \frac{p_i p_j}{\Delta(\omega_i, \omega_j \mid W)} \right]^{-1} \quad \text{(HMSS)}. \tag{3.6}$$

We give an mathematical analysis to interpret how criteria (3.5) and (3.6) work in dealing with the class separation problem, and why criterion (3.6) is even better than criterion (3.5). Consider a general criterion below

$$\max_W J(W) = f\big(\Delta(\omega_1, \omega_2 \mid W), \Delta(\omega_1, \omega_3 \mid W), \dots, \Delta(\omega_{c-1}, \omega_c \mid W)\big). \tag{3.7}$$

In order to reduce the class separation problem, the objective $J(W)$ must has the ability to balance all the pairwise distances. We claim that this ability relies on the partial derivative of $J(W)$ with respect to the pairwise distances. Apparently, an increment of any $\Delta(\omega_i, \omega_j|W)$ will enlarge $J(W)$, and for this an small one should have bigger inference, because from the classification point of view when the distance between two classes is small then any increment of the distance will significantly improve the classification accuracy, but when the distance is large enough then the improvement of accuracy will be ignorable (it is well known that for Gaussian distribution the probability out the range of $\pm 3\sigma$ is less than 0.01%). Besides, the partial derivatives must vary as the varying of the pairwise distances so as to take account of the current values of the pairwise distances in the procedure of subspace selection, but not only the initial distances in the original high dimensional space. According to the discussion above, the partial derivatives must be monotone decreasing functions of $\Delta(\omega_i, \omega_j \mid W)$. In the cases of criteria (3.4) and (3.5), we set $J(W) = \log G(W)$ and $J(W) = -H^{-1}(W)$, and then the derivatives are calculated as below

$$\frac{\partial \log G(W)}{\partial \Delta(\omega_i, \omega_j \mid W)} = \frac{q_i q_j}{(\Delta(\omega_i, \omega_j \mid W))^{-1}} \tag{3.8}$$

and

$$\frac{\partial - H^{-1}(W)}{\partial \Delta(\omega_i, \omega_j \mid W)} = \frac{q_i q_j}{(\Delta(\omega_i, \omega_j|W))^{-2}}. \tag{3.9}$$

We can see that in both cases the partial derivative monotonically decreases with respect to the pairwise distance and thus provides the ability to reduce the class separation problem. However, note that the order of decreasing for HMSS is higher than that for GMSS (−2 vs −1), which implies that HMSS is more powerful than GMSS in reducing the class separation problem. Besides, as $\Delta(\omega_i, \omega_j \mid W)$ increases, we have

$$\log\big(\Delta(\omega_i, \omega_j \mid W)\big) \to \infty \tag{3.10}$$

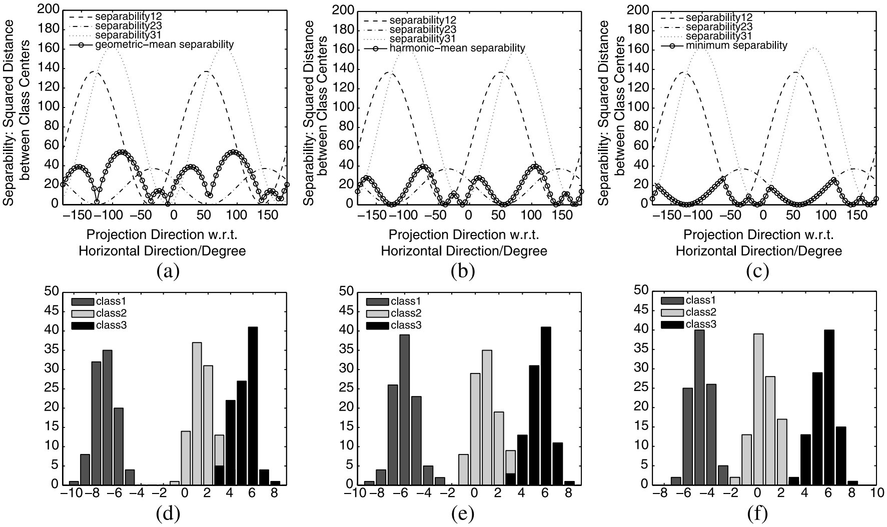


Fig. 3.2 GMSS, HMSS and MMDA for the same three-class problem in Fig. 3.1: *first column*, GMSS; *second column*, HMSS; *third column*, MMDA. *Top row* shows plots of pairwise separations and the separations by different criteria, i.e., GMSS, HMSS and MMDA. *Bottom row* shows the histograms of the three classes projected onto the GMSS, HMSS and MMDA directions, which are around 93 degree, 104 degree and 115 degree, respectively

but

$$-\Delta(\omega_i, \omega_j \mid W)^{-1} \to 0. \tag{3.11}$$

The logarithm value (3.10) is unbounded, and thus in GMSS a large pairwise distance still possibly affects small ones. In contrast, the bounded result (3.11) makes HMSS is more favorable. To solve the maximization problems of (3.5) and (3.6), [39] provides a gradient descent algorithm with a projection onto the orthogonal constraint set. Further, [3] suggests exploiting the structure of orthogonal constraint and optimizing the subspace on the Grassmann manifold [12]. For details of these optimization algorithms, please refer to [39] and [3]. The corresponding results of GMMS and HMSS on the illustrative example in Fig. 3.1 are shown in Fig. 3.2. One can see that the merged class pair in the FLDA subspace is better separated by using the more sophisticated methods.

3.2.2 Max–Min Distance Analysis

Previous discussions show that GMSS and HMSS are able to reduce the class separation problem of FLDA. Such merits come from the inherence of geometric or harmonic means in adaptively emphasizing small pairwise distance between classes. A further question is: can we select a subspace that mostly considers small pairwise distance? Namely, we may intend to find an optimal subspace which gives the maximized minimum pairwise distance. Generally, such aim cannot be achieved by GMSS or HMSS, neither other subspace selection methods. To this end, [5] proposed the max-min distance analysis (MMDA) criterion,

$$\max_{W} \min_{1\leq i<j\leq c} \Delta(\omega_i, \omega_j \mid W) \tag{3.12}$$

where the inner minimization chooses the minimum pairwise distance of all class pairs in the selected subspace, and the outer maximization maximizes this minimum distance. Let the optimal value and solution of (3.12) be Δ_{opt} and W_{opt}, and then we have

$$\Delta(\omega_i, \omega_j \mid W_{\mathrm{opt}}) \geq \Delta_{\mathrm{opt}}, \quad \text{for all } i \neq j, \tag{3.13}$$

which ensures the separation (as best as possible) of any class pairs in the selected low dimensional subspace. Furthermore, by taking the prior probability of each class into account, the MMDA criterion is given by

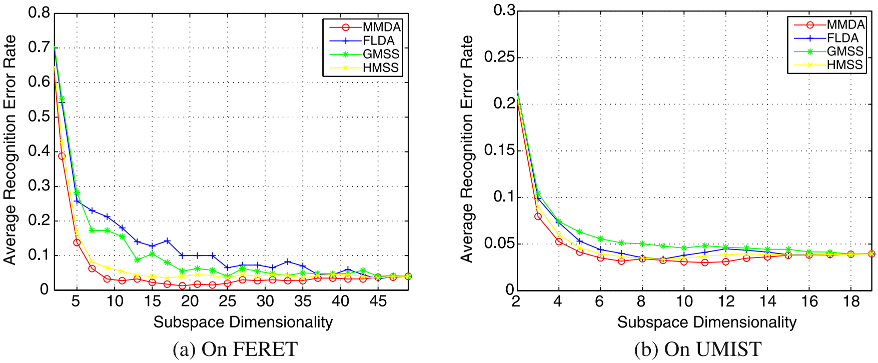
$$\max_{W} \min_{1\leq i<j\leq c} \bigl\{(p_i p_j)^{-1} \Delta(\omega_i, \omega_i \mid W)\bigr\}. \tag{3.14}$$

Note that, the use of $(p_i p_j)^{-1}$ as weighting factor is an intuitive choice. In order to obtain a relatively high accuracy, it has to put more weight on classes with high prior probabilities; however, because the minimization in the max-min operation has a negative effect, we need to put a smaller factor, for example, the inverse factor $(p_i p_j)^{-1}$, on the pairwise distance between high-prior probability classes so that it has a greater chance to be maximized.

The solving of MMDA criteria (3.12) and (3.14) can be difficult. The inner minimizations there are over discrete variables i and j, and thus it makes the objective function for the outer maximization nonsmooth. To deal with this nonsmooth max-min problem [5] introduced the convex relaxation technique. Specifically, the authors proposed a sequential semidefinite programming (SDP) relaxation algorithm, with which an approximate solution of (3.12) or (3.14) can be obtained in polynomial time. Refer to [5] for details of the algorithm. The MMDA result on the illustrative example in Fig. 3.1 is shown in Fig. 3.2, from which one can see that MMDA gives the best separation between blue and green classes among the four criteria.

3.2.3 Empirical Evaluation

The evaluation of general mean criteria, including GMSS and HMSS, and the MMDA are conducted on two benchmark face image datasets, UMIST [1] and FERET [32]. The UMIST database consists of 564 face images from 20 individuals. The individuals are a mix of race, sex and appearance and are photographed in a range of poses from profile to frontal views. The FERET database contains 13 539 face images from 1565 subjects, with varying pose, facial expression and



(a) On FERET

(b) On UMIST

Fig. 3.3 Face recognition by subspace selection and nearest neighbor classification in the selected low dimensional subspace

age. 50 subjects with 7 images for each are used in the evaluation. Images from both databases are cropped with reference to the eyes, and normalized to 40 by 40 pixel arrays with 256 gray levels per pixel. On UMIST, 7 images for each subject are used for training and the rest images are used for test, while on FERET, a 6 to 1 split is used for training/test setup. The average recognition performances over ten random trials are shown in Fig. 3.3. One can see that, on FERET, the general mean criterion (GMSS and HMSS) and MMDA show significant improvements on recognition rate compared with FLDA, while on UMIST, though GMSS gives slight inferior performance to FLDA, HMSS and MMDA still improve the performance in certain extent.

3.2.4 Related Works

In addition to the general mean criteria and max-min distance analysis, there are also some methods proposed in recent years to deal with the class separation problem of FLDA. Among these methods, approximate pairwise accuracy criterion (aPAC) [28] and fractional step LDA (FS-LDA) [30] are the most representative ones, and both of them use weighting schemes to emphasize close class pairs during subspace selection. Besides, the Bayes optimality of FLDA is further studied when the dimensionality of subspace is less than class number minus 1. In particular, it is shown that the one dimensional Bayes optimal subspace can be obtained by convex optimization given the information of the order of class centers projected onto the subspace [18]. Such result generalizes the early result of Bayes optimal one dimensional Bayes optimal subspace on a special case of three Gaussian distributions [36]. Further, the authors of [18] suggested selecting a general subspace by greedy one dimensional subspace selection and orthogonal projection. The homoscedastic Gaussian assumption is another limitation of FLDA. Various methods have been developed to extend FLDA to heteroscedastic Gaussian cases, e.g., the using of information theoretic divergences such as Kullback–Leibler divergence [10, 39], and Chernoff [29] or Bhattacharyya distance [35] to measures the discrimination among heteroscedastic Gaussian distributions. Besides, nonparametric and semiparametric method provide alternative ways for extensions of FLDA, by which classic work includes Fukunaga's nonparametric discriminant analysis (NDA) [16], its latest extension to multiclass case [27] and subclass discriminant analysis [57]. In addition, recent studies show that FLDA can be converted to a least square problem via a proper coding of class labels [49, 50]. The advantages of such least square formulation are that the computational speed can be significantly improved and also regularizations on the subspace are more readily imposed.

3.3 Subspace Learning—A Local Perspective

It has shown that the global linearity of PCA and FLDA prohibit their effectiveness for non-Gaussian distributed data, such as face images. By considering the local geometry information, a dozen of manifold learning algorithms have been developed, such as locally linear embedding (LLE) [34], ISOMAP [40], Laplacian eigenmaps (LE) [2], Hessian eigenmaps (HLLE) [11], and local tangent space alignment (LTSA) [53]. All of these algorithms have been developed intuitively and pragmatically, that is, on the base of the experience and knowledge of experts for their own purposes. Therefore, it will be more informative to provide some a systematic framework for understanding the common properties and intrinsic differences in the algorithms. In this section, we introduce such a framework, that is, "patch alignment", which consists of two stages: part optimization and whole alignment. The framework reveals (i) that algorithms are intrinsically different in the patch optimization stage and (ii) that all algorithms share an almost identical whole alignment stage.

3.3.1 Patch Alignment Framework

The patch alignment framework [55] is composed of two ingredients, first, part optimization and then whole alignment. For part optimization, different algorithms have different optimization criteria over patches, each of which is built by one measurement associated with its related ones. For whole alignment, all part optimizations are integrated into together to form the final global coordinate for all independent patches based on the alignment trick. Figure 3.4 illustrates the patch alignment framework.

Given an instance x_i and its k nearest neighbors $[x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(k)}]$, the part optimization at x_i is defined by

$$\arg\min_{Y_i} \operatorname{tr}\bigl(Y_i L_i Y_i^{\mathrm{T}}\bigr) \tag{3.15}$$

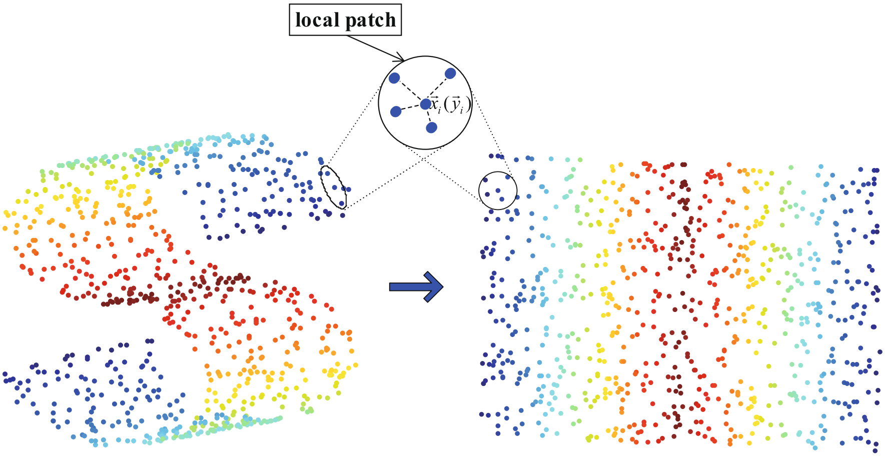


Fig. 3.4 Patch alignment framework

where $Y_i = [y_i, y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(k)}]$ is projection of the local patch $X_i = [x_i, x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(k)}]$ onto the low dimensional subspace, and L_i encodes the local geometry information at instance x_i and is chosen algorithm-specifically. By summarizing part optimizations over all instances, we get

$$\arg \min_{Y_1, Y_2, \ldots, Y_n} \sum_{i=1}^{n} \operatorname{tr}\left(Y_i L_i Y_i^{\mathrm{T}}\right). \tag{3.16}$$

Let $Y = [y_1, y_2, \ldots, y_n]$ be the projection of all instances $X = [x_1, x_2, \ldots, x_n]$. As for each local patch Y_i should be a subset of the whole alignment Y, the relationship between them can be expressed by

$$Y_i = Y S_i \tag{3.17}$$

where S_i is a proper 0-1 matrix called the selection matrix. Thus,

$$\begin{aligned}
&\arg\min_{Y} \sum_{i=1}^{N} \operatorname{tr}\left(Y_i L_i Y_i^{\mathrm{T}}\right) \\
&= \arg\min_{Y} \sum_{i=1}^{N} \operatorname{tr}\left(Y S_i L_i S_i^{\mathrm{T}} Y^{\mathrm{T}}\right) \\
&= \arg\min_{Y} \operatorname{tr}\left(Y L Y^{\mathrm{T}}\right)
\end{aligned} \tag{3.18}$$

with

$$L = \left(\sum_{i=1}^{N} S_i L_i S_i^{\mathrm{T}} \right) \tag{3.19}$$

Table 3.1 Manifold learning algorithms filled in the patch alignment framework

Algorithm	Patch X_i	Representation of part optimization L_i	Objective function
LLE NPE ONPP	Given instance and its neighbors	$\begin{bmatrix} 1 & -c_i^{\mathrm{T}} \\ -c_i & c_i c_i^{\mathrm{T}} \end{bmatrix}$	Nonlinear Linear Orthogonal linear
ISOMAP	Given instance and the rest ones	$(1/N) \cdot \tau(D_G^i)$	Nonlinear
LE LPP	Given instance and its connected ones in the undirected graph	$\begin{bmatrix} \sum_{j=1}^{l} (w_i)_j & -w_i^{\mathrm{T}} \\ -w_i & \mathrm{diag}(w_i) \end{bmatrix}$	Nonlinear Linear
LTSA LLTSA	Given instance and its neighbors	$R_{k+1} - V_i V_i^{\mathrm{T}}$, where V_i denotes d largest right singular vectors of $X_i R_{k+1}$	Nonlinear Linear
	Given instance and its neighbors	$H_i H_i^{\mathrm{T}}$	Nonlinear

called the alignment matrix. Further by letting $Y = U^{\mathrm{T}} X$, that is, a linear projection, (3.18) is rewritten as

$$\arg\min_{U} \operatorname{tr}\bigl(U^{\mathrm{T}} X L X^{\mathrm{T}} U\bigr). \tag{3.20}$$

Further, we can impose the orthogonal constraint $U^{\mathrm{T}} U = I$ on the projection matrix U, or the constraint $Y^{\mathrm{T}} Y = I$ on the Y, which leads to $U^{\mathrm{T}} X X^{\mathrm{T}} U = I$. In both cases, (3.20) is solved by eigen- or generalized eigen-decomposition.

Among all the manifold learning algorithms, the most representatives are locally linear embedding (LLE) [34], ISOMAP [40], Laplacian eigenmaps (LE) [2]. LLE uses linear coefficients to represent local geometry information, and find a low-dimensional embedding such that these coefficients are still suitable for reconstruction. ISOMAP preserves geodesic distances between all instance pairs. And LE preserves proximity relationships by manipulations on an undirected weighted graph, which indicates neighbor relations of pairwise instances. It has been shown that all these algorithms can be filled into the patch alignment framework, where the difference among algorithms lies in the part optimization stage while the whole alignment stage is almost the same. There are also other manifold learning algorithms, for example, Hessian eigenmaps (HLLE) [11], Generative Topographic Mapping (GTM) [6] and local tangent space alignment (LTSA) [53]. We can use the patch alignment framework to explain them in a unified way. Table 3.1 summarizes these algorithms in the patch alignment framework.

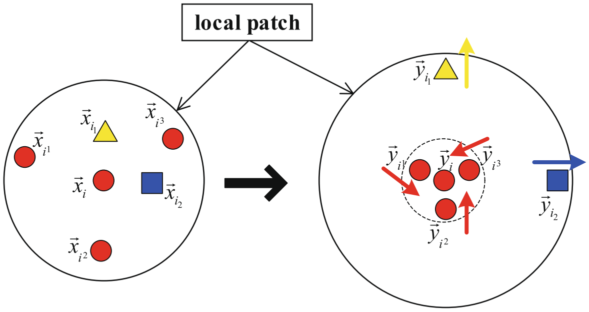


Fig. 3.5 The motivation of DLA. The measurements with the same shape and color come from the same class

3.3.2 Discriminative Locality Alignment

One representative subspace selection method based on the patch alignment framework is the discriminative locality alignment (DLA) [54]. In DLA, the discriminative information, encoded in labels of samples, is imposed on the part optimization stage and then the whole alignment stage constructs the global coordinate in the projected low-dimensional subspace.

Given instance x_i and its k nearest neighbors $[x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(k)}]$, we divide the k neighbors into two groups according to the label information, that is, belonging to the same class with x_i or not. Without losing generality, we can assume the first k_1 neighbors $[x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(k_1)}]$ having the same class label with x_i and the rest $k - k_1$ neighbors $[x_i^{(k_1+1)}, x_i^{(k_1+2)}, \dots, x_i^{(k)}]$ having different class labels (otherwise, we just have to resort the indexes properly). And their low dimensional representations are y_i, $[y_i^{(1)}, y_i^{(2)}, \dots, y_i^{(k_1)}]$ and $[y_i^{(k_1+1)}, y_i^{(k_1+2)}, \dots, y_i^{(k)}]$, respectively. The key idea of DLA is enforcing y_i close to $[y_i^{(1)}, y_i^{(2)}, \dots, y_i^{(k_1)}]$ while pushing it apart from $[y_i^{(k_1+1)}, y_i^{(k_1+2)}, \dots, y_i^{(k)}]$. Figure 3.5 illustrates such motivation.

For instance, x_i and its same class neighbors, we expect the summation of squared distance in the low dimensional subspace to be as small as possible, that is,

$$\arg\min_{y_i} \sum_{p=1}^{k_1} \left\| y_i - y_i^{(p)} \right\|^2 \tag{3.21}$$

However, for x_i and its different class neighbors, we want the corresponding result to be large, that is,

$$\arg\max_{y_i} \sum_{p=k_1+1}^{k} \left\| y_i - y_i^{(p)} \right\|^2 \tag{3.22}$$

A convenient tradeoff between (3.21) and (3.22) is

$$\arg\min_{Y_i} \left(\sum_{p=1}^{k_1} \left\| y_i - y_i^{(p)} \right\|^2 - \gamma \sum_{p=k_1+1}^{k} \left\| y_i - y_i^{(p)} \right\|^2 \right) \tag{3.23}$$

where γ is a scaling factor between 0 and 1 to balance the importance between measures of the within-class distance and the between-class distance. Let

$$\omega_i = \Bigl[\overbrace{1, \ldots, 1,}^{k_1}\ \overbrace{-\gamma, \ldots, -\gamma}^{k-k_1}\Bigr]^{\mathrm{T}}, \tag{3.24}$$

then (3.23) is readily rewritten as

$$\arg\min_{Y_i} \operatorname{tr}\bigl(Y_i L_i Y_i^{\mathrm{T}}\bigr), \tag{3.25}$$

where

$$L_i = \begin{bmatrix} \sum_{j=1}^{k} \omega_i & -\omega_i^{\mathrm{T}} \\ -\omega_i & \operatorname{diag}(\omega_i) \end{bmatrix}. \tag{3.26}$$

To obtain the projection mapping $y = U^{\mathrm{T}} x$, we just substitute (3.26) into the whole alignment formula (3.18), and solve the eigen-decomposition problem with constraint $U^{\mathrm{T}} U = I$. It is worth emphasizing some merits of DLA here: (1) it exploits local geometry information of data distribution; (2) it is ready to deal with the case of nonlinear boundaries for class separation; (3) it avoids the matrix singularity problem.

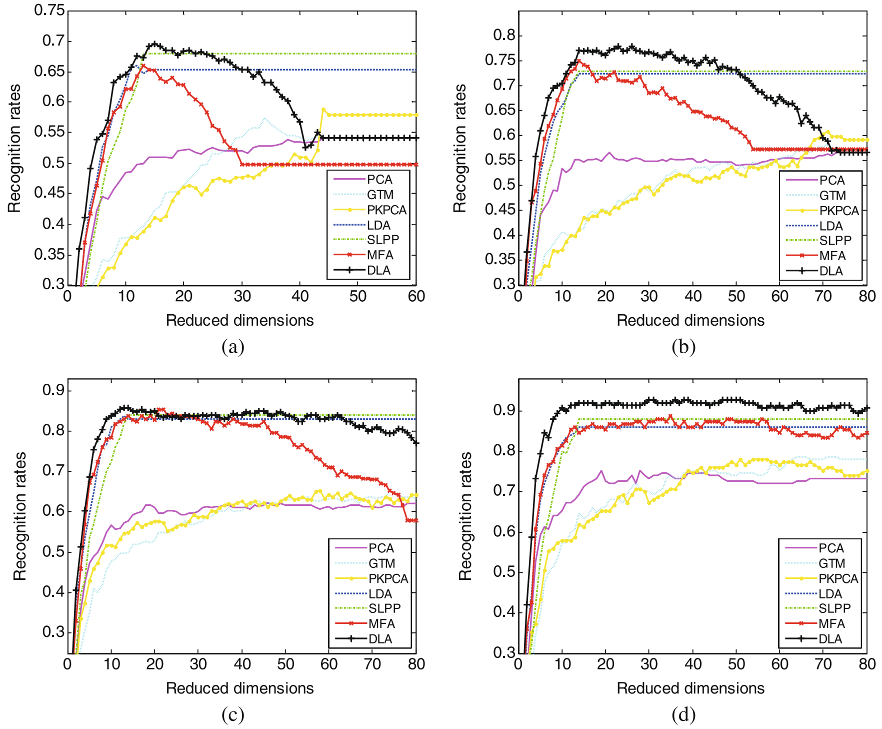
Now we evaluate the performance of the proposed DLA in comparison with six representative algorithms, that is, PCA [23], Generative Topographic Mapping (GTM) [6], Probabilistic Kernel Principal Components Analysis (PKPCA) [42], LDA [14], SLPP [7] and MFA [48], on Yale face image dataset [1]. For training, we randomly selected different numbers (3, 5, 7, 9) of images per individual, used 1/2 of the rest images for validation, and 1/2 of the rest images for testing. Such trial was independently performed ten times, and then the average recognition results were calculated. Figure 3.6 shows the average recognition rates versus subspace dimensions on the validation sets, which help to select the best subspace dimension. It can be seen that DLA outperforms the other algorithms.

3.3.3 Manifold Elastic Net

Manifold elastic net (MEN) [56] is a subspace learning method built upon the patch alignment framework. However, the key feature of MEN is that it is able to achieve sparse basis (projection matrix) by imposing the popular elastic net penalty (i.e., the combination of the lasso penalty and the L2 norm penalty). As sparse basis are more interpretable both psychologically and physiologically, MEN is expected to give more meaningful results on face recognition, which will be shown in experiments later.

First, MEN uses the same part optimization and whole alignment as in DLA, that is, the following minimization is considered

$$\arg\min_{Y} \operatorname{tr}\bigl(Y L Y^{\mathrm{T}}\bigr). \tag{3.27}$$



(a) (b) (c) (d)

Fig. 3.6 Recognition rate vs. subspace dimension on Yale dataset. **a** 3 images per subject for training; **b** 5 images per subject for training; **c** 7 images per subject for training; **d** 9 images per subject for training

However, rather than substituting $Y = U^{\mathrm{T}}X$ directly, (3.27) is reformed equivalently as below

$$\arg\min_{Y,U} \operatorname{tr}\left(YLY^{\mathrm{T}}\right) + \beta\left\|Y - U^{\mathrm{T}}X\right\|^2. \tag{3.28}$$

Note that (3.28) indeed will lead to $Y = U^{\mathrm{T}}X$. Given the equivalence between the two formulations, the latter is more convenient to incorporate the minimization of classification error. Specifically, letting stores the response or prediction result, which are proper encodings of the class label information, we expect $U^{\mathrm{T}}X$ to be close to T, that is,

$$\arg\min_{U}\left\|T - U^{\mathrm{T}}X\right\|^2. \tag{3.29}$$

By combing (3.28) and (3.29), we get the main objective of MEN

$$\arg\min_{Y,U}\left\|T - U^{\mathrm{T}}X\right\|^2 + \alpha \operatorname{tr}\left(Y^{\mathrm{T}}LY\right) + \beta\left\|Y - U^{\mathrm{T}}X\right\|^2 \tag{3.30}$$

where α and β are trade-off parameters to control the impacts of different terms.

To obtain a sparse projection matrix U, an ideal approach is to restrict the number of nonzero entries in it, that is, using the L_0 norm as a penalty over (3.30). However, the L_0 norm penalized (3.30) is an NP-hard problem and thus intractable practically. One attractive way of approximating the L_0 norm is the L_1 norm, i.e., the Lasso penalty [41], which is convex and actually the closet convex relaxation of the L_0 norm. Various efficient algorithms exist for solving Lasso penalized least square regression problem, including the LARS [13]. However, the lasso penalty has the following two disadvantages: (1) the number of variables to be selected is limited by the number of observations and (2) the lasso penalized model can only selects one variable from a group of correlated ones and does not care which one should be selected. These limitations of Lasso are well addressed by the so-called elastic net penalty, which combines the L_2 and L_1 norm together. MEN adopts the elastic net penalty [58]. In detail, the L_2 of the projection matrix is helpful to increase the dimension (and the rank) of the combination of the data matrix and the response. In addition, the combination of the L_1 and L_2 of the projection matrix is convex with respect to the projection matrix and thus the obtained projection matrix has the grouping effect property. The final form of MEN is given by

$$\begin{aligned}\arg\min_{Y,U} &\left\| T - U^{\mathrm{T}} X \right\|^2 + \alpha \operatorname{tr}\bigl(Y^{\mathrm{T}} L Y\bigr) + \beta \left\| Y - U^{\mathrm{T}} X \right\|^2 \\ &+ \lambda_1 \|U\|_1 + \lambda_2 \|U\|^2. \end{aligned} \tag{3.31}$$

We report an empirical evaluation of MEN on the FERET dataset. From in total 13 539 face images of 1565 individuals, 100 individuals with 7 images per subject are randomly selected in the experiment. 4 or 5 images per individual are selected as training set, and the remaining is used for test. All experiments are repeated five times, and the average recognition rates are calculated. Six representative dimension reduction algorithms, that is, principal component analysis (PCA) [23], Fisher's linear discriminant analysis (FLDA) [14], discriminative locality alignment (DLA) [54], supervised locality preserving projection (SLPP) [7], neighborhood preserving embedding (NPE) [19], and sparse principal component analysis (SPCA) [9], are also performed for performance comparison.

The performance of recognition is summarized in Fig. 3.7. Apparently, the seven algorithms are divided into 3 groups according to their performance. The baseline level methods are PCA and SPCA, which is because they are both unsupervised methods and thus may not give satisfying performance due to the missing of label information. LPP, NPE and LDA only show moderate performance. In contrast, DLA and MEN give rise to significant improvements. Further, the sparsity of MEN makes it outperform DLA. The best performance of MEN is actually not surprising, since it considers the most aspects on data representation and distribution, including the sparse property, the local geometry information and classification error minimization.

Figure 3.8 shows the first ten bases selected by different subspace selection methods. One can see that the bases selected by LPP, NPE and FLDA are contaminated by considerable noises, which explains why they only give moderate recognition

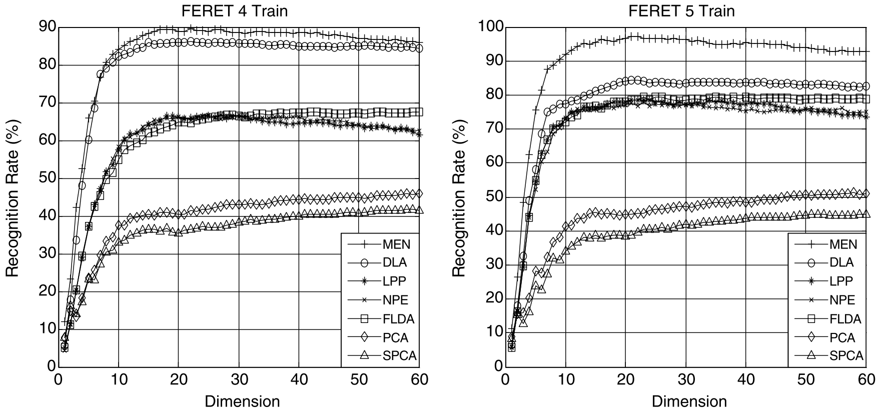


Fig. 3.7 Performance evaluation on the FERET dataset

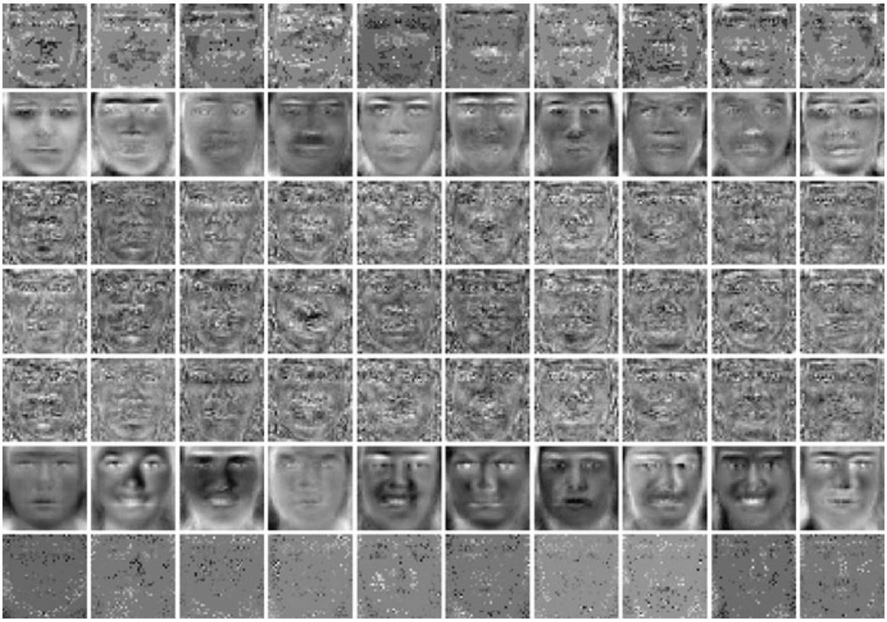


Fig. 3.8 Plots of first 10 bases obtained from 7 dimensionality reduction algorithms on FERET for each column, from top to bottom: MEN, DLA, LPP, NPE, FLDA, PCA, and SPCA

performance. The bases from PCA, that is, Eigenfaces, are smooth but present relatively few discriminative information. In terms of sparsity, SPCA gives the desired bases; however, the problem is that the patterns presented in these bases are not grouped so that cannot provide meaningful interpretation. The bases from MEN, which we call "MEN's faces", have a low level of noise and are also reasonably sparse. And more importantly, thanks to the elastic net penalty, the sparse patterns

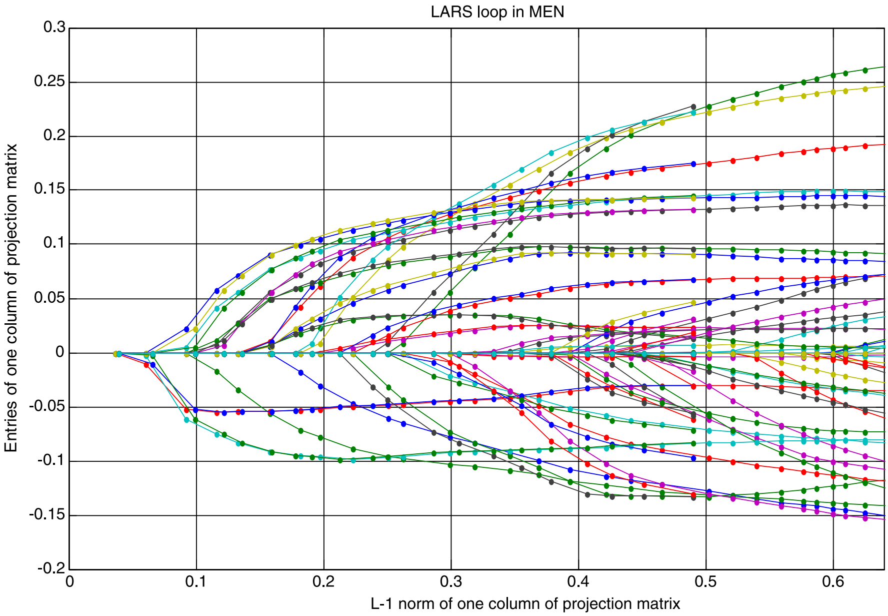


Fig. 3.9 Entries of one column of projection matrix vs. its L_1 norm in one LARS loop of MEN

of MEN's bases are satisfying grouped, which gives meaningful interpretations, for example, most discriminative facial features are obtained, including eyebrows, eyes, nose, mouth, ears and facial contours.

The optimization algorithm of MEN is built upon LARS. In each LARS loop of the MEN algorithm, all entries of one column in the projection matrix are zeros initially. They are sequentially added into the active set according to their importance. The values of active ones are increased with equal altering correlation. In this process, the L_1 norm of the column vector is augmented gradually. Figure 3.9 shows the altering tracks of some entries of the column vector in one LARS loop. These tracks are called "coefficient paths" in LARS. As shown by these plots, one can observe that every coefficient path starts from zero when the corresponding variable becomes active, and then changes its direction when another variable is added into the active set. All the paths keep in the directions which make the correlations of their corresponding variables equally altering. The L_1 norm is increasing along the greedy augment of entries. The coefficient paths proceed along the gradient decent direction of objective function on the subspace, which is spanned by the active variables.

In addition, Fig. 3.10 shows 10 of the 1600 coefficient paths from LAPS loop. It can be seen that MEN selects ten important features sequentially. For each feature, its corresponding coefficient path and the "MEN face" when the feature is added into active set are assigned the same color which is different with the other 9 features. In each "MEN face", the new added active feature is marked by a small circle, and all the active features are marked by white crosses. The features selected by

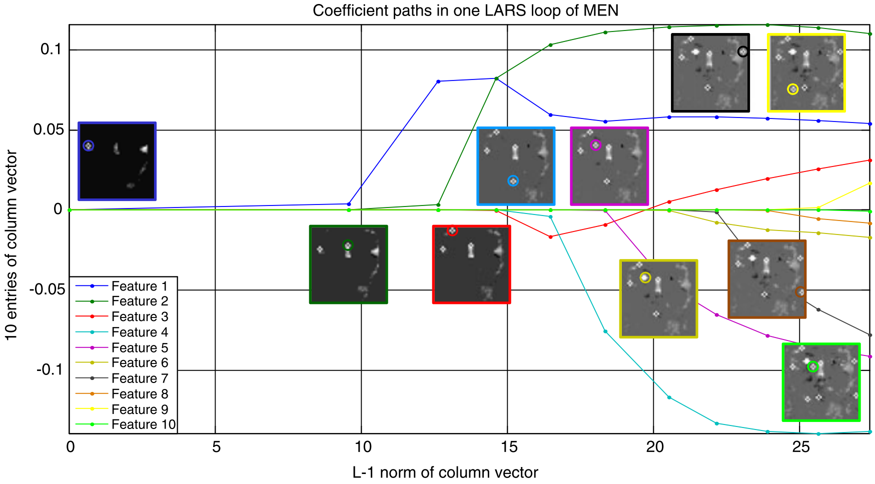


Fig. 3.10 Coefficient paths of 10 features in one column vector

MEN can produce explicit interpretation of the relationship between facial features and face recognition: feature 1 is the left ear, feature 2 is the top of nose, feature 3 is on the head contour, feature 4 is the mouth, feature 5 and feature 6 are on the left eye, feature 7 is the right ear, and feature 8 is the left corner of mouth. These features are already verified of great importance in face recognition by many other famous face recognition methods. Moreover, Fig. 3.10 also shows MEN can group correlated features, for example, feature 5 and feature 6 are selected sequentially because they are both on the left eye. In addition, features which are not very important, such as feature 9 and feature 10 in Fig. 3.10, are selected after the selection of the other more significant features and assigned smaller value than those more important ones. Therefore, MEN is a powerful algorithm in feature selection.

3.3.4 Related Works

Applying the idea of manifold learning, that is, exploring local geometry information of data distribution, into semisupervised or transductive subspace selection leads to a new framework of dimension reduction by manifold regularization. One example is the recently proposed manifold regularized sliced inverse regression (MRSIR) [4]. Sliced inverse regression (SIR) was proposed for sufficient dimension reduction. In a regression setting, with the predictors X and the response Y, the sufficient dimension reduction (SDR) subspace B is defined by the conditional independency $Y \perp X \mid B^{\mathrm{T}}X$. Under the assumption that the distribution of X is elliptic symmetric, it has been proved that the SDR subspace B is related to the inverse regression curve $E(X \mid Y)$. It can be estimated at least partially by a generalized eigendecomposition between the covariance matrix of the predictors $\mathrm{Cov}(X)$ and

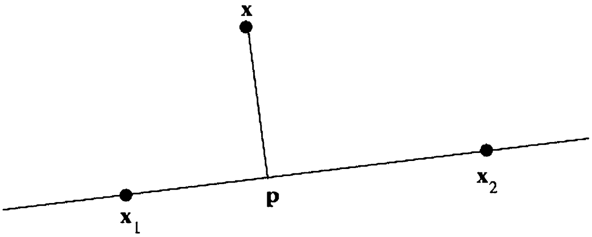


Fig. 3.11 Point $\mathbf{p}$ is the projection of query $\mathbf{x}$ onto the feature line $\overline{\mathbf{x}_1\mathbf{x}_2}$

the covariance matrix of the inverse regression curve $\mathrm{Cov}(E(X \mid Y))$. If Y is discrete, this is straightforward. While Y is continuous, it is discretized by slicing its range into several slices so as to estimate $E(X \mid Y)$ at each slice.

Suppose Γ and Σ are respectively the empirical estimates of $\mathrm{Cov}(E(X \mid Y))$ and $\mathrm{Cov}(X)$ based on a training data set. Then, the SDR subspace B is given by

$$\max_B \operatorname{trace}\big(\big(B^{\mathrm{T}}\Sigma B\big)^{-1}B^{\mathrm{T}}\Gamma B\big). \tag{3.32}$$

To construct the manifold regularization, [4] uses the graph Laplacian L of the training data $X = [x_1, x_2, \dots, x_n]$. Letting $Q = \frac{1}{n(n-1)}XLX^{\mathrm{T}}$ and $S = \frac{1}{n(n-1)}XDX^{\mathrm{T}}$, with D being the degree matrix, then MRSIR is defined by

$$\max_B \operatorname{trace}\big(\big(B^{\mathrm{T}}\Sigma B\big)^{-1}B^{\mathrm{T}}\Gamma B\big) - \eta \operatorname{trace}\big(\big(B^{\mathrm{T}}SB\big)^{-1}B^{\mathrm{T}}QB\big), \tag{3.33}$$

where η is a positive weighting factor. The use of manifold regularization extends SIR in many ways, that is, it utilizes the local geometry that is ignored originally and enables SIR to deal with the tranductive/semisupervised subspace selection problems.

So far we have introduced subspace selection methods that exploit local geometry information of data distribution. Based on these methods, classification can be performed in the low dimensional embedding. However, as the final goal is classification, an alternative approach is to do classification directly using the local geometry information. This generally leads to nonparametric classifiers, for example, nearest neighbor (NN) classifier. The problem is that simple NN classifier cannot provide satisfying recognition performance when data are of very high dimensions as in face recognition. To this end, Li and Liu proposed the nearest feature line (NFL) for face recognition [25, 26]. In NFL, a query is projected onto a line segment between any two instances within each class, and the nearest distance between the query and the projected point is used to determine its class label. Figure 3.11 shows an example of projecting a query $\mathbf{x}$ onto the feature line spanned by instances $\mathbf{x}_1$ and $\mathbf{x}_2$, where the projected point $\mathbf{p}$ is given by

$$\mathbf{p} = \mathbf{x}_1 + \mu * (\mathbf{x}_2 - \mathbf{x}_1), \tag{3.34}$$

with

$$\mu = \frac{(\mathbf{x} - \mathbf{x}_1)^{\mathrm{T}}(\mathbf{x}_2 - \mathbf{x}_1)}{\|\mathbf{x}_2 - \mathbf{x}_1\|^2}. \tag{3.35}$$

One extension of NFL is the nearest linear combination (NLC) [25]. There a query is projected onto a linear subspace spanned by a set of basis vectors, where the basis vectors can be any form from a subspace analysis or a set of local features, and the distance between the query and the projection point is used as the metric for classification. Empirical studies shown that NFL and NLC produces significantly better performance than the simple nearest neighborhood (NN) when the number of prototype templates (basis vectors representing the class) is small.

Another method related to NLC and NS approach is the sparse representation classifier (SRC) [47], which treats the face recognition problem as searching for an optimal sparse combination of gallery images to represent the probe one. SRC differs from the standard NLC in the norm used to define the projection distance. Instead of using the 2-norm as in NLC [25], SRC uses the 1- or 0-norm, such that the sparsity emerges.

3.4 Transfer Subspace Learning

Conventional algorithms including subspace selection methods are built under the assumption that training and test samples are independent and identically distributed (i.i.d.). For practical applications, however, this assumption cannot be hold always. Particularly, in face recognition, the difference of expressions, postures, aging problem and lighting conditions makes the distributions of training and test face different. To this end, a transfer subspace learning (TSL) framework is proposed [38]. TSL extends conventional subspace learning methods by using a Bregman divergence based regularization, which encourages the difference between the training and test samples in the selected subspace to be minimized. Thus, we can approximately assume the samples of training and test are almost i.i.d. in the learnt subspace.

3.4.1 TSL Framework

The TSL framework [38] is presented by the following unified form

$$\arg\min_{U} F(U) + \lambda D_U(P_l||P_u) \tag{3.36}$$

where $F(U)$ is the objective function of a subspace selection method, for example, FLDA or PCA et al., and $D_U(P_l||P_u)$ is the Bregman divergence between the training data distribution P_l and the test data distribution P_u in the low dimension subspace U, and parameter λ controls the balance between the objective function and the regularization. Note that generally the objective function $F(U)$ only depends on the training data.

For example, when $F(U)$ is chosen to be FLDA's objective, (32) will give a subspace in which the training and test data distributions are close to each other and

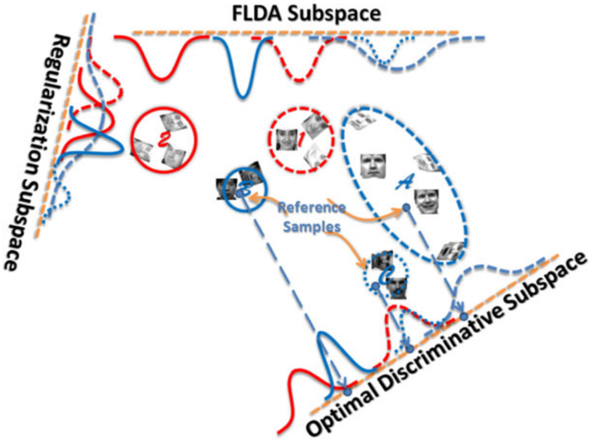


Fig. 3.12 Two classes of training samples are marked as 1 and 2, while three classes of test samples are marked as A, B and C. *Blue circles* A and C are merged together in the FLDA subspace, where discrimination of the training samples can be well preserved. *Blue circles* A and B are mixed in the regularization subspace, where there exists the smallest divergence between training domain (1, 2) and test domain (A, B and C). *Blue circles* A, B and C can be well separated in the discriminative subspace, which is obtained by optimizing the combination of the proposed regularization (the divergence between training sets 1, 2 and test sets A, B, C) and FLDA

the discriminative information in the training data is partially preserved. In particular, suppose we have two classes of training samples, represented by two red circles (1 and 2, e.g., face images in the FERET dataset), and three classes of test samples, represented by three blue circles (A, B and C, e.g., face images in the YALE dataset), as shown in Fig. 3.12. On one hand, FLDA finds a subspace that fails to separate the test circle A from the test circle C, but the subspace is helpful to distinct different subjects in the training set. On the other hand, the minimization of the Bregman divergence between training and test distributions would give a subspace that makes the training data and test data almost i.i.d., but give little discriminative power. Apparently, neither of them individually can find a best discriminative subspace for test. However, as shown in the figure, a combination of FLDA and the Bregman regularization does find the optimal subspace for discrimination, wherein A, B and C can be well separated and samples in them can be correctly classified with given references. It is worth emphasizing that the combination works well because the training and test samples are coming from different domains but both domains share some common properties.

The authors suggest solving (3.36) by gradient descent method [38],

$$U \leftarrow U - \tau\left(\frac{\partial F(U)}{\partial U} + \lambda\frac{\partial D_U(P_l||P_u)}{\partial U}\right) \tag{3.37}$$

where τ is the learning rate, that is, step size for updating. As $F(U)$ is usually known, so is its derivative. The problem remaining is how to estimate $D_U(P_l||P_u)$ and its derivatives.

Definition 1 (Bregman divergence regularization) Let $f : S \to R$ be a convex function defined on a closed convex set $S \in R^+$. We denote the first order derivative of f as f', whose inverse function as $\xi = (f')^{-1}$. The probability density for the training and test samples in the projected subspace U are $p_l(y)$ and $p_u(y)$ respectively, wherein $y = U^{\mathrm{T}}x$ is the low-dimensional representation of the sample x. The difference at $\xi(p_l(y))$ between the function f and the tangent line to f at $(\xi(p_l(y)), f(\xi(p_l(y))))$ is given by:

$$\begin{aligned} & d\big(\xi\big(p_l(y)\big), \xi\big(p_u(y)\big)\big) \\ & \quad = \big\{f\big(\xi\big(p_u(y)\big)\big) - f\big(\xi\big(p_l(y)\big)\big)\big\} - p_l(y)\big\{\xi\big(p_u(y)\big) - \xi\big(p_l(y)\big)\big\}. \end{aligned} \tag{3.38}$$

Based on (3.38), the Bregman divergence regularization, which measures the distance between $p_l(y)$ and $p_u(y)$, is a convex function given by

$$D_U(P_l||P_u) = \int d\big(\xi\big(p_l(y)\big), \xi\big(p_u(y)\big)\big)\, d\mu \tag{3.39}$$

where $d\mu$ is the Lebesgue measure.

By taking a special form $f(y) = y^2$, $D_U(P_l||P_u)$ can be expressed as [38]

$$\begin{aligned} & D_W(P_l||P_u) \\ & \quad = \int \big(p_l(y) - p_u(y)\big)^2\, dy \\ & \quad = \int \big(p_l(y)^2 - 2p_l(y)p_u(y) + p_u(y)^2\big)\, dy. \end{aligned} \tag{3.40}$$

Further, the kernel density estimation (KDE) technique is used to estimate $p_l(y)$ and $p_u(y)$. Suppose there are n_l training instances $\{x_1, x_2, \ldots, x_{n_l}\}$ and n_u test instances $\{x_1, x_2, \ldots, x_{n_l}\}$, then through projection $y_i = U^{\mathrm{T}}x_i$, we have the estimates [38]

$$p_l(y) = (1/n_l)\sum_{i=1}^{n_l} G_{\Sigma_1}(y - y_i) \tag{3.41}$$

and

$$p_u(y) = (1/n_u)\sum_{i=n_l+1}^{n_l+n_u} G_{\Sigma_2}(y - y_i) \tag{3.42}$$

where $G_{\Sigma_1}(y)$ is a Gaussian kernel with covariance Σ_1, so is $G_{\Sigma_2}(y)$. With these estimates, the quadratic divergence (3.40) is rewritten as [38]

$$D_W(P_l||P_u) = \frac{1}{n_l^2}\sum_{s=1}^{n_l}\sum_{t=1}^{n_l} G_{\Sigma_{11}}(y_t - y_s) + \frac{1}{n_u^2}\sum_{s=n_l+1}^{n_l+n_u}\sum_{t=n_l+1}^{n_l+n_u} G_{\Sigma_{22}}(y_t - y_s) - \frac{2}{n_l n_u}\sum_{s=1}^{n_l}\sum_{t=n_l+1}^{n_l+n_u} G_{\Sigma_{12}}(y_t - y_s) \tag{3.43}$$

where $\Sigma_{11} = \Sigma_1 + \Sigma_1$, $\Sigma_{12} = \Sigma_1 + \Sigma_2$ and $\Sigma_{22} = \Sigma_2 + \Sigma_2$. Further, by basis matrix calculus, we have

$$\begin{aligned}\frac{\partial D_U(P_l||P_u)}{\partial U} &= \frac{2}{n_l^2}\sum_{i=1}^{n_l}\sum_{t=1}^{n_l} G_{\Sigma_{11}}(y_i - y_t)(\Sigma_{11})^{-1}(y_t - y_i)x_i^{\mathrm{T}} \\ &- \frac{2}{n_l n_u}\sum_{i=1}^{n_l}\sum_{t=n_l+1}^{n_l+n_u} G_{\Sigma_{12}}(y_t - y_i)(\Sigma_{12})^{-1}(y_t - y_i)x_i^{\mathrm{T}} \\ &+ \frac{2}{u^2}\sum_{i=n_l+1}^{n_l+n_u}\sum_{t=n_l+1}^{n_l+n_u} G_{\Sigma_{22}}(y_i - y_t)(\Sigma_{22})^{-1}(y_t - y_i)x_i^{\mathrm{T}} \\ &- \frac{2}{n_l n_u}\sum_{i=l+1}^{n_l+n_u}\sum_{t=1}^{n_l} G_{\Sigma_{12}}(y_t - y_i)(\Sigma_{12})^{-1}(y_t - y_i)x_i^{\mathrm{T}}.\end{aligned} \tag{3.44}$$

3.4.2 Cross Domain Face Recognition

Based on the YALE, UMIST and a subset of FERET datasets, cross-domain face recognition is performed by applying the TSL framework. In detail, we have (1) Y2F: the training set is on YALE and the test set is on FERET; (2) F2Y: the training set is on FERET and the test set is on YALE; and (3) YU2F: the training set is on the combination of YALE and UMIST and the test set is on FERET. In the training stage, the labeling information of test images is blind to all subspace learning algorithms. However, one reference image for each test class is preserved so that the classification can be done in the test stage. The nearest neighbor classifier is adopted for classification, i.e., we calculate the distance between a test image and every reference image and predict the label of the test image as that of the nearest reference image.

We compare TSL algorithms, for example, TPCA, TFLDA, TLPP, TMFA, and TDLA, with conventional subspace learning algorithms, for example, PCA [23], FLDA [14], LPP [20], MFA [48], DLA [54] and the semi-supervised discriminant analysis (SDA) [8]. Table 3.2 shows the recognition rate of each algorithm with the corresponding optimal subspace dimension. In detail, conventional subspace learning algorithms, for example, FLDA, LPP and MFA, perform poorly because they assume training and test samples are i.i.d. variables and this assumption is unsuitable for cross-domain tasks. Although SDA learns a subspace by taking test samples into account, it assumes samples in a same class are drawn from an identical

Table 3.2 Recognition rates of different algorithms under three experimental settings. The number in the parenthesis is the corresponding subspace dimensionality

	Y2F	F2Y	YU2F
LDA	39.71(70)	36.36(30)	29.57(30)
LPP	44.57(65)	44.24(15)	45.00(35)
MFA	40.57(65)	34.54(60)	27.85(70)
DLA	50.43(80)	50.73(15)	50.86(65)
SDA	44.42(65)	41.81(40)	32.00(35)
MMDR	45.60(60)	42.00(75)	49.75(80)
TLDA	57.28(15)	50.51(20)	55.57(45)
TLPP	58.28(30)	53.93(25)	58.42(30)
TMFA	63.14(70)	56.96(35)	65.42(70)
TDLA	63.12(60)	61.82(30)	65.57(70)

underlying manifold. Therefore, SDA is not designed for the cross-domain tasks. Although MMDR considers the distribution bias between the training and the test samples, it ignores the discriminative information contained in the training samples. We have given an example in the synthetic data test to show that the training discriminative information is helpful to separate test classes. Example TSL algorithms perform consistently and significantly better than others, because the training discriminative information can be properly transferred to test samples by minimizing the distribution distance between the training and the test samples. In particular, TDLA performs best among all TSL examples because it inherits the merits of DLA in preserving both the discriminative information of different classes and the local geometry of samples in an identical class.

Acknowledgements The authors thank Prof. Stan Z. Li for insightful discussions on nearest feature line.

References

1. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 711–720 (1997)
2. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**(6), 1373–1396 (2003)
3. Bian, W., Tao, D.: Harmonic mean for subspace selection. In: 19th International Conference on Pattern Recognition, pp. 1–4 (2008)
4. Bian, W., Tao, D.: Manifold regularization for sir with rate root-n convergence (2010)
5. Bian, W., Tao, D.: Max-min distance analysis by using sequential sdp relaxation for dimension reduction. IEEE Trans. Pattern Anal. Mach. Intell. 99(PrePrints) (2010)
6. Bishop, C.M., Svensén, M., Williams, C.K.I.: GTM: The generative topographic mapping. Technical Report NCRG/96/015, Neural Computing Research Group, Dept of Computer Science & Applied Mathematics, Aston University, Birmingham B4 7ET, United Kingdom, April 1997

7. Cai, D., He, X., Han, J.: Using graph model for face analysis. Technical report, Computer Science Department, UIUC, UIUCDCS-R-2005-2636, September 2005
8. Cai, D., He, X., Han, J.: Srda: An efficient algorithm for large-scale discriminant analysis. IEEE Trans. Knowl. Data Eng. **20**(1), 1–12 (2008)
9. D'aspremont, A., Ghaoui, L.E., Jordan, M.I., Lanckriet, G.R.G.: A direct formulation for sparse PCA using semidefinite programming. SIAM Rev. **49**(3), 434–448 (2007)
10. Decell, H., Mayekar, S.: Feature combinations and the divergence criterion. Comput. Math. Appl. **3**(4), 71–76 (1977)
11. Donoho, D.L., Grimes, C.: Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proc. Natl. Acad. Sci. USA **100**(10), 5591–5596 (2003)
12. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**, 303–353 (1998)
13. Efron, B., Hastie, T., Johnstone, L., Tibshirani, R.: Least angle regression. Ann. Stat. **32**, 407–499 (2004)
14. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**, 179–188 (1936)
15. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, San Diego (1990)
16. Fukunaga, K., Mantock, J.: Nonparametric discriminant analysis. IEEE Trans. Pattern Anal. Mach. Intell. **5**, 671–678 (1983)
17. Graham, D.B., Allinson, N.M.: Characterizing virtual eigensignatures for general purpose face recognition. In: Wechsler, H., Phillips, P.J., Bruce, V., Fogelman-Soulie, F., Huang, T.S. (eds.) Face Recognition: From Theory to Applications. NATO ASI Series F, Computer and Systems Sciences, vol. 163, pp. 446–456 (1998)
18. Hamsici, O.C., Martinez, A.M.: Bayes optimality in linear discriminant analysis. IEEE Trans. Pattern Anal. Mach. Intell. **30**(4), 647–657 (2008)
19. He, X., Cai, D., Yan, S., Zhang, H.-J.: Neighborhood preserving embedding. In: Proc. Int. Conf. Computer Vision (ICCV'05) (2005)
20. He, X., Niyogi, P.: Locality preserving projections. In: Thrun, S., Saul, L., Scholkopf, B. (eds.) Advances in Neural Information Processing Systems, vol. 16. MIT Press, Cambridge (2004)
21. He, X., Yan, S., Hu, Y., Niyogi, P.: Face recognition using Laplacianfaces. IEEE Trans. Pattern Anal. Mach. Intell. **27**(3), 328–340 (2005)
22. Huang, J., Smola, A.J., Gretton, A., Borgwardt, K.M., Schölkopf, B.: Correcting sample selection bias by unlabeled data. In: NIPS, pp. 601–608 (2006)
23. Jolliffe, I.: Principal Component Analysis, 2nd edn. Springer Series in Statistics, Springer, New York (2002)
24. Li, L.: Sparse sufficient dimension reduction. Biometrika **94**(3), 603–613 (2007)
25. Li, S.Z.: Face recognition based on nearest linear combinations. In: CVPR, pp. 839–844 (1998)
26. Li, S.Z., Lu, J.: Face recognition using the nearest feature line method. IEEE Trans. Neural Netw. **10**(2), 439–443 (1999)
27. Li, Z., Lin, D., Tang, X.: Nonparametric discriminant analysis for face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(4), 755–761 (2009)
28. Loog, M., Duin, R., Haeb-Umbach, R.: Multiclass linear dimension reduction by weighted pairwise Fisher criteria. IEEE Trans. Pattern Anal. Mach. Intell. **23**(7), 762–766 (2001)
29. Loog, M., Duin, R.P.W.: Linear dimensionality reduction via a heteroscedastic extension of lda: The Chernoff criterion. IEEE Trans. Pattern Anal. Mach. Intell. **26**, 732–739 (2004)
30. Lotlikar, R., Kothari, R.: Fractional-step dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. **22**(6), 623–627 (2000)
31. Pan, S.J., Kwok, J.T., Yang, Q.: Transfer learning via dimensionality reduction. In: Proc. of the Twenty-Third AAAI Conference on Artificial Intelligence (2008)
32. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The Feret evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1090–1104 (2000)
33. Rao, C.R.: The utilization of multiple measurements in problems of biological classification. J. R. Stat. Soc., Ser. B, Methodol. **10**(2), 159–203 (1948)

34. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**, 2323–2326 (2000)
35. Saon, G., Padmanabhan, M.: Minimum Bayes error feature selection for continuous speech recognition. In: Advances in Neural Information Processing Systems, vol. 13, pp. 800–806. MIT Press, Cambridge (2001)
36. Schervish, M.: Linear discrimination for three known normal populations. J. Stat. Plan. Inference **10**, 167–175 (1984)
37. Shakhnarovich, G., Moghaddam, B.: Face recognition in subspaces. In: Handbook of Face Recognition, pp. 141–168 (2004)
38. Si, S., Tao, D., Geng, B.: Bregman divergence-based regularization for transfer subspace learning. IEEE Trans. Knowl. Data Eng. **22**(7), 929–942 (2010)
39. Tao, D., Li, X., Wu, X., Maybank, S.J.: Geometric mean for subspace selection. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 260–274 (2009)
40. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)
41. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B, Stat. Methodol. **58**, 267–288 (1996)
42. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. J. R. Stat. Soc., Ser. B, Stat. Methodol. **61**(3), 611–622 (1999)
43. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cogn. Neurosci. **3**, 71–86 (1991)
44. Wang, X., Tang, X.: A unified framework for subspace face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **26**, 1222–1228 (2004)
45. Wang, X., Tang, X.: Subspace analysis using random mixture models. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 574–580 (2005)
46. Wang, X., Tang, X.: Random sampling for subspace face recognition. Int. J. Comput. Vis. **70**, 91–104 (2006)
47. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 210–227 (2009)
48. Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., Lin, S.: Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. **29**(1), 40–51 (2007)
49. Ye, J.: Least squares linear discriminant analysis. In: Proceedings of the 24th International Conference on Machine Learning, ICML '07, pp. 1087–1093 (2007)
50. Ye, J., Ji, S.: Discriminant analysis for dimensionality reduction: An overview of recent developments. In: Boulgouris, N., Plataniotis, K.N., Micheli-Tzanakou, E. (eds.) Biometrics: Theory, Methods, and Applications. Wiley-IEEE Press, New York (2010). Chap. 1
51. Ye, J., Li, Q.: A two-stage linear discriminant analysis via qr-decomposition. IEEE Trans. Pattern Anal. Mach. Intell. **27**(6), 929–941 (2005)
52. Ye, J., Li, Q.: A two-stage linear discriminant analysis via qr-decomposition. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 929–941 (2005)
53. Zhang, Z., Zha, H.: Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM J. Sci. Comput. **26**, 313–338 (2005)
54. Zhang, T., Tao, D., Yang, J.: Discriminative locality alignment. In: Proceedings of the 10th European Conference on Computer Vision, pp. 725–738, Berlin, Heidelberg, 2008
55. Zhang, T., Tao, D., Li, X., Yang, J.: Patch alignment for dimensionality reduction. IEEE Trans. Knowl. Data Eng. **21**, 1299–1313 (2009)
56. Zhou, T., Tao, D., Wu, X.: Manifold elastic net: A unified framework for sparse dimension reduction. Data Min. Knowl. Discov. (2010)
57. Zhu, M., Martinez, A.M.: Subclass discriminant analysis. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1274–1286 (2006)
58. Zou, H., Hastie, T.: Regularization and variable selection via the Elastic Net. J. R. Stat. Soc. B **67**, 301–320 (2005)
59. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. J. Comput. Graph. Stat. **15** (2004)