Pattern Recognition 45 (2012) 2708–2718

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Robust classification using $\ell_{2,1}$-norm based regression model

Chuan-Xian Ren [a,b], Dao-Qing Dai [a,*], Hong Yan [b,c]

[a] Center for Computer Vision and Department of Mathematics, Sun Yat-Sen University, Guangzhou 510275, PR China
[b] Department of Electric Engineering, City University of Hong Kong, Kowloon, Hong Kong
[c] School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 5 July 2011
Received in revised form
3 January 2012
Accepted 6 January 2012
Available online 17 January 2012

Keywords:
$\ell_{2,1}$-norm
Sparsity regularization
Nearest subspace
Multiple task learning
Dummy variables

ABSTRACT

A novel classification method using $\ell_{2,1}$-norm based regression is proposed in this paper. The $\ell_{2,1}$-norm based loss function is robust to outliers or large variations distributed in the given data, and the $\ell_{2,1}$-norm regularization term selects correlated samples across the whole training set with grouped sparsity. A probabilistic interpretation under the multiple task learning framework presents theoretical foundation for the optimal solution. Complexity analysis of our proposed classification algorithm is also presented. Several benchmark data sets including facial images and gene expression data are used for evaluating the effectiveness of the new proposed algorithm, and the results show competitive performance particularly better than those using dummy matrix as the response variables. This result is very useful since it is important for selecting appropriate response variables in classification oriented regression models.

1. Introduction

In general supervised learning, each training data sample consists of a vector of observations x and a class-label z. The classification problem can be stated as follows: given training data, we need to produce a rule (or *classifier*) h, so that $h(x)$ can be evaluated for any possible value of x and that the class attributed to any new observation, specifically

$$\hat{z} = h(x)$$

is as close as possible to the true class label z. For the training data set, the true labels z_i are known but would not necessarily match their in-sample approximations $\hat{z}_i = h(x_i)$. For new observations, the true labels z_j are unknown, but it is a prime target for the classification procedure that the approximation $\hat{z}_j = h(x_j) \approx z_j$ is as close to the correct one as possible, in which the quality of this approximation needs to be judged on the basis of the statistical or probabilistic properties of the overall population from which future observations will be drawn, therefore it can be viewed as a regression problem.

Under the assumption that samples from a specific object class lie on a linear subspace, an efficient linear regression-based classification (LRC) is presented in [1] for the problem of face recognition. The class-specific models of the registered users are developed using the downsampled gallery images, thereby defining the task of face recognition as a problem of linear regression. It is worth noting that in this sense, the response variables are test samples instead of binary class labels. Least-square criterion is used to estimate the vectors of the parameters for a given probe against all class models. Finally, the decision rules are given in favor of the class with the minimum error estimation.

Aiming to generalize the representative capacity of the training samples, Li and Lu [2] propose a novel nearest feature line (NFL) method to virtually enlarge the training set for face recognition. The basic idea of NFL is to use a linear way to interpolate and extrapolate each pair of prototypes belonging to the same class and to model the possible variants of the training samples. Because of its generalization capacity, NFL has been widely used in many computer vision and pattern recognition applications [2,3]. Chien and Wu [4] proposed a new nearest feature space (NFS) method, in which the feature space of each class is constructed in the manner similar to the FL constructed in NFL to model the variants of the training samples, thus NFS needs to construct only one FS for each class of the training samples. Lu and Tan [5] proposed the use of the nearest feature space metric to seek a subspace analysis to improve the discriminant power of the subspace for classification.

Another simple classifier is the nearest centroid classifier [6,7], in which each class is represented by a single centroid, usually the mean of all training vectors within a class. During testing, the distance between a test vector and each class centroid is found and the vector is assigned the label of the nearest class.

On the other hand, the extremely high dimensionality and low sample size of given data create problems when applying

* Corresponding author. Tel.: +86 20 8411 0141; fax: +86 20 8403 7978.
E-mail address: stsddq@mail.sysu.edu.cn (D.-Q. Dai).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2012.01.003

traditional classifiers to the tasks of facial images recognition and cancer data classification, referred to as the *curse of dimensionality* in pattern recognition literature [4]. Nevertheless, the solution appears to lie in utilizing support vector machines (SVM) [8–10], which uses a *kernel trick* to transform data into a kernel-space with an even greater separability. Although originally formulated to deal with binary-class problems, SVM has since been adapted successfully to perform multi-class classification [10,9]. Recently, sparsity regularization has been investigated and applied to classification studies. Sparsity has become the mainstay among data analysis requirements owing to important practical benefits such as: (i) simpler, more robust models; (ii) countering of overfitting; (iii) discovery of the most-relevant features; and (iv) potentially better modeling of prior information. Based on these benefits, as well as other motivations, sparsity continues to enjoy great interest in various models and applications. ℓ_1-SVM was proposed to perform classification using the ℓ_1-norm regularization that tends to give a sparse solution [8].

In this paper, we propose an efficient and robust classification method to exploit $\ell_{2,1}$-norm minimization of both loss function and regularization. Instead of using an ℓ_2-norm based loss function that is sensitive to outliers, the $\ell_{2,1}$-norm based loss function is adopted in our work to decrease the negative impact of outliers. Meanwhile, the $\ell_{2,1}$-norm regularization is performed to select class-specific samples across all data points with some sparsity, i.e., the test sample will have small residues in the correct class-subspace and simultaneously has large residues in the false subspace. Besides, the loss function has a rotational invariant property [11], thus our proposed new method is called RRC (*Rotational invariant norm based Regression for Classification*) for short.

Most similar to our work is the RFS of Nie et al. [12] which uses joint $\ell_{2,1}$-norm minimization for feature selection, then SVM classifier is exploited for classification. Thus it is distinct to our method as we aim to directly obtain the class labels for the test samples using the regression model. However, the main objective of RFS is dimensionality reduction, i.e., selecting the important variables. Meanwhile, the response variables of RFS is the binary class label $\{0,1\}$, not the test samples as shown in our novel model. Finally, the coefficient matrix obtained in RFS method is used for detecting non-zero elements and then selecting out the associated features, however, it is exploited for finding the correct subject with minimal error in our model.

Sun et al. [13] and Liu et al. [14] developed a similar model for $\ell_{2,1}$-norm regularization to couple feature selection across multiple tasks feature learning. Such regularization has close connections to group Lasso [15–17]. Specifically, Sun et al. proposed an equation constrained $\ell_{2,1}$-norm minimization model for the multiple measurement vector problem, then they presented a dual reformulation of the convex optimization problem and developed a new algorithm based on the prox-method in [13]. Liu et al. considered the $\ell_{2,1}$-norm regularized least square regression model for joint feature selection from multiple tasks in [14]. Then they proposed that the computation can be accelerated by reformulating the objective function as two equivalent smooth convex optimization problems which are then solved via the Nesterovs method.

Another related work is [18] which emphasizes that a test sample can be represented as a sparse linear combination of the whole training set, so that the non-zero parts of the combination coefficients imply the correct subject. This algorithm produces an ℓ_1-norm minimized least square problem. However, for a small data set or in the small sample size case, if the number of training samples in each class do not adequately span a subspace, then the combination coefficient will not be *sparse enough* and thus will decrease the classification performance [19,20]. The main information with primate goals, objective functions, inputs and outputs of these works has been summarized in Table 1.

The rest of this paper is organized as follows. In Section 2, we summarize the notations and the definitions of norms before presenting the classification model and its detailed algorithm. A probabilistic interpretation is provided for the regularized optimization criterion. Complexity analysis is also presented in this section. Extensive experiment results on facial images and gene expression data sets are shown in Section 3, and the RRC algorithm is compared with other *state-of-the-art* methods. Section 4 concludes this paper.

2. Main algorithm using $\ell_{2,1}$-norm based regression

In this section, we will present a robust classification algorithm using the $\ell_{2,1}$-norm based regression model. The corresponding loss function can be expected to reduce the effect of outliers. The solved coefficients will be directly used to determine the identity of the test samples.

2.1. Basic model and the matrix norms

The formulation of the basic regression model for classification is explained as below. Let $Y=[y_1,y_2,\ldots,y_N]\in\mathbb{R}^{d\times N}$ be an input matrix representing N test samples and $A=[a^1;a^2;\ldots;a^n]\in\mathbb{R}^{n\times d}$ be the matrix containing the n training vectors for all classes, concatenated row-by-row and class labels. we aim to obtain matrix $X\in\mathbb{R}^{n\times N}$ such that

$$Y\approx A^TX.$$

Table 1
Comparison of related works.

Method	Goal	Objective function	Input	Output
LRC [1]	Classification	$\min_{x_k}\|A_k^Tx_k-y\|_2^2$ $(k=1,2,\ldots,C)$	$A_k\in\mathbb{R}^{n_k\times d}$ is a sub-dictionary; $y\in\mathbb{R}^d$ is a test sample;	x_k is the *comb. coeff.* for *samples*
SRC [18]	Classification	$\min_x\|x\|_1$ s.t. $A^Tx=y$	$A\in\mathbb{R}^{n\times d}$ is the dictionary; $y\in\mathbb{R}^d$ is a test sample;	x is the *sparse represent.* for *samples*
aMTFL [14]	Feature selection	$\min_X\sum_{j=1}^{t}\|A_j^Tx_j-y_j\|_2^2+\rho\|X\|_{2,1}$	$A_j\in\mathbb{R}^{n\times m_j}$ is a sub-dictionary; $y_j\in\mathbb{R}^{m_j}$ is the j-th response;	x_j is the *weight vector* for *features*
RFS [12]	Feature selection	$\min_W\|X^TW-Y\|_{2,1}+\rho\|W\|_{2,1}$	$X\in\mathbb{R}^{d\times n}$ is the dictionary; $Y\in\mathbb{R}^{n\times C}$ is the dummy matrix;	W is the *weight matrix* for *features*
RRC	Classification	$\min_X\|A^TX-Y\|_{2,1}+\rho\|X\|_{2,1}$	$A\in\mathbb{R}^{n\times d}$ is the dictionary; $Y\in\mathbb{R}^{d\times N}$ denotes test samples;	X is the *comb. coeff.* for *samples*

Beginning from this formulation we may now consider several alternatives, each of which incorporates some desirable characteristics such as sparsity or group sparsity on the matrix X.

We summarize the notations and definitions of norms here. Matrices are written in uppercase letters while vectors are written in lowercase letters. For matrix $A=(a_{ij})$, its ith row and jth column are denoted by a^i, a_j respectively.

The ℓ_p-norm of the vector $v\in\mathbb{R}^n$ is defined as $\|v\|_p=(\sum_{i=1}^n |v_i|^p)^{1/p}$, where $p\geq 1$ is a real number. The vector norms treat an $n\times m$ matrix as a vector of size nm, and use one of the familiar vector norms. For example, using the p-norm for vectors, we get

$$\|A\|_p=\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^p\right)^{1/p}.$$

The special case $p=2$ is the *Frobenius* norm, which is defined as

$$\|A\|_F=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}a_{ij}^2}=\sqrt{\sum_{i=1}^{n}\|a^i\|_2^2}. \tag{1}$$

The $\ell_{2,1}$-norm of a matrix is first introduced in [11] as a rotational invariant ℓ_1-norm (R_1-norm) and also used for multi-kernel learning [16], multi-task learning [14] and high order tensor factorization [21]. It is defined as

$$\|A\|_{2,1}=\sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m}a_{ij}^2}=\sum_{i=1}^{n}\|a^i\|_2, \tag{2}$$

which is rotational invariant for rows $\|AR\|_{2,1}=\|A\|_{2,1}$ for any rotational matrix R [11,12].

The $\ell_{2,1}$-norm can be generalized to $\ell_{p,q}$-norm

$$\|A\|_{p,q}=\left(\sum_{i=1}^{n}\left(\sum_{j=1}^{m}|a_{ij}|^p\right)^{q/p}\right)^{1/q}. \tag{3}$$

Note that $\ell_{p,q}$-norm is a valid norm because it satisfies the three norm conditions, including the triangle inequality $\|A\|_{p,q}+\|B\|_{p,q}\geq\|A+B\|_{p,q}$ [12]. However, The symmetric property is not held for the $\ell_{p,q}$-norm for an arbitrary matrix $A\in\mathbb{R}^{n\times m}$, i.e., $\|A\|_{p,q}\neq\|A^T\|_{p,q}$ (Fig. 1).

2.2. Objective criterion as a constrained problem

We focus on the following problem:

$$\arg\min_X J(X)=\|A^TX-Y\|_{2,1}, \tag{4}$$

in which $X=[x_1,x_2,\ldots,x_N]$ is the *representation* matrix and $x_i\in\mathbb{R}^n$ $(i=1,2,\ldots,N)$ is the *representation* of test sample y_i using the entire training vectors.

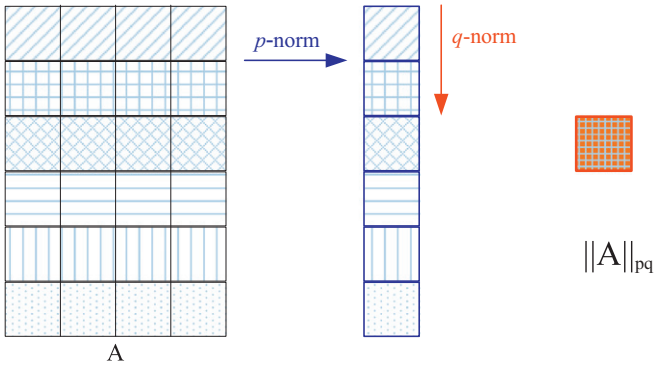

Fig. 1. Demonstration of computing the $\ell_{p,q}$-norm for matrix A.

We also add a regularization term $R(X)$ with parameter ρ. Regularization methods have been widely used in the literature of pattern recognition and machine learning [22,23]. By incorporating prior information into the formulation, the regularization techniques have been shown to be powerful in making the stable solution. The $\ell_{2,1}$ penalty is used in this paper since it is convex and the corresponding objective criterion can be easily optimized.

Let $R(X)=\|X\|_{2,1}$ and ρ is the penalty coefficient, the regularized form of Eq. (4) is equivalent to

$$\begin{aligned}\arg\min_X J(X)+\rho R(X)&=\arg\min_X\|A^TX-Y\|_{2,1}+\rho\|X\|_{2,1}\\&=\arg\min_X\frac{1}{\rho}\|A^TX-Y\|_{2,1}+\|X\|_{2,1}.\end{aligned}$$

Suppose $E=(1/\rho)(Y-A^TX)$ and $U=\binom{X}{E}$, we have

$$\frac{1}{\rho}\|A^TX-Y\|_{2,1}+\|X\|_{2,1}=\|E\|_{2,1}+\|X\|_{2,1}=\left\|\begin{pmatrix}X\\E\end{pmatrix}\right\|_{2,1}=\|U\|_{2,1},$$

subject to $A^TX+\rho E=Y$.

Let $B=(A^T,\rho I)$, then the objective criterion can be formulated as

$$U^*=\arg\min_U\|U\|_{2,1}\quad\text{s.t. } BU=Y. \tag{5}$$

This optimization problem has been widely used in the multiple measurement vector model in the signal processing community, and recently an efficient algorithm was proposed to solve this specific problem [12], in which the gradient-descent approach is implemented on the *Lagrangian* function to obtain the iterative solution

$$U=D^{-1}B^T(BD^{-1}B^T)^{-1}Y, \tag{6}$$

in which D is a diagonal matrix with the ith diagonal element as $d_{ii}=1/(2\|u^i\|_2)$. Usually, the matrix $BD^{-1}B^T$ may be singular thus it is added by a diagonal matrix ζI in the computation, where I is the unit matrix and ζ is a very small positive constant. The reported results show that the algorithm is more efficient than other existing algorithms. Theoretical analysis guarantees that the gradient descent based method will converge to the global optimum.

Since the problem in Eq. (5) is a convex problem, U is a global optimum solution to the problem and the convergence has been proven in [12].

2.3. Classification

The optimal representation $\hat{X}$ is the first n rows of matrix U^*. When the matrix $\hat{X}$ is used for classification, we can define C functions $\delta_s:[0,\ldots,0,\hat{x}_s^T,0,\ldots,0]\in\mathbb{R}^n$ where $\hat{x}\in\mathbb{R}^n$ is the representation for the test sample y and the δ_s selects only those $\hat{x}$ that correspond to class s as [18] does. The approximation in terms of the coefficients associated with the sth class is then $\hat{y}=A^T\delta_s(\hat{x})$, and classification can be achieved by assigning y to the identity (ID) that minimizes $\|y-A^T\delta_s(\hat{x})\|_2$. Fig. 2 presents a simple demonstration for the classification rule, where the subject s^* with minimal reconstruction residue r_s is considered to be the true ID.

The pseudo-code of our classification method is described in Algorithm 1. In each iteration, U is calculated with the current D, and then D is updated based on the current calculated U. The iteration procedure is repeated until the algorithm converges. It is worth noting that the Algorithm 1 monotonically decreases the objective of the problem in Eq. (5) within each iteration, which can be theoretically proven as [12] does (thus we omit it here).

Particularly, in the matrices $U\in\mathbb{R}^{(n+d)\times N}$ and $X\in\mathbb{R}^{n\times N}$, d is the dimensionality of training samples, n and N are the sizes of training set and test set respectively. However, a similar matrix obtained by the RFS method belongs to $\mathbb{R}^{(n+d)\times C}$, in which C is the

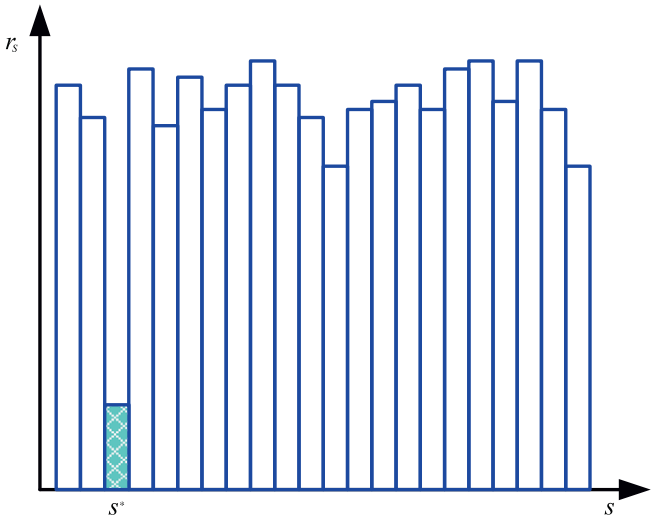

Fig. 2. Demonstration of classification rule using regularized $\ell_{2,1}$-norm minimization, where the subject s^* with minimal reconstruction residue is considered to be the true class.

number of classes and a dummy matrix $Y \in \mathbb{R}^{n\times C}$ is used as the response variables in the RFS approach. Then, the matrix X is exploited to determine the reconstruction error and the optimal subject identity, i.e., classification in our algorithm; nevertheless, the coefficient matrix obtained by RFS is used for choosing the important variables corresponding to the significant entries. Thus, both of the properties and purposes of the coefficient matrices are different between Nie's RFS and our RRC.

Algorithm 1. Classification using $\ell_{2,1}$-norm based regression model.

- **Input:** $A \in \mathbb{R}^{n\times d}$; $Y \in \mathbb{R}^{n\times N}$; ρ;
- **Output:** ID(Y)

1. Set $B = [A^T, \rho I]$;
2. Initialize $D_t \in \mathbb{R}^{(n+d)\times(n+d)}$ as an identity matrix;
3. Iteratively compute U using Eq. (6);
4. **For** $j = 1, 2, \ldots, N$
   - Get $r_s(y_j) = \|y - A^T\delta_s(x_j)\|_2$, $s = 1, 2, \ldots, C$;
   - $\mathrm{ID}(y_j) = \min_{s=1,2,\ldots,C} r_s(y_j)$;

   End
5. **Return** $\mathrm{ID}(Y) = (\mathrm{ID}(y_1), \mathrm{ID}(y_2), \ldots, \mathrm{ID}(y_N))$;

2.4. A probabilistic interpretation for the optimal solution

From a linear regression point of view, classification model $Y=AX$ is a *multi-features matching* problem, which will obtain linear representations X through the simultaneously multiple features interaction of many samples. As a result, our new classification method is a generalized and regularized regression model using $\ell_{2,1}$-norm as the loss function and penalty function. In this section, we intend to interpret this multi-features matching objective function under the framework of *multi-task learning* problem [24], namely, each individual feature matching subproblem can be viewed as a corresponding task learning problem and the prior on X embodies the correlation between different tasks, thus it is convenient to present a probabilistic interpretation for the regularized optimization problem.

Let $a^j = [a_1^j, a_2^j, \ldots, a_n^j] \in \mathbb{R}^{1\times n}$ be the vector for the jth task, $A = [a^1; a^2; \ldots; a^n]^T \in \mathbb{R}^{n\times d}$, $y^j = [y_1^j, y_2^j, \ldots, y_N^j] \in \mathbb{R}^{1\times N}$, and $Y = [y_1, y_2, \ldots, y_N] \in \mathbb{R}^{d\times N}$. In the previous section, we consider linear models

$$y^j = (A^T)^j X, \quad j = 1, 2, \ldots, d, \tag{7}$$

in which $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{n\times N}$ is composed of the coefficient matrix of equation $Y = A^T X$.

Assume that, given the dictionary A, the corresponding target y^j for the jth task has a distribution of

$$p(y_j|X, A, \sigma_j) = \sqrt{\frac{\sigma_j}{2\pi}} \exp\left(-\frac{\sigma_j \|y^j - (A^T)^j X\|_2}{2}\right). \tag{8}$$

Denote $\sigma = [\sigma_1, \sigma_2, \ldots, \sigma_d]^T \in \mathbb{R}^d$ and assume that data $\{y^j\}_{j=1}^d$ are drawn independently from the distribution in Eq. (8), then the likelihood function can be written as

$$p(Y|X, A, \boldsymbol{\sigma}) = \prod_{j=1}^{d} p(y^j|X, A, \sigma_j). \tag{9}$$

To capture the task relatedness, a prior on X is defined as follows. The ith row of X, denoted as $x^i \in \mathbb{R}^{1\times N}$, corresponds to the ith representation for the whole dictionary. Assume that x^i is generated according to the exponential prior

$$p(x^i|\varsigma_i) \propto \exp(-\|x^i\|\varsigma_i), \quad i = 1, 2, \ldots, n, \tag{10}$$

in which $\varsigma_i > 0$ is the hyper-parameter.

Denote $\boldsymbol{\varsigma} = [\varsigma_1, \varsigma_2, \ldots, \varsigma_n]^T \in \mathbb{R}^n$, and assume that $x^1, x^2, \ldots, x^n$ are drawn independently from the prior in Eq. (10). Then the prior for X can be expressed as

$$p(X|\delta) = \prod_{i=1}^{n} p(x^i|\varsigma_i). \tag{11}$$

It follows that the posterior distribution for X, which is proportional to the product of the prior and the likelihood function [25], is given by

$$p(X|A, Y, \sigma, \boldsymbol{\varsigma}) \propto p(Y|X, A, \sigma) p(X|\varsigma). \tag{12}$$

Taking the negative logarithm of Eq. (12) and combining with Eqs. (8)–(11), we obtain the maximum posterior estimation of X by minimizing

$$\sum_{j=1}^{d} \sigma_j \|y^j - (A^T)^j X\|_2 + \sum_{i=1}^{n} \varsigma_i \|x^i\|_2. \tag{13}$$

For the convenience of discussion, we assume that $\sigma = \sigma_j, (j = 1, 2, \ldots, d)$ and $\varsigma = \varsigma_i, (i = 1, 2, \ldots, n)$. We then obtain from Eq. (13) the following joint $l_{2,1}$-norm minimized regression problem:

$$\arg\min_X \|A^T X - Y\|_{2,1} + \rho \|X\|_{2,1}, \tag{14}$$

in which $\rho = \varsigma/\sigma$ and $\|\cdot\|_{2,1}$ is the $\ell_{2,1}$-norm of the given matrix.

When the number of tasks equals one, the prior in Eq. (10) reduces to the Laplace prior distribution [26]. It is easy to show that in this case the problem in Eq. (14) reduces to the ℓ_1-norm regularized optimization problem. It is also worth noting that the independent assumption of X is only concerned for the rows rather than columns, thus it has little influence on the *group effect* between features.

2.5. Complexity analysis

The computation of our RRC algorithm for classification involves two main steps: coefficient matrix computation and least residue classification. We use the term *flam* [27,28], a compound operation consisting of one addition and one multiplication, to measure the operation counts.

The complexity of the first step focusses on obtaining the inverse matrix of $BD_t^{-1}B^T \in \mathbb{R}^d$, which requires $\mathcal{O}(d^3)$ flams. Noting that the parameter d is the reduced dimensionality obtained by some dimensionality reduction methods such as Fisherfaces [29–31], random projection [18] or down-sampling [1]. Thus the real value of d used for computing the representation matrix X is usually much smaller than the number of training samples n. The classification step is composed of some subspace projection and nearest subspace search procedures and thus it will require NCd^2 flams, in which N is the number of test samples and C is the number of classes.

In a similar way, we can summarize the complexity for the LRC algorithm. The computation also includes subspace regression and nearest subspace search steps. In the first step, all the subspace projections (the hat matrices) can be predetermined once and for all, so the complexity can be approximated by $\sum_{i=1}^{C} n_i^3$, in which n_i denotes the number of training samples in the ith class.

For the SRC method, it is shown in [18] that the complexity of sparse representation for test sample y is $\mathcal{O}(n^2)$. In the classification stage, the complexity is completely identical to that of LRC and our proposed approach. These detailed results are summarized in Table 2.

3. Experiment results

In this section, we will firstly present a numerical instance using synthetic data to evaluate the regularization parameter selection of RRC algorithm. To validate the real classification performance of our method, we systematically compare it with nearest neighborhood (1NN), SVM, SRC, RFS and LRC methods with the benchmark databases. We believe the results are helpful in supporting our novel viewpoint that the new proposed method has a competitive performance with other methods for facial images and gene expression data classification. The important statistics of these image databases are summarized in Table 3. The recognition accuracy is exploited to measure the performance of the classifier. The training process is randomly repeated for 100 times and then the average values are computed to obtain the final results.

3.1. Parameter selection and a synthetic example

Empirical results show the classification performance of our new method is stable if the parameter ρ is restricted in a small range such as (0, 1), and the corresponding coefficient matrix X is indeed discriminant for the purpose of classification. Therefore, the regularization parameter can be determined by cross-validation [32] from a small range of positive values throughout the computation.

As shown in Fig. 3, assume matrix A is composed of 10 samples belonging to two different classes

$$A = \begin{pmatrix} -0.30 & 0.83 & -0.21 & -1.03 & 0.14 & 0.93 & 1.97 & 1.92 & 2.04 & 2.23 \\ 0.02 & 1.53 & 0.63 & 0.95 & 0.52 & 2.93 & 2.18 & 3.60 & 1.27 & 2.43 \\ 0.05 & 0.47 & 0.18 & 0.31 & 0.26 & 2.35 & 0.43 & 2.10 & 1.97 & 1.63 \end{pmatrix},$$

Table 2
The complexities of some regression based classification methods.

Item	LRC	SRC	RRC
Coefficients computation	$\sum_{i=1}^{C} n_i^3$	$\mathcal{O}(n^2)$	$\mathcal{O}(d^3)$
Nearest subspace search	NCd^2	NCd^2	NCd^2

Table 3
Statistics for experiment databases.

Database	n_class	n_sample*/class	N_sample	Dimension
Yale	15	11	165	112×92
Extended Yale B	38	64	2432	112×92
PIE	68	43	2924	112×92
Georgia Tech.	50	15	750	112×92
Leukemia	2	(47,25)	72	3571
Colon	2	(22,40)	62	2000
Brain	5	(10,10,10,4,8)	42	5597
Lymphoma	3	(42,9,11)	62	4026

*: The notation n_sample/class for the gene expression data are different among the classes, thus they are reported in this table using a vector rather than a scale.

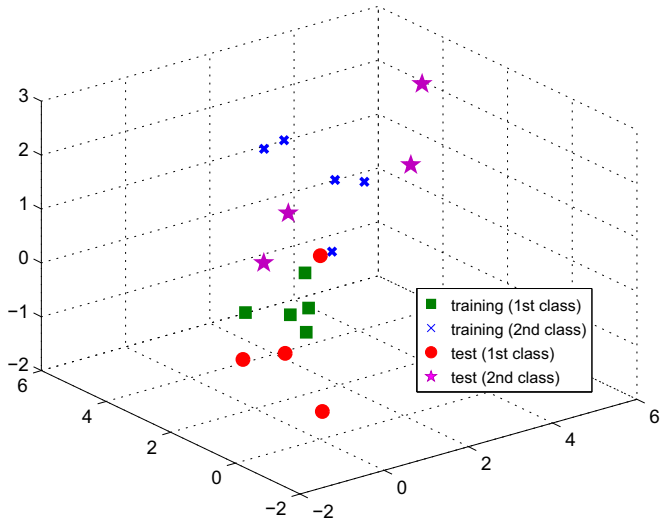

Fig. 3. Synthetic data for parameter selection and evaluation.

in which the first five columns (*green squares*) belong to the first class while the last five columns (*blue crosses*) belong to the second class. Now let matrix T be composed of the test data

$$T = \begin{pmatrix} -0.94 & -0.53 & -0.48 & -0.19 & 1.76 & 4.23 & 0.34 & 2.42 \\ -0.16 & 1.68 & -0.71 & -0.27 & 4.02 & 2.34 & 1.41 & 0.33 \\ -0.15 & -0.88 & -1.17 & 1.53 & -0.26 & 3.00 & 1.72 & 2.47 \end{pmatrix},$$

with the first four columns (*red circles*) belonging to the first class while the last four columns (*purple pentagrams*) belong to the second class.

We exploit the iterative procedure and set ρ to be 0.02, 0.2, 0.8, 1.2, 2, 5, respectively, to get the corresponding coefficient matrices and show the respective numerical results in Fig. 4. Obviously, the results with $\rho = 0.02$, 0.2 and 0.8 are almost identical to each other, and all the parameter values except for $\rho = 5$ give the correct classification results, thus the stability and effectiveness of our algorithm for classification is validated on this synthetic data. More results on the real data are presented as follows.

3.2. Facial image recognition

In the experiments, all facial images are aligned at the centers of the eyes and mouth, then cropped into resolution 92×112, and some dimensionality reduction methods are used to reduce the number of features and the computation complexity.

Firstly, to show the classification ability of the obtained coefficient matrix from our RRC method, we use the Yale and Extended Yale B databases to compute the coefficient matrices respectively, then

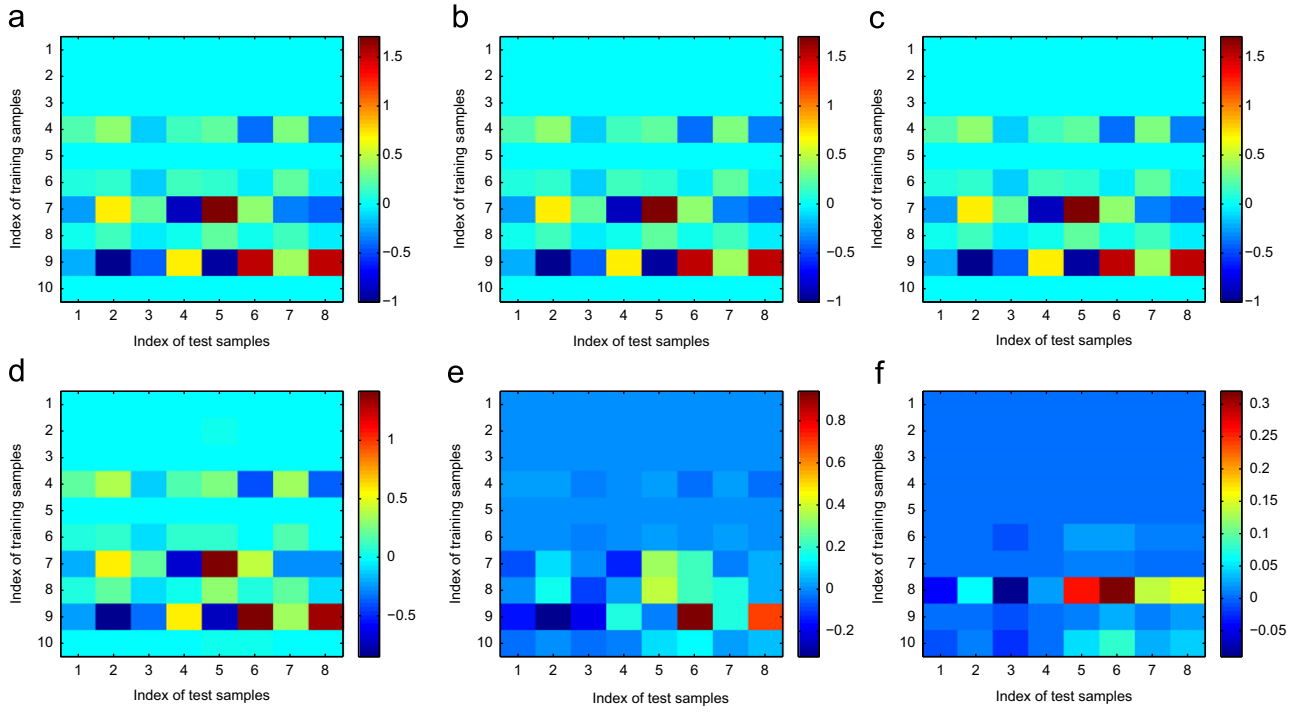

Fig. 4. Density maps of the computed coefficient matrices using different regularization parameters 0.02, 0.2, 0.8, 1.2, 2 and 5, respectively.

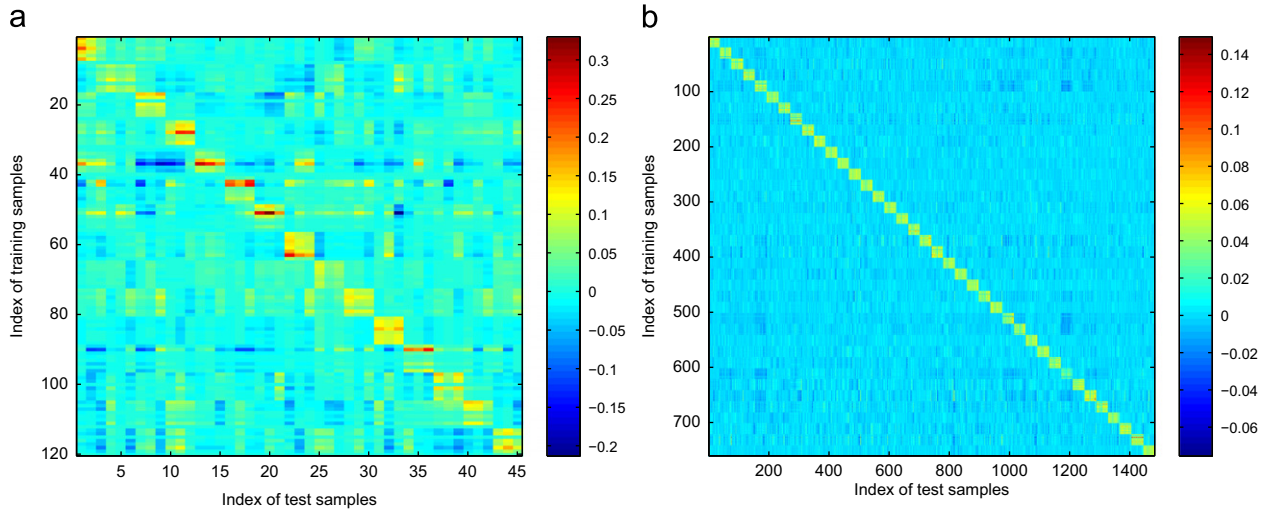

Fig. 5. Discriminant distributions of the coefficient matrices, in which (a) and (b) are associated with the results on the Yale and Extended Yale B databases, respectively.

present the scale data in Fig. 5. The Yale database[1] contains 165 gray scale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink. The Extended YaleB database[2] consists of 2432 frontal-face images of 38 individuals. The cropped and normalized images were captured under various laboratory-controlled lighting conditions [33].

The subfigure (a) corresponds to the scale coefficients from the Yale database, in which eight samples per subject are used to construct the dictionary A. We can see that the scale data show a clear profile with their block distribution depending on the columns and thus we believe it is indeed discriminant and helpful for classification. The subfigure (b) exhibits a more explicit presentation for the sample distribution since the larger training set and test set provide more interactive and discriminant information in the coefficient matrix.

3.2.1. *Results for the Yale data*

Fig. 6 presents the recognition results on the Yale database, in which Fisher features are uniformly exploited for a fair comparison. We can see that, when the six samples per class are used for training, 1NN and SVM obtain very similar results, 71.5% and 71.2% respectively. The SRC approach obtains a higher accuracy of 73.5%. The accuracy of the LRC approach is below 65% perhaps because the distribution of these data does not closely satisfy the assumption of being class-specific related. However, due to the robustness to the outliers and large variations, our RRC algorithm shows the best performance among these methods. The right

[1] Available at: http://cvc.yale.edu/projects/yalefaces/yalefaces.html.
[2] Available at: http://vision.ucsd.edu/leekc/ExtYaleDatabase/ExtYaleB.html.

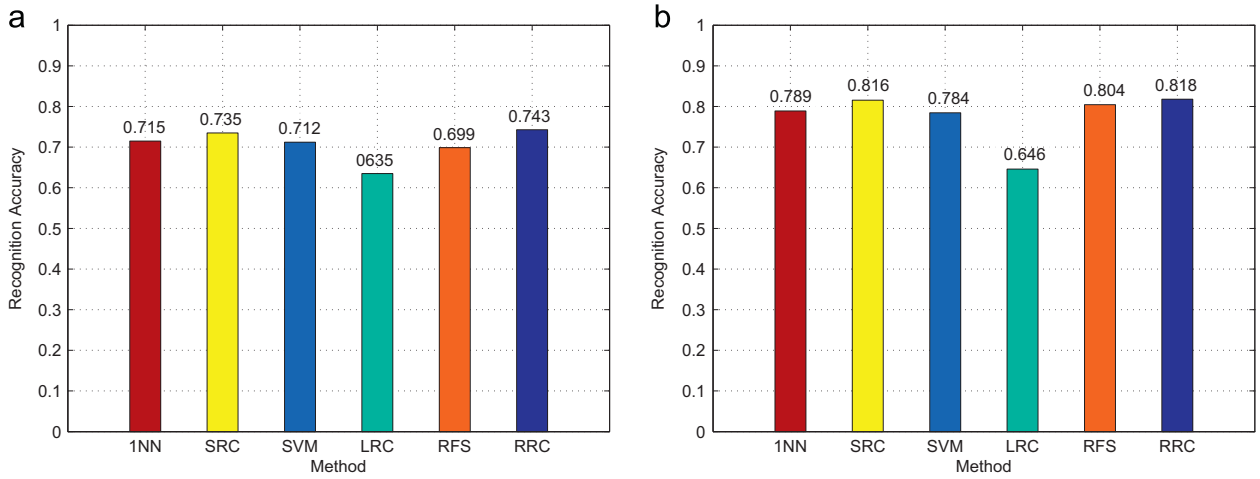

Fig. 6. The classification results on the Yale database, where the numbers of training samples in each subject are 6 (a) and 8 (b), respectively.

plane corresponds to the case that eight samples per class are used for training, and it shows similar recognition results to that of the left plane. The accuracy of the 1NN approach is close to that of SVM, and the SRC approach is more robust since the error impact is sufficiently taken account of in the objective model. The performance of the LRC method is not as robust as other methods, perhaps due to its ℓ_2-norm based computation of the projection matrix. Here, our method still exhibits competitive results as shown in the left plane. It is easy to see that our RRC method performs better than the RFS approach, which shows the difference of using test samples and dummy variables.

3.2.2. Results for the CMU data

The CMU PIE (Pose, Illumination, and Expression) images[3] were captured by 13 synchronized cameras and 21 flashes, under varying poses, illumination, and expressions [34]. For each subject, we manually select 43 images from five near frontal poses (C05, C07, C09, C27, C29) under different illuminations and expressions.

Fig. 7 presents the recognition performance on the CMU-PIE database. It is easy to see that, although only six samples in each class are used for training (the left plane), all the mentioned methods obtain satisfying results. Particularly the performances of the SRC and our Fisher-features based RRC methods are more robust and more competitive. When the number of training samples per class becomes 10, all these methods achieve a result close to 100%.

3.2.3. Results for the extended Yale B data

The recognition results of the Extended Yale B database are shown in Fig. 8, in which the left plane corresponds to the case of 10 samples per class used for training. Still, the Fisher features are used in each classification method. The results of the LRC and 1NN methods are close to each other, 78.83% and 81.50%, respectively. SVM preserves the robust characteristic and the accuracy is 83.24%. Here, the SRC and RRC methods obtain almost identical recognition results, which are higher than that of RFS approach, and they are outlier-resistant to the illumination variations because both of them use very robust loss function rather than the simple ℓ_2-norm. The right plane shows the results of 20 images per class used for training. The accuracy of LRC is not improved as the training samples are increased, thus it does not capture more discriminant information using class-specific model on this database. However, the results of other methods are enhanced with different degrees, for example, the accuracies of 1NN and SVM achieve 90.23% and 93.01%. Meanwhile, SRC and RRC methods obtain an identical result of 94.6%, still higher than that of RFS.

We also implement another group experiments to evaluate the performances of algorithms LRC, RRC and SRC. The Extended YaleB database is divided in five subsets as [1] does, subset 1 consisting of 266 images (7 images per subject) under nominal lighting conditions was used as the gallery while all others were used for validation. Subsets 2 and 3, each consisting of 12 images per subject, characterize slight-to-moderate luminance variations, while subset 4 (14 images per person) and subset 5 (19 images per person) depict severe light variations. The SRC and LRC methods were implemented with random projection and random down-sampling approaches, respectively. The results are shown in Table 4. The proposed RRC approach showed excellent performance for moderate light variations yielding 100% recognition accuracy for subsets 2 and 3. The recognition success however falls to 82.98% and 31.87% for subsets 4 and 5, respectively. The proposed RRC approach has shown better tolerance for considerable illumination variations compared to SRC method. The three algorithms, however, could not withstand severe luminance alterations.

3.2.4. Results for the Georgia Tech. data

The Georgia Tech (GT) database[4] consists of 50 subjects with 15 images per subject [35]. It characterizes several variations such as pose, expression, cluttered background and illumination.

The first eight images of each subject of Georgia Tech. data are used for training while the remaining seven served as probes [1]. Table 5 shows the comparison of the RRC with LRC and SRC, for the sake of fair comparison the random projection and down-sampling are used in the SRC and LRC approaches respectively. The proposed RRC algorithm outperforms the SRC approach by a margin of 9% and achieves a recognition accuracy of 71.35%, although it is slightly less than that of LRC.

Both Tables 4 (results for the subsets based EYaleB data) and 5 (results for the Georgia Tech. data) show that, when there are few training samples in each subject, or there exist some large variations such as pose, illumination and expression in the test samples, the robustness of SRC will be inferior to that of LRC and RRC. LRC method projects any test sample onto the subspace of each individual subject, and then the minimal reconstruction error criterion is used to determine the ID of the test samples, thus LRC presents a better recognition performance.

[3] Available at: http://www.ri.cmu.edu/projects/project_418.html.

[4] Available at: http://www.anefian.com/research/face_reco.htm.

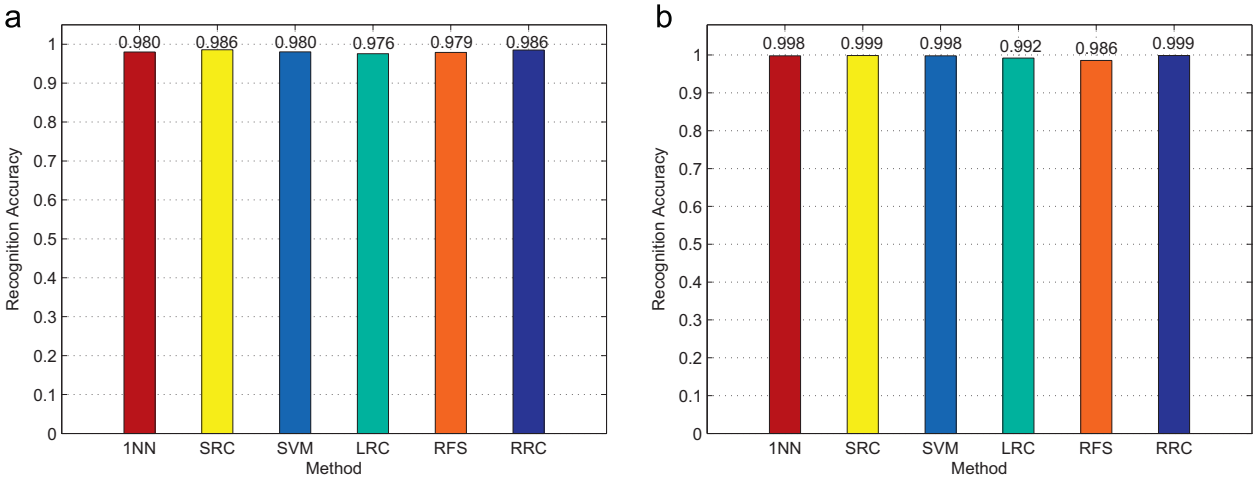

Fig. 7. The classification results on the CMU database, where the numbers of training samples in each subject are 6 (a) and 10 (b), respectively.

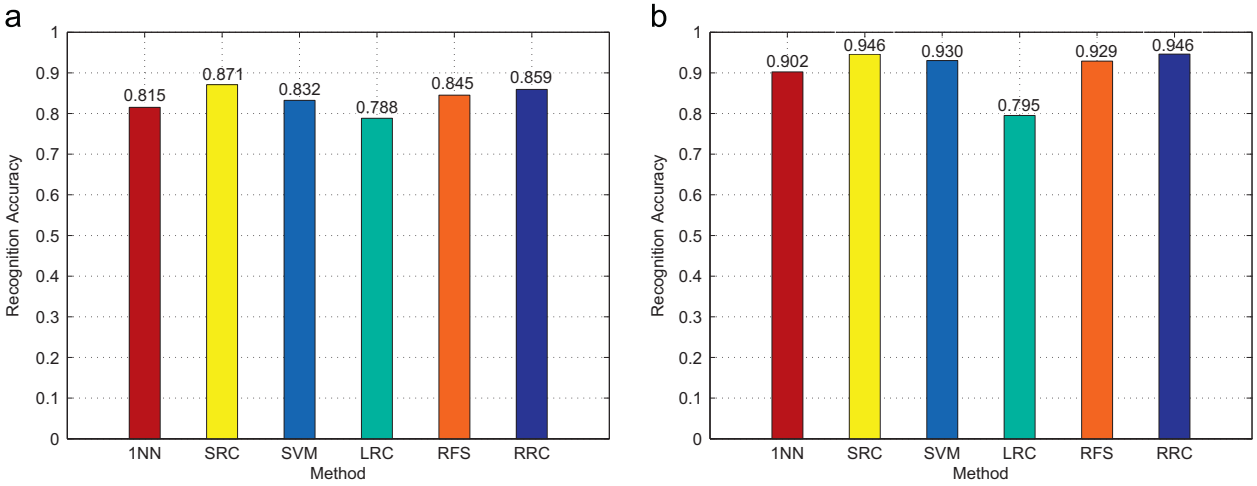

Fig. 8. The classification results on the EYaleB database, where the numbers of training samples in each subject are 10 (a) and 20 (b), respectively.

Table 4
Results for the extended Yale B database.

Approach	Subset2 (%)	Subset3 (%)	Subset4 (%)	Subset5 (%)
LRC	100	100	83.27	33.61
RRC	100	100	82.98	31.87
SRC	100	92.31	74.11	26.74

3.3. Gene expression data classification

The Leukemia gene expression data set[5] is a widely used benchmark data set. It consists of gene expression profiles of two classes of leukemia: acute lymphoblastic leukemia (ALL) and acute myeloblastic leukemia (AML). The data set consists of 7129 genes and 72 samples (47 ALL and 25 AML) [36].

The Colon cancer data set[6] has been frequently used in previous studies in gene selection and classification. It consists of the gene expression profiles of 2000 genes for 62 tissue samples among which 40 are colon cancer tissues and 22 are normal tissues [37].

The Brain data set[7] consists of gene expression profiles of 5597 genes for 10 medulloblastomas, 5 CNS AT/RTs, 5 renal and extrarenal rhabdoid tumors, and 8 supratentorial PNETs, as well as 10 non-embryonal brain tumors (malignant glioma) and 4 normal human cerebella [38].

The Lymphoma data set[8] consists of gene expression profiles of 4026 genes for 62 lymphoma tissue samples, among which 42 are of diffuse large B-cell lymphoma (DLBCL), 9 are of chronic lymphocytic leukaemia (CLL) and 150 are of follicular lymphoma (FL) [39].

For these gene expression data, the BSS/WSS method [40–42] is used for dimensionality reduction and feature selection. In particular, random projection is then used in the SRC method to guarantee the over-completeness of the dictionary matrix A, and the dimensionality is further reduced to a constant $(\sharp rows/\sharp columns)r$, in which r is a ratio parameter.

Table 6 presents the results of the LRC, 1NN, SRC, SVM, RFS and RRC methods on the Leukemia database. It should be noted that, all the gene expression data used in this section are randomly divided into 2/3 part into a training set and the remaining 1/3 part into a test set. The first row in the table denotes the number of selected features, 100, 200, 300, 400, 500, respectively. We can

[5] Available at: http://www.meds.com/leukemia/.
[6] Available at: http://www.molbio.princeton.edu/colondata.
[7] Available at: http://www.genome.wi.mit.edu/MPR/CNS.
[8] Available at: http://genome-www.stanford.edu/lymphoma/.

Table 5
Results for the Georgia tech. database.

Approach	LRC	RRC	SRC
Recognition rate (%)	72.83	71.35	62.14

Table 6
Results on the *Leukemia* data.

♯ variables	100	200	300	400	500
LRC	0.9258	0.9458	0.9496	0.9496	0.9446
1NN	0.9454	0.9625	0.9583	0.9600	0.9583
SRC	0.9096	0.8975	0.8812	0.8742	0.8742
SVM	0.9642	0.9638	0.9529	0.9533	0.9529
RFS	0.9575	0.9650	0.9604	0.9546	0.9529
RRC	0.9729	0.9704	0.9675	0.9662	0.9646

Table 7
Results on the *colon* data.

♯ variables	100	200	300	400	500
LRC	0.7338	0.7557	0.7638	0.7690	0.7524
1NN	0.7952	0.8110	0.8214	0.8252	0.8305
SRC	0.7971	0.7862	0.7819	0.7538	0.7624
SVM	0.8281	0.8510	0.8429	0.8548	0.8333
RFS	0.7848	0.8095	0.8190	0.8095	0.8010
RRC	0.8567	0.8681	0.8714	0.8719	0.8576

easily see that, when 100 features are selected for classification, the LRC method obtains an accuracy of 92.58% and then keeps it at a stable level of 95% as the number of features increases to 500. The 1NN, SVM and RFS methods obtain a better accuracy (96%) than LRC. The SRC approach ($r=0.5$) performs poorly on this database. However, we can see from the last row of the table, our new method shows the best classification performance and it is generally improved for all the cases with a different number of features. Particularly, the classification accuracies always remain close to 97% and thus it seems very robust for the leukemia data.

Table 7 shows the classification results for the colon data. Obviously, the performance of the LRC approach is not as good as that of other methods, and its best accuracy is only 76.90% when 400 features are selected for classification. The highest accuracy for the 1NN method is 83.05% when 500 features are preserved, however, it has no distinct improvement for the case in which 400 features are used. RFS and SRC ($r=0.3$) show very similar performances as 1NN does. SVM obtains a better performance and its accuracy achieves 85.48% when the number of features is 400. In the last row of this table, it can be seen that our new method obtains the best results for all the different numbers of features that are used for classification. Notably, the accuracy achieves 87.19% when 400 features are selected. As a result, Table 7 also shows some characteristics in the classification: (1) the classification results of gene expression data is not always improved as the selected features increases; (2) all the methods mentioned above except for SRC produce good results when 400 features are selected for classification; (3) only the BSS/WSS approach is exploited for dimensionality reduction and features selection as the preprocessing step, since our main objective is to validate and compare the effectiveness of classification algorithms.

Tables 8 and 9 show the results on the brain and lymphoma data, respectively. We can see that, SRC method performs poorly on these two data and the optimal results only achieve 76.36% ($r=0.5$) and 92.62% ($r=0.4$) respectively. RFS performs slightly better than that of SRC. For the brain database, the highest accuracy of 1NN approach is 92.71%. In the case in which 200 features are selected for classification, other three methods have obtains close to 90% and 94% results when 100 and 200 features are preserved, respectively. We also notice that, even if more features (300, 400, 500) being selected for classification, it does not obtain more competitive results for this data, and that the accuracies will decrease to a lower value. On the other hand, for the lymphoma data, when only 100 features are selected for classification, the accuracies of the LRC and RRC methods are 96.43% and 98.14%, respectively. These are lower than the results of the 1NN and SVM methods, 99.33% and 99.38% respectively. However, when more features are selected for classification, the LRC, SVM and our RRC methods keep the accuracies at stable rates above 99% except for 1NN. Therefore, our new algorithm presents robustness and effectiveness, especially shows better performance than the results of dummy matrix based RFS approach.

Table 8
Results on the *brain* data.

♯ variables	100	200	300	400	500
LRC	0.8957	0.9400	0.9079	0.9100	0.9143
1NN	0.8493	0.9271	0.8686	0.8614	0.8693
SRC	0.7164	0.7636	0.7064	0.7229	0.6993
SVM	0.9078	0.9421	0.9114	0.9185	0.9200
RFS	0.7307	0.8179	0.7943	0.7986	0.7986
RRC	0.9000	0.9386	0.9157	0.9143	0.9221

Table 9
Results on the *lymphoma* data.

♯ variables	100	200	300	400	500
LRC	0.9643	0.9919	0.9952	0.9957	0.9957
1NN	0.9933	0.9990	0.9886	0.9886	0.9871
SRC	0.9100	0.9186	0.9262	0.9010	0.8867
SVM	0.9938	0.9995	0.9952	0.9981	0.9910
RFS	0.9171	0.9238	0.9319	0.9314	0.9429
RRC	0.9814	0.9971	0.9938	0.9914	0.9952

4. Conclusions

This paper proposes a novel classification method and formulates the objective in a regularized $\ell_{2,1}$-norms minimization model. The main advantages of using this $\ell_{2,1}$-norm includes two aspects: the loss function is robust to outliers or large variations within samples, and the regularization item is able to select group-specific samples across the whole training set and with reasonable sparsity. A probabilistic interpretation under the multiple task learning framework is presented to enable better understanding of the group-specific regularized objective. The differences between several related works are presented and compared in the paper. Extensive experiments are performed on the benchmark data sets including facial images (Yale, Extended Yale B, CMU-PIE and Georgia Tech.) and gene expression data (Leukemia, Colon, Brain and Lymphoma) to demonstrate the competitive performance of our RRC method. Specially, the results of our RRC are better than the results of dummy matrix based approach. It is our future work to explore how to select optimal response variables in classification oriented regression models.

Acknowledgments

This project is supported in part by NSF of China (10771220, 90920007), the Ministry of Education of China (SRFDP-20070558043),

the Postdoctoral Science Foundation of China (2011M501361) and the City University of Hong Kong (Projects 9610034 and 7008094).

References

[1] I. Naseem, R. Togneri, M. Bennamoun, Linear regression for face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (11) (2010) 2106–2112.

[2] S. Li, J. Lu, Face recognition using the nearest feature line method, IEEE Transactions on Neural Networks 10 (2) (1999) 439–443.

[3] S. Li, K. Chan, C. Wang, Performance evaluation of the nearest feature line method in image classification and retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (11) (2000) 1335–1339.

[4] J. Chien, C. Wu, Discriminant waveletfaces and nearest feature classifiers for face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (12) (2002) 1644–1649.

[5] J. Lu, Y. Tan, Nearest feature space analysis for classification, IEEE Signal Processing Letters 18 (1) (2011) 55–58.

[6] R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Diagnosis of multiple cancer types by Shrunken centroids of gene expression, Proceedings of the National Academy of Sciences of the United States of America 99 (10) (2002) 6567–6572.

[7] A. Dabney, Classification of microarrays to nearest centroids, Bioinformatics 21 (22) (2005) 4148–4154.

[8] G. Grinblat, L. Uzal, H. Ceccatto, P. Granitto, Solving nonstationary classification problems with coupled support vector machines, IEEE Transactions on Neural Networks 22 (1) (2011) 37–51.

[9] Y. Artan, M. Haider, D. Langer, et al., Prostate cancer localization with multispectral MRI using cost-sensitive support vector machines and conditional random fields, IEEE Transactions on Image Processing 19 (9) (2010) 2444–2455.

[10] M. Davenport, R. Baraniuk, C. Scott, Tuning support vector machines for minimax and Neyman–Pearson classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (10) (2010) 1888–1898.

[11] C. Ding, D. Zhou, X. He, H. Zha, R_1-PCA: rotational invariant L_1-norm principal component analysis for robust subspace factorization, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 281–288.

[12] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization, in: Twenty-Fourth Annual Conference on Neural Information Processing Systems, 2010.

[13] L. Sun, J. Liu, J. Chen, J. Ye, Efficient recovery of jointly sparse vectors. in: Advances in Neural Information Processing Systems, vol. 23, 2009.

[14] J. Liu, S. Ji, J. Ye, Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, 2009, pp. 339–348.

[15] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B 68 (1) (2006) 49–67.

[16] F. Bach, Consistency of the group Lasso and multiple kernel learning, The Journal of Machine Learning Research 9 (2008) 1179–1225.

[17] J. Friedman, T. Hastie, R. Tibshirani, A Note on the Group Lasso and a Sparse Group Lasso, Arxiv preprint arXiv:1001.0736, 2010.

[18] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2) (2009) 210–227.

[19] C.X. Ren, D.Q. Dai, Sparse representation by adding noisy duplicates for enhanced face recognition: an elastic net regularization approach, in: Chinese Conference on Pattern Recognition, CCPR'09, Nanjing, 5–6 November 2009, pp. 513–517.

[20] V. Patel, R. Chellappa, M. Tistarelli, Sparse representations and random projections for robust and cancelable biometrics, in: The 11th International Conference on Control Automation Robotics & Vision (ICARCV), 2010, pp. 1–6.

[21] H. Huang, C. Ding, Robust tensor factorization using R_1 norm, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2008.

[22] P. Schneider, K. Bunte, H. Stiekema, et al., Regularization in matrix relevance learning, IEEE Transactions on Neural Networks 21 (5) (2010) 831–840.

[23] T. Pock, D. Cremers, H. Bischof, A. Chambolle, Global solutions of variational models with convex regularization, SIAM Journal on Imaging Sciences 3 (2010) 1122–1145.

[24] A. Argyriou, T. Evgeniou, M. Pontil, Multi-task feature learning, Advances in Neural Information Processing Systems 19 (2007) 41.

[25] T. Glasmachers, C. Igel, Maximum likelihood model selection for 1-norm soft margin SVMs with multiple parameters, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (8) (2010) 1522–1528.

[26] S. Babacan, R. Molina, A. Katsaggelos, Bayesian compressive sensing using Laplace priors, IEEE Transactions on Image Processing 19 (1) (2010) 53–63.

[27] G.W. Stewart, Matrix Algorithms Volume I: Basic Decompositions, SIAM, 1998.

[28] D. Cai, X. He, J. Han, SRDA: an efficient algorithm for large scale discriminant analysis, IEEE Transactions on Knowledge and Data Engineering 20 (1) (2008) 1–12.

[29] 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 19 (1997) 711–720.

[30] X.S. Zhuang, D.Q. Dai, Inverse fisher discriminate criteria for small sample size problem and its application to face recognition, Pattern Recognition 38 (11) (2005) 2192–2194.

[31] D.Q. Dai, P.C. Yuen, Face recognition by regularized discriminant analysis, IEEE Transactions on System, Man and Cybernetics, Part B 37 (4) (2007) 1080–1085.

[32] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection. in: International Joint Conference on Artificial Intelligence, vol. 14, 1995, pp. 1137–1145.

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

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

[35] L. Chen, H. Man, A. Nefian, Face recognition based on multi-class mapping of fisher scores, Pattern Recognition 38 (6) (2005) 799–811.

[36] T. Golub, D. Slonim, P. Tamayo, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (5439) (1999) 531.

[37] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, A. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proceedings of the National Academy of Sciences of the United States of America 96 (12) (1999) 6745–6750.

[38] S. Pomeroy, P. Tamayo, M. Gaasenbeek, et al., Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature 415 (6870) (2002) 436–442.

[39] A. Alizadeh, M. Eisen, R. Davis, et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403 (6769) (2000) 503–511.

[40] S. Dudoit, J. Fridlyand, T. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, Journal of the American Statistical Association 97 (457) (2002) 77–87.

[41] Y. Shi, D. Dai, C. Liu, H. Yan, Sparse discriminant analysis for breast cancer biomarker identification and classification, Progress in Natural Science 19 (11) (2009) 1635–1641.

[42] W. Yang, D. Dai, H. Yan, Finding correlated biclusters from gene expression data, IEEE Transactions on Knowledge and Data Engineering 23 (4) (2010) 568–584.

Chuan-Xian Ren received the BS degree from Fuyang Normal University, Fuyang, China, in 2005, PhD degree at the Mathematics Department, Faculty of Mathematics and Computing, Sun Yat-Sen(Zhongshan) University, Guangzhou, China, in 2010. During 2010 and 2011, he was with the Department of Electronic Engineering, City University of Hong Kong, as a senior research associate. He is currently an assistant professor of the Faculty of Mathematics and Computing, Sun Yat-Sen University, Guangzhou, China.

His research interests include face recognition, statistical data analysis.

Dao-Qing Dai received the PhD degree in mathematics from Wuhan University, Wuhan, China, in 1990.

He is currently a Professor and Associate Dean of the Faculty of Mathematics and Computing, Sun Yat-Sen (Zhongshan) University, Guangzhou, China. He was with the Free University, Berlin, Germany, as an Alexander von Humboldt Research Fellow from 1998 to 1999. He served as programm cochair of Sinobiometrics'2004 and program committee members for several international conferences. His research interests include image processing, wavelet analysis, face recognition and gene expression data analysis.

He is author or coauthor of over 100 refereed technical papers. He received the outstanding research achievements in mathematics award from International Society for Analysis, Applications, and Computation (ISAAC) in Fukuoka, Japan, in 1999.

Hong Yan (S'88–M'89–SM'93–F'06) received his BS degree from Nanking University of Posts and Telecommunications, Nanking, China, in 1982, MS degree from the University of Michigan, Ann Arbor, in 1984, and PhD degree from Yale University, New Haven, CT, in 1989, all in electrical engineering. During 1982 and 1983, he worked

on signal detection and estimation as a graduate student and research assistant at Tsinghua University, Beijing, China. From 1986 to 1989, he was a Research Scientist at General Network Corporation, New Haven, where he worked on design and optimization of computer and telecommunications networks. He joined the University of Sydney, Australia in 1989 and became Professor of Imaging Science in 1997. He is currently professor of electronic engineering at City University of Hong Kong.

His research interests include image processing, pattern recognition and bioinformatics. He is author or coauthor of over 300 refereed technical papers in these areas. Prof. Yan is a Fellow of the Institute of Electrical and Electronic Engineers (IEEE), the International Association for Pattern Recognition (IAPR), and the Institution of Engineers, Australia (IEAust), and a member of the International Society for Computational Biology (ISCB).