Adaptive Class Preserving Representation for Image Classification

Jian-Xun Mi[1, 2], Qiankun Fu[1, 2], Weisheng Li[1, 2]
[1] Chongqing Key Laboratory of Computational Intelligence,
Chongqing University of Posts and Telecommunications, Chongqing, China
[2] College of Computer Science and Technology,
Chongqing University of Posts and Telecommunications, Chongqing, China
mijianxun@gmail.com, S150201028@stu.cqupt.edu.cn, liws@cqupt.edu.cn

Abstract

In linear representation-based image classification, an unlabeled sample is represented by the entire training set. To obtain a stable and discriminative solution, regularization on the vector of representation coefficients is necessary. For example, the representation in sparse representation-based classification (SRC) uses L1 norm penalty as regularization, which is equal to lasso. However, lasso overemphasizes the role of sparseness while ignoring the inherent structure among samples belonging to a same class. Many recent developed representation classifications have adopted lasso-type regressions to improve the performance. In this paper, we propose the adaptive class preserving representation for classification (ACPRC). Our method is related to group lasso based classification but different in two key points: When training samples in a class are uncorrelated, ACPRC turns into SRC; when samples in a class are highly correlated, it obtains similar result as group lasso. The superiority of ACPRC over other state-of-the-art regularization techniques including lasso, group lasso, sparse group lasso, etc. are evaluated by extensive experiments.

1. Introduction

SRC is a widely used method in many computer vision applications including face recognition and other image classification problems [1]. In SRC, a test sample is collaboratively represented by a dictionary consisting of all the training samples, which is formulated as

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{1}$$

where $\mathbf{y}$ is the test sample, $\mathbf{A}$ is the dictionary matrix in which each column represents a sample, and $\mathbf{x}$ is the corresponding representation vector for $\mathbf{y}$. A model fitting procedure is used to produce the solution of $\mathbf{x}$. Ordinary least squares (OLS) is a simple way to estimate $\mathbf{x}$ and it is well known that to incorporate regularization on $\mathbf{x}$ into OLS results in more stable and interpretable solution. The linear regression model used in SRC is built under the principle of parsimony which is also important in human perception. Representation of $\mathbf{y}$ by a small subset of entire training samples is a practice of parsimony [2]. In SRC, the lasso regression is a promising approach to select a subset of training samples by imposing a L1 norm penalty on $\mathbf{x}$ which produces a sparse vector [3]. This constrained least squares model is given by:

$$\mathbf{x} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{2}$$

where $\|\cdot\|_2$ is L2 norm (vector norm), $\|\cdot\|_1$ is L1 norm, and λ balances the fidelity term with the regularization term. However, there is a limitation to lasso regression. If there are some samples whose correlation are high, lasso prefers to select only one sample from them and ignores others [4]. Such phenomenon leads to inconsistent selection of training samples [5]. Elastic net (EN) regression provides a flexible model to address problem above [4]. Elastic net criterion is defined as the following optimization problem:

$$\mathbf{x} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda(1-\alpha)\|\mathbf{x}\|_1 + \lambda\alpha\|\mathbf{x}\|_2^2. \tag{3}$$

The complex regularization function $(1-\alpha)\|\mathbf{x}\|_1 + \alpha\|\mathbf{x}\|_2^2$ is a convex combination of the lasso and ridge penalty where $\alpha \in [0,1]$. Besides the sparsity of $\mathbf{x}$, the solution of EN regression has the effect of grouping that highly corrected samples share similar regression coefficients. Many other improvement studies of lasso regression also tend to address the correlation problem. In [6], sample de-correlation is considered in the lasso regression. A covariate-correlated lasso is proposed to select covariates most strongly correlating with response variable in [7]. Another recent study [8] also considers correlation among training samples which proposes adaptive model balancing the L1 norm with L2 norm guided by training set itself.

There is an obvious drawback for the mentioned methods above in classification. They do not incorporate the class

information into the representation procedure, that is, the regression is not aware of any intra-class structure. For example, in SRC, the solution of problem (2) is computed first and the followed decision is to determine which class has the smallest representative residual:

$$Label(\mathbf{y}) = \arg\min_i \|\mathbf{y} - \mathbf{A}_i\mathbf{x}_i\|_2^2 \quad (\ i = 1, \cdots, c\ ), \qquad (4)$$

where c is the number of classes, $\mathbf{A}_i$ denotes the samples from the i th class and $\mathbf{x}_i$ is a vector consisting of coefficients associated with $\mathbf{A}_i$ (the entire training set $\mathbf{A}=[\mathbf{A}_1, \mathrm{A}_2, \cdots, \mathbf{A}_c]$ and the representing vector $\mathbf{x}=\left[\mathbf{x}_1^T, \mathrm{x}_2^T, \cdots, \mathbf{x}_c^T\right]^T$). Only at this moment, the label of training set is used. Besides this drawback, let us carefully look at the parsimony principle for classification. A sparse $\mathbf{x}$ means only a few training samples are involved in representing a test sample. These training samples can be distributed across different classes or come from only one or two classes. Surely the second case is favorable for classification. Hence it could be more discriminative to constrain the number of classes but not the number of samples. In some cases, a good representation of a test sample could be a dense one by samples from the correct class. So, strong sample-wise sparse constraint on $\mathbf{x}$ can result in a poor solution.

The group lasso (GL) regression, proposed in [9], produces a representation vector of group-wise sparsity. GL predefines sample groups before regression and is to let groups compete with each other during the regression, which means all the members of a particular group are either used or not used. In [10] class specific sparse representation classification is proposed based on the group lasso regression. In this method training samples from a certain class are defined as a group, and a group sparse solution of $\mathbf{x}$ can be obtained by solving the following regularized optimization problem:

$$\mathbf{x} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \sum_{i=1}^{c} \|\mathbf{x}_i\|_2 \qquad (5)$$

Further improvement of group lasso can be found in [11] where the sparse group lasso (SGL) regression is proposed:

$$\mathbf{x} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda(1-\alpha)\|\mathbf{x}\|_1 + \lambda\alpha \sum_{i=1}^{c} \|\mathbf{x}_i\|_2 \ . \qquad (6)$$

SGL can be also seen as a variant of elastic net. SGLR not only considers the group-wise sparseness but also the sample-wise sparseness. A manually selected value of α determines the ratio of these two kinds of sparseness. Note that all classes are equally constrained despite of possible different intra-class sample structure.

In this paper, we propose an adaptive class preserving sparse representation (ACPR) method and apply it to perform image classification. Our proposed method takes three factors into account. First, the class information should be preserved during regression; Second, the sparseness of the representation vector is constrained; Third, the extent of the sample-wise sparseness of a class is adaptive to intra-class correlation structure. Although previous studies of lasso-type regression consider class information, there is no study, to our knowledge, which exploits the intra-class correlation to build an adaptive model. In ACPR, classes receive separate constraint, i.e., the constraint on vector $\mathbf{x}_i$ is determined by correlation structure of the i th class. We find that our regularizer on each class can adaptively balance the L1 norm with L2 norm. Thus, regression with ACPR can obtain a fine representation result that all three important factors are implemented.

In the following, we first briefly review and summarize the related lasso-type regression methods. Then our regression model is presented. Next, we give an effective iterative algorithm to solve the optimization problem. Experiments on different datasets are preformed to evaluate the method.

2. Related lasso-type regression techniques

That a test image is collaboratively represented by training set is a key phase in representation based image classification algorithms. They assume the test sample can be represented well as a linear combination of the training samples belonging to its correct class. Under this linear regression framework, regularization on representation vector plays an important role in obtaining a correct solution when data set is large. A general constrained regression model employs the following optimization form:

$$\mathbf{x} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\, \mathrm{g}(\mathbf{x}) \qquad (7)$$

where $\mathrm{g}(\cdot)$ is a penalty to regularize $\mathbf{x}$. Variations and noises are highly likely to be seen in a test image. Without an effective regularizer, the performance of regression could be degraded seriously. Many lasso-type regularizers receive increasing attention. They tend to compel the solution of $\mathbf{x}$ to be sparse. There are at least two merits to do this. First sparse representation has discriminative nature to perform classification. Second noises on test image are more likely to be detected. When $\mathrm{g}(\mathbf{x})=\|\mathbf{x}\|_1$, (7) is lasso regression problem which is very popular to perform sparse representation. Lasso mainly considers sample-wise sparsity that only a very small subset of samples represent a test sample. If we define samples from a same class as a group, group sparse representation constrains the number of class used in the regression which results in group-wise sparsity. When each class consists of only one or two samples, group sparse representation obtains a similar result as lasso. According to the definition of group lasso that $\mathrm{g}(\mathbf{x})=\sum_{i=1}^{c}\|\mathbf{x}_i\|_2$, the intra-class representation

Methods	Property			
	Sample-wise sparsity	Group-wise sparsity	Correlation	Adaptiveness
Lasso	✓			
EN	✓		✓	
GL		✓		
SGL	✓	✓		
ACPR	✓	✓	✓	✓

Table 1. A summary of four aspects of the five related regression methods. 'Correlation' stands for that the corre-lation among samples are considered. 'Adaptiveness' means the model has the capability of self-adjustment for regularization according to the structure of data set.

sub-vector $\mathbf{x}_i$ is constrained by L2 norm which produces a dense solution of $\mathbf{x}_i$. By using naive combination of both sample-wise sparsity and group-wise sparsity together $g(\mathbf{x})=(1-\alpha)\|\mathbf{x}\|_1+\alpha\sum_{i=1}^{c}\|\mathbf{x}_i\|_2$, SGL balances the two kinds of sparseness by a pre-assigned parameter. The solution of EN regression shows grouping effect, however, a group of samples which are highly correlated can be across classes. Nevertheless, EN has its merit over lasso that correlation among samples is considered. Here, we summarize the mentioned lasso-type regressions from four aspects in Table 1.

3. Model of ACPR

In this section, we consider a regression model which preserves class information and has self-adjustable constraint according to the specific intra-class inherent structure. For a certain class, we prefer L2 norm to L1 norm constraint when samples of the class are highly correlated. On the contrary, L1 norm based constrain is a better choice when samples are uncorrelated.

3.1. Formulation

The formulation of the proposed regression model is expressed as follows:

$$\min_{\mathbf{x}}\sum_{i=1}^{c}\Phi(\mathbf{x}_i) \quad s.t. \quad \mathbf{y}=\mathbf{A}\mathbf{x} \tag{8}$$

where $\Phi(\mathbf{x}_i)=\|\mathbf{A}_i\mathrm{Diag}(\mathbf{x}_i)\|_*$, the regularizer $\|\cdot\|_*$ denotes the nuclear norm of a matrix which is the sum of the singular values of the matrix, and $\mathrm{Diag}(\cdot)$ transforms a vector into a diagonal matrix. Suppose there are n_i samples in the i th class and each sample is preprocessed by normalization of zero mean and unit vector length. The representation vector for the i th class is a column vector $\mathbf{x}_i=[x_{i,1},\cdots,x_{i,n_i}]^T\in R^{n_i\times 1}$.

Here we firstly discuss the property of $\Phi(\mathbf{x}_i)$. Assume that samples belonging to the i th class are distinct from each other, which means samples are orthogonal $\mathbf{A}_i^T\mathbf{A}_i=\mathbf{I}$. Then the regularizer is decomposed into

$$\Phi(\mathbf{x}_i)=\mathrm{Tr}\left(\sqrt{\mathrm{Diag}(\mathbf{x}_i)\mathbf{A}_i^T\mathbf{A}_i\mathrm{Diag}(\mathbf{x}_i)}\right)=\|\mathbf{x}_i\|_1. \tag{9}$$

Thus, for the i th class, the regularizer $\Phi(\mathbf{x}_i)$ turns into L1 norm penalty. This is reasonable because when there is no intra-class structure L1 norm is a better choice to prevent overfitting. In the special case that every class does not have intra-class structure, (8) is equivalent to lasso regression.

If we assume that the samples in the i th class are highly correlated that $\mathbf{A}_i^T\mathbf{A}_i\approx\mathbf{1}\mathbf{1}^T$ ($\mathbf{1}$ is a column vector of size n_i and every element is equal to one). Then we have

$$\Phi(\mathbf{x}_i)=\sqrt{\mathbf{x}_i^T\mathbf{x}_i}=\|\mathbf{x}_i\|_2. \tag{10}$$

So, if a class has very similar samples, its corresponding regularizer approximates L2 norm. Under the circumstances that every class has highly correlated samples, (8) turns into GL regression. In general, samples from a class are neither too independent nor totally identical. The regularizer balances the L1 norm with L2 norm adaptively according to the intra-class structure of a class:

$$\|\mathbf{x}_i\|_2\le\Phi(\mathbf{x}_i)\le\|\mathbf{x}_i\|_1. \tag{11}$$

Let us consider a special case that one sample, say the first one, is mistakenly grouped into a different class and it is independent to all other samples. Here we define two block matrices:

$$\mathbf{A}_i^T\mathbf{A}_i=\begin{bmatrix}1 & \mathbf{0}^T\\ \mathbf{0} & \mathbf{A}_{i/1}^T\mathbf{A}_{i/1}\end{bmatrix}\text{ and }\mathrm{Diag}(\mathbf{x}_i)=\begin{bmatrix}x_{i,1} & \mathbf{0}^T\\ \mathbf{0} & \mathrm{Diag}(\mathbf{x}_{i,/1})\end{bmatrix} \tag{12}$$

where $\mathbf{0}$ is a column vector of size n_i-1 and in which every element is equal to zero, $\mathbf{A}_{i/1}$ is sample matrix without the first independent sample, and $\mathbf{x}_{i,/1}$ is the representation vector for $\mathbf{A}_{i/1}$. Then we can see the regularizer has following property:

$$\begin{aligned}&\Phi(\mathbf{x}_i)\\ &=\mathrm{Tr}\left(\sqrt{\begin{bmatrix}x_{i,1} & \mathbf{0}^T\\ \mathbf{0} & \mathrm{Diag}(\mathbf{x}_{i,/1})\end{bmatrix}\begin{bmatrix}1 & \mathbf{0}^T\\ \mathbf{0} & \mathbf{A}_{i/1}^T\mathbf{A}_{i/1}\end{bmatrix}\begin{bmatrix}x_{i,1} & \mathbf{0}^T\\ \mathbf{0} & \mathrm{Diag}(\mathbf{x}_{i,/1})\end{bmatrix}}\right)\\ &=\mathrm{Tr}\left(\begin{bmatrix}\|x_{i,1}\|_1 & \mathbf{0}^T\\ \mathbf{0} & \sqrt{\mathrm{Diag}(\mathbf{x}_{i,/1})\mathbf{A}_{i/1}^T\mathbf{A}_{i/1}\mathrm{Diag}(\mathbf{x}_{i,/1})}\end{bmatrix}\right)\\ &=\|x_{i,1}\|_1+\Phi(\mathbf{x}_{i,/1}).\end{aligned} \tag{13}$$

So the independent sample is separated from others which is constrained by L1 norm alone. Furthermore, it is not hard to prove that if there are some subsets in a class which are independent of each other, the regularizer is transformed into subgroups:

$$\Phi(\mathbf{x}_i)=\Phi(\mathbf{x}_i^1)+\cdots+\Phi(\mathbf{x}_i^p) \tag{14}$$

where p denotes the number of subgroups in the class. It indicates that, compared with lasso and group lasso, the proposed regularizer can exploit the intra-class structure further. That is, although sample group is determined by the class label, adaptation of (8) is strong enough to reflect the intra-class structure.

In real-world applications, a test image is more or less contaminated by variations and noises. So, we solve the following optimization problem instead of (8)

$$\arg\min_{\mathbf{x}} \|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2+\lambda\sum_{i=1}^{c}\Phi(\mathbf{x}_i)\ . \tag{15}$$

3.2. Optimization

An iterative optimization algorithm is proposed to solve (8) , which is summarized in Algorithm 1. In each iteration of updating $\mathbf{x}$, we decompose (15) into c sub-problems:

$$\arg\min_{\mathbf{x}_i} \|\hat{\mathbf{y}}_i-\mathbf{A}_i\mathbf{x}_i\|_2^2+\lambda\|\mathbf{A}_i\ \mathrm{Diag}(\mathbf{x}_i)\|_*\ ,\quad i=1,\cdots,c\ . \tag{16}$$

which is known as trace lasso problem [12]. In this way, Algorithm 1 updates $\mathbf{x}$ one class at a time.

Algorithm 1: Procedure of Adaptive Class Preserving Sparse Representation

1: **Input:** training set $\mathbf{A}=[\mathbf{A}_1\cdots\mathbf{A}_c]$, the query image $\mathbf{y}$, parameter $\lambda>0$, $\varepsilon>0$, initial representation vector $\mathbf{x}^{(0)}=\left[(\mathbf{x}_1^{(0)})^T\cdots(\mathbf{x}_c^{(0)})^T\right]^T$, $t=1$.

2: Generate a random permutation vector $\mathbf{m}\in R^{1\times c}$ filled with integers from 1 to the number of class c .

3: Define a temporary variable $\mathbf{x}^{new}=\mathbf{x}^{(t)}$.

For $i=1$ *to* c

4: Compute $\hat{\mathbf{y}}_{\mathbf{m}(i)}=\mathbf{y}-\mathbf{A}\mathbf{x}^{new}+\mathbf{A}_{\mathbf{m}(i)}\mathbf{x}_{\mathbf{m}(i)}^{(t)}$.

5: Update only $\mathbf{x}_{\mathbf{m}(i)}^{(t)}$ by solving (16) with Algorithm 2. The input includes $\hat{\mathbf{y}}_{\mathbf{m}(i)}$ and $\mathbf{A}_{\mathbf{m}(i)}$.

6: Let $\mathbf{x}_{\mathbf{m}(i)}^{new}=\mathbf{x}_{\mathbf{m}(i)}^{(t+1)}$.

End;

7: Let $\mathbf{x}^{(t+1)}=\mathbf{x}^{new}$.

8: If $\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_\infty\le\varepsilon$, go to step 9, otherwise, let $t=t+1$ and go to step 2;

9: **Output:** The optimal representation vector $\mathbf{x}^*$.

In optimization problem (8), the objective function consists of a fidelity term and a regularization term. It is well known that OLS is convex and in [13] nuclear norm is proved to be convex as well. So, linear combination of them is also a convex function. And then we let $L(\mathbf{x})=L(\mathbf{x}_1,\cdots,\mathbf{x}_c)$ denote the objective function of (8). We break $\mathbf{x}$ into small sub-vectors and update each sub-vector one by one at random order to avoid priority. The corresponding sub-vector of $\mathbf{x}$ is replaced by new sub-vector immediately. So, if the sub-problem (16) is solved effectively, the original problem will converge to the unique global optimum solution.

To solve (16), we first convert it to the following equivalent problem:

$$\min_{\mathbf{J},\mathbf{x}_i}\|\hat{\mathbf{y}}_i-\mathbf{A}_i\mathbf{x}_i\|_2^2+\lambda\|\mathbf{J}\|_*\quad s.t.\ \ \mathbf{J}=\mathbf{A}_i\mathrm{Diag}(\mathbf{x}_i). \tag{17}$$

It can be solved by minimizing the corresponding augmented Lagrange multiplier formulation:

$$\begin{aligned}L(\mathbf{J},\mathbf{x}_i)=&\|\hat{\mathbf{y}}_i-\mathbf{A}_i\mathbf{x}_i\|_2^2+\lambda\|\mathbf{J}\|_*+\mathrm{Tr}(\mathbf{Y}^T(\mathbf{J}-\mathbf{A}_i\ \mathrm{Diag}(\mathbf{x}_i)))\\&+\frac{\mu}{2}\|\mathbf{J}-\mathbf{A}_i\ \mathrm{Diag}(\mathbf{x}_i)\|_F^2\end{aligned} \tag{18}$$

where $\mathbf{Y}$ is a Lagrange multiplier matrix and μ is a penalty parameter. We solve (18), which now is an unconstrained problem, with alternating direction method (ADM) [14, 15]. $\mathbf{J}$ and $\mathbf{x}_i$ are updated respectively by fixing the other one, and then $\mathbf{Y}$ is updated. First, we fix $\mathbf{x}_i$ to solve

$$\mathbf{J}=\arg\min_{\mathbf{J}}\frac{\lambda}{\mu}\|\mathbf{J}\|_*+\frac{1}{2}\left\|\mathbf{J}-(\mathbf{A}_i\ \mathrm{Diag}(\mathbf{x}_i)-\frac{1}{\mu}\mathbf{Y}\ )\right\|_F^2\ . \tag{19}$$

Singular Value Thresholding operator can provide a close form solution to (19)[14]. Then we fix $\mathbf{J}$, and $\mathbf{x}_i$ is obtained by solving

$$\begin{aligned}\mathbf{x}_i=\arg\min_{\mathbf{x}_i}\ &\|\hat{\mathbf{y}}_i-\mathbf{A}_i\mathbf{x}_i\|_2^2+\mathrm{Tr}(\mathbf{Y}^T(\mathbf{J}-\mathbf{A}_i\ \mathrm{Diag}(\mathbf{x}_i)))\\&+\frac{\mu}{2}\|\mathbf{J}-\mathbf{A}_i\ \mathrm{Diag}(\mathbf{x}_i)\|_F^2\ .\end{aligned} \tag{20}$$

It also has a closed form solution:

$$\begin{aligned}\mathbf{x}_i&=\mathbf{P}(2\mathbf{A}_i^T\hat{\mathbf{y}}_i+\mathrm{diag}(\mathbf{A}_i^T(\mathbf{Y}+\mu\mathbf{J})))\\\mathbf{P}&=(2\mathbf{A}_i^T\mathbf{A}_i+\mu\,\mathrm{Diag}(\mathrm{diag}(\mathbf{A}_i^T\mathbf{A}_i)))^{-1}\end{aligned} \tag{21}$$

where $\mathrm{diag}(\cdot)$ takes the diagonal elements of a matrix and stores them in a vector. The multiplier is updated by gradient descent method:

$$\mathbf{Y}=\mathbf{Y}+\mu(\mathbf{J}-\mathbf{A}_i\ \mathrm{Diag}(\mathbf{x}_i))\ . \tag{22}$$

The procedure of solving the sub-problem is summarized in Algorithm 2. Since the whole procedure of solving optimization (16) is derived by the theory of ADM, the convexity of this sub-problem guarantees its global convergence. After computing optimal $\mathbf{x}^*$, we make the

decision of label by (4) which is widely used in representation based classifications.

Algorithm 2: Solving sub-problem (17)

1: **Input:** training set of the i th class $\mathbf{A}_i$, the regressand $\hat{\mathbf{y}}_i$, parameter λ.

2: **Initialize:** $\mathbf{x}_i^{(0)}$, $\mathbf{Y}^{(0)}$, $\mathbf{J}^{(0)}$, $\mu > 0$, $\varepsilon > 0$, $k = 0$.

3: Fix the others and compute $\mathbf{J}^{(k+1)}$ by solving (19).

4: Fix the others and compute $\mathbf{x}_i^{(k+1)}$ with (21).

5: Update the multiplier $\mathbf{Y}^{(k+1)}$ according to (22)

6: If $\|\mathbf{J}^{k+1} - \mathbf{J}^k\|_\infty \le \varepsilon$, $\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\|_\infty \le \varepsilon$, and $\|\mathbf{J}^{k+1} - \mathbf{A}_i \,\mathrm{Diag}(\mathbf{x}_i^{k+1})\|_\infty \le \varepsilon$, go to step 7, otherwise, let $k = k+1$ and go to step 3;

7: **Output:** The optimal vector $\mathbf{x}_i^*$.

3.3. Time complexity

Like conventional trace lasso, iterative optimization is used in solving ACPR model. In one round of updating, the time complexity of trace lasso is $O(n^3)$, where n is the number of the entire training samples. The sub-problem of ACPR is also a trace lasso problem whose time complexity is $O(p^3)$ (here we simply assume that each class has p training samples). And for each update of the entire $\mathbf{x}$, the time complexity of proposed ACPR is $O(cp^3)$. Since $n=cp$, $O(cp^3) \ll O(c^3p^3) = O(n^3)$.

4. Sparsity vs correlation

For image classification, many papers argue that the computation of a representation vector by collaboratively representing a test sample is better than by representing it separately [3, 16]. The reason is that constrained linear regression compels samples to compete with each other, which makes the decision more reliable. Since it is inevitable that there are variations and noises on test samples, to constrain the sparsity of the representation vector prevents it from overfitting. Some studies suggest using lasso regression with L1 norm can perform robust classification. In lasso regression, competition is among entire training samples. However, lasso ignores the correlation of the training set. If some samples are correlated, it is very likely that only one of them represents a test sample. In [10, 17], it is proved that competition among classes may obtain better performance. These papers use GL regression to represent a test sample where samples belonging to a same class form a group. It is common that images from a same class may be correlated to some extent.

For a specific class, the representation sub-vector is constrained by L2 norm which allows correlated samples from a class to represent the test sample together and makes the representation adequate. However, if there are plenty of training samples in a class, the risk of overfitting will increase when the intra-class structure is complex. SGL regression and EN both consider the correlation and sparsity, but a manually chosen parameter is required to balance these two factors.

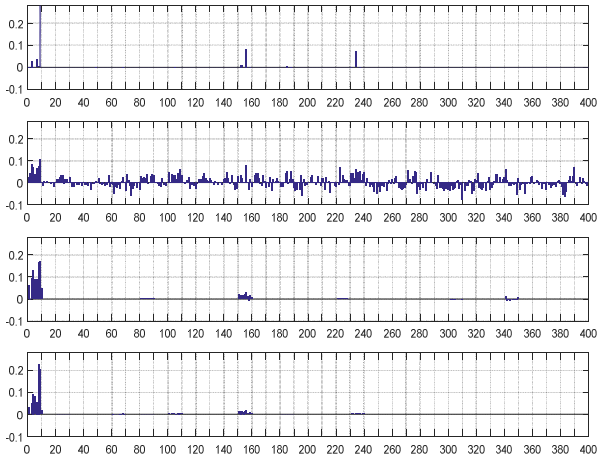

Figure 1. Illustration of the representation vectors by lasso regression, ridge regression, group lasso regression, and ACPR. The ORL face data set consisting of 40 subjects is used and each subject has 10 frontal face images. An artificial test sample is created by combining 5 images (from the first class) and some Gaussian noises together.

An obvious drawback of previous lasso-type regressions is that it is impossible to assign different constraint for each specific class. In general, the intra-class structure can vary a lot in terms of correlation. The proposed ACPR constrains each specific class differently by balancing L1 norm with L2 norm. ACPR also exploits the potential rich intra-class structure of a specific class. If there are mutually independent sub-groups, ACPR regroups them adaptively so as to provide a more-refined representation.

We provide a visualization of the representation vectors computed by different regression approaches in Figure 1. Lasso regression uses the least training samples to represent the test sample, which apparently ignores the correlation. Ridge regression which uses a L2 norm as its regularizer, produces a dense vector where many correlated samples responding together. ACPR performs similarly to GL regression. However, the intra-class sparsity of ACPR is stronger than GL regression and weaker than lasso regression. It indicates that ACPR not only takes into account intra-class sample-wise sparsity but also group-wise sparsity in an adaptive way. For EN regression and SGL regression, if two parameters in them are carefully chosen case by case, they may obtain an approximate vector as ACPR. But it is almost impossible to implement in practice.

Method	$t=2$	$t=5$	$t=8$	$t=10$
Lasso	48.00	65.88	74.46	77.20
Ridge	30.09	52.44	65.83	69.52
GL	43.35	60.52	65.49	68.88
SGL	48.52	62.62	70.00	69.84
EN	51.88	67.44	70.86	75.36
ACPRC	53.05	71.36	78.23	81.96

Table 2. The recognition accuracy (%) corresponding to t samples from each class on the GT database.

Method	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$
Lasso	80.00	87.39	90.04	92.80	93.75
Ridge	75.56	82.11	87.17	91.80	93.69
GL	82.00	89.50	92.63	94.80	96.31
SGL	85.84	90.71	94.42	96.10	96.75
EN	85.13	91.36	93.79	96.35	96.81
ACPRC	87.03	92.07	94.75	96.50	97.56

Table 3. The mean accuracy (%) corresponding to t samples from each class on the ORL database.

Figure 2. Some samples from FERET, LFW, GT, ORL, COIL20, and MNIST databases (from top to bottom).

5. Experiments

In this section, we evaluate the effectiveness of the proposed ACPRC by comparing with five state-of-the-art regressions including Ridge, Lasso, GL, SGL, and EN. Some well-known image datasets are used: two common face databases (GT and ORL), two large face databases (LFW and FERET), MNIST handwritten digits database, and COIL20 object recognition database. Some samples are shown in Figure 2 where intra-class variations can be seen. All used images are normalized to have zero mean and unit L2 norm.

5.1. Georgia Tech (GT) database

The GT database, which can be downloaded from http://www.anefian.com/research/face_reco.htm, consists of 15 images per subject from 50 individuals taken in different times. It characterizes some variations such as facial expressions, cluttered backgrounds, and lighting conditions. Images were processed to an order of 15×15. For each subject, we selected t (= 2, 5, 8, 10) images to form the training set and the rest for testing. Table 2 shows the recognition rates and standard deviations of all the algorithms over 10 random splits.

Taking into account such a complex situation, all the used methods faced a huge challenge. The performances of all algorithms, especially ridge regression, are not satisfactory because of the lack of training samples and various noises in the case of $t=2$. The recognition accuracies rise along with increase of training samples for the most methods. ACPRC has advantage of performance in all cases, which obviously indicate that our method has shown better tolerance for illumination and pose variations. Especially, ACPRC is the only one whose accuracy exceeds 80% when $t=10$.

5.2. ORL database

The ORL face database is composed of 40 distinct subjects with 10 images per subject sampled at different times with sorts of variations such as facial expressions, varying illuminations and facial details (glasses or not). We selected the first 2, 3, 4, 5, and 6 images from each subject for training and the remaining for testing. Images were resized to 15×12. The recognition accuracy rates are reported in the Table 3.

We can figure out that the accuracy rates of all the algorithms increase along with adding more training samples. GL and SGL perform better than Lasso, which is different from the results on GT database. The cause may be that GT has lower intra-class correlation, which weakens role of group structure. Although methods with grouping effect performed better, ACPRC outperformed the second best around 1% on average.

5.3. LFW database

The LFW is a large-scale database composed of more than 13000 images of unconstrained faces with variations of pose, illumination, expression, misalignment and occlusion, and so on. In the first test scheme (LFW11), a subset of 143 subjects with no less than 11 images per subject was chosen. We randomly selected 10 images as training dataset and the

Method	LFW11	LFW16
Lasso	72.21 ± 1.02	76.83 ± 2.44
Ridge	72.79 ± 0.86	78.25 ± 1.01
GL	71.95 ± 1.38	79.29 ± 2.05
SGL	70.70 ± 1.32	76.96 ± 1.68
EN	70.16 ± 1.38	74.00 ± 1.86
ACPRC	74.84 ± 1.04	80.64 ± 1.57

Table 4. The mean accuracy (%) with standard deviations corresponding to two test schemes respectively on the LFW database.

Method	$t = 2$	$t = 3$	$t = 4$	$t = 5$
Lasso	53.00	46.13	57.00	64.00
Ridge	56.20	45.38	51.33	62.00
GL	59.60	50.50	63.83	66.75
SGL	59.50	50.88	63.83	65.00
EN	54.70	43.13	54.50	60.25
ACPRC	60.00	52.63	64.17	70.50

Table 5. The mean accuracy (%) corresponding to t samples from each class on the FERET database.

remaining images as testing dataset. In the second subset (LFW16), which consists of 85 subjects with no less than 16 images per subject, we randomly used 15 images as training dataset and the remaining images as testing dataset. Images were processed to an order of 15×12.

The recognition accuracy rates of each method were reported in Table 4. We can easily figure out that the advantage of our proposed ACPRC by the competitive results. Especially, in the subset of LFW16, only our method achieves the accuracy of 80% up. The method of EN just reaches 70.16% and 74.00% on the two subsets, and the reason is that EN does not consider the group-wise sparsity.

5.4. FERET database

In this experiment, a subset with 1400 images from 200 subjects is selected from the FERET database, which is arguably one of the largest publicly available database. Each subject consists of 7 facial images with mainly expression variations and posture changes. We used the first 2, 3, 4, and 5 images in each individual for training and the remaining images for testing. Images were processed to the resolution of 15×12. The experiment results are listed in Table 5.

It can be seen from Table 5 that our method still obtains the best results in all cases, although all the method perform not better when the number of images per subject is 3. More specifically, the recognition accuracy of our method reaches 70.50% when the number of images per subject is 5. The improvement over other algorithms are respectively 6.50%, 8.50%, 3.75%, 4.5%, and 9.25%.

Method	MNIST	COIL20
Lasso	86.68 ± 2.00	98.08 ± 0.55
Ridge	82.36 ± 2.17	94.61 ± 1.08
GL	84.16 ± 1.22	97.50 ± 0.93
SGL	86.48 ± 1.28	96.67 ± 0.52
EN	85.20 ± 1.67	96.69 ± 0.76
ACPRC	88.88 ± 1.65	99.03 ± 0.60

Table 6. The mean accuracy (%) with standard deviations on the MNIST and COIL20 databases.

5.5. MNIST and COIL20 databases

The MNIST database of handwritten has been widely adopted in the field of pattern recognition. It contains 10 images classes, corresponding to 10 handwritten digits from $0 \sim 9$ and each class has more than 5000 images. We selected the first 50 samples of each subject and randomly chose a half as training set and the rest as test set.

The COIL20 database contains 20 objects and each object has 72 images taken at pose interval of 5 degrees. We randomly selected a half of each object as training dataset and the other half as test set respectively.

For these two databases, images were resized to 15×15. We reported the means and standard deviations of the accuracy rates in Table 6. On the both databases, the state-of-the-art regularization methods worked similar except Ridge. ACPRC outperformed the second best over 2% on the MNIST database and around 1% on the COIL20 database respectively.

5.6. Discussion on experiments

A full summary of our extensive experimental results is illustrated in Figure 3. Without imposing sparsity, performance of Ridge regression is not comparable with others in all cases, which confirms the fact that sparsity is essential to achieve robust recognition.

On the ORL database, lasso regression performs only a little better than Ridge regression because there are small variations within class. GL achieves the second best result on GT database. The reason could be that the change within each class of GT database is larger and group-wise sparsity benefits the representation.

Group-wise sparsity of GL shows instability with the larger standard deviation than ACPRC on the LFW database because of the different intra-class correlation for each group. However, our proposed ACPRC considers both the group-wise sparsity and the correlation of data which is self-adjustable to specific with-in class structure. For the roles of group-wise sparsity and sample-wise sparsity, one is not necessarily better than another. Especially, our method performs the perfect competitive on GT database because of the superiority of adaptive class preserving regularization term which can balance the L1 regularization

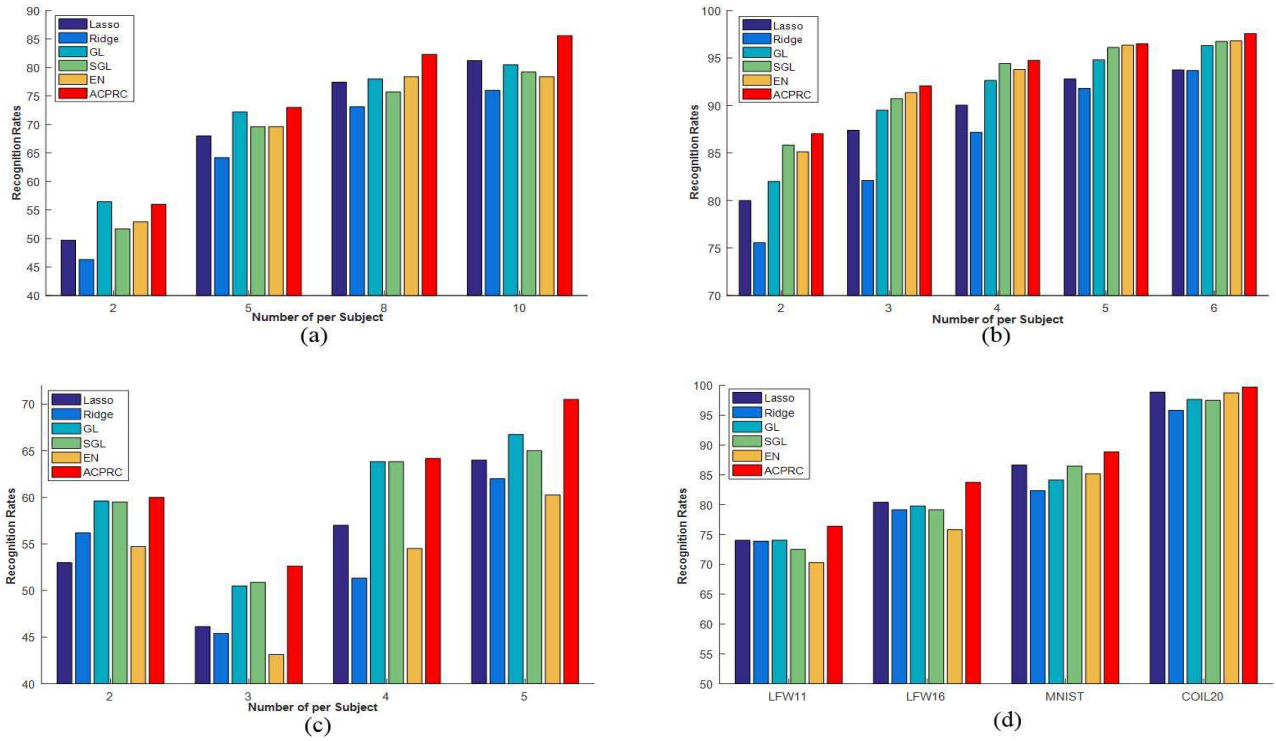

Figure 3. Recognition accuracy rates on different databases. (a) GT database with 2, 5, 8, and 10 images of each subject for training. (b) ORL database with 2, 3, 4, 5, and 6 images of each subject for training. (c) FERET database with 2, 3, 4, and 5 images of each subject for training. (d) LFW, MNIST and COIL20 database.

and group sparse regularization according to the correlation.

None of these algorithms including our ACPRC performs perfect on the MNIST database, although ACPRC obtains the highest accuracy. The main reason may lie in the fact that the number of images for experiment is not enough and the handwritten digit data do not fit the subspace structure well. ACPRC get the best accuracies 99.03% on the COIL20 database, which indicate that the adaptive class preserving sparse regularization term also helps handle general pattern recognition problems.

6. Conclusion

In this paper, we propose adaptive class preserving sparse representation for image classification. In the procedure of collaboratively representing a test sample, ACPR penalties the representation vector of each class differently. Comparing with previous lasso-type regressions, ACPR balances lasso regression with group lasso regression adaptively. To solve optimization problem of ACPR, we decompose the model into sub-problems and solve it by ADM. The superiority of ACPR based image classification is validated by extensive experiments.

Acknowledgments: This work was supported partially by the National Nature Science Foundation of China under Grant Nos. 61472055, U1401252, and 61672120. And partially this research is funded by the National Key Research and Development Program of China under Grant Nos. 2016YFB1000905.

References

[1] J. Wright, M. Yi, J. Mairal, G. Sapiro, T. S. Huang and Y. Shuicheng. Sparse Representation for Computer Vision and Pattern Recognition. Proceedings of the IEEE, 98(6) :1031-1044, 2010.

[2] M. H. Hansen and B. Yu. Model Selection and the Principle of Minimum Description Length. Journal of the American Statistical Association, 96(454) :746-774, 2001.

[3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma. Robust Face Recognition via Sparse Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2) :210-227, 2009.

[4] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. Journal of The Royal Statistical Society Series B-statistical Methodology, 67(2) :301-320, 2005.

[5] H. Zou. The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476) 2012.

[6] S.-B. Chen, C. Ding, B. Luo and Y. Xie. Uncorrelated Lasso. AAAI Conference on Artificial Intelligence. 2013.

[7] B. Jiang, C. Ding and B. Luo. Covariate-Correlated Lasso for Feature Selection. Machine Learning and Knowledge Discovery in Databases: European Conference. 2014.
[8] J. Wang, C. Lu, M. Wang, P. Li, S. Yan and X. Hu. Robust Face Recognition via Adaptive Sparse Representation. IEEE Transactions on Systems, Man, and Cybernetics, 44(12) :2368-2378, 2014.
[9] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. Journal of The Royal Statistical Society Series B-statistical Methodology, 68(1) :49-67, 2006.
[10] S. Huang, Y. Yang, D. Yang, L. Huangfu and X. Zhang. Class specific sparse representation for classification. Signal Processing, 116(38-42, 2015.
[11] N. Simon, J. Friedman, T. Hastie and R. Tibshirani. A Sparse-Group Lasso. Journal of Computational and Graphical Statistics, 22(2) :231-245, 2013.
[12] E. Grave, G. R. Obozinski and F. R. Bach. Trace lasso: a trace norm regularization for correlated designs. Advances in Neural Information Processing Systems. 2187-2195 2011.
[13] B. Recht, M. Fazel and P. A. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. Siam Review, 52(3) :471-501, 2010.
[14] J. Cai, E. J. Candes and Z. Shen. A Singular Value Thresholding Algorithm for Matrix Completion. Siam Journal on Optimization, 20(4) :1956-1982, 2010.
[15] R. Liu, Z. Lin and Z. Su. Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning. Asian Conference on Machine Learning. 116-132 2013.
[16] L. Zhang, M. Yang and X. Feng. Sparse Representation or Collaborative Representation: Which Helps Face Recognition? IEEE International Conference on Computer Vision. 2011.
[17] N. Rao, R. Nowak, C. Cox and T. Rogers. Classification With the Sparse Group Lasso. IEEE Transactions on Signal Processing, 64(2) :448-463, 2016.