Xudong Jiang

Linear Subspace Learning-Based Dimensionality Reduction

A feature extraction module in the pattern recognition system

© DIGITAL STOCK & LUSHPIX

The ultimate goal of pattern recognition is to discriminate the class membership of the observed novel objects with the minimum misclassification rate. An observed object is often represented by a high-dimensional real-valued vector after some preprocessing while its class membership can be represented by a much lower dimensional binary vector. Thus, in the discriminating process, a pattern recognition system intrinsically reduces the dimensionality of the input data into the number of classes. In fact, dimensionality reduction often occurs implicitly in all modules of a recognition system: preprocessing, feature extraction, and classification. In some applications such as visual object detection and recognition, bioinformatics, and data mining, high data dimensionality imposes great burdens on the robust and accurate recognition due to insufficient knowledge about the data population and limited number of training samples. Dimensionality reduction thus becomes a separate and maybe the most critical module of such recognition systems. Linear subspace analysis is a powerful tool for dimensionality reduction. It also provides a solid foundation for various nonlinear approaches. This is evidenced by numerous techniques published in the past two decades. While some of them, such as sparse representation [1], [2] and subspace arrangements [3], directly solve the classification and clustering problems, most approaches such as the principal component analysis (PCA) [4], linear discriminant analysis (LDA) [4], null-space LDA (NLDA) [5], locality preserving projections (LPP) [6], [7], marginal Fisher analysis (MFA) [8], and their numerous variants serve as a means of feature extraction.

Dimensionality reduction functioning as a feature extraction has two objectives. One objective is to reduce the computational complexity of the subsequent classification with the minimum loss of

Digital Object Identifier 10.1109/MSP.2010.939041
Date of publication: 17 February 2011

1053-5888/11/$26.00©2011IEEE

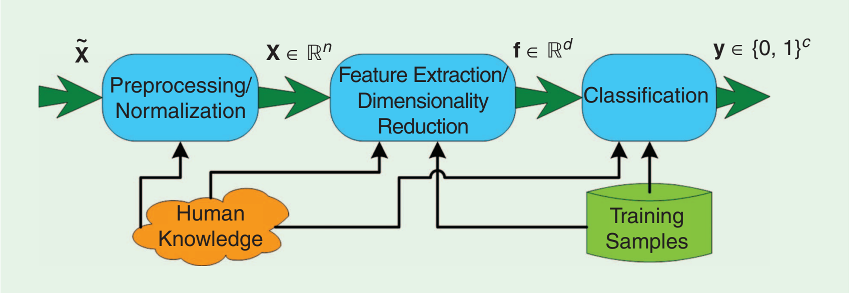
information needed for classification. The second objective is to circumvent the generalization problem of the subsequent classification and hence enhance its accuracy and robustness. To achieve the first objective, it is straightforward that we should maximize the information carried by the data in the extracted low-dimensional subspace. Although PCA does maximize the data structure information in the principal space and hence is optimal for data reconstruction, it is the discriminative information that plays roles in pattern recognition. Thus, most researchers prefer discriminant analysis to the principal component analysis, as evidenced by the fact that the vast majority of the published approaches are based on some kind of the "most discriminative" criteria. There is no doubt that various discriminant analyses can effectively achieve the first objective. The second objective of the dimensionality reduction is, however, far from straightforward. The most discriminative subspace may not be an effective criterion for it because any dimensionality reduction causes a loss of information, including the discriminative information. Any subspace cannot contain more discriminative information than any larger one that includes the former. Why can the dimensionality reduction boost the classification accuracy if the discriminative information is the most critical for classification? Although some general phenomena, such as the curse of dimensionality, small sample size problem, noise removal effect of dimensionality reduction, and better generalization in a lower dimensional space, are well known in the pattern recognition community, they have not indicated what dimensions should be extracted or what else should be removed for a more robust classification. We cannot develop an effective dimensionality reduction technique to maximize the classification accuracy just based on these general phenomena.

> ALTHOUGH THE ULTIMATE GOAL OF ALL MODULES OF A PATTERN RECOGNITION SYSTEM IS TO EXTRACT THE MOST DISCRIMINATIVE INFORMATION, IT IS THE MOST DISCRIMINATIVE INFORMATION ABOUT THE WHOLE DATA POPULATION, NOT ON A SPECIFIC TRAINING SET.

It is thus necessary to study the underlying principles and insights of why and how the dimensionality reduction can enhance the generalization accuracy and robustness of the subsequent classification. This is critical because the second objective of the dimensionality reduction is more important than the first one in most applications with the rapid growth of computation power. The study will also help us find the commonalities and differences of various dimensionality reduction techniques and their pros and cons. Without a thorough analysis and gaining an in-depth understanding of the underlying principles, it is difficult to bring the research in this area to a significantly higher level. This article studies the linear subspace learning-based dimensionality reduction as a feature extraction module in the pattern recognition system. Hopefully, some doubts, misunderstandings, ambiguities, and paradoxes in this area can be resolved by this study. For an in-depth analysis, we need to start from some fundamental yet critical issues in pattern recognition and then explore some problems of the statistical classification.

FUNCTIONALITIES OF PATTERN RECOGNITION MODULES

To study how the dimensionality reduction enhances the recognition accuracy, we need to explore the roles of different modules of the recognition system. A statistical pattern recognition system can be partitioned into three modules as shown in Figure 1. The preprocessing/normalization module segments the object of interest from the background, removes noise, and normalizes its representation. This module is usually designed based on some human knowledge to reduce the intraclass variation of patterns with minimum loss of their interclass distinction, i.e., to extract the most discriminative information from the pattern. Although its input $\tilde{\mathbf{x}}$ and output $\mathbf{x}$ may lie in the same domain, e.g., both are images, dimensionality reduction implicitly occurs at this early stage. Among various pattern representations after the first module, we consider the most widely applied vector format $\mathbf{x} \in \mathbb{R}^n$ in an n-dimensional Euclidian space, called data space. The feature extraction/dimensionality-reduction module transfers the pattern from the data space $\mathbf{x} \in \mathbb{R}^n$ to a feature space $\mathbf{f} \in \mathbb{R}^d$. Some approaches are based on the human knowledge about the pattern, e.g., extracting image local structures such as corner, blob, and local orientation [9], [10], and global structures such as Fourier transform and various moments [11]. For many difficult recognition tasks, human beings lack sufficient knowledge about the discriminative features hidden in the data, and hence machine learning from training samples becomes more prevalent. Obviously, the objective of this module is the same as the first one: extracting the most discriminative information. Dimensionality reduction ($d < n$) often explicitly occurs at this intermediate stage. The last module, classification, establishes decision boundaries in the feature space that separate patterns of different classes. As the extracted features are often abstract with little physical interpretation, this module is mainly designed based on the



$\tilde{\mathbf{x}}$
Preprocessing/ Normalization
$\mathbf{x} \in \mathbb{R}^n$
Feature Extraction/ Dimensionality Reduction
$\mathbf{f} \in \mathbb{R}^d$
Classification
$\mathbf{y} \in \{0, 1\}^c$
Human Knowledge
Training Samples

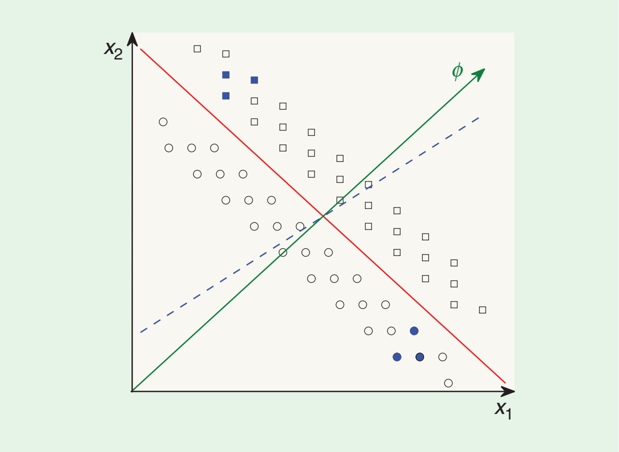
[FIG1] A general model of the statistical pattern recognition system.

machine learning with limited human interference such as some assumptions of the data distribution model, class prior probability, and loss function. The class label can be represented by a c-dimensional binary vector for a c-class problem. Thus, classification transforms the feature vector, $\mathbf{f} \in \mathbb{R}^d$, into the class label vector, $\mathbf{y} \in \{0, 1\}^c$, which again extracts the most discriminative information and, in most applications, implicitly reduces the dimensionality ($c < d$).

TO BOOST THE CLASSIFICATION ACCURACY, THE DIMENSIONALITY REDUCTION SHOULD BE TARGETED AT REMOVING THE DIMENSIONS UNRELIABLE FOR THE CLASSIFICATION.

We see from above that all three modules in fact have a common objective but are realized in different ways based on different rules because one way or one rule cannot fully achieve the challenging objective. This common objective in all modules is to extract the most discriminative pattern representations or equivalently, to discard the redundant representations. This is some kind of dimensionality reduction based on some rules generated by human knowledge or machine learning (or both). To understand how the dimensionality reduction in the first two modules helps the final classification, let's explore a simple classification example graphically illustrated in Figure 2.

Suppose the circles and squares in Figure 2 represent the whole data population of two classes, respectively. A classifier can be easily trained by them to form a decision boundary shown by the red solid line, which perfectly classifies all data. Obviously, the dimension spanned by its normal vector ϕ (the green arrow) contains the most discriminative information and the one orthogonal to ϕ has hardly discriminative information. Nevertheless, this redundant dimension causes no harm to the classification because it is ignored by the classifier trained to extract the most discriminative information. Why do we need the first two modules to reduce the dimensionality or to extract the most discriminative pattern representations? It is well known that the probability of misclassification decreases or at least does not increase as the data dimensionality increases, as long as the decision is based on the knowledge about the whole data population. This was theoretically proven in [4], [12], and [13]. However, it is also well known that high dimensionality often degrades the classification performance in practice (curse of dimensionality) [4], [13]. This paradox can be resolved by distinguishing the discriminative information about the data population from that on the training set. The trained classifier can only capture the most discriminative information on the training data. If some statistics estimated on the training data deviate from those of the data population, the misclassification rate on the novel data increases. This is always the case in the practice. The question is only how severe it is. For example, if the available training data are only the blue solid points as shown in Figure 1, the decision boundary of the trained classifier will be the blue dashed line. The misclassification rate on the data population or on the novel data can approach the maximum 50%. The increasing probability of misclassification along with the increase of the data dimensionality for a fixed number of training samples was theoretically proven in [12] with a simple example.



[FIG2] A simple example showing the problem of classification with unrepresentative training samples. The decision boundary (the red solid line) trained by the circles and squares largely deviates from that (the blue dashed line) trained by the solid points.

If the first two modules can extract only the dimension ϕ based on some human knowledge about the whole data population, the classifier can easily perform a perfect classification in this one-dimensional subspace even if the solid points are the only available training data. This dimensionality reduction is quite possible if proper human knowledge such as some physical characteristics of the pattern is applied in the segmentation and feature extraction. However, if the dimensionality reduction is based on the machine learning from the training samples (the solid points), it cannot extract the right dimension ϕ based on any kind of the "most discriminative" criterion because it in principle just duplicates the classification process. Therefore, some criterion other than the most discriminative should be developed for the dimensionality reduction via machine learning. As the classifier is trained to capture some statistics on the training samples, a problem occurs if they are unreliable in some dimensions (largely deviating from those on the data population). To boost the subsequent classification accuracy or robustness, the dimensionality reduction should be targeted at circumventing this problem. Although the ultimate objective of all modules of a pattern recognition system is to extract the most discriminative information, it is the most discriminative information about the whole data population, not on a specific training set. A classifier is trained to capture the most discriminative information on the training samples. Therefore, to boost the classification accuracy, the dimensionality reduction should be targeted at removing the dimensions unreliable for the classification. Hence, to develop effective techniques of dimensionality reduction via machine learning, we need to study where the possible problem of a statistical classification lies.

PROBLEMS OF CLASSIFICATION, REGULARIZATION, AND SEMIDIMENSIONALITY REDUCTION

Classification is to assign a given novel pattern, here represented by a column vector $\mathbf{x} \in \mathbb{R}^n$ if no feature extraction is imposed, to one of the c categories, ω_i. The minimum probability of misclassification is achieved by assigning the pattern to the class that has the maximum probability after the pattern $\mathbf{x}$ has been observed, called a posteriori probability $P(\omega_i|\mathbf{x})$. This maximum a posterior (MAP) rule is a Bayes decision rule with the 0/1 loss function. It leads to the optimal classification called Bayes classification. As $P(\omega_i|\mathbf{x}) = P(\omega_\mathrm{i})p(\mathbf{x}|\omega_\mathrm{i})p^{-1}(\mathbf{x})$ and $p(\mathbf{x})$ is not a function of ω_i, the Bayes classification is to evaluate the discriminant functions that can be defined as

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_\mathrm{i}) + \ln P(\omega_\mathrm{i}) \qquad (1)$$

and find the class ω_i that has the maximum value of the discriminant function for a given pattern $\mathbf{x}$. Here, a natural logarithm ln is applied as it is a monotonically increasing function that does not affect the decision result but will simplify its evaluation if $p(\mathbf{x}|\omega_\mathrm{i})$ is an exponential function.

THE LARGE DEVIATIONS OF THE SMALL EIGENVALUES FROM THE POPULATION VARIANCES RESULT IN A SEVERE OVER-FITTING PROBLEM OF THE CLASSIFIER THAT GREATLY AFFECTS THE CLASSIFICATION ACCURACY ADVERSELY.

Further quantitative analysis needs an analytical form of the class-conditional probability function $p(\mathbf{x}|\omega_\mathrm{i})$. We take the multivariate Gaussian distribution as an example due to several reasons. First, it is the most natural distribution and the sum of a large number of independent random distributions obeys Gaussian distribution. It has the maximum uncertainty of all distributions having a given mean and variance. Moreover, it is an appropriate model for many situations, from handwritten characters to some speech sounds, where the data can be viewed as some prototype corrupted by a large number of random processes [4]. Multiprototype distribution can be well approximated by Gaussian mixture, the weighted sum of a number of Gaussian distributions. Last, dimensionality reduction techniques such as PCA and LDA and many classifiers are only specified by the second-order statistics, and so is the Gaussian distribution. Although LDA, Mahalanobis distance, and many classifiers are proven optimal only under Gaussian assumption, they are successfully employed in many applications. Under the Gaussian assumption

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i)\right], \qquad (2)$$

the discriminant function (1) becomes

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i) + \mathrm{b}_i. \qquad (3)$$

In practice, b_i is often not strictly determined by (1) and (2) but used as a threshold for users to control the error rate of class ω_i at a price of the other classes, e.g., to compromise between the false acceptance and false rejection rates in a biometric verification or object detection application.

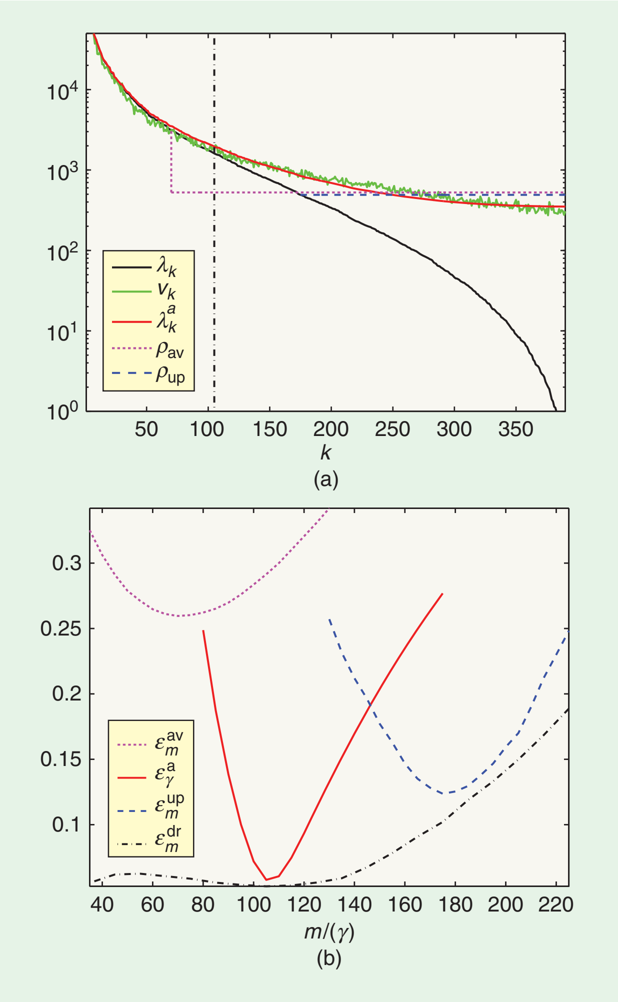
The problem is that human knowledge cannot provide the class mean $\bar{\mathbf{x}}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ of the data population, which can only be estimated or learned by machine from the available training samples. If some estimates largely deviate from those of the data population, we will face a large misclassification rate. From (3) we see that the discriminant function is very sensitive to the covariance matrix $\boldsymbol{\Sigma}_i$ because the data vector is multiplied by its inverse. However, it is very difficult to study the problems of $\boldsymbol{\Sigma}_i$ directly as it carries two different kinds of information by n^2 estimates: data variations and correlations. Eigen-decomposition provides an effective tool to simplify the problem. As the covariance matrix is symmetric, its eigenvectors provide an orthogonal basis for n-space. After applying eigen-decomposition, $\boldsymbol{\Phi}_i^T\boldsymbol{\Sigma}_i\boldsymbol{\Phi}_i = \boldsymbol{\Lambda}_i = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, the discriminant function (3) is simplified as

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)^T\boldsymbol{\Phi}_i\boldsymbol{\Lambda}_i^{-1}\boldsymbol{\Phi}_i^T(\mathbf{x}-\bar{\mathbf{x}}_i) + b_i \\ &= -\frac{1}{2}\sum_{k=1}^{n}\frac{(z_k-\bar{z}_k)^2}{\lambda_k} + b_i, \end{aligned} \qquad (4)$$

where z_k and $\bar{z}_k$ are respectively the projections of $\mathbf{x}$ and $\bar{\mathbf{x}}_i$ on the orthonormal eigenvector Φ_k corresponding to the eigenvalue λ_k of $\boldsymbol{\Sigma}_i$. For symbolic simplicity, the class index i is omitted where the index k is necessary. As an eigenvalue λ_k is the variance of the training samples of a class projected on the eigenvector Φ_k, it is an estimate of the class population variance based on the available training data. If it deviates from the population variance, the decision rule (3) or (4) overfits the training samples and hence leads to a poor generalization or prediction on the novel testing data. This problem will become very severe if some eigenvalues largely deviate from the population variances.

The black curve of Figure 3(a) shows an eigen-spectrum (λ_k sorted in descending order) obtained from 400 face images of size 20 × 20 and the green curve shows the variances v_k of other 8,500 face images (representing the face population) projected on the eigenvectors Φ_k. They are plotted in logarithm scale for comparison because we see from (4) that it is not the amount of the difference but the amount of the ratio between λ_k and v_k that affects the accuracy of the discriminant function (4). All images are taken from a face detection database used in [14]. Other sets of training images produce results very similar to Figure 3. It shows deviations between the eigenvalues and the population variances. One way to quantify this disparity over the range space is to compute $e(\lambda) = \mu\{(\ln v_k - \ln \lambda_k)^2\}_{1\le k\le r}$, where $\mu\{\cdot\}_{1\le k\le r}$ is a mean operator over $1 \le k \le r$ and r is the rank of $\boldsymbol{\Sigma}_i$.

Figure 3 shows significantly larger deviations of the smallest eigenvalues. This phenomenon was elucidated in [14], [15], and [16], where more examples on several other real data sets can be



[FIG3] Problems of eigenvalues and their regularization. Part (a) shows the eigen-spectrum λ_k and its regularized versions computed from 400 face images, and variances v_k of other 8,500 face images projected on the eigenvectors Φ_k. Part (b) shows the normalized disparity between the regularized eigen-spectrum and the variances v_k.

found. It seems to be a general problem verified in [14] by synthetic data with known true population variances. Although, in general, the largest sample-based eigenvalues are biased upwards and the smallest ones are biased downwards, the bias is more pronounced when the population variances tend toward equality, and it is correspondingly less severe when their values are highly disparate [15]. In most applications, population variances often first decay very rapidly and then stabilize so that the smallest eigenvalues are biased much more severely than the largest ones [14], [16]. This is evidenced by Figure 3. The large deviations of the small eigenvalues from the population variances result in a severe overfitting problem of the classifier that greatly affects the classification accuracy adversely.

One solution is to regularize the covariance matrix $\boldsymbol{\Sigma}_i$. A common practice in classification and data regression is to add a constant to its diagonal elements, $\boldsymbol{\Sigma}_i^a = \boldsymbol{\Sigma}_i + a\mathbf{I}$. We can let $a = \gamma\,\mathrm{trace}(\boldsymbol{\Sigma}_i)/r$ so as to select γ invariably to the data scale. The normalized disparity of the regularized eigen-spectrum $\varepsilon_\gamma^a = e(\lambda^a)/e(\lambda)$ against γ is shown by the red curve of Figure 3(b). Its minimum is $\varepsilon_{0.08}^a = 0.06$. The regularized eigen-spectrum λ_k^a with $\gamma = 0.08$ is shown by the red curve of Figure 3(a). Although this method was originally proposed to circumvent the singularity of $\boldsymbol{\Sigma}_i$ and the numerical instability of its inverse, we see from Figure 3 that the regularized eigen-spectrum can be very close to the population variances. It is thus not a surprise that numerous algorithms for classification, data regression, dimensionality reduction, and manifold learning adopt this classical technique [15], [17]–[19]. The underlying principle of $\boldsymbol{\Sigma}_i^a = \boldsymbol{\Sigma}_i + a\mathbf{I}$ can been seen by its equivalence to adding the constant to all eigenvalues $\lambda_k^a = \lambda_k + a$. From $(\lambda_k + a)/v_k = (1 + a/\lambda_k)\lambda_k/v_k$, we see that the factor $(1 + a/\lambda_k)$ is larger for smaller λ_k and smaller for larger λ_k. Therefore, the regularized eigen-spectrum can be very close to the population variances as shown in Figure 3. Problems of this method are the increased disparity of large eigenvalues and no dimensionality reduction effect. Either the $n \times n$ covariance matrix or the $n \times n$ eigenvector matrix is needed to compute the discriminant function (3) or (4).

Another solution, called probabilistic subspace learning [20], [21], decomposes the discriminant function (4) into two parts and replaces the small eigenvalues by a constant as

$$g_i(\mathbf{x}) = -\frac{1}{2}\left[\sum_{k=1}^{m}\frac{(z_k - \bar{z}_k)^2}{\lambda_k} + \sum_{k=m+1}^{n}\frac{(z_k - \bar{z}_k)^2}{\rho}\right] + b_i. \quad (5)$$

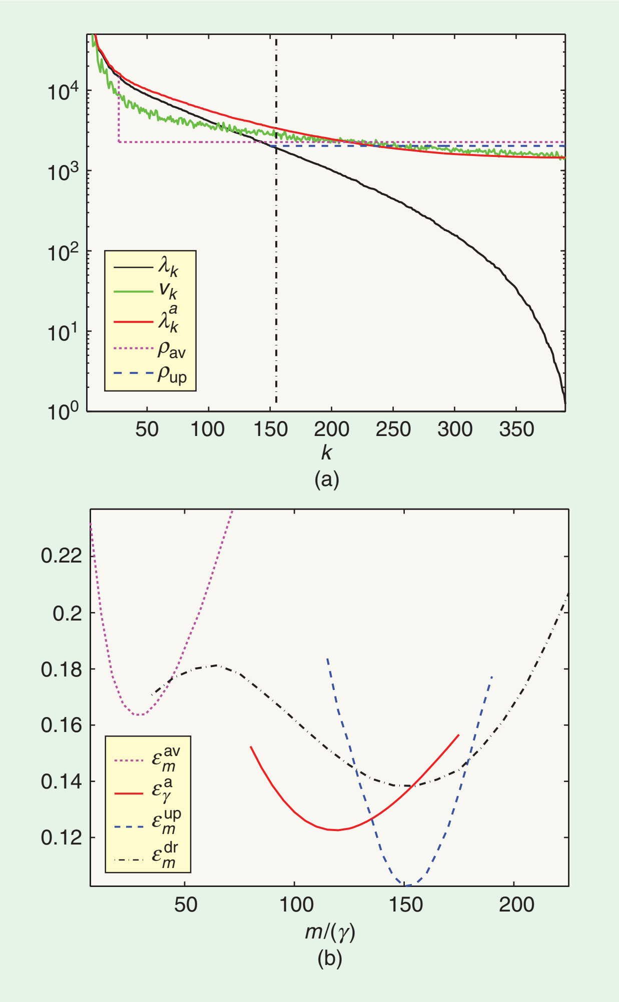
The constant is computed by $\rho_{av} = \mu\{\lambda_k\}_{m<k\le r}$ in [20] and [21] as it is the optimal approximation to λ_k for $m < k \le r$. This method leads to one of the best performers, called the Bayesian algorithm [22], in the face recognition community and is adopted in many other approaches of visual object recognition [23]–[25]. In fact, this method regularizes the eigen-spectrum by setting $\lambda_k^{\rho_{av}} = \rho_{av}$ for $m < k \le n$. The normalized disparity $\varepsilon_m^{av} = e(\lambda^{\rho_{av}})/e(\lambda)$ against m is shown by the magenta dotted curve of Figure 3(b). Its minimum is $\varepsilon_{70}^{av} = 0.26$. The regularized eigen-spectrum $\lambda_k^{\rho_{av}}$ for $70 < k \le n$ is shown by the magenta dotted line in Figure 3(a). We see a much greater disparity than λ_k^a. The problem is the computation of the constant ρ. The purpose of the regularization is not best approximating to the eigenspectrum but to the population variances. Eigenvalues in the subspace $m < k \le n$ are replaced by a constant ρ because they are unreliable, and so is their arithmetic average ρ_{av}. As they are biased downwards, it is proposed in [26] to use their upper bound as the constant $\rho_{up} = \max\{\lambda_k\}_{k>\mathrm{m}}$, which is also adopted in [27]. The normalized disparity $\varepsilon_m^{up} = e(\lambda^{\rho_{up}})/e(\lambda)$ against m is shown by the blue dashed curve of Figure 3(b). Its minimum is $\varepsilon_{175}^{up} = 0.12$. The regularized eigen-spectrum $\lambda_k^{\rho_{up}}$ for $175 < k \le n$ is shown by the blue dashed line in Figure 3(a). We see a much smaller disparity than λ_k^{av}, which is greater than ε^a in this example but smaller than it in another (Figure 4). The upper bound ρ_{up} leads to significantly higher face recognition accuracy than the average ρ_{av} [26].

In fact, this regularization has some role of dimensionality reduction as it is not necessary to project the data to the eigenvectors Φ_k for $k > m$. As we choose orthonormal eigenvectors, Euclidian distance between two vectors in the eigen-space is identical to that in the data space and hence,

$$\sum_{k=m+1}^{n} \frac{(z_k - \bar{z}_k)^2}{\rho} = \frac{1}{\rho}\left[(\mathbf{x} - \bar{\mathbf{x}}_i)^T (\mathbf{x} - \bar{\mathbf{x}}_i) - \sum_{k=1}^{m} (z_k - \bar{z}_k)^2 \right]. \qquad (6)$$

Thus, we only need $n \times m$ eigenvector matrix to compute (5) for classification. However, the n-dimensional class mean vectors $\bar{\mathbf{x}}_i$ are still required. We call it semidimensionality reduction.

Besides adding a constant to all eigenvalues or replacing the unreliable eigenvalues by a constant as discussed above, another regularization technique [16] replaces the unreliable eigenvalues $\lambda_k, m < k \leq r$ by a model $\alpha(k+\beta)^{-1}$, where α and β are two constants determined by the reliable eigenvalues. The rationale behind it is that the population variance is not constant in the unreliable subspace but decays much slower than the eigenvalue does. This decaying nature can be modeled by $\alpha(k+\beta)^{-1}$, which will be certainly closer to the population variances than the constant ρ_{av} or ρ_{up} if proper values of α and β are chosen.



[FIG4] Parts (a) and (b) show results of the same program as of Figure 3 but using 400 nonface training images and 8,500 nonface test images.

DIMENSIONALITY REDUCTION FOR REMOVING UNRELIABLE DIMENSIONS

Various regularization techniques that greatly improve the classification accuracy are evidenced by a large amount of publications. As analyzed in the last section, the underlying principle behind the regularization is that it reduces the disparity between the eigenvalues and the population variances and hence attenuates the overfitting problem. Obviously, we can also remove the unreliable dimensions to reduce the disparity in the remaining subspace. The normalized disparity of the eigen-spectrum in the subspace $(1 \leq k \leq m)$ against m, $\varepsilon_m^{dr} = e(\lambda^{dr})/e(\lambda)$, is shown by the black dot-dashed curve in Figure 3(b). The minimum is $\varepsilon_{105}^{dr} = 0.05$. The extracted and removed subspaces resulting in the minimum ε^{dr} are separated by a vertical black dot-dashed line in Figure 3(a). It shows that the dimensionality reduction effectively reduces the disparity because large disparity occurs at small eigenvalues. Therefore, similar to various regularization techniques that modify the smallest eigenvalues, removing the subspace spanned by the eigenvectors corresponding to the smallest eigenvalues improves the inference of the classifier, i.e., reduces the misclassification rate on the novel testing data.

However, this dimensionality reduction may also reduce the interclass distinction and the discriminant functions (3) or (4) of different classes in general should be evaluated in a common feature subspace for comparison. To extract a common subspace reliable for all classes and to prevent possible significant loss of the interclass distinction, we combine all class-conditional covariance matrices plus the covariance matrix of class mean to create a covariance mixture as

$$\boldsymbol{\Sigma}_\alpha = \sum_{i=1}^{c} \alpha_i \boldsymbol{\Sigma}_i + \eta \boldsymbol{\Sigma}_\mu, \qquad (7)$$

where α_i and η are weights and

$$\boldsymbol{\Sigma}_\mu = \sum_{i=1}^{c} \frac{q_i}{q} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^2. \qquad (8)$$

The covariance matrix of class mean $\boldsymbol{\Sigma}_\mu$ is also called interclass scatter matrix, where q_i is the sample size of class ω_i, and $\bar{\mathbf{x}}$ and q are respectively the mean and the sample size of the whole training set. Eigen-decomposition is then applied to the constructed covariance mixture

$$\boldsymbol{\Phi}^T \boldsymbol{\Sigma}_\alpha \boldsymbol{\Phi} = \boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}. \qquad (9)$$

If we remove a subspace spanned by eigenvectors corresponding to the smallest eigenvalues of $\boldsymbol{\Sigma}_\alpha$, it tends to remove unreliable dimensions of all class-conditional covariance matrices $\boldsymbol{\Sigma}_i$ and retain large interclass distinction residing in a subspace that has large eigenvalues of $\boldsymbol{\Sigma}_\mu$. Therefore, classification on the m-dimensional feature vector

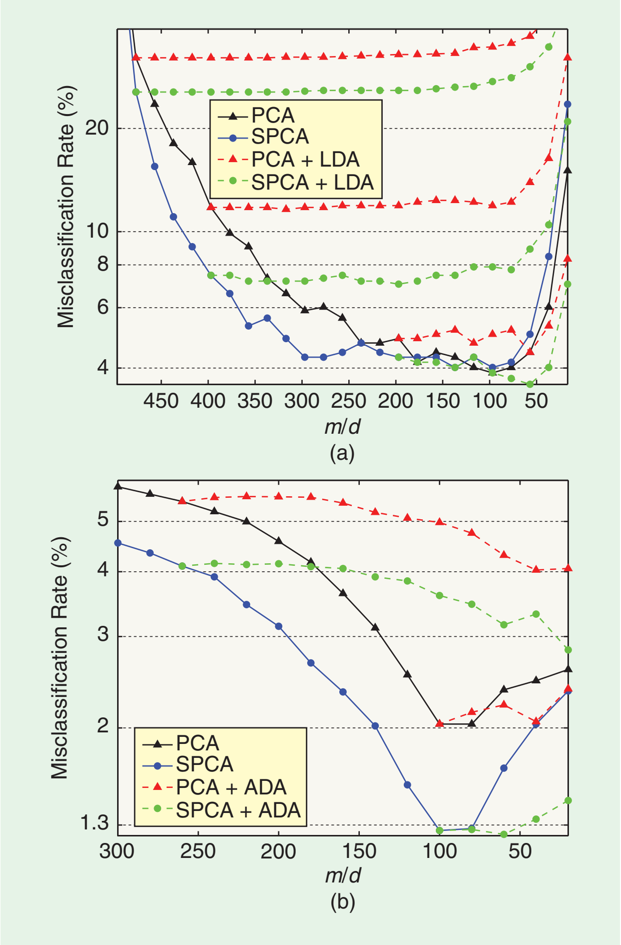
$$\mathbf{f} = \boldsymbol{\Phi}_m^T \mathbf{x} \qquad (10)$$

is most likely to perform better than on the n-dimensional data vector $\mathbf{x}$, where $\mathbf{\Phi}_m$ consists of m eigenvectors corresponding to the m largest eigenvalues of the covariance mixture $\mathbf{\Sigma}_\alpha$. If we regard (7), (9), and (10) as a separate module called dimensionality reduction, the subsequent classification (3) is simplified in the m-dimensional subspace

$$g_i(\mathbf{f}) = -\frac{1}{2}(\mathbf{f}-\bar{\mathbf{f}}_i)^T \mathbf{\Sigma}_{fi}^{-1}(\mathbf{f}-\bar{\mathbf{f}}_i) + b_i. \qquad (11)$$

It is not necessary to project training samples into the subspace as $\bar{\mathbf{f}}_i = \mathbf{\Phi}_m^T \bar{\mathbf{x}}_i$ and $\mathbf{\Sigma}_{fi} = \mathbf{\Phi}_m^T \mathbf{\Sigma}_i \mathbf{\Phi}_m$. The objective of the dimensionality reduction by (7), (9), and (10) is to facilitate an effective removal of the unreliable dimensions and hence boost the classification accuracy (11). Thus, larger values of α_i should be assigned to less reliable covariance matrices so that more dimensions characterized by the smallest eigenvalues of less reliable classes can be removed by the eigen-decomposition of $\mathbf{\Sigma}_\alpha$.

The weights α_i are critical in some applications because different classes may have different characteristics and hence the reliability of the estimated covariance matrices can be significantly different. Figure 4 is generated by the same program as Figure 3 but uses 400 and 8,500 nonface training and testing images, respectively, from a face detection database used in [14]. Other partitions of the training and testing sets produce very similar results to Figure 4. It shows much larger disparity between the eigenvalue and the population variance than that of Figure 3. More examples can be found in [14]. This is a general problem caused by the different characteristics of different classes. In the applications of biometric verification and object detection, for example, the positive and negative classes are highly asymmetric because the former represents only one particular person or object while the latter represents the whole "rest of the world" that contains all other people or objects. Thus, it is much more difficult to collect a representative training set for the negative class than for the positive one. This often results in a larger eigenvalue bias of the negative class. Furthermore, as pointed out in [15] and further evidenced by Figures 3 and 4, the bias is more pronounced when population variances tend toward equality, and less severe when their values are highly disparate. This is also verified in [14] by synthetic data with known true population variances. As the negative class occupies a much larger subspace and hence has flatter eigen-spectrum, in general, we need to assign a larger weight to the negative class than to the positive one.



[FIG5] Misclassification rate against the reduced dimensionality by PCA/SPCA and PCA/SPCA+LDA/ADA of (a) face identification and (b) face detection problems. The left-most point of each dashed curve indicates the dimensionality *m* of the PCA/SPCA subspace in which LDA/ADA further reduces it to *d* indicated by the other points on the same dashed curve.

It is very interesting to see that if we set $\eta = 1$ and $\alpha_i = q_i/q$, the constructed covariance mixture $\mathbf{\Sigma}_\alpha$ will be identical to the covariance matrix $\mathbf{\Sigma}_t$ of all training data without considering their class labels. It is also called a total scatter matrix. This shows that the well-known PCA is a specific case of the aforementioned dimensionality reduction method. Therefore, this study also reveals the underlying principle of why PCA, though an unsupervised method that minimizes the data reconstruction error rather than maximizes the class discrimination, can improve the classification accuracy. Although many approaches apply PCA only aimed at circumventing the singularity problem of the intraclass scatter matrix for the subsequent discriminant analysis, as analyzed above, the role of PCA for classification is in fact far beyond that. Figure 5 (refer to the experimental section) demonstrates the significant gains in classification accuracy by using PCA to reduce the dimensionality much lower than the rank of the intraclass scatter matrix. More evidence can be found in the experimental results of [8], [14], [19], [25], and [28].

Nevertheless, PCA is not optimized for classification. The weights $\eta = 1$ and $\alpha_i = P(\omega_i)$ or $\alpha_i = q_i/q$ are required for PCA to achieve the least-mean-square data reconstruction error, which is irrelevant to classification. Our objective for classification is to remove the unreliable dimensions in which the sample-based class-conditional variances are largely deviate from the population variances. The reliability of a covariance matrix does not depend on the class prior probability. More training samples of a class may result in a more reliable covariance matrix if they are properly collected. However, it is the less reliable covariance matrix that should be heavier weighted in the covariance mixture so that more dimensions characterized by the small variances of this class can be removed. From the analysis, we see that PCA helps improve the classification accuracy, not because it minimizes the data reconstruction error, but because it has some roles in removing the unreliable dimensions. As its objective is not from the classification point of view, PCA may not effectively remove the unreliable dimensions. In sharp contrast to PCA that weights $\boldsymbol{\Sigma}_i$ proportionally to q_i, it is suggested in [14] to pool $\boldsymbol{\Sigma}_i$ with weights inversely proportional to q_i if there is no prior knowledge about the class characteristics and the data collection procedure. Even $\eta = 1$ in PCA may not be optimal for classification. Although a larger value of η ensures less loss of the interclass distinction, it leads to less effective removal of the unreliable dimensions. Hence, more dimensions have to be removed, which in turn results in more loss of the interclass distinction. The aforementioned limitations of PCA are verified by the experimental results shown in Figure 5.

PCA is an unsupervised technique, as no class label is needed. For a two-class problem, dimensionality reduction (7), (9), (10) is called asymmetric principal component analysis (APCA) [14] due to the asymmetric treatment of the two covariance matrices. For a multiple-class problem, more generally, we call it supervised principal component analysis (SPCA), as it utilizes the class label and other class-specific information by imposing different weights on the covariance matrices. The optimal values of weights are application dependent. The objective of the SPCA (7), (9), (10) is to effectively remove the unreliable dimensions and hence boost the classification accuracy (11). Thus, it may not greatly reduce the dimensionality for a fast classification in some applications. In addition, APCA or SPCA may not work well for a classifier that neither explicitly nor implicitly weights the feature by the inverse of its variance, such as the classical nearest-neighbor classifier (NNC) with Euclidian distance and the sparse representation-based classifier (SRC) where the ℓ^1-minimization is applied [1], [2].

DIMENSIONALITY REDUCTION BY RETAINING DISCRIMINATIVE DIMENSIONS

As discussed in the last two sections, if the dimensionality reduction is aimed at enhancing the inference accuracy of the subsequent classification, it should be targeted at removing the unreliable dimensions. In some applications, we need to reduce a very high dimensional data vector to a very low dimensional feature vector to facilitate a simple and fast classification. This can be effectively achieved by extracting the most discriminative dimensions, which ensures the minimum loss of the discriminative information in the extracted subspace among all other subspaces of the same dimensionality. Linear discriminant analysis and its various variants are the most widely studied approaches.

In the identification applications, we often have a large number of classes with only a few samples per class for training so that each individual $\boldsymbol{\Sigma}_i$ is extremely unreliable. One solution to regularize them is to pool them together to form a common covariance matrix, $\boldsymbol{\Sigma}_w = \sum_{i=1}^{c} q_i \boldsymbol{\Sigma}_i / q$, which is also called intraclass scatter matrix. The discriminant function (3) is thus simplified as

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)^T \boldsymbol{\Sigma}_w^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i) + \mathrm{b}_i = \mathbf{x}^T \boldsymbol{\Sigma}_w^{-1} \bar{\mathbf{x}}_i + t_i, \quad (12)$$

where t_i absorbs all terms that is either constant to $\mathbf{x}$ or constant to i. We see that it is a linear function of $\mathbf{x}$ and hence the decision boundary $g_i(\mathbf{x}) = g_j(\mathbf{x})$ between any two classes ω_i and ω_j is a hyperplane specified by its normal vector $\boldsymbol{\psi}_{ij} = \boldsymbol{\Sigma}_w^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)$ and the threshold $t_i - t_j$. This means that for the optimal classification between two classes ω_i and ω_j, only one dimension spanned by $\boldsymbol{\psi}_{ij}$ is necessary. Thus, under the constraint of the linear classification, this dimension contains the most (in fact, all) discriminative information to differentiate class ω_i and ω_j. It is easy to see that the training data in this dimension have the maximum ratio κ between the interclass and intraclass variances. Therefore, we can define this ratio κ as a discriminant value to assess the discriminating power of a dimension. Although we need $(c-1)!$ hyperplanes to classify c classes, their normal vectors $\boldsymbol{\psi}_{ij} = \boldsymbol{\Sigma}_w^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)$ only span a $(c-1)$-dimensional subspace as only $c-1$ of them are linear independent. Therefore, we can reduce the n-dimensional data space to this $(c-1)$-dimensional subspace without losing any discriminative information as the linear classification (12) produces exactly the same results in the two spaces. However, if the dimensionality is reduced to d, $d < c-1$, some discriminative information will be lost. The subspace spanned by the eigenvectors corresponding to the d largest eigenvalues of the matrix $\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_\mu$ contains the most discriminative information among all possible d-dimensional subspaces for the linear classification (12) because an eigenvalue of $\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_\mu$ is the ratio κ between the interclass and intraclass variances in the dimension spanned by the corresponding eigenvector. This is the well-known LDA that performs the eigen-decomposition

$$\boldsymbol{\Psi}^T \boldsymbol{\Sigma}_w^{-1} \boldsymbol{\Sigma}_\mu \boldsymbol{\Psi} = \boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_\mathrm{n}\}. \quad (13)$$

We see from the above analysis that the objective of LDA is to find the one among all possible d-dimensional subspaces in which the linear classification (12) achieves the closest result to that in the original n-space. It is undoubtedly an effective method to largely reduce the data dimensionality with the minimum loss of the classification capability in a linear sense.

For a two-category classification problem, LDA can only extract one dimension. It is insufficient for a reasonable classification for some problems such as various tasks of verification and object detection because the two class-conditional covariance matrices are significantly different and hence the optimal classification is obviously not linear. To apply the discriminant analysis in such problems, an asymmetric discriminant analysis (ADA) is proposed in [14] to extract a rich number of features. It solves the following eigen-decomposition problem:

$$\boldsymbol{\Psi}^T(\boldsymbol{\Sigma}_1 + \beta\boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\Sigma}_1 + \gamma\boldsymbol{\Sigma}_\mu)\boldsymbol{\Psi} = \boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\} \quad (14)$$

in the APCA subspace. The underlying principle is that the discriminative information is not only carried by the distinction of the two class means but also by the distinction of the two class variances. The constant γ weights the discriminative information about the class mean against that about the variance. The asymmetry of the two classes is balanced by the constant β. It is proven in [14] that the ADA with $\gamma = \beta = 1$ maximizes the Bhattacharyya distance [29] between two classes in the subspace spanned by the eigenvectors corresponding to the largest $\max(\lambda_k, 1 - \lambda_k)$. Note that, different from those in the last two sections, eigenvectors of LDA or ADA are not orthogonal. The Euclidian distance in a space using LDA/ADA eigenvectors as a base will be different from that using an orthogonal base. However, it is easy to show that the Mahalanobis distance is not affected by the orthogonality of the base.

As the rank of $\boldsymbol{\Sigma}_w$ is at most $\min(n, q - c)$, $\boldsymbol{\Sigma}_w$ is often singular in some applications so that the discriminant value of LDA (13) and ADA (14) cannot be evaluated. Numerous variants or generalizations of LDA have been proposed to circumvent this problem, which are summarized under a common framework graphically [8] and algebraically [19]. A popular approach called Fisherface or Fisher LDA (FLDA) [30] applies PCA so as to make $\boldsymbol{\Sigma}_w$ nonsingular before LDA. Another approach called direct LDA (DLDA) [17] removes null space of $\boldsymbol{\Sigma}_\mu$ and extracts the eigenvectors corresponding to the smallest eigenvalues of $\boldsymbol{\Sigma}_w$. This is under the assumption that the most discriminative information resides in the range space of $\boldsymbol{\Sigma}_\mu$. NLDA [5] extracts features from the null space of $\boldsymbol{\Sigma}_w$. Interestingly, this appears to contradict the popular FLDA that only uses the range space and discards the null space of $\boldsymbol{\Sigma}_w$. A common aspect of all these methods is that they all remove some dimensions, either in the principal or the null space, before the LDA process. It is difficult to compare the effectiveness of the aforementioned LDA variants because we see from (13) that both $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Sigma}_\mu$ contribute to the discriminant value κ in a dimension. NLDA and DLDA appear to retain more discriminative information as any dimension in the intersection of the null space of $\boldsymbol{\Sigma}_w$ and the range space of $\boldsymbol{\Sigma}_\mu$ has infinite discriminant value κ according to (13). DLDA ensures the class mean distinction $\boldsymbol{\Sigma}_\mu$ untouched in the first stage. However, small and zero eigenvalues of $\boldsymbol{\Sigma}_w$ are unreliable that may cause severe problem as we analyzed in the last two sections. Just a small decrease or increase in the number of training samples may greatly change them. Furthermore, the most discriminative dimensions are not restricted within the range space of $\boldsymbol{\Sigma}_\mu$ or the null space of $\boldsymbol{\Sigma}_w$. Therefore, the above LDA variants are criticized in the literature [27], [31], [32] as a significant amount of discriminative information could be lost before the LDA process.

To avoid losing discriminative information before the LDA process, the dual-space LDA approach (DSL) [31], [32] performs LDA on the principal space of $\boldsymbol{\Sigma}_w$ and its complementary space separately and combines the two sets of the extracted LDA features. Obviously, it is suboptimal to extract features separately from the two subspaces. Furthermore, how to fuse the two feature sets properly is an open problem as they do not share the same metric measurement. Features from the principal space, $k \leq m$, are weighted by the inverse of their intraclass variance and those from the complementary space, $k > m$, are equally weighted by some constant. From the last two sections, we see that this feature weighting is problematic in the principal space for a large value of m and is problematic in the complementary subspace for a small value of m. One solution to these problems is first to partition the data space into three subspaces: reliable, unreliable, and null space of $\boldsymbol{\Sigma}_w$, then to regularize the eigenvalues differently in these three subspaces and finally to apply LDA in the whole space [16]. Consistent gains in face recognition accuracy of this approach were reported in [16]. Another way [33] to avoid losing information of $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Sigma}_\mu$ before the discriminant evaluation is to modify the LDA (13) to

$$\boldsymbol{\Psi}^T \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\Sigma}_\mu \boldsymbol{\Psi} = \boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}. \tag{15}$$

As $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_w + \boldsymbol{\Sigma}_\mu$ and hence the null space of $\boldsymbol{\Sigma}_t$ is the intersection of the null spaces of $\boldsymbol{\Sigma}_\mu$ and $\boldsymbol{\Sigma}_w$, no discriminative information is lost by evaluating (15) in the range space of $\boldsymbol{\Sigma}_t$. However, (15) deviates from the LDA (13) and hence the extracted subspace may not be the most discriminative in a sense of LDA or of the classification (12). Moreover, it puts an undue emphasis on the null space of $\boldsymbol{\Sigma}_w$ as the discriminant value from $\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Sigma}_\mu$ in the range space of $\boldsymbol{\Sigma}_w$ ($\kappa < 1$) is always smaller than that in its null space ($\kappa = 1$). In addition, there is also a problem of how to properly scale the features from the principal and null spaces of $\boldsymbol{\Sigma}_w$, which may not be of full rank even in the reduced subspace.

Most aforementioned approaches focus on the singularity problem of $\boldsymbol{\Sigma}_w$. In fact, as analyzed in the last two sections, the unreliability, bias, and instability of the small eigenvalues of $\boldsymbol{\Sigma}_w$ cause great problems wherever its inverse is applied in the discriminant evaluation (13), (14) or the classification (3), (12). Any regularization or dimensionality reduction technique discussed in the last two sections can be applied to attenuate this problem before applying discriminant analysis to further reduce the dimensionality for a fast classification. Significant gains in classification accuracy were reported by applying various regularization techniques in the LDA approaches [15], [16], [19], [27], [34]. Also, great gains in classification accuracy were reported by applying PCA, APCA, or SPCA to reduce the data dimensionality much smaller than the rank of the $\mathbf{S}_w$ before applying LDA or other discriminative methods [8], [14], [19], [25], [28].

EXPERIMENTAL STUDIES

The different roles of dimensionality reduction by PCA, SPCA, and LDA/ADA for pattern recognition are further explored in two experiments. One is a face identification problem on a data set [16] extracted from the facial recognition technology (FERET) database with many classes (1,194 people) and only two samples per class, and the other is a face detection problem in the database used in [14] with only two classes (face and nonface) and many (9,000) samples per class. Images are cropped into the size of 33 × 38 for the identification problems and 20 × 20 for the detection problem. In the identification experiment, 497 people are randomly selected for training, the remaining 697 people are used for testing, and the linear classifier (12) is applied in the feature space. In the detection experiment, four experiments, each with a distinct 25% images as testing set and the remaining images as training set, are conducted and the average misclassification rate over the four distinct testing sets is computed. The detection applies the quadratic classifier (11) for PCA-related approaches and its asymmetric version with $\beta = 0.75$ [14] for SPCA-related approaches, where b_i is set so that the two classes have the same misclassification rate. For the identification problem, as there is no ground for significantly different distributions of different persons, the same parameter $\alpha_i = 1/497$ is chosen for $\mathbf{\Sigma}_\alpha$. We choose $\eta = 1/4$ to differentiate the covariance mixture $\mathbf{\Sigma}_\alpha$ significantly from the total scatter matrix $\mathbf{\Sigma}_t$ where $\eta = 1$. For the detection problem, we choose $\eta = 1$ (same as in PCA) but $\alpha_1 = 1/5$ (for face class) and $\alpha_2 = 4/5$ (for nonface class) to remove significantly more unreliable dimensions of the nonface class in the SPCA stage as discussed in the last section. For ADA, $\gamma = 10$ and $\beta = 0.75$ is chosen [14]. Figure 5 shows the misclassification rates against the dimensionality m reduced by PCA and SPCA and d reduced by PCA+LDA/ADA and SPCA+LDA/ADA. The most left point of each dashed curve indicates the dimensionality m of the PCA/SPCA subspace, in which the LDA/ADA further reduces it to d indicated by the other points on the same dashed curve.

The experimental results shown in Figure 5 further verify the analysis of this article. It is the regularization technique or the dimensionality reduction by the supervised principal component analysis (including PCA) that plays the most vital role in boosting the classification accuracy while the discriminative method can greatly reduce the dimensionality with the minimum loss of the discriminative information. The question may arise as to why NLDA can work well in some applications if the smallest and zero eigenvalues are the most unreliable. The reason behind it is that the classification of the NLDA features does not use the variance due to zero eigenvalues in all dimensions of the null space. Thus, it implicitly circumvents the problem of the unreliable small eigenvalues to a certain extent by evenly weighting all features. Another question is why some approaches using LDA alone can also work well on some data sets. The underlying causes include the avoidance of feature scaling in the classification and the linearity of LDA but the nonlinearity of the classifier. These approaches, though applying LDA (13) for feature extraction, do not apply its origin (12) as classifier. Most of them apply the NNC with Euclidian distance. While the simple Euclidian distance ignores the data variance and hence circumvents the problem of the unreliable small eigenvalues to a certain extent, the complex data distribution is captured by the NNC that computes all distances from a novel pattern to all training samples. The NNC, though very simple, is highly nonlinear, can form arbitrary complex, nonlinear decision boundary and classifies all training samples without error. LDA restricts such highly nonlinear classifier to a subspace, which is, though the most discriminative, only in a linear sense. This restriction has similar role to the regularization. Therefore, the improvement of the classification accuracy by LDA is most likely contributed by its linearity constraint rather than its most discriminative nature. However, LDA that represents the class distinction by using the difference of class mean only may impose too strict constraint on some complex data structure. Therefore, some approaches that utilize the locality and neighborhood of the training samples such as LPP [6], [7] and MFA [8] extract more discriminative features than LDA. Nevertheless, experiments in [7] and [8] still show that a PCA stage either is necessary to "remove the noise" [7] or significantly improves the performance [8] of these discriminative approaches.

CONCLUSIONS

To recognize unknown data, a pattern recognition system is designed based on the human knowledge about the data population and the machine learning from the known training samples. The difficult recognition task is performed in several stages. Classification as the last stage is mainly trained by the available training samples. Thus, it extracts the most discriminative information on the training data, which in general deviates from that about the whole data population as only a finite set of training samples is applicable. This deviation increases the misclassification rate on the novel data. The problem becomes very severe if the data lie in a high-dimensional space. Moreover, high dimensionality also makes it difficult to apply sophisticated classifiers. Linear subspace learning-based dimensionality reduction provides a powerful tool to circumvent these problems. It also serves as a solid foundation for various nonlinear approaches. Dimensionality reduction as an intermediate stage of a pattern recognition process has two objectives. One is to reduce the computational complexity of the subsequent classification with the minimum loss of the discriminative information, and the other is to circumvent the over-fitting problems of the classification and hence enhance its inference accuracy and robustness.

To achieve the first objective, we need to maximize the discriminative information in the reduced low-dimensional space. Discriminative approaches such as LDA, NLDA, DLDA, ADA, LPP, MFA, and their various variants can undoubtedly reduce the data dimensionality in large scale with the minimum loss of the discriminative information. Since these approaches in general have similar objective to that of classification, i.e., extracting the most discriminative information on the training samples, problems of misclassification on novel data or poor generalization/inference capability caused by the high dimensionality of the data may not be effectively circumvented. However, some constraints on these discriminative approaches such as the linearity and the limitation to the zero intraclass variation, which are not imposed on the subsequent classification, play some roles in improving the classification accuracy.

The second objective cannot be effectively achieved only based on the consideration of some general phenomena, such as the curse of dimensionality, small sample size problem, noise removal effect of the dimensionality reduction and better generalization in a lower dimensional space. For an effective dimensionality reduction, we have to find out which dimensions are more problematic or harmful than others for a robust classification and hence should be removed. It is shown that the smallest eigenvalues of the class-conditional covariance matrix have the largest deviation from the population variances and hence cause the most severe problem in classification and LDA/ADA evaluation.

Therefore, regularization of these unreliable statistics or removal of the corresponding dimensions by SPCA greatly enhances the classification accuracy. They also help the discriminant evaluation of LDA, ADA, LPP, and MFA to find a portable set of reliable and most discriminative dimensions. However, they may not be effective for a classifier that neither explicitly nor implicitly weights the feature by the inverse of its variance, such as the classical NNC with Euclidian distance and the sparse representation-based classifier SRC.

As regularization does not reduce or fully reduce the data dimensionality and the removal of the unreliable dimensions by SPCA may not lead to a portable feature vector, discriminative approaches such as LDA, ADA, LPP, MFA, and their variants can be followed to greatly reduce the dimensionality for a simple and fast classification. Although various regularization techniques are also applied in many classifiers, they should be applied before the dimensionality reduction because the regularization in the classification stage cannot recover the improperly removed dimensions in the dimensionality reduction stage. With the in-depth understanding of the roles of dimensionality reduction for pattern recognition and the underlying principles revealed in this article, it is not a surprise that most top performers of the state-of-the-art techniques either apply various regularized discriminative analyses or apply two-stage approaches, such as PCA+LDA, PCA+LPP, SPCA+ADA, and PCA+MFA to accomplish the both objectives of the dimensionality reduction.

ACKNOWLEDGMENT

This work was supported in part by Singapore National Science and Engineering Research Council Thematic Strategic Research Programme Grant 062 130 0056.

AUTHOR

Xudong Jiang (exdjiang@ntu.edu.sg) received the B.Sc. and M.Sc. degrees from the University of Electronic Science and Technology of China in 1983 and 1986, respectively, and the Ph.D. degree from Helmut Schmidt University, Hamburg, Germany, in 1997, all in electrical and electronic engineering. From 1998 to 2004, he was a lead scientist and head of the Biometrics Laboratory at the Institute for Infocomm Research, A*Star, Singapore. He has been a faculty member since 2003 and is currently a tenured associate professor and director of the Centre for Information Security at Nanyang Technological University, Singapore. He has published over 90 papers in journals and conferences. His research interests include pattern recognition, computer vision, signal and image processing, and biometrics. He is a Senior Member of the IEEE.

REFERENCES

[1] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[2] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, June 2010.

[3] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 12, pp. 1945–1959, Dec. 2005.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2001.

[5] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 1, pp. 4–13, Jan. 2005.

[6] M. Belkin and P. Niyogi, "Towards a theoretical foundation for Laplacian-based manifold methods," *J. Comput. Syst. Sci.*, vol. 74, no. 8, pp. 1289–1308, 2008.

[7] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using faces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[8] S. Yan, D. Xu, B. Zhang, Q. Yang, H. Zhang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[9] X. D. Jiang, "On orientation and anisotropy estimation for online fingerprint authentication," *IEEE Trans. Signal Processing*, vol. 53, no. 10, pp. 4038–4049, Oct. 2005.

[10] X. D. Jiang, "Extracting image orientation feature by using integration operator," *Pattern Recogni.*, vol. 40, no. 2, pp. 705–717, Feb. 2007.

[11] P. Yap, X. D. Jiang, and A. Kot, "Two dimensional polar harmonic transforms for invariant image representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 7, pp. 1259–1270, July 2010.

[12] G. V. Trunk, "Problem of dimensionality: A simple example," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 1, no. 3, pp. 306–307, July 1979.

[13] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[14] X. D. Jiang, "Asymmetric principal component and discriminant analyses for pattern classification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 5, pp. 931–937, May 2009.

[15] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, Mar. 1989.

[16] X. D. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 3, pp. 383–394, Mar. 2008.

[17] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recogn. Lett.*, vol. 26, no. 2, pp. 181–191, 2005.

[18] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Machine Learn. Res.*, vol. 7, no.11, pp. 2399–2434, 2006.

[19] S. Ji and J. Ye, "Generalized linear discriminant analysis: A unified framework and efficient model selection," *IEEE Trans. Neural Netw.*, vol. 19, no. 10, pp. 1768–1782, Oct. 2008.

[20] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 696–710, July 1997.

[21] B. Moghaddam, "Principal manifolds and probabilistic subspace for visual recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 6, pp. 780–788, June 2002.

[22] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognit.*, vol. 33, no. 11, pp. 1771–1782, Nov. 2000.

[23] K. K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 1, pp. 39–51, 1998.

[24] C. Liu, "A Bayesian discriminating features method for face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 6, pp. 725–740, 2003.

[25] X. Wang and X. Tang, "A unified framework for subspace face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 9, pp. 1222–1228, Sept. 2004.

[26] X. D. Jiang, B. Mandal, and A. Kot, "Enhanced maximum likelihood face recognition," *Electron. Lett.*, vol. 42, no. 19, pp. 1089–1090, Sept. 2006.

[27] X. D. Jiang, B. Mandal, and A. C. Kot, "Complete discriminant evaluation and feature extraction in kernel space for face recognition," *Mach. Vis. Appl.*, vol. 20, no. 1, pp. 35–46, Jan. 2009.

[28] A. Martinez and A. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 2, pp. 228–233, 2001.

[29] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.

[30] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 711–720, July 1997.

[31] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: A complete kernel fisher discriminant framework for feature extraction and recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 2, pp. 230–244, Feb. 2005.

[32] W. Zheng and X. Tang, "Fast algorithm for updating the discriminant vectors of dual-space LDA," *IEEE Trans. Inform. Forensics Security*, vol. 4, no. 3, pp. 418–427, Sept. 2009.

[33] J. Ye, R. Janardan, C. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 8, pp. 982–994, Aug. 2004.

[34] B. Mandal, X. D. Jiang, H. Eng, and A. Kot, "Prediction of eigenvalues and regularization of eigenfeatures for human face verification," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 717–724, June 2010.

[SP]