Pattern Recognition 43 (2010) 1269–1281

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A regularization framework for robust dimensionality reduction with applications to image reconstruction and feature extraction

Zhizheng Liang [a,]*, Youfu Li [b]

[a] School of Computer Science and Technology, China University of Mining and Technology, China
[b] Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 26 March 2009
Received in revised form
12 October 2009
Accepted 16 October 2009

Keywords:
Regularization framework
Nonlinear eigenvalue problem
SCF iteration
Robust
Feature extraction
Image reconstruction

ABSTRACT

Dimensionality reduction has many applications in pattern recognition, machine learning and computer vision. In this paper, we develop a general regularization framework for dimensionality reduction by allowing the use of different functions in the cost function. This is especially important as we can achieve robustness in the presence of outliers. It is shown that optimizing the regularized cost function is equivalent to solving a nonlinear eigenvalue problem under certain conditions, which can be handled by the self-consistent field (SCF) iteration. Moreover, this regularization framework is applicable in unsupervised or supervised learning by defining the regularization term which provides some types of prior knowledge of projected samples or projected vectors. It is also noted that some linear projection methods can be obtained from this framework by choosing different functions and imposing different constraints. Finally, we show some applications of our framework by various data sets including handwritten characters, face images, UCI data, and gene expression data.

1. Introduction

In data analysis problems where there are a large number of input variables, it is often beneficial to reduce the dimension of data in order to improve the efficiency and accuracy of data analysis. Consequently, dimensionality reduction becomes one of key techniques in data analysis. Dimensionality reduction aims at reducing the dimensionality of data such that the extracted features are as representable as possible. During the past several decades, a variety of algorithms and techniques [1–8] for dimensionality reduction have been developed. Among them, principal component analysis (PCA) and linear discriminant analysis (LDA) are regarded as the most powerful tools of dimensionality reduction. In general, PCA is to find an orthogonal set of vectors by maximizing the variance of the projected data, whereas LDA is to seek discriminant vectors by maximizing the ratio of the between-class distance to the within-class distance. It is shown that LDA is a more effective method for extracting features in the classification problem as compared to PCA in general cases. However, LDA often suffers from the small sample size (3S) problem when the dimension of data is much larger than the number of data points.

In recent years, many approaches [10–16] have been proposed to deal with high dimensional data and the 3S problem. For example, the Fisherface method [2] first applies PCA to reduce the dimension of samples to obtain a full-rank within-class scatter matrix. Then standard LDA is used to extract features. In [15], Chen et al. proposed the null space-based LDA, where the between-class scatter is maximized in the null space of the within-class scatter matrix. In [12], Howland and Park proposed the LDA/GSVD algorithm which circumvents the singularity problem by using the generalized singular value decomposition. Direct LDA [17] first removes the null space of the between-class scatter matrix and then seeks the projection to minimize the within-class scatter. In order to reduce the computational cost of LDA, Ye and Li [5] proposed a two-stage LDA extension (LDA/QR). Their method first applies the QR decomposition on a small matrix, and then followed by LDA. Further, Zhang and Sim [10] analyzed LDA via the Fukunaga–Koontz transform, which provides a unified framework for understanding some variants of LDA. In [14], Li et al. proposed an efficient and stable method to calculate discriminant vectors based on the maximum margin criterion (MMC). The difference between Fisher's criterion and MMC is that the former maximizes the Fisher quotient while the latter maximizes the average distance. In [18], the authors proposed a unified framework for generalized LDA via a transfer function. It is shown that uncorrelated LDA is a special case of PCA plus LDA and regularized LDA.

Although PCA and LDA have been successfully used in solving some problems in pattern recognition and machine learning, they are prone to the presence of outliers due to the fact they do not involve robust functions in the cost function. In order to deal with this problem, some researchers proposed robust algorithms [19–24] for dimensionality reduction in recent years. In [24], the

* Corresponding author.
E-mail address: cuhk_liang@yahoo.cn (Z. Liang).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.10.012

authors formulated matrix factorization as an L1 norm minimization problem, which can be efficiently solved by alternate convex programming. However, the solution does not have rotational invariance. Considering this point, the authors [21] proposed rotational invariant L1 norm principal component analysis which combines some merits of PCA and L1 PCA [24]. Their method can suppress the effect of outliers by defining a modified covariance matrix which softens contributions from outliers. In [19], the authors proposed a method of principal component analysis based on a new L1 norm optimization technique. The L1 norm optimization algorithm is robust to outliers and is easy to implement.

Note that the algorithms for robust PCA are to minimize the error between original data and reconstructed data in terms of different objective functions. However, they may produce undesirable classification performances due to the fact they are devised from the viewpoint of data reconstruction. Furthermore, they do not make full use of prior knowledge of data points such as the geometrical structure of data points. To this end, we develop a regularization framework of discriminant analysis by using prior knowledge of data points. In this framework, one can flexibly choose robust functions to suppress the presence of outliers. Moreover, a regularization parameter is used to control the tradeoff between the data reconstruction error and prior knowledge of data points. It is found that the optimization problem can be formulated as a nonlinear eigenvalue problem under proper conditions. Further, we propose a projected nonlinear eigenvalue problem. In addition, we also conduct extensive experiments to evaluate the proposed framework on various data sets including handwritten numerals, UCI data sets, face images and gene expression data. Overall, the main contributions of this paper include

(1) We develop a regularization framework of discriminant analysis for dimensionality reduction. In this framework, one can choose robust functions to suppress the presence of outliers. Moreover, we are also capable of using this framework to implement the data reconstruction problem.
(2) We give the detailed analysis on the relationship among some linear projected methods. In particular, we show that regularized MMC is a special case of our framework, which helps explain why regularized MMC is a robust feature extraction method, and also point out the range of the regularization parameter in regularized MMC.
(3) We conduct extensive experiments on various data sets to evaluate the effectiveness of our framework and compare it with some linear projected methods.

The rest of this paper is organized as follows. Section 2 overviews linear projection methods including PCA, LDA, MMC, and regularized MMC. In Section 3, we give a regularization framework of discriminant analysis for dimensionality reduction and show how to solve the optimization problem. In Section 4, links to some existing linear projected methods are given. Section 5 gives the detailed experimental results. Section 6 contains some concluding remarks and further directions.

2. PCA, LDA, MMC and regularized MMC

Assume that $x_1,\ldots,x_m$ are a set of n-dimensional samples of size m, $x_i \in \Re^n (i=1,\ldots,m)$. Each sample belongs to exactly one of c object classes $\{l_1,\ldots,l_c\}$ and the number of samples in the ith class is m_i. The between-class scatter matrix, the within-class scatter matrix, and the total scatter matrix are defined as:

$$S_b = \sum_{i=1}^{c} m_i(\mu_i-\mu)(\mu_i-\mu)^T,$$

$$S_w = \sum_{i=1}^{c}\sum_{x\in l_i}(x-\mu_i)(x-\mu_i)^T, \quad S_t = \sum_{i=1}^{m}(x_i-\mu)(x_i-\mu)^T,$$

where $_i$ is the centroid of the ith class and μ is the global centroid of the sample set.

2.1. PCA

Principal component analysis, also called Karhumen–Loeve transform in some sense, extracts the desired number of principal components for data by minimizing the mean squared error criterion. The optimal linear transformation $U \in \Re^{n\times k}$ for PCA is the one that maximizes the total scatter in a reduced dimensional space. The matrix U can be obtained by performing the eigendecomposition on S_t and the columns of U are eigenvectors of S_t corresponding to the first k largest eigenvalues. It is easy to verify that the ith eigenvalue is the variance of data that is projected onto the ith eigenvector. A good property of PCA is that it decorates the data.

2.2. Classical LDA

Classical LDA seeks the direction on which data points of different classes are far from each other while requiring data points of the same class to be close to each other. To be specific, LDA is to find the optimal projection by optimizing the objective function in the following:

$$\max trace((U^TS_wU)^{-1}(U^TS_bU)). \tag{1}$$

The optimal transformation U can be obtained by solving the generalized eigenproblem: $S_bu=\lambda S_wu$. In general, there are at most $c-1$ eigenvectors corresponding to nonzero eigenvalues since the rank of the matrix S_b is not bigger than $c-1$. When S_w is singular, one can overcome it by applying some methods such as LDA/QR [6], PCA plus LDA [2], LDA/GSVD [9], and LDA/FKT [10].

2.3. MMC and regularized MMC

MMC aims at maximizing the average margin between different classes. To be specific, MMC is to optimize the objective function: $trace(U^T(S_b-S_w)U)$ under the proper constraint. The optimal transformation U can be obtained by performing the eigen-decomposition on the matrix (S_b-S_w). The matrix U is composed of the first k eigenvectors of S_b-S_w corresponding to the first k largest eigenvalues. The regularized MMC is to maximize $trace(U^T(S_b-\gamma S_w)U)$ with a nonnegative regularization parameter γ. As pointed out in [25], the MMC or regularized MMC can also be performed within the range space of S_t since the null space of S_t does not contain any discriminant information. As a result, the computational complexity of MMC or regularized MMC can be further reduced.

3. The regularization framework of discriminant analysis

3.1. The regularization framework

In this section, we assume that the data is centralized without loss of generality. In fact, this is easily obtained by a translation of data. It is shown [26] that the standard PCA is equivalent to solving the following optimization problem:

$$\min \quad \sum_{i=1}^{m} \|x_i - UU^T x_i\|^2, \tag{2}$$

$$\text{s.t.} \quad U^T U = I_k,$$

where $\| \ \|$ denotes a 2-norm and I_k is an $k \times k$ identity matrix.

From Eq. (2), one can see that $UU^T x_i$ denotes the reconstructed data of x_i in a subspace spanned by the column vectors of U and PCA makes the reconstruction error of m data points as small as possible. In some sense, this corresponds to recovering m data points in a subspace. Further, motivated by some ideas of signal and image recovery [27–31], we propose the general framework to obtain the projection matrix U by minimizing the regularized cost function of the form in the following:

$$F(U) = \Psi(U) + \lambda\Phi(U) \quad = \sum_{i=1}^{m} \psi(J_{p,q}(x_i - UU^T x_i)) + \lambda \sum_{i=1}^{r} \phi(J_{p,q}(U^T g_i)), \tag{3}$$

$$\text{s.t.} \quad U^T U = I_k,$$

where $\Psi(U)$ is the data reconstruction term, $\Phi(U)$ is the regularization term that penalizes the roughness or smoothness of projected data or the projected matrix, and λ is a non-negative parameter controlling the tradeoff between $\Psi(U)$ and $\Phi(U)$. In a statistical framework, $\Psi(U)$ denotes the distortion between the original data and the reconstructed data. In the Bayesian estimation framework, $\Phi(U)$ is prior knowledge of projected data or the projected matrix U. The linear operators $g_i : \Re \to \Re^n$ in Eq. (3), for $i = 1, \ldots, r$, are to produce the difference between neighboring samples as done in [28], which will be discussed in Section 3.2. The function $J_{p,q}()$ in Eq. (3) denotes a measure for the vector. In general, the vector norm can be used as the measure. Here, we define the following generalized measure for a vector y:

$$J_{p,q}(y) = (\sum_{i=1}^{n} |y_i|^p)^{q/p}, \quad p > 0, \ \ q \geq 0. \tag{4}$$

where y_i is the ith component of the vector y and n is the dimension of the vector y. Note that the definition of Eq. (4) obviously contains some widely used vector norms such as L1 norms and L2 norms. Further, one can observe that $J_{p,q}()$ may become a non-smooth function by choosing p and q. For simplicity, in this paper we set p=2. Thus $J_{2,q}()$ is a smooth function and this will lead to a low computational complexity for dealing with Eq. (3). Moreover, $J_{2,q}()$ gets penalized as q is reduced. In other words, when q approaches zero, a big 2-norm of y yields a small value of $J_{2,q}()$. This may suppress data points where they have big reconstruction errors and may decrease the effect of outliers in data analysis. In addition, the functions $\phi()$ and $\psi()$ in Eq. (3) need to be defined. As discussed in [28], the function $\psi() : \Re \to \Re$ is a continuous function which decreases on $(-\infty, 0]$, increases on $[0, +\infty)$, and satisfies $\psi(0) = 0$; the function $\phi() : \Re \to \Re$ is a potential function. Note that the functions $\phi()$ and $\psi()$ may be the same in real applications. In order to make outliers be smoothed and preserve non-outliers effectively, we hope to adopt robust functions in the regularized cost function in Eq. (3). Table 1 lists some functions that can be used in our regularization framework. In fact, these functions have been widely used in signal and image recovery [27] and have been shown that outliers can be suppressed by using these functions. Overall, the objective function in Eq. (3) means that the data may lie in a potential subspace and we hope to find this subspace by minimizing the regularized objective function which considers prior knowledge of data points.

Table 1
The functions in our regularization framework where $v > 0$ is a parameter.

(1) $f(t) = |t|^v$ [27]
(2) $f(t) = 1 - \exp(-vt^2)$ [35]
(3) $f(t) = \frac{vt^2}{1+vt^2}$ [33]
(4) $f(t) = \frac{v|t|}{1+v|t|}$ [32]
(5) $f(t) = t^2$ if $|t| \leq v$, $f(t) = v(v+2|t-v|)$ if $|t| > v$ [28]
(6) $f(t) = \log(vt^2+1)$ [34]
(7) $f(t) = \log(v|t|+1)$ [28]

To solve the above optimization framework of Eq. (3), we need to introduce the Lagrangian as follows:

$$L(U) = F(U) + trace(\Lambda^T(U^T U - I_k)), \tag{5}$$

where Λ is an $k \times k$ matrix containing the Lagrange multipliers associated with the constraint specified by $U^T U = I_k$.

The Lagrangian L has to be minimized with respect to the variables U and Λ. If one differentiates L with respect to the variables U and Λ, one can obtain the following first-order necessary condition of Eq. (3):

$$\nabla_U L(U) = 0, \tag{6}$$

$$U^T U = I_k,$$

where $\nabla_U L$ denotes an $n \times k$ matrix whose (i,j)th entry is the partial derivative of L with respect to the (i,j)th entry of U.

In the case that p=2 in Eq. (3), it is easy to obtain

$$\nabla_U L(U) = \sum_{i=1}^{m} \psi'((x_i^T x_i - x_i^T UU^T x_i)^{q/2}) \frac{q}{2} (x_i^T x_i - x_i^T UU^T x_i)^{(q/2)-1} (-2x_i x_i^T) U$$
$$+ \sum_{i=1}^{r} \phi'((g_i^T UU^T g_i)^{q/2}) \frac{q}{2} (g_i^T UU^T g_i)^{(q/2)-1} 2(g_i g_i^T) U + 2U\Lambda,$$

where $\psi'()$ denotes the derivative of $\psi()$ and $\phi'()$ denotes the derivative of $\phi()$.

For notational simplicity, let

$$H(U) = \sum_{i=1}^{m} \psi'((x_i^T x_i - x_i^T UU^T x_i)^{q/2}) \frac{q}{2} (x_i^T x_i - x_i^T UU^T x_i)^{(q/2)-1} (x_i x_i^T)$$
$$- \sum_{i=1}^{r} \phi'((g_i^T UU^T g_i)^{q/2}) \frac{q}{2} (g_i^T UU^T g_i)^{(q/2)-1} (g_i g_i^T). \tag{7}$$

From Eqs. (6) and (7), one has

$$H(U)U = U\Lambda. \tag{8}$$

One can find that the eigenvalue problem of Eq. (8) is nonlinear since $H(U)$ contains the matrix variable U in the general case. It is noted that the eigenvalue problem of Eq. (8) becomes linear when $\psi(t) = \phi(t) = t^2$ and $q = 1$. Also note that UQ is a solution for any orthogonal matrix $Q \in \Re^{k \times k}$ if U is a solution. As a result, the solution of the nonlinear eigenvalue problem is a k-dimension invariant space rather than a specific matrix. As pointed out in [36–38], the most widely used technique for solving the nonlinear eigenvalue problem is to reduce it to a sequence of linear eigenvalue problems. Based on this idea, the method called the self consistent field (SCF) iteration [36] is used to deal with the nonlinear eigenvalue problem. The basic idea of the SCF iteration is the following. Given an initial $U^{(0)}$, one can compute $H(U^{(0)})$ and then obtains $U^{(1)}$ which consists of eigenvectors corresponding to the first k biggest eigenvalues by solving the linear eigenvalue problem $H(U^{(0)})U = U\Lambda$. From $U^{(1)}$, one continues to obtain the next projection matrix and the process is repeated until the stopping criterion is met. For completeness, we briefly outline the main steps of the SCF iteration for solving the nonlinear eigenvalue problem in the following.

SCF iteration for solving Eq. (8)

```
(1) Given an initial matrix $U^{(0)}$
(2) For $i=1,2,\ldots$ until convergence
(3) Compute $H^{(i)}=H(U^{(i-1)})$ from Eq. (7)
(4) Obtain $U^{(i)}$ such that $H^{(i)}U^{(i)}=U^{(i)}\Lambda^{(i)}$ and $\Lambda^{(i)}$ contains the first $k$ largest eigenvalues of $H^{(i)}$.
(5) end
```

Note that the SCF iteration will locally converge under some conditions [38]. Since the SCF iteration may not be converge, it is common to keep track of the objective value found so far, i.e., the one with the smallest objective value. At each iteration, we set $F_{best}^{(i)}=\min\{F_{best}^{(i-1)},F(U^{(i)})\}$, where $F_{best}^{(i-1)}$ is the best objective value found in previous $i-1$ iterations and $F(U^{(i)})$ is the objective value of the ith iteration. Since $F_{best}^{(i)}$ is a decreasing sequence, it has a limit point. In addition, from Eq. (7), it is found that $H(U)$ is an $n\times n$ matrix and it is necessary to perform the eigen-decomposition on $H(U)$ many times in the above algorithm. It is obvious that directly performing step 4 in the above algorithm is not effective or impractical in real applications when the dimension of data is large. To deal with this problem, the authors in [36] developed a strategy by restricting the solution of the nonlinear eigenvalue problem to a subspace. To be specific, at each iteration, the solution of the nonlinear eigenvalue problem is restricted to the subspace spanned by the solution in the previous iteration, the gradient of the Lagrangian, and the search direction provided in the previous iteration. In some sense, this corresponds to solving a projected nonlinear eigenvalue problem.

Based on similar ideas in [36], we restrict the solution of Eq. (8) to a subspace spanned by the samples in the training set. That is, we can obtain the approximate solution, denoted by

$$U=XA, \tag{9}$$

where X is the data matrix and A is an $m\times k$ matrix.

Substituting Eq. (9) into (3) and setting $p=2$, one can obtain:

$$F(XA)=\sum_{i=1}^{m}\psi(J_{2,q}(x_i-XAA^TX^Tx_i))+\lambda\sum_{i=1}^{r}\phi(J_{2,q}(A^TX^Tg_i)) \tag{10}$$

$$\text{s.t.}\quad A^TX^TXA=I_k.$$

In a similar way, the first order necessary condition of Eq. (10) can be obtained by setting the gradient of the Lagrangian associated with Eq. (10) to zero with respect to the variable A. Then it is not difficult to obtain

$$\hat{H}(A)A=X^TXA\Lambda$$

where

$$\begin{aligned}\hat{H}(A)=&\sum_{i=1}^{m}\psi'((x_i^Tx_i-x_i^TXAA^TX^Tx_i)^{q/2})\\&\times\frac{q}{2}(x_i^Tx_i-x_i^TXAA^TX^Tx_i)^{(q/2)-1}(X^Tx_ix_i^TX)\\&-\sum_{i=1}^{r}\phi'((g_i^TXAA^TX^Tg_i)^{q/2})\frac{q}{2}(g_i^TXAA^Tg_i)^{(q/2)-1}(X^Tg_ig_i^TX)X.\end{aligned}$$

As a result, solving Eq. (10) is equivalent to solving the following problem

$$\hat{H}(A)A=X^TXA\Lambda, \tag{11}$$

$$A^TX^TXA=I_k.$$

In general, when the number of the training samples is smaller than the dimension of samples, the generalized nonlinear eigenvalue problem of Eq. (11) is much smaller than the nonlinear eigenvalue problem in Eq. (8) since $\hat{H}(A)$ is an $m\times m$ matrix. Consequently, solving the nonlinear eigenvalue problem of Eq. (11) is much more efficient than solving Eq. (8) in the small sample size problem in terms of the computational complexity. Similar to the algorithm of the nonlinear eigenvalue problem, Eq. (11) can also be solved by reducing it to a sequence of generalized linear eigenvalue problems. Here we also summarize the main steps for solving Eq. (11).

SCF iteration for solving Eq. (11)

```
(1) Given an initial matrix $A^{(0)}$.
(2) For $i=1,2,\ldots$ until convergence
(3) Compute $\hat{H}^{(i)}=\hat{H}(A^{(i-1)})$ from Eq. (11)
(4) Obtain $A^{(i)}$ such that $\hat{H}^{(i)}A^{(i)}=X^TXA^{(i)}\Lambda^{(i)}$ and $\Lambda^{(i)}$ contains the first $k$ largest eigenvalues.
(5) end
```

From Eq. (11), it is observed that the inner product of data points is involved when one computes $\hat{H}^{(i)}$ and X^TX is a Gram matrix. Based on these facts, one can use a kernel function to replace the inner product of data points. Thus one can easily extend the above algorithm to its kernel version. As a result, the above algorithm also gives us a strategy to extend the regularization framework of Eq. (3) to its kernel version. In addition, when the dimension of data and the number of samples are relatively large, directly adopting the sample space may involve solving a large eigenvalue problem. In order to further avoid this problem, one can further restrict the solution of Eq. (3) to a small subspace that is smaller than the sample space. To be specific, one partitions the data into many clusters by k-means algorithm [25] and obtains the centroid of each cluster. Then we use the centroid to approximate the structure of clusters since the centroids are a good approximation of the original data. Finally, we restrict the solution to the subspace spanned by the centroids of clusters. It is obvious that this subspace belongs to the sample space. In an extreme case of supervised learning, one can restrict the solution to the subspace spanned by the centroids of classes if the number of classes and the dimension of data are relatively large. As a result, one can deal with large-scale data sets by using this simple strategy in our regularization framework.

3.2. The selection of regularization operators

In this subsection, we show how to choose the regularization operators in the general case. Let G denote an $n\times r$ matrix whose ith column is the linear operator g_i in Eq. (3). Typically, G can be simply set as an identity matrix, i.e., $G=I_n$, where I_n is an identity matrix and $r=n$. This simple setting corresponds to providing prior knowledge for the projection matrix U. In general, the regularization operators in signal and image recovery [28] are often defined as the difference between neighboring samples in the original space. In a similar way, we define the regularization operators in our framework to produce the difference between neighboring samples in the projected space. For clarity, we will discuss them in two cases: supervised learning and unsupervised learning.

In supervised learning, one knows that one sample must belong to one class. In such a case, we have prior knowledge of class information. Similar to the setting in signal recovery, we set U^Tg_i to be the difference between x_i and its neighboring samples in the projected space. Assume that $x_1,\ldots,x_s,x_i$ belong to the same class. We regard $x_1,\ldots,x_s$ as the neighboring samples of x_i. In such a case, we define the difference between x_i and its neighboring samples in the projected space as follows:

$$U^T g_i = \sum_{j=1}^{s} (U^T x_i - U^T x_j),$$

where $U^T x_i$ is the projected data of x_i. Further, we also give another definition of linear operators. Since x_i belongs to some class, one can obtain the mean of the samples whose labels are the same as that of x_i, denoted by μ. Then one regards μ as a virtual sample and further takes it as the neighborhood of x_i. In such a case, the first-order difference between x_i and μ in the projected space is defined as $U^T g_i = U^T x_i - U^T \mu^{(i)}$, where $\mu^{(i)}$ denotes the mean of the class to which x_i belongs. Note that there are m linear operators since there are m data points in such a case. In addition, a more reasonable definition for linear operators is to use the weighted first-order difference between x_i and the other samples in the class to which x_i belongs.

In unsupervised learning, one does not have any class information. However, in real applications, one hopes to preserve the local structure of data points when they are projected into a new space. Specifically, it requires that the projected data should be close to each other if the data points x_i and x_j are close to each other. In general, the local structure of data points is characterized by the neighborhood of data points. Usually, one obtains the neighborhood of x_i with the property: $O(x_i, \varepsilon) = \{x \mid \|x - x_i\| < \varepsilon\}$, where ε is a positive constant, or simply defines s nearest-neighbor points of x_i. Then one can define the difference between x_i and its neighboring points in the projected space. Assume that there are s neighboring points of x_i. Here we give the weighted difference between x_i and its neighboring points in the projected space, denoted by

$$U^T g_i = \sum_{j=1}^{s} w(x_i, x_j)(U^T x_i - U^T x_j),$$

where $w(x_i, x_j) = \exp(-\|x_i - x_j\|^2/\sigma)$ is the weight of the samples x_i and x_j, and σ is a parameter. Particularly, when the parameter σ approaches the positive infinity, then $w(x_i, x_j) = 1$ holds.

Finally, it should be pointed out that the regularization operators in unsupervised learning can be used in supervised learning and combing them may yield the regularization operators in semi-supervised learning. In practice, the linear operators defined above are very similar to those defined in signal and image recovery [28]. In order to enforce more strong smoothness, one can use the high-order difference between the samples in the projected space which are similar to the high-order difference in signal and image recovery. The difference between the operators in our framework and the operators in signal and image recovery is that the former is used in a vector space and the latter is mainly used in a scalar space. More specifically, some methods in image recovery [28] are mainly used to recover an image. However, the proposed framework in this paper can be used to simultaneously recover some images or the vector-valued image such as a color image in a subspace.

4. Links to linear projection techniques

In this section, we will show the relationship between our framework and classical linear projected methods. It is of interest to note that some linear projected methods belong to our framework by choosing different functions in Eq. (3).

(i) When $\lambda = 0$, $\psi(t) = t^2$, p=2, q=1, Eq. (3) can be written as

$$\sum_{i=1}^{m} (x_i^T x_i - x_i^T U U^T x_i), \tag{12}$$

s.t. $U^T U = I_k$.

It is straightforward to verify that minimizing Eq. (12) is equivalent to obtaining the first k eigenvectors of XX^T corresponding to the first k largest eigenvalues. In the case that the data is centralized, this is equivalent to classical PCA [26].

(ii) When $\lambda = 0$, p=2, q=1, Eq. (3) can be written as

$$\sum_{i=1}^{m} \psi\left(\sqrt{x_i^T x_i - x_i^T U U^T x_i}\right), \tag{13}$$

s.t. $U^T U = I_k$.

Thus, minimizing Eq. (13) is equivalent to rotational invariant L1-norm PCA [21].

(iii) When $\psi(t) = \phi(t) = t^2$, $p = 2, q = 1$, and $g_i^T U = (U^T x_i - U^T \mu^{(i)})^T$, where $\mu^{(i)}$ denotes the mean of samples in the class to which x_i belongs, Eq. (3) can be written as

$$\begin{aligned}&\sum_{i=1}^{m} (x_i^T x_i - x_i^T U U^T x_i) + \lambda \sum_{i=1}^{n} (U^T x_i - U^T \mu^{(i)})^T (U^T x_i - U^T \mu^{(i)}) \\ &= trace(XX^T) - trace(U^T XX^T U) + \lambda\, trace(U^T S_w U).\end{aligned} \tag{14}$$

Since $trace(XX^T)$ is a constant, minimizing Eq. (14) is equivalent to the following optimization problem under proper constraints:

$$\max[trace(U^T XX^T U) - \lambda\, trace(U^T S_w U)]. \tag{15}$$

Note that Eq. (15) can be further written as follows in the case of the centralized data:

$$\max[trace(U^T S_b U) - (\lambda - 1) trace(U^T S_w U)]. \tag{16}$$

Thus, the optimization problem of Eq. (16) is equivalent to the regularized MMC. Consequently, regularized MMC belongs to our framework. Further, one remarkable difference between our framework and regularized MMC is that the former can be used in unsupervised learning and the latter is only available for supervised learning where each class at least has two samples. In addition, note that the regularization parameter γ in regularized MMC often takes positive values. However, it is reasonable that the parameter in the regularized MMC takes negative values since regularized MMC becomes PCA in the case of $\gamma = -1$. It is not pointed out in previous literature that which range of the parameter γ in regularized MMC is reasonable when the parameter λ takes negative values. From our regularization framework, one knows that the parameter γ in our framework takes values in $[0, +\infty)$. From this point, we can infer that the parameter γ in regularized MMC should be chosen from the interval of $[-1, +\infty)$.

Finally, it should be pointed out that the constraint $U^T U = I_k$ is imposed in our framework. In fact, one also imposes the constraint $U^T B U = I_k$ in our framework, where B is a positive definite matrix. Thus a class of generalized algorithms can be induced from our framework and this will further extend our framework in some sense. For example, if one hopes to extract uncorrelated feature vectors, one can replace $U^T U = I_k$ with $U^T S_t U = I_k$ in our regularization framework.

5. Experimental results

In this section, we carry out experiments on various data sets to explore the performance of the proposed regularization framework. In the regularization framework, the functions and parameters we use are set as follows. $\psi(t) = t^2$ if $|t| \le \nu$, $\psi(t) = \nu(\nu + 2|t - \nu|)$ if $|t| > \nu$, where the parameter ν is set as the median value of the reconstructed error as done in [21], $\varphi(t) = t$, and $q = 1$. We also assume that the samples in the same class are close to each other. Further, when the classification problem is involved, the nearest neighbor distance rule with the Euclidean distance measure is used as the classifier. In addition, note that our experiments are implemented on a Pentium 1.6-G computer with 1024M RAM and all the algorithms are programmed using Matlab language.

5.1. Experiments on handwritten numerical characters

In this section, we report experimental results using the well-known character dataset: the United States Postal services (USPS) database, in which there are 9298 handwritten character measurements divided into 10 classes. The entire USPS dataset is divided into two parts, a training set with 7291 measurements and a test set with 2007 measurements. All digits are 16×16 images which represent as 256 feature vectors and are collected from mail envelopes in buffalo, NY. In our framework, the linear operators in Eq. (3) are set as follows:

$$U^T g_i = \sum_{j=1}^{s} w(x_i, x_j)(U^T x_i - U^T x_j),$$

where $w(x_i, x_j) = \exp(-\|x_i - x_j\|^2/\sigma)$ is the weight of the samples x_i and x_j, and σ is set as the standard derivation of training sample pairs.

In the first set of experiments, we demonstrate the effectiveness of the regularization framework in different regularization parameters for data visualization. First, we choose 300 samples from 1, 6, 9 numerals, each class having 100 samples. As a result, we have 300 samples to train our method and then obtain the projection matrix U. Thus these 300 samples are projected into a two-dimension space. Fig. 1 shows the projection of the training samples in different regularization parameters $\lambda(0, 1, 100, 10\,000)$. Note that our method with $\lambda = 0$ is equivalent to rotational invariant L1 PCA [21]. As can be seen from Fig. 1, incorporating prior knowledge in the data reconstruction can change the distribution of projected data points. Thus one can obtain better visual results by tuning the regularization parameter. Further, this shows that the regularization parameter may have a potential effect on the data classification problem.

In the second set of experiments, we evaluate our method on handwritten numerals. We randomly choose 100 samples per class to form the training set and then use all the 2007 testing samples to form the testing set. First, we carry out experiments to show the effect of parameters in our framework. Fig. 2(a) shows error rates of our method versus various parameters (log) on handwritten numerals. As can be seen from Fig. 2(a), the performance of our method becomes worse when the parameter λ is larger than some value and this shows that choosing an improper parameter may result in performance degradation. As a result, choosing a proper parameter is very important for our framework. In addition, for comparison, we perform regularized MMC, orthogonal LDA [5] and LDA/FKT [10]. LDA/FKT and orthogonal LDA are two effective methods for solving LDA. The regularization parameter γ in RMMC and the regularization parameter λ in our framework are chosen from {0, 0.0001, 0.01,

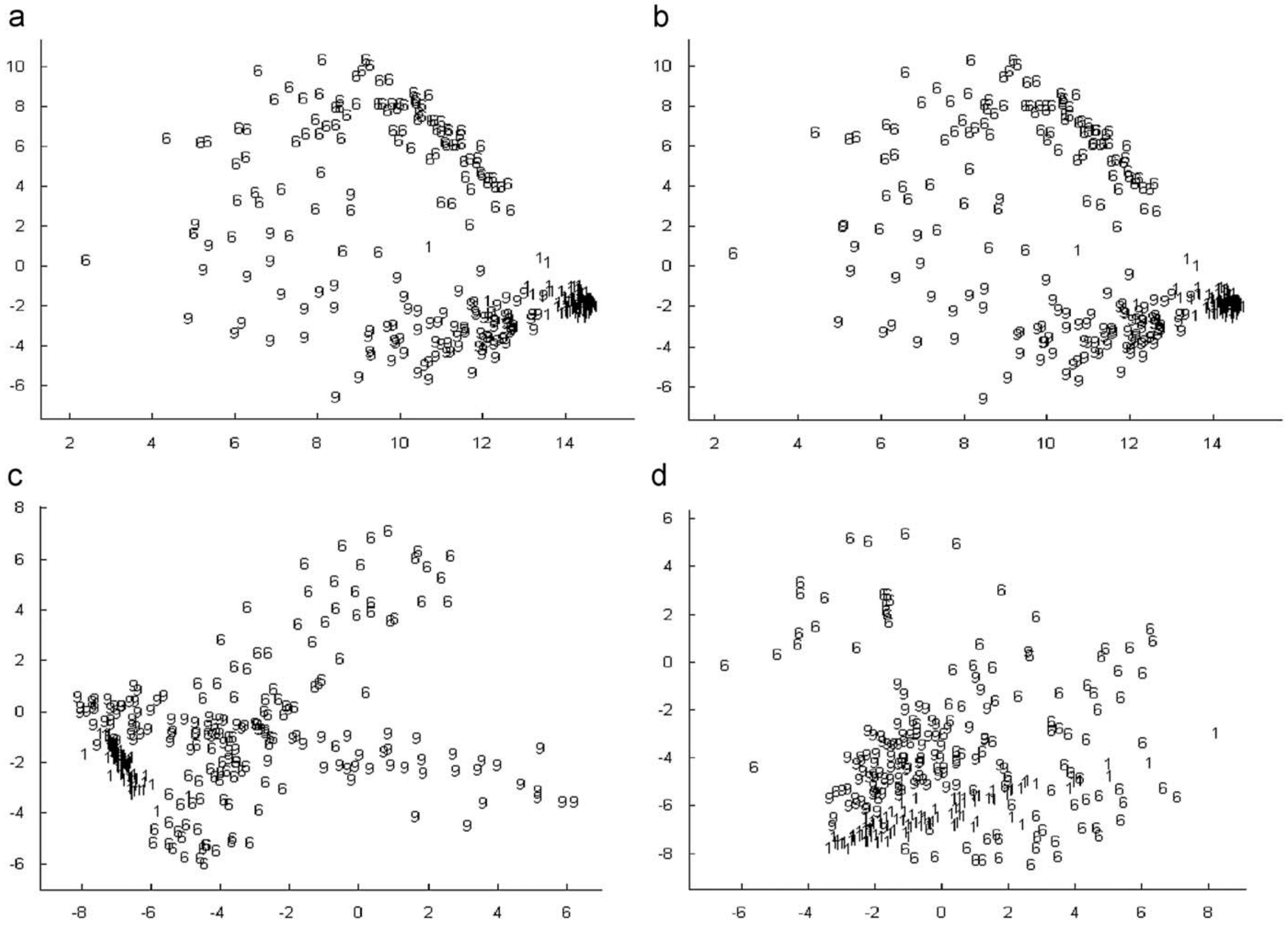

Fig. 1. Visualization of handwritten character images in different regularization parameters: (a) λ=0; (b) λ=1; (c) λ=100; (d) λ=10 000.

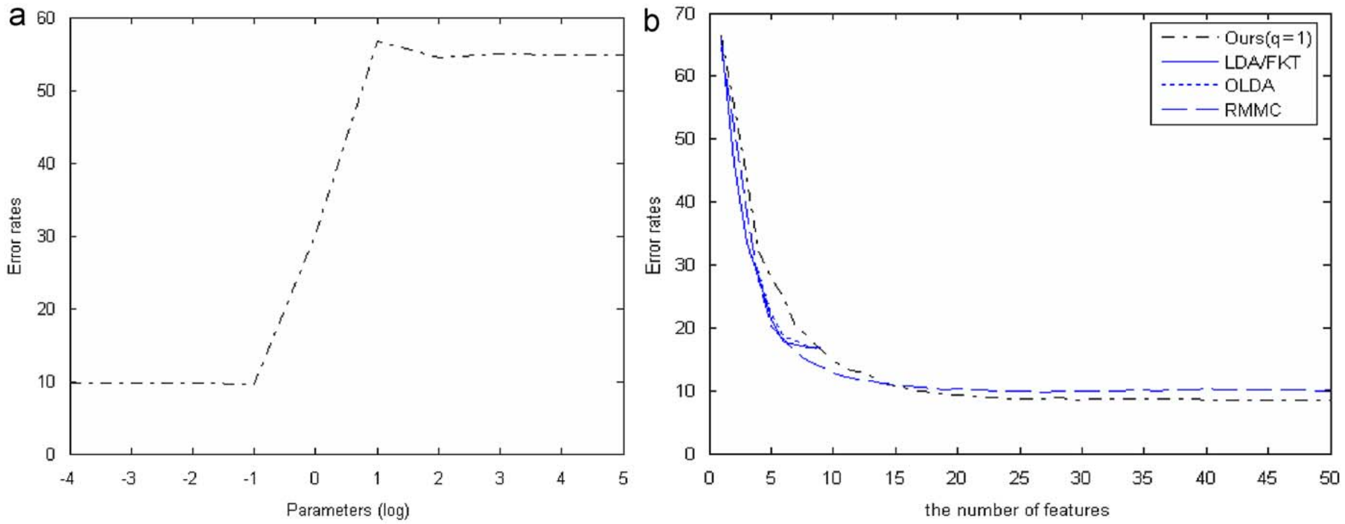

Fig. 2. Experimental results on handwritten numerals: (a) error rates (%) versus various parameters; (b) error rates (%) versus reduced dimensions.

$\ldots, 1000, 10\,000\}$. In our experiments, we randomly perform 5 runs to choose optimal parameters and perform additional 5 runs to obtain the experimental result. Fig. 2(b) shows the average error rate of each method versus the reduced dimensions from five runs. It is observed from Fig. 2(b) that the best classification performance of LDA/FKT, RMMC or OLDA is lower than that of our method. This may come from the fact that the RMMC method does not use robust functions and the number of extracted features in LDA/FKT or OLDA is limited to nine. Our method, however, adopts robust functions to overcome the influence of outliers such as characters with the bad image quality. Moreover, the projected dimension of our framework is not limited by the number of classes. In short, these experiments show that discriminant information can be effectively obtained by using our framework in terms of robust functions.

5.2. Experiments on UCI data sets

To further demonstrate the performance of the proposed regularization framework, we continue to carry out experiments on a collection of benchmark data sets that can be obtained from UCI machine learning repository [39]. These data sets have been widely used in testing and evaluating the performance of some machine learning algorithms. The attributes of each data set are normalized to the interval of $[-1, 1]$. In order to evaluate the performance of algorithms, 10-fold cross validation is performed. The 10-fold cross-validation is a widely used technique in machine learning and involves partitioning the whole data set into 10 roughly equalized parts. The classifier is trained on nine of the partitions and tested on the remaining partition. This is repeated until each partition has been tested on a new classifier built with the remainder of the data set. For comparison, we also perform LDA/FKT, OLDA regularized MMC (RMMC), Chernoff LDA(CLDA) [40]. The regularization operators in Eq. (3) are set the same as done in Section 5.1. We also perform double 10-fold cross-validation on these data sets. One is to choose regularization parameters in our framework or regularized MMC and the other is to report our experimental results. Fig. 3(a) shows the performance comparison on the breast data, which consists of 683 measurements from 2 classes in a 10-dimensional space. It is shown that the best performance of all methods is obtained by our method. Fig. 3(b) shows the performance comparison on the diabetes data, which consists of 768 measurements from 2 classes in an eight-dimensional space. It is shown that our method is superior to RMMC when the number of features is bigger than 4 and the CLDA method does not perform well on this data set. Fig. 3(c) shows the performance comparison on the glass data, which consists of 214 measurements from 6 classes in a nine-dimensional space. It is shown that our method outperforms CLDA and OLDA is superior to LDA/FKT in most cases. Fig. 3(d) shows the performance comparison on the heart data, which consists of 270 measurements from 2 classes in a 13-dimensional space. It is shown that RMMC is superior to our method and the performance of OLDA is the same as LDA/FKT since both only extract a feature. Fig. 3(e) shows the performance comparison on the iris data, which consists of 150 measurements from 3 classes in a 4-dimensional space. It is shown that our method is superior to RMMC when the number of features is bigger than 1. Fig. 3(f) shows the performance comparison on the Ionosphere data, which consists of 351 measurements from 2 classes in a 34-dimensional space. It is shown that our method achieves the best performance among all methods and the performances of our method and CLDA alternately change with the change of features. Fig. 3(g) shows the performance comparison on the sonar data, which consists of 208 measurements from 2 classes in a 60-dimensional space. It is shown that the performance of our method is better than that of RMMC in most cases and our method still achieves the best performance among all methods. On this data set, the CLDA method performs worse. Fig. 3(h) shows the performance comparison on the vehicle data, which consists of 846 measurements from 4 classes in an 18-dimensional space. It is shown that our method outperforms RMMC with the increase of features and the CLDA method can obtain the best performance. Fig. 3(i) shows the performance comparison on the wine data, which consists of 178 measurements from 3 classes in a 12-dimensional space. It is shown that our method achieves competitive performance with RMMC and OLDA in the case of the best performance and the performance of CLDA becomes better with the increase of features. These experimental results show our method can achieve better classification performances than other methods in most cases due to the fact we adopt robust functions in Eq. (3). In the following, we carry out experiments to compare the running time of various dimensionality reduction methods. Here the convergent condition of our method is set if the difference between the norms of Λ in Eq. (8) in successive iterations is less than 10^{-3} or the maximal number of iterations is 100. Table 2 shows the average time consumed by different methods and average numbers of iterations of our method is also reported. From Table 2, it is found that the time of our method is much longer than that of other methods (RMMC, LDA/FKT, OLDA, and Chernoff LDA) since our method involves a series of linear eigenvalue problems and other methods only solve an eigenvalue problem. It is also found that the average numbers of iterations for our method do not exceed the maximal number of iterations in the case that the difference between the norms of Λ in Eq. (8) in successive iterations is less than 10^{-3}. This shows that our method locally converges on these data sets based on our defined cost function.

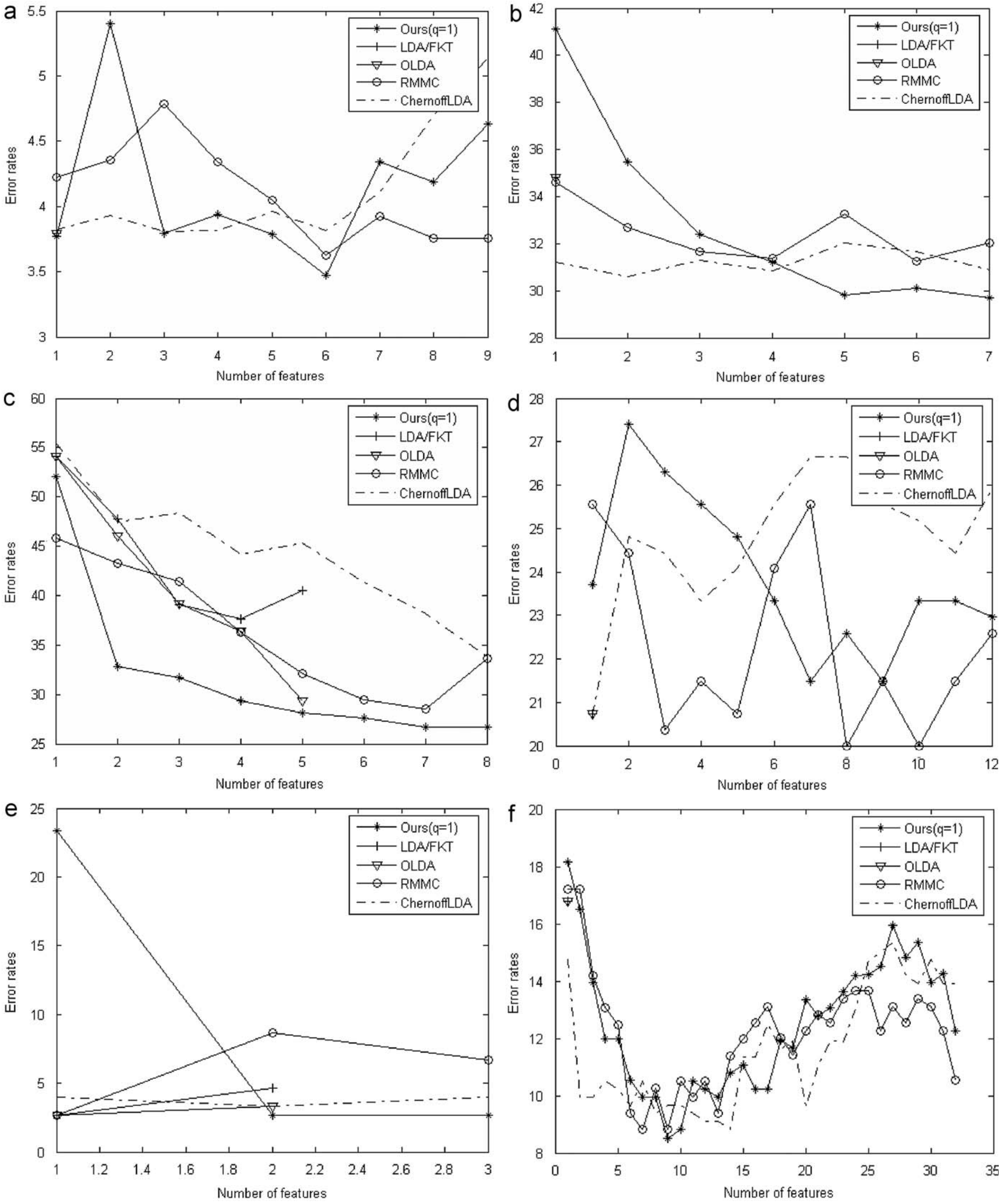

Fig. 3. Error rates (%) versus reduced dimensions on UCI data sets: (a) Breast; (b) Diabetes; (c) Glass; (d) Heart; (e) Iris; (f) Ionosphere; (g) Sonar; (h) Vehicle; (i) Wine.

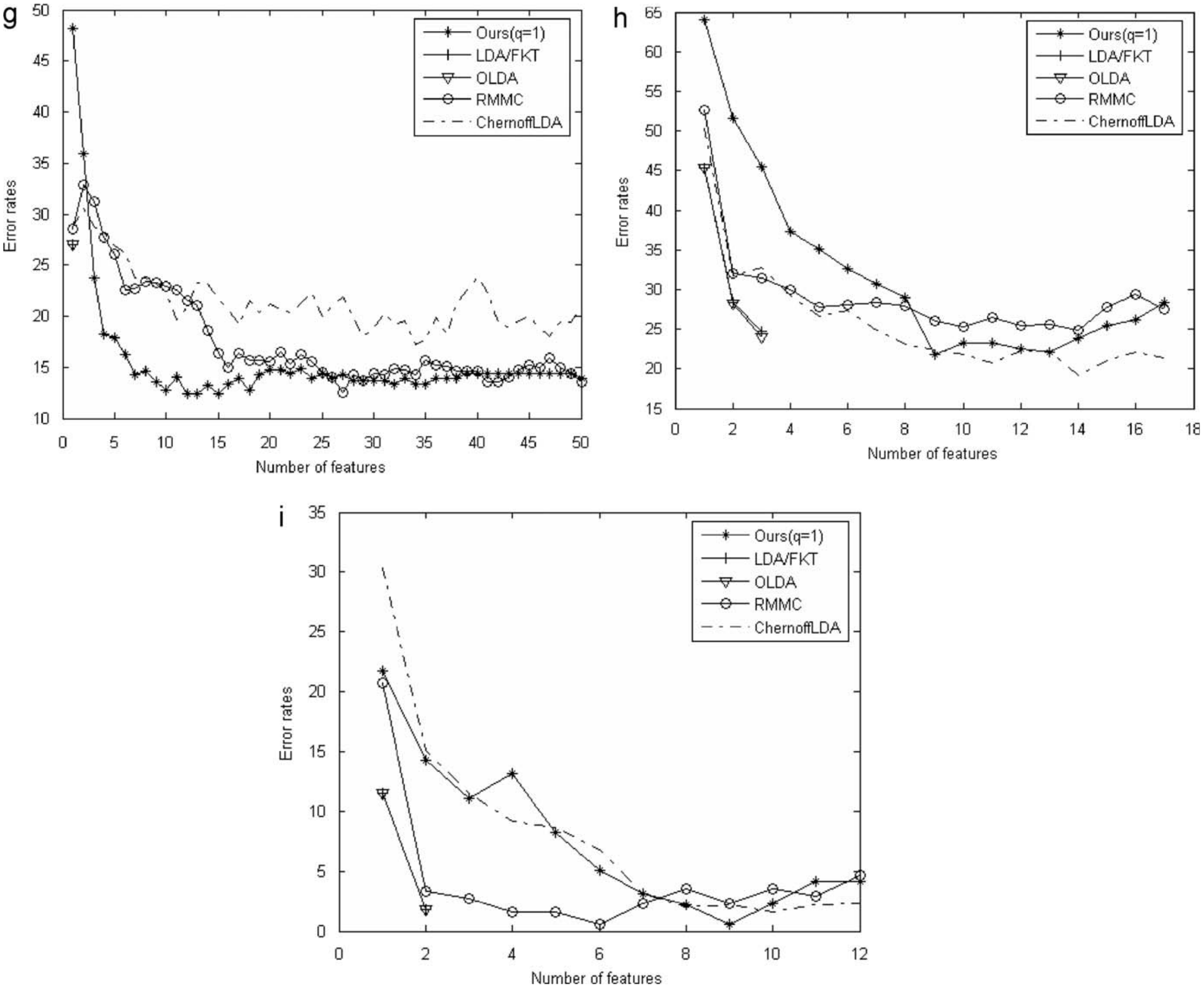

Fig. 3. (*Continued*)

Table 2
Time and iterations of different methods on some UCI data sets.

	Average time (s)					Iterations
	LDA/FKT	OLDA	RMMC	ChernoffLDA	Ours	Ours
Breast	0.064	0.076	0.065	0.064	4.75	4
Diabetes	0.073	0.078	0.072	0.075	8.05	45.2
Glass	0.001	0.014	0.001	0.015	0.57	50.5
Heart	0.006	0.014	0.003	0.109	0.62	23.6
Iris	0.004	0.003	0.003	0.007	0.24	34.3
Ionosphere	0.028	0.028	0.034	0.046	2.49	49.6
Sonar	0.026	0.05	0.003	0.084	0.98	17.7
Vehicle	0.195	0.203	0.201	0.214	14.15	79.3
Wine	0.004	0.009	0.006	0.008	0.33	25.6

5.3. Experiments on face images

In this section, we investigate the performance of the proposed regularization framework for image reconstruction and image classification on the UMIST database [41]. The UMIST database contains 20 persons with totally 564 images. There are variations of race, sex and appearance with different subjects. The size of each image is approximately $220*210$ pixels with 256 grey levels per pixel. Precropped images with a size of $112*96$ may also be made available from the database. For computational simplicity, we downsample each image into $56*46$ pixels in our experiments.

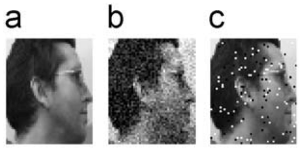

Fig. 4. The image and its noisy images: (a) the original image; (b) the image with Gaussian noise; (c) the image with salt and pepper noise.

In the first set of experiments, we show the effectiveness of our regularization framework for the image reconstruction problem. Since we focus on the image reconstruction problem, the operator g_i in our regularization framework is simply set as $g_i = e_i (i = 1, \ldots, n)$, where e_i is an $n \times 1$ vector in which the ith component is 1 and others are zero, and the regularized parameter λ takes values in {0.01, 1, 100}. This corresponds to providing prior knowledge on the projection matrix. First, we use all the samples as the training samples to obtain the projected matrix U. Since $UU^T x_i$ denotes the reconstructed data of x_i, we reshape $UU^T x_i$ into the image. Fig. 5(a1)–(a3) shows 10 reconstructed images of the first image in Fig. 4

obtained by our method with different numbers of features $k=1,\ldots,10$ and different parameters. It is observed that the reconstructed images become clearer as the dimension of the subspace is increased. For comparison, the PCA method is also performed to represent and reconstruct the same face image. Fig. 5(a4) shows the reconstructed images obtained by PCA. It is observed that the images obtained by our method are smoother than the images obtained by PCA due to the fact we add prior knowledge of the projected matrix. In addition, two noisy images are constructed by adding white Gaussian noise with the mean 0 and the variance 0.01 and the salt and pepper noise with noise level 0.05 to the original images. Fig. 4(b) and (c) show the noisy images

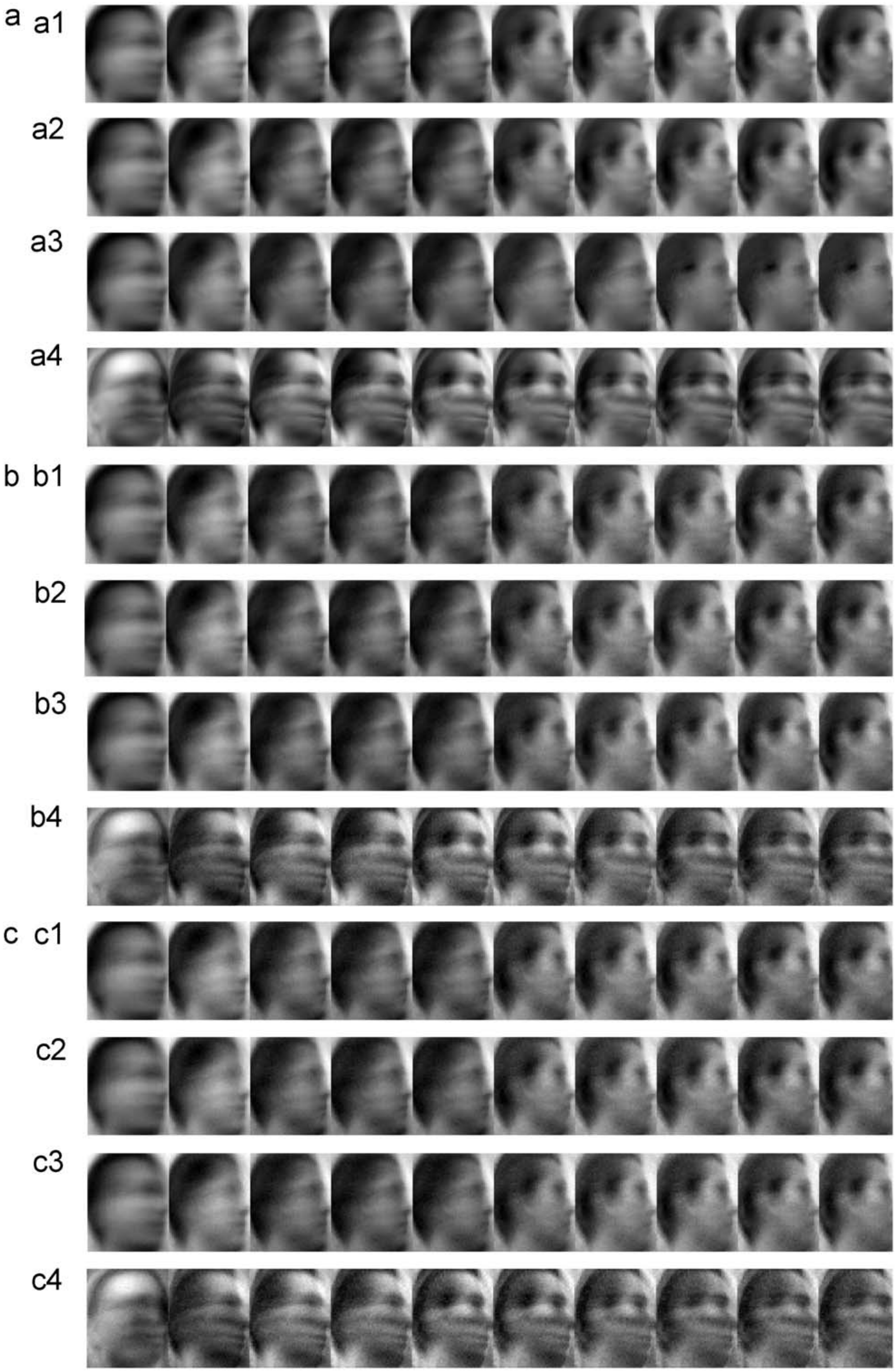

Fig. 5. The reconstructed images in different scenarios: (a) the reconstructed images without noise: (a1) the reconstructed images by our method $\lambda=0.01$; (a2) the reconstructed images by our method $\lambda=1$; (a3) the reconstructed images by our method $\lambda=100$; (a4) the first 10 reconstructed images by classical PCA; (b) the reconstructed images with Gaussian noise: (b1) the reconstructed images by our method $\lambda=0.01$; (b2) the reconstructed images by our method $\lambda=1$; (b3) the reconstructed images by our method $\lambda=100$; (b4) the first 10 reconstructed images by classical PCA; (c) the reconstructed images with salt and pepper noise: (c1) the reconstructed images by our method $\lambda=0.01$; (c2) the reconstructed images by our method $\lambda=1$; (c3) the reconstructed images by our method $\lambda=100$; (c4): the first 10 reconstructed images by classical PCA.

obtained by adding noise to the first image in Fig. 4(a). The reconstructed images obtained by our method with different numbers of features and different parameters are shown in Fig. 5(b1–b3) and (c1–c3). The reconstructed images obtained by classical PCA are displayed in Fig. 5(b4) and (c4). As can be seen from these figures, the PCA method does not perform well though it is to minimize the mean squared error. The reconstructed images obtained by our method have slightly better visual quality than those obtained by classical PCA method. The possible reason may lie in the fact we use robust functions and add smoothing constraints to the objective function. This further shows that choosing a proper function and adding smoothing constraints in our framework can improve the quality of the reconstruction images in some sense.

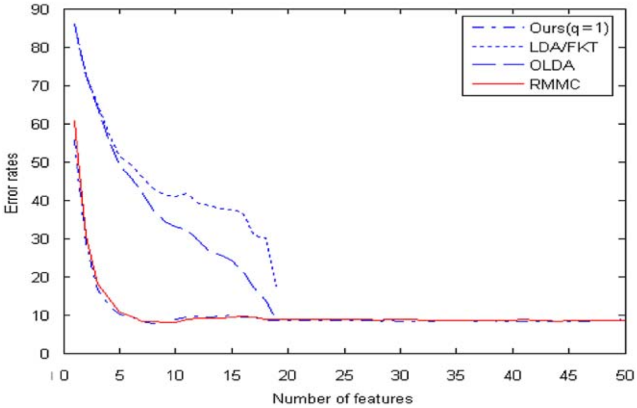

Fig. 6. Average error rates (%) of various methods versus reduced dimensions.

In the second set of experiments, we devise experiments to show the performance of our regularization framework on the image classification problem. In this set of experiments, 50% images are randomly chosen and added to white Gaussian noise with the mean 0 and the variance 0.01. Then a training sample set is formed by randomly choosing 5 images from each individual and the remaining images are used for testing. To enhance the accuracy of performance, the classification performance reported in the experiment is averaged over 20 runs. In other words, 20 different training and testing sets are used for performance evaluation. Note that the regularization operators are set the same as done in Section 5.1 and we also perform LDA/FKT, OLDA and regularized MMC (RMMC) for comparison purposes. The regularization parameters in our framework and RMMC are chosen from additional 5 runs. Fig. 6 shows experimental results of different projected methods versus reduced dimensions. As can be seen from Fig. 6, our method can obtain similar performances with RMMC since these two methods involve the regularization technique and the regularization method is an effective technique for overcoming data fitting. It is also observed that OLDA is superior to LDA/FKT due to the fact that orthogonalization contributes to noise reduction as discussed in [5]. Overall, this experiment shows that our framework is a stable and robust method for feature extraction.

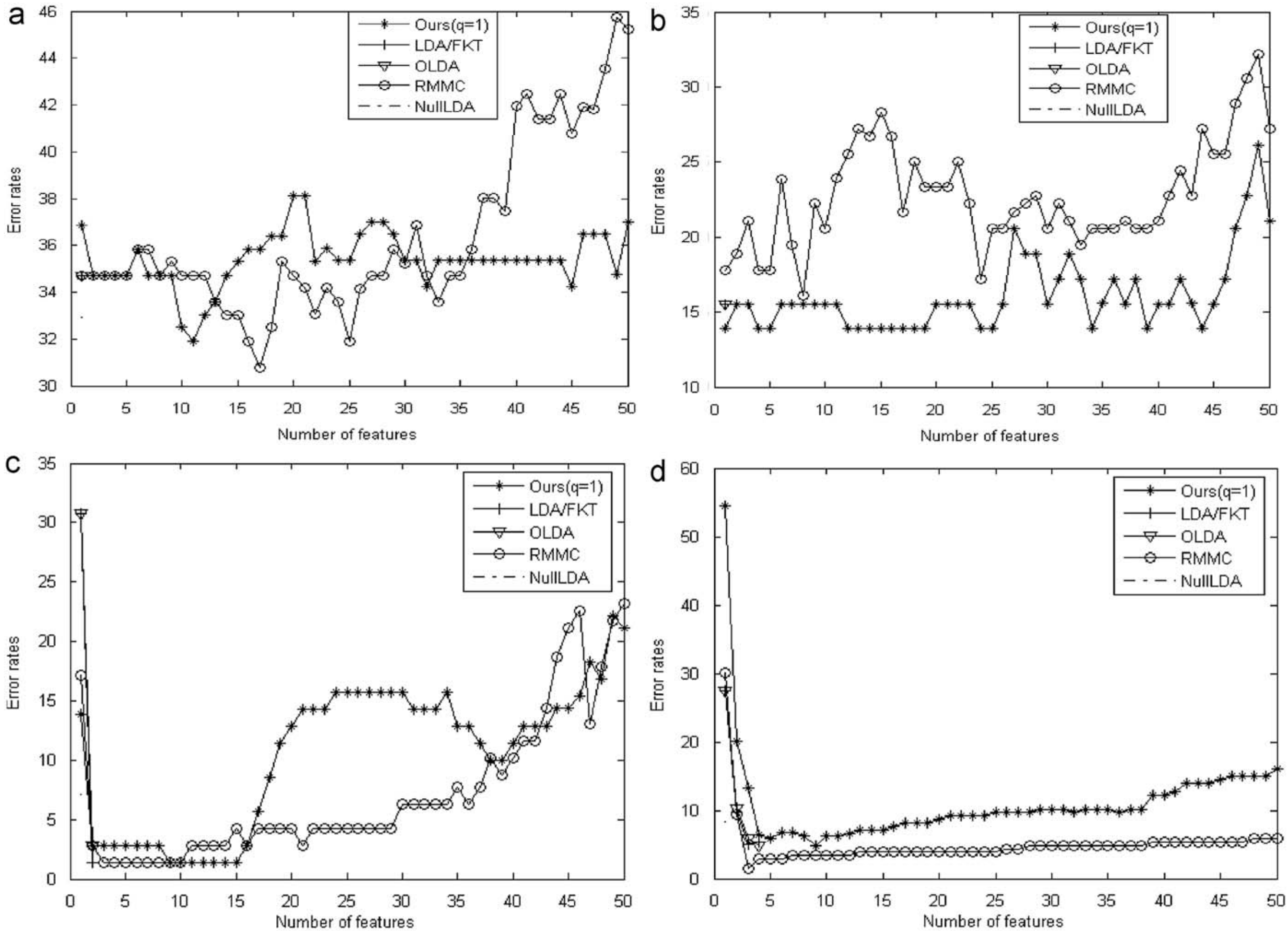

Fig. 7. Error rates (%) of various methods versus reduced dimensions on gene data sets: (a) Breast cancer; (b) Colon; (c) Leukemia 2; (d) SRBCT.

5.4. Experiments on gene expression data

The recently developed microarray technology is expected to contribute significantly to progress in cancer treatment. In general, the dimension of gene data is huge and the number of samples is relatively small. How to extract the effective features is very important for gene data classification and this will affect cancer diagnosis. Here we further explore the performance of our optimization framework on four available data sets: Breast cancer(2 classes/24481genes/97samples) [42], Colon(2/2000/6) [43], Leukemia2(3/11225/72) [44], and SRBCT(4/2308/83) [45]. Note that gene expression values of all gene data are normalized to the interval of $[-1, 1]$.

In this set of experiments, we also use 10-fold cross validation to evaluate the performance of our regularization framework since these data sets are relatively small. That is, the classification performance is averaged over 10 runs. Since these data sets belong to the small sample size, we carry out Null LDA [15] except OLDA, LDA/FKT, and RMMC for comparison. It is also noted that the regularization operators are set the same as done in Section 5.1. Fig. 7 shows the error rate of each data set with various numbers of extracted features. From Fig. 7(a), one can see that the best performance of RMMC is superior to that of OLDA, LDA/FKT or Null LDA on breast cancer data set. It is also noted that our framework adopting the robust function can achieve competitive performance with RMMC. From Fig. 7(b), one can see that our method consistently outperforms RMMC with the change of features and the performance of OLDA is consistent with that of LDA/FKT or Null LDA on the colon data set. From Fig. 7(c), it is observed that the best performance of our method is the same as that of RMMC and LDA/FKT is superior to OLDA on the Leukemia data set. From Fig. 7(d), it is noted that the best classification performance of our framework is superior to that of LDA/FKT, OLDA or Null LDA and our method is competitive with RMMC on the SRBCT data set. Overall, these experiments further show that our regularization framework is an effective and robust feature extraction method for high-dimensional data.

6. Conclusions and further directions

In this paper, we propose a family of dimensionality reduction algorithms based on a new form of regularization. The objective function of the proposed framework contains the data reconstruction term and the regularization term which allows us to exploit prior knowledge of data points. Different from some previous dimensionality reduction methods, the proposed framework can suppress the presence of outliers of data when robust functions are chosen. That is, the regularized cost function allows outliers to be smoothed in the general case. Moreover, it is found that different discriminant algorithms are characterized by different functions and different types of prior knowledge can be included in our regularization framework. More specifically, some linear projection methods such as PCA, L1-PCA, and RMMC could be derived from the regularization framework. This explains why RMMC is a robust feature extraction method and also provides new insights for us to understand the problem of dimensionality reduction. In addition, we also conduct extensive experiments to demonstrate the effectiveness of our framework by choosing robust functions.

It should be pointed out that the performance of our framework might be improved by using different cost functions or regularization operators. However, we do not attempt to find better functions or regularization operators since our aim here is to develop a regularization framework of robust dimensionality reduction, which are the directions of our future work. Specifically, based on this regularization framework, we plan to explore discriminant algorithms in the future by applying special functions under different conditions and to study the effectiveness of various functions including nonsmooth cost functions and nonsmooth regularization operators occurring in image and signal recovery [27]. In addition, another further work is to implement our regularization framework in the reproducing kernel Hilbert space induced by a nonlinear function. We hope to study these problems in the near future.

Acknowledgments

This work described in this paper was supported by the Research Grants of Council of Hong Kong (No. CityU117507) and China University of Mining and Technology.

References

[1] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, New York, 1990.

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

[3] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence (2007) 40–51.

[4] S. Yan, D. Xu, B. Zhang, H.J. Zhang, Graph embedding: a general framework for dimensionality reduction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 830–837.

[5] J. Ye, Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems, Journal of Machine Learning Research (2005) 483–502.

[6] J. Ye, Q. Li, A two-stage discriminant analysis via QR decomposition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2005) 929–941.

[7] X. He, S. Yan, Y. Hu, P. Niyogi, H.J. Zhang, Face recognition using Laplacianfaces, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (3) (2005) 328–340.

[8] Y. Pang, L. Zhang, Z. Liu, Neighbourhood preserving projections (NPP): a novel linear dimension reduction method, Lecture Notes in Computer Science, vol. 3644, Springer, Berlin, 2005, pp. 117–125.

[9] G.H. Folub, C.F. Van Loan, Matrix Computation, third ed., The Johns Hopkins University Press, MD, USA, 1996.

[10] S. Zhang, T. Sim, Discriminant subspace analysis: a Fukunaga–Koontz approach, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 1732–1745.

[11] H. Cevikalp, M. Neamtu, M. Wilkes, A. Barkana, Discriminative common vectors for face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (1) (2005) 4–13.

[12] P. Howland, H. Park, Generalized discriminant analysis using the generalized singular value decomposition, IEEE Transactions on Pattern Analysis and Machine Intelligence 8 (2004) 995–1006.

[13] J. Ye, R. Janardan, C.H. Park, H. Park, An optimization criterion for generalized discriminant analysis on undersampled problems, IEEE Transactions on Pattern Analysis and Machine Intelligence 8 (2004) 982–994.

[14] H. Li, K. Zhang, T. Jiang, Efficient and robust feature extraction by maximum margin criterion, IEEE Transactions on Neural Networks 17 (1) (2006) 157–165.

[15] L.F. Chen, Y.M. Liao, M.T. Ko, J.C. Lin, G.J. Yu, A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recognition (2000) 1713–1726.

[16] R. Huang, Q. Liu, H. Lu, S. Ma, Solving the small sample problem of LDA, in: Proceedings of the IEEE International Conference on Pattern Recognition, vol. 3, 2002, pp. 29–32.

[17] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data with application to face recognition, Pattern Recognition 34 (10) (2001) 2067–2070.

[18] S. Ji, J. Ye, A unified framework for generalized discriminant analysis for generalized linear discriminant analysis, 2008, pp. 1–7.

[19] N. Kwak, Principal component analysis based on L1-norm maximization, IEEE Transactions on Pattern Analysis and Machine Intelligence (2008) 1672–1680.

[20] H. Aanas, R. Fisker, K. Astrom, J. Carstensen, Robust factorization, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (9) (2002) 1215–1225.

[21] C. Ding, D. Zhou, X. He, H. Zha, R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization, in: Proceedings of the 23rd International Conference on Machine Learning, June 2006.

[22] A. Baccini, P. Besse, A.D. Falguerolles, A L1-norm PCA and a heuristic approach, in: E. Diday, Y. Lechevalier, P. Opitz (Eds.), Ordinal and Symbolic Data Analysis, Springer, Berlin, 1996, pp. 359–368.

[23] Q. Ke, T. Kanade, Robust subspace computation using L1 norm, Technical Report CMU-CS-03-172, Carnegie Mellon University, 〈http://citeseer.ist.psu.edu/ke03robust.html〉, August 2003.
[24] Q. Ke, T. Kanade, Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2005.
[25] J. Liu, S. Chen, X. Tan, A study on three linear discriminant analysis based methods in the small sample size problem, Pattern Recognition 41 (2008) 102–116.
[26] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning—Data Mining, Inference, and Prediction, Springer-Verlag, Berlin, 2001.
[27] M. Nikolova, Analytical bounds on the minimizers of (nonconvex) regularized least-squares, AIMS Journal on Inverse Problems and Imaging 1 (4) (2007) 661–677.
[28] M. Nikolova, Minimizers of cost-functions involving non-smooth data-fidelity terms. Application to the processing of outliers, SIAM Journal on Numerical Analysis 40 (3) (2002) 965–994.
[29] M. Nikolova, Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization, SIAM Journal on Imaging Sciences 1 (1) (2008) 2–25.
[30] S. Durand, M. Nikolova, Denoising of frame coefficients using L1 data-fidelity term and edge-preserving regularization, SIAM Journal on Multiscale Modeling and Simulation 6 (2007) 547–576.
[31] M. Nikolova, M. Ng, Analysis of half-quadratic minimization methods for signal and image recovery, SIAM Journal on Scientific Computing 27 (3) (2005) 937–966.
[32] D. Geman, G. Reynolds, Constrained restoration and recovery of discontinuities, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (1992) 367–383.
[33] S. Geman, D.E. McClure, Statistical methods for topographic image reconstruction, in: Proceedings of the 46th Session of the ISI, Bulletin of the ISI, vol. 52, 1987, pp. 22–26.
[34] T. Hebert, R. Leahy, A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors, IEEE Transactions on Medical Imaging 8 (1989) 194–202.
[35] Y.G. Leclerc, Constructing simple stable description for image partitioning, International Journal on Computer Vision 3 (1989) 73–102.
[36] C. Yang, J.C. Meza, L. Wang, A trust region direct constrained minimization algorithm for the Kohn–Sham equation, SIAM Journal on Scientific Computing 29 (2007) 1854–1875.
[37] Z. Bai, C. Yang, From self-consistency to SOAR: solving nonlinear eigenvalue problems, SIAM News, April, 2006.
[38] C. Yang, W. Gao, J. Meza, On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems, with LBNL Report 63037, 2007.
[39] D.J. Newman, S. Hettich, C.L. Blake, C.J. Merz, UCI repository of machine learning databases, 〈http://www.ics.uci.edu/mlearn/MLRepository.html〉, 1998.
[40] M. Loog, R.P. Duin, Linear dimensionality reduction via a heteroscedasitc extension of LDA: the Chernoff criterion, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (2004) 732–739.
[41] 〈http://images.ee.umist.ac.uk/danny/database.html〉.
[42] L. Veer, H. Dai, M. Vijver, A.M. Hart, M. Mao, et al., Gene expression profiling predicts clinical outcome of breast cancer, Letters to Nature 415 (2002) 530–536.
[43] N. Barkai, D.A. Tterman, et al., Broad patterns of gene expression revealed by clustering analysis of tumour and normal colon tissues probed by oligonucleotide arrays, in: Proceedings of National Academy of Sciences of the United States of American, vol. 96, 1999, pp. 6745–6750.
[44] S. Armstrong, et al., MLL translocations specify a distinct gene expression profile that distinguishes a unique leukaemia, Nature Genetics 30 (2002).
[45] J. Khan, et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Medicine 7 (6) (2001).

About the Author—ZHIZHENG LIANG graduated from the Department of Automation at TianJin University of Technology and Education in 1999. He received his M.Sc. from the Department of Automation in Shandong University in 2001 and his Ph.D. in pattern analysis and intelligent systems from Shanghai Jiaotong University, (China) in 2005. Then he was a postdoctoral researcher at Shenzhen Graduate School in Harbin Institute of Technology. And then he was a research fellow at City University of Hong Kong. Now he works at School of Computer Science and Technology, China University of Mining and Technology. His current interests include image processing, pattern recognition and machine learning.

About the Author—Y.F. LI (SM'01) received the Ph.D. degree in robotics from the Department of Engineering Science, University of Oxford, Oxford, UK, in 1993. From 1993 to 1995, he was a Postdoctoral Research Associate in the AI and Robotics Research Group, Department of Computer Science, University of Wales, Aberystwyth, UK. In 1995, he joined the City University of Hong Kong, Kowloon, Hong Kong, where currently he is an associate professor in the Department of Manufacturing Engineering and Engineering Management. His research interests include robotics, machine vision, robot sensing, and sensor-based control. Dr. Li is an associate editor of the IEEE Transactions on Automation Science and Engineering.