IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2018.DOI

Efficient Low Rank Regression Based on Nuclear Norm Regularization

LINLU WU[1], XIANGDONG ZHANG[1], QIANQIAN WANG[1], QUANXUE GAO[1]

[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China

Corresponding author: Xiangdong Zhang (e-mail: xdchen@mail.xidian.edu.cn), QianQian Wang(qianqian174@foxmail.com).

ABSTRACT In real world application, data often has a low rank structure as it contains much redundant information, such as face image. Low rank regression method has been proved to be an effective learning mechanism by exploring the low rank structure of the real world data. However, the majority of the existing low rank regression methods just impose the low rank constraint on the coefficient matrix so that they do not take into account the global information of the data. In addition, those methods need much computing time, especially when the data dimension is high. In this paper, we proposed a novel model called Nuclear Norm Low Rank Regression (NNLR) which directly imposes low rank constraint on projected data. In this way, our method can preserve the global geometric structure of data and obtain an explicit solution. An efficient iterative algorithm is proposed to solve the optimize problem. Moreover, we extend the application of NNLR from single view data to multi-view data. Extensive experiments on some well-known datasets containing single-view and multi-view datasets show that NNLR is superior to existing low rank regression method.

INDEX TERMS Low rank regression, Nuclear norm, Efficient solution

I. INTRODUCTION

MULTIVARIABLE linear regression is a basic and effective technique in data mining and machine learning. It models the relationship between predictors and responses by a linear approach. Ridge regression can achieve better results by adding a Frobenius norm based regularization on linear regression loss objective [1], [2]. Other classical methods include lasso regression [3], least angle regression [4], and elastic net regression [5].

In many applications, such as gene expression, document classification and face recognition, predictor and response are high dimensional, the performance of linear regression methods may degrade when deal with this kind of data. The reason may be that the ordinary linear regression method is equivalent to separately regressing each response on the set of predictors, which ignores the underlying correlations between responses and that may be infeasible in the case of high-dimension. In general, the curse-of-dimensional can be mitigated by dimension reduction.

An efficient mechanism to solve the problem of curse-of-dimensional in linear regression method is low rank regression, which means the rank of the coefficient matrix is less than both the dimension of predictor and response. In low rank regression method, dimension reduction is achieved by imposing low rank constraint on the coefficient matrix. Anderson first introduced the low rank regression in [6]. Later many related literature, such as [7], [8], [9], were proposed. For more description and discussion of reduced-rank regression, we refer the reader to the books [10], [11]. All these works can be roughly divided into two categories.

The first category is rank regularization methods, which can be viewed as a L0-norm regularizer in terms of the singular value vector of the coefficient matrix. The rank regularization was first proposed in [12]. Adding a ridge regularization was proposed in [13]. Although the general rank minimization is a non-convex and NP-hard problem, the objectives with rank regularization are solvable in [7] by using the singular value shrinkage estimator. After that, Cai et al. [14] introduced a low rank regression method named low rank ridge regression (LRRR), and they solved the problem by expressing the coefficient matrix as the product of two lower rank matrices. Then, Zheng et al. [8] extended the low rank regression method to a multi-view low rank regression method and provided a closed form solution. [15] solved the problem when the responses have a tensor structure. But the discontinuity of rank in rank regularization method results in inflexible bias-variance trade-off in model fitting of application.

The second category is nuclear norm regularization methods, which can also be viewed as a L1-norm of the singular

value vector of the coefficient matrix. Nuclear norm based methods are a hot topic of research in recent years, such as [15], [16] which introduced nuclear norm based regression models as classifiers used in face recognition. Yuan et al. [17] proposed a nuclear norm regularization regression(NNRR) method, in which the regularizer is defined as the nuclear norm of the coefficient matrix. The solution of the nuclear norm method can be found in [18], [9], from which we can find that the solution is computationally intensive. As we know, by adding a ridge regularization [19] on the linear regression loss function, i.e. ridge regression, can achieve better performance than linear regression, It is regrettable that the method in [17] did not consider this point.

More recently, low-rank robust regression (denoted as LR-RR) [20] has been proposed to learn a robust regression model in the clean low-rank sample space highly correlated to output variables. Although LR-RR can reduce most arbitrary sparse outliers/noise both within the domain subspace and outside of it, it tends to be sensitive to outliers/noise among a set of disjoint subspaces. Then, Zhang et al. [21] proposed low-rank-sparse representation for robust regression(LRS-RR) which can handle the outliers/noise lying on the sparsely disjoint subspaces. Moreover, Li et al. [22] proposed a low rank Extreme learning Machine(LR-ELM) method. ELM is a least-square based learning algorithm and LR-ELM captures the underlying relationship between features and preserves the global geometric structure by imposing low-rank constraints with the extracted features.

However, those work improves the performance of linear regression by imposing low rank constraint on the coefficient matrix, they do not take into account the global information of projected data. In addition, they need much computing-time especially for high dimensional data. To tackle these problems, in this paper, we propose a new nuclear norm linear regression method (NNLR). We compared our method with other similar methods in Table 1, and the contributions of this paper are summarized as follows:

- Firstly, we impose low rank constraint on the projected data which means the product of the data matrix and the coefficient matrix rather than the coefficient matrix. This set-up can preserve the global geometric structure of data while imposing low rank constraint on coefficient. Another advantage of this set-up is the optimize problem can be solved efficiently.
- Secondly, our method includes a ridge regularization which makes the method more robust [23], [24]. Therefore, NNLR allows the low rank regression problem to be solved stably and efficiently. Compared to the discontinuous rank regularization methods, our method results flexible bias-variance trade-off in the model fitting. Compared to the computationally intensive nuclear norm method introduced in Yuan et at. [17], our method is more robust and efficient.
- Thirdly, we developed an efficient iterative algorithm to solve NNLR. The proposed optimizer algorithm consumes little training time, especially when the data dimension is high.
- Finally, we extended the NNLR method to multi-view case in which the result is decided by the sum of each view. Extensive experiment on some well-known datasets shows that NNLR outperforms LRRR and other related method.

TABLE 1: Comparison with related methods.

Methods	Advantage	Disadvantage
NNRR [17]	(1) Capture the underlying correlations between responses.	(1) Computationally intensive. (2) Poor robustness.
LRRR [14]	(1) Capture the underlying correlations between responses. (2) Include L2-regularizer.	(1) Inflexible bias-variance trade-off in model fitting. (2) Computationally intensive when dimension is high.
LR-RR [20]	(1) Robust to noise.	(1) Neglect the underlying correlations between responses.
NNLR	(1) Capture the underlying correlations between responses. (2) Preserve the global geometric structure. (3) Include L2-regularizer.	(1) Two continuous parameters need to adjust.

II. BACKGROUND

Multi-variable linear regression is a basic regression method due to it is simple and effective in data mining and machine learning. Suppose there are n observations, response $y_i \in R^c$ and predictor $x_i \in R^p$, the multi-variable linear regression model

$$Y = X^T C + E \tag{1}$$

where indicator matrix $Y = [y_1, y_2, ..., y_n]^T$ and data matrix $X = [x_1, x_2, ..., x_n]$, $C \in R^{p\times c}$ is a coefficient matrix, E is the regression noise. To solve this problem, Ordinary least squares approach assumes a linear dependence among responses and boils down to finding a matrix C that minimizes the squared error $\left\|Y - X^T C\right\|_F^2$. But this method usually have low performance in the high-dimension case as it ignores the underlying correlations between responses(classes) and which can be mitigated by dimension reduction. In [14], we can see that low rank regression is equivalent to perform Linear Discriminant Analysis (LDA) and linear regression simultaneously. In low rank regression method, dimension reduction is achieved by imposing low rank constraint on the coefficient matrix.

To improve the performance of linear regression method, in recent work [14] researchers introduced a method named low rank ridge regression (LRRR), which is shown in following

$$\left\|Y - X^T AB\right\|_F^2 + \lambda \left\|AB\right\|_F^2 \tag{2}$$

where $\|\|_F$ denotes the Frobenius norm, $A \in R^{p\times s}$, $B \in R^{s\times c}$ and $s < \min(p, c)$, s is the rank of C and λ is the

nonnegative parameter. Similarly, we can see that the LRRR objective function is equivalent to the following problem

$$\left\|Y - X^T C\right\|_F^2 + \lambda \left\|C\right\|_F^2 \qquad s.t. \quad rank(C) \leq s \tag{3}$$

Therefore, we can see LRRR is a robust rank regularization method. Because rank is discontinuity, i.e. s can only take a finite number of integers, LRRR results in an inflexible bias-variance trade-off in the model fitting. Fortunately, nuclear norm regularization regression methods result in a continuous solution path. A nuclear norm regularization method introduced in [17] is shown in following

$$\left\|Y - X^T C\right\|_F^2 + \lambda \|C\|_* \tag{4}$$

where $\|\|_*$ denotes the nuclear norm. The nuclear norm is the L1-norm of the singular value vector which encourages sparsity among the singular values and achieves low rank. The solution to problem (4) generally does not have an explicit form. Extensive works have been devoted to this minimization problem [18] [14], but they are computationally intensive for large-scale data. In order to make the method more robust [23], [24], we incorporate a ridge regularizer, i.e.$\|C\|_F^2$, into our method.

III. NUCLEAR NORM BASED LOW RANK REGRESSION

A. MOTIVATION AND NNLR METHOD

As above analyses, rank regularization results in inflexible bias-variance trade-off in the model fitting, the solution to nuclear norm regularization problem is computationally intensive especially in high dimension case. In addition, both of them do not take into account the global information of data which is important for classification.

According to Sylvester Inequality, we know that the rank of the product of the data matrix and the coefficient matrix smaller than the rank of the coefficient matrix, i.e. $rank(X^T C) \leq rank(C)$. So we can see that low rank constraint on C impairs low rank constraint on $X^T C$ and the converse is not necessarily true. The following theorem introduced in [25] showed that low rank constraint on $X^T C$ is equal to constraint on C.

Theorem 1: If $rank(X^T C) \leq r$ then there is a matrix C' with $rank(C') \leq r$ and $X^T C = X^T C'$.

Proof: Suppose $rank(X^T C) = r' \leq r$, we can write $X^T C = BA$ where $B \in R^{(p+1)\times r'}$ and $A \in R^{r'\times q}$ with $rank(B) = rank(A) = r'$. Since the dimensions of $span\{BA\}$ (the space spanned by the columns of the matrix) and $span\{B\}$ are noth r', we have $span\{B\} = span\{BA\} = span(X^T C) \subseteq span\{X\}$ and thus there is a matrix D such that $B = X^T D$ This implies $X^T C = X^T DA$ and the matrix $C' := DA$ has rank at most r' since $rank(A) = r'$

To improve low rank regression method, we need to preserve the global geometric structure, i.e. low rank structure, in the subspace. Therefore we use the nuclear norm of $X^T C$ archive the low rank constraint on C. There are three advantages of this set-up. First, $\left\|X^T C\right\|_*$ can describe the global geometric structure. of data which is important for classification, but $\|C\|_*$ can not. Second, supposing $rank(X^T C) = a$, according to theorem 1, it is possible that there is a matrix C' satisfying $rank(X^T C)$<a. In this case, if we impose low rank constraint on the coefficient matrix C, we will get an inaccurate solution, but if we impose low rank constraint on $X^T C$, the solution will be more explicit. Third, the problem, $\min_C \left\|Y - X^T C\right\|_F^2 + \lambda \left\|X^T C\right\|_*$, can be solved more explicitly and efficiently by singular value shrinkage estimator which will be introduce in the next section. Therefore, this set-up can preserve more useful information which may improve the performance and result in a more explicit and efficient solution.

B. NUCLEAR NORM BASED LOW RANK REGRESSION FOR SINGLE-VIEW PROBLEM

As introduced in [23], [24], by adding a Frobenius norm based regularization on the linear regression loss, ridge regression can achieve better performance and more robust than linear regression. Therefore, it is important and necessary to add the ridge regularization into low rank regression formulation. In addition, to improve the low rank regression method, we impose the low rank constraint on $X^T C$ rather than C for more efficient solution and better performance. So, we propose the following nuclear norm based low rank regression(NNLR) as

$$\min_C \left\|Y - X^T C\right\|_F^2 + \lambda_1 \left\|C\right\|_F^2 + \lambda_2 \left\|X^T C\right\|_* \tag{5}$$

where $C = [C_1, C_2, ..., C_n] \in R^{p\times c}$ is the coefficient matrix. λ_1 is the regularization parameter, λ_2 is the positive scalar.

To make the problem easy to solve, Eq. (5) can be written as

$$\begin{aligned} \min_{C,M} \left\|X^T C - Y\right\|_F^2 + \lambda_1 \left\|C\right\|_F^2 + \lambda_2 \|M\|_*, \\ s.t. \quad M = X^T C, \end{aligned} \tag{6}$$

1) Iterative solution

According to Augmented Lagrangian method(ALM) introduced in [26], we can replace the constrained optimization problem showed in Eq. (6) by the following unconstrained problem

$$\begin{aligned} \min_{C,M,A} \left\|X^T C - Y\right\|_F^2 + \lambda_1 \left\|C\right\|_F^2 + \lambda_2 \|M\|_* \\ + \frac{\mu}{2} \left\| \frac{A}{\mu} + M - X^T C \right\|_F^2 \end{aligned} \tag{7}$$

where μ is a positive scalar, A is an estimate of the Lagrange multiplier. (7) is a deformation of standrad ALM , and it is equivalent to the form of standrad ALM when we update M and C and can be solved more convenient.

Fix A and C, update M When A and C is fixed, the first and second item of problem (7) could be seen as constant, then the problem in Eq. (7) transforms into

$$\min_M \lambda_2 \|M\|_* + \frac{\mu}{2} \left\| \frac{A}{\mu} + M - X^T C \right\|_F^2 \tag{8}$$

To solve this problem, we need the theorem of singular value shrinkage operator [9], Consider the singular value decomposition (SVD) of $H \in R^{m\times n}$ with rank r:

$$H = U\Sigma V^T \qquad \Sigma = diag(\{\sigma_i\}_{1\le i\le r}) \tag{9}$$

Algorithm 1 Nuclear norm low rank regression

Input: Data matrix $X \in R^{p\times n}$, regularization weight parameterλ_1 and λ_2, parameter ρ, $1.1 \le \rho \le 1.2$, label matrix $Y \in R^{n\times c}$.
Repeat:
Compute the optimal solution M of Eq. (8) by using Eq. (13).
Compute the optimal solution C of Eq. (14) by using Eq. (16).
Update A and μ by using rule (17).
Until converge
Output: Coefficient matrix $C \in R^{p\times c}$.

where $U \in R^{m\times r}$, $V \in R^{n\times r}$ and the singular value σ_i are positive. For each $\tau \ge 0$ define the singular value shrinkage operator as follows

$$\begin{aligned} D_\tau(H) &:= UD_\tau(\Sigma)V^T \\ D_\tau(\Sigma) &= diag(\max(\sigma_i - \tau, 0)) \end{aligned} \tag{10}$$

Theorem 2: For each $\tau \ge 0$ and $Q \in R^{m\times n}$, the singular value shrinkage operator obeys:

$$D_\tau(Q) = \arg\min_H \frac{1}{2}\|H - Q\|_F^2 + \tau\|H\|_* \tag{11}$$

Then, the problem in Eq. (8) can be written as following:

$$\min_M \frac{1}{2}\left\|M - \left(X^TC - \frac{A}{\mu}\right)\right\|_F^2 + \frac{\lambda_2}{\mu}\|M\|_* \tag{12}$$

So, M^*, the optimal solution to Eq. (8), can be formed by

$$M^* = D_\tau(X^TC - \frac{A}{\mu}) \tag{13}$$

Fix M and A, update C Since M and A is fixed, the third item of Eq. (7) can be seen as constant. Therefore, the problem (7) becomes

$$\min_C \left\|X^TC - Y\right\|_F^2 + \lambda_1\|C\|_F^2 + \frac{\mu}{2}\left\|\frac{A}{\mu} + M - X^TC\right\|_F^2 \tag{14}$$

Let $J(C)$ be the object function of problem in Eq. (14), Taking derivative of J w.r.t. C, we have

$$\frac{\partial J}{\partial C} = [(2+\mu)X^TX + 2\lambda_1 I]C - X^T(2Y + \mu M - A) \tag{15}$$

where $I \in R^{n\times n}$ is an identity matrix. Setting Eq. (15) to zero, the optimal solution to the problem (14) can be written as follows

$$C^* = [(2+\mu)X^TX + 2\lambda_1 I]^{-1}X^T(2Y + \mu M - A) \tag{16}$$

Fix M and C, update AAccording to Augmented Lagrangian method, the variable matrix A and coefficient μ can be updated by the following rule

$$A \leftarrow A + \mu(M - X^TC) \qquad \mu \leftarrow \rho\mu \tag{17}$$

where ρ is a constant which is between 1.1 and 1.2 empirically.

Algorithm 2 Multi-view low rank regression

Input: Data matrix $X_v \in R^{p_v\times n}$, $v = 1, 2, ..., V$ regularization weight parameterλ_1 and λ_2, parameter ρ, $1.1 \le \rho \le 1.2$, label matrix $Y \in R^{n\times c}$.
For view v:
Calculate coefficient matrix of each viewC_v by Algorithm 1.
end
Output: Coefficient matrix of each view $C_v \in R^{p_v\times c}$, $v = 1, 2, ..., V$.

2) Algorithm

We present Algorithm 1 to sum up the steps of nuclear norm low rank regression model for single-view problem. The input of the algorithm are data matrix $X \in R^{p\times n}$ which is centered and normalized, where P is the dimension of data and N is the number of sample, class label indicator matrix $Y \in R^{n\times c}$, where C is the number of class, the regularization weight parameterλ_1, the parameter of nuclear norm item and the parameter of Augment Lagrange method. The output of this algorithm is matrix $C \in R^{p\times c}$. We need to repeat updating the variable in Eq. (7) until the result of Eq. (7) converge, and then we get the optimal solution C^*.

For single-view situation, after getting optimal solution, the following decision function is used to decide the class of sample x

$$\arg\max_{1\le j\le c}(y)_j \tag{18}$$

where vector $y = x^TC^* \in R^{1\times c}$, $(y)_j$ is the j-th element of y. The classification of sample x corresponds to the index of maximum value in vector y.

3) Computational complexity of NNLR

In the iterations of NNLR, the computational costs are mainly SVD, eigen decomposition and matrix multiplication. For the data matrix $X \in R^{p\times n}$, the computation complexity of full SVD is $o(p * n^2)$. Each iteration of NNLR mainly includes the SVD of X^TC, so the computational cost is $o(n * c^2)$. In general, c is is a small value and so the solution is efficient. In addition, we need to compute the covariance matrix X^TX, which is up to $o(p*n^2)$ in each iteration, and then we need to do eigen decomposition,which needs $o(n^3)$. Hence, the total complexity for NNLR is $o(p * c^2 + p * n^2 + n^3)$ and t s the iteration number.

C. NUCLEAR NORM BASED LOW RANK REGRESSION FOR MULTI-VIEW PROBLEM

In many tasks [27], [28], [29], the data include information from many views. Therefore, our method should be extended

TABLE 2: The summary of the single-view datasets.

Dataset	n	c	d
AR	1680	12	2000
Extended Yale B	1984	31	1024
Cal101-HOG	441	7	620
Cal101-LBP	441	7	1160
Cal101-Sift	441	7	2560

TABLE 3: The summary of the multi-view datasets.

Dataset	n	c	v	d_v
MSRC	210	7	5	24,576,512,256,254
Cal101	441	7	3	620,1160,2560
Cal101-HOG&LBP	441	7	2	620,1160
Digits	2000	10	6	76,216,64,240,47,6

to work for multi-view problem. When the data includes many views, the objective function becomes

$$\min_{C} \sum\nolimits_{v} \{\left\|Y - X_v^T C_v\right\|_F^2 + \lambda_1 \left\|C_v\right\|_F^2 + \lambda_2 \left\|X_v^T C_v\right\|_*\} \tag{19}$$

where $X_v \in R^{p_v \times n}$, $C_v \in R^{p_v \times c}$, p_v is the dimension of view v, $\sum_v p_v = p$, $v = 1, 2, ..., V$.

In multi-view case, we predict a class of sample x using algorithm 1 for each view and then obtain the C_v of each view. Then, we learn label vector $y \in R^{1\times c}$ that minimizes the difference between itself and projected data of each view $x_v^T C_v$

$$\min_{y} \sum\nolimits_{v} \left\|x_v^T C_v - y\right\|_F^2 \tag{20}$$

The optimum solution of Eq. (20) be shown as

$$y = \left(\sum\nolimits_{v} x_v^T C_v\right) / V \tag{21}$$

After obtaining y, we can get the class indicator of sample x by using Eq. (18). The Algorithm to solve multi-view low rank regression problem is described in Algorithm 2.

IV. EXPERIMENTS

In this section, we perform experiments on 4 different datasets including single-view data and multi-view-data. We compare the recognition rate of NNLR with the classical method and related methods which are rank regularized low rank regression method LRRR, sparse regression based on L21 norm method L21LR, ridge regression RR and linear regression LR. In addition, as suggested by referees, we compare our method with two new methods, LR-RR [21] and LR-ELM [22]. The solution to these methods in multi-view case is similar to the method introduced in section 3.3, i.e. we calculate the coefficient matrix of each view separately and then get the last result of classification by Eq. (21) and Eq. (18).

A. DETAILS OF DATASETS

Extended Yale B face dataset includes 1984 frontal-face pictures of 31 people with different lighting conditions dataset, where each person has 64 pictures and the size of each picture is 32×32.

A subset of the AR face dataset [30] containing 1680 images from 12 individuals were selected and used in our experiments. The pixel values were normalized on 0-255 and the size of images is 50 rows and 40 columns.

USPS Handwrite Digits contain 10 classes, 0 through 9, and each class has 1100 images which is 8 bit gray image. We randomly select 110 images from each class for our single-view experiment.

MSRC [31]is an image scene data including trees, buildings, planes, cows, faces, cars and so on. It has 210 images from 7 classes and every class has 30 images. We extract 5 features from the original picture making up new dataset for our multi-view experiment.

Digits consists of the features of handwritten numbers (0-9) extracted from the Netherlands utility graph collection. Each number (category) digitized in the binary image has 200 examples that are represented in the six feature sets (views) in this data set. View-1 is a feature set of 76 Fourier coefficients of the character shapes; View-2 is a featureset of 216 profile correlations; View-3 is a feature set of 64 Karhunen-Love coefficients; View-4 is a feature set of 240 pixel averages in a 2×3 window; View-5 is a feature set of 47 Zernike moments; View-6 is a feature set of 6 Morphological features.

The Caltech101 database is a dataset with 101 image classes. In our experiments, we selected seven of these classes, namely Face, Dolla-Bill, Garfield, Motorbikes, Stop-Sign, Snoopy and Windsor-Chair, a total of 441 image. Then, we extract three features from each image. View-1 is a Cal101-HOG feature of 620 dimensions, view-2 is an 1160-dimensional Cal101-LBP feature and view-3 is a 2560-dimensional Cal101-Sift feature. We use the 3 features separately in the single-view experiment and sets (Cal101-HOG, Cal101-LBP) and (Cal101-HOG, Cal101-LBP, Cal101-Sift) in the multi-view experiments.

Some sample images are shown in FIGURE 1.

B. SINGLE-VIEW EXPERIMENTS

In the single-view experiments, we choose AR, PIE, USPS and features of cal101 dataset to complete the experiment. In order to show the influence of dimensions on the efficiency of the algorithm, we separately use the Cal101-HOG, Cal101-LBP and Cal101-Sift in experiments. For AR, PIE and USPS, 50% images of each individual randomly selected are used as the training set and the rest as the testing set. Summary of the single-view datasets attributes are presented in TABLE 2, where n is sample number, c is class number, d is the dimension of features. For the cal101 dataset, we randomly select 290 images for training and the rest for test. For all the experiments, we average the results over 10 random splits. We also computed the training time (in seconds) of all the

FIGURE 1: Some samples from AR, Digits, USPS and Extended Yale B.

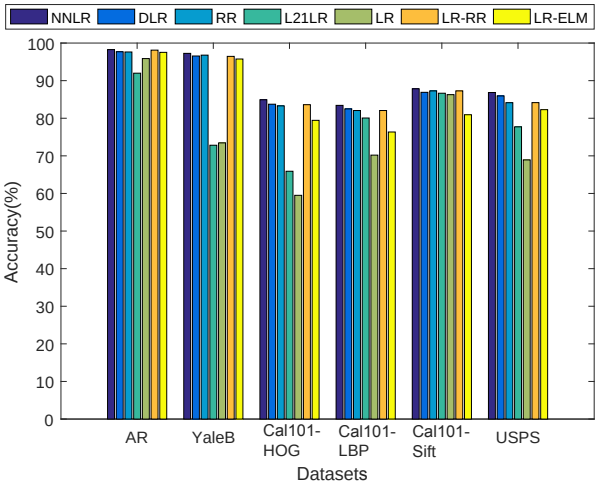

FIGURE 2: Recognition rate of single-view datasets.

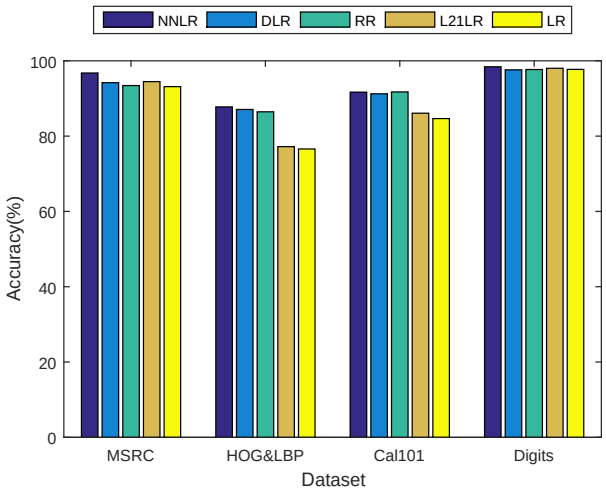

FIGURE 3: Recognition rate of multi-view datasets.

methods on the datasets used in the experiments. All the experiments were run on a PC (CPU: Intel Core i7-4790, 3.6 GHz; RAM: 16 GB; 64-bit operation system).

FIGURE 2 shows the recognition rate histogram, TABLE 4 shows the mean and standard deviation of the accuracy over 10 random split. TABLE 5 describes the mean and standard deviation of computing time.

C. MULTI-VIEW EXPERIMENTS

Previous work [23], [24] show that ridge regression will have better performance than linear regression. However, all existing work is based on a single view. Does multi-view ridge regression produce better results than multi-view linear regression? Therefore, we will examine the performance of all methods on multi-view datasets.

We use 4 datasets in the multi-view experiment, which are MSRC, Digits, Cal101(using 3 features as 3 views respectively) and Cal101-HOG&LBP (using Cal101-HOG and Cal101-LBP of Cal101 as 2 views respectively). Summary of the datasets attributes are presented in TABLE 3, where n is sample number, c is class number, v is view number and d_v lists the dimensions of different views. For Digits and MSRC, 50% images of each individual randomly selected are used as the training set and the rest as the testing set. For cal101 and Cal101-HOG&LBP dataset, we randomly select 290 images for training and the rest for test. Similarly, we computed the training time (in seconds) of all the methods on the datasets.

To help researchers easily compare the efficiency of two kind low rank regression method, we define a variable r,

$$r = \frac{t_{LRRR}}{t_{NNLR}} \tag{22}$$

where t_{LRRR} is the training time of LRRR and t_{NNLR} is the training time of NNLR. We use the training time of NNLR and LRRR on Cal101-HOG, Cal101-LBP, Cal101-Sift, Cal101-HOG&LBP and Cal101 to calculate r and plot the curve of r versus dimension of those datasets. The curve is showed in FIGURE 4. In addition, FIGURE 3 shows that the recognition rate histogram, TABLE 6 shows the mean and standard deviation of accuracy over 10 random splits. TABLE 6 describes the mean and standard deviation of computing time.

D. EXPERIMENT RESULTS AND ANALYSIS

From FIGURE 2, FIGURE 3 and TABLE 4, TABLE 5, we can obtain some interesting observations.

1) **Comparison of linear regression(LR) and the regularized linear regression.** The regularized linear regression methods achieve better recognition rate than LR in single-view datasets and multi-view datasets. FIGURE 2 and TABLE 4 show that NNLR, LR-RR, LR-ELM, LRRR and RR produce better recognition rate than LR in all single-view datasets, L21LR outperforms

TABLE 4: Recognition rate of different methods on single-view datasets (mean±std%).

Dataset	AR	Extended Yale B	Cal101-HOG	Cal101-LBP	Cal101-Sift	USPS
NNLR	**98.25±0.55**	**97.25±2.31**	**84.91±2.36**	**83.42±2.43**	**87.83±2.18**	**86.84±1.5**
LR-RR	98.11±0.61	96.45±2.67	83.60±2.43	82.05±2.40	87.27±2.25	84.15±1.65
LR-ELM	97.52±2.78	95.75±2.78	79.44±2.55	76.34±3.13	80.93±3.19	82.28±2.15
LRRR	97.68±0.67	96.55±2.42	83.73±2.23	82.48±2.36	86.89±2.02	85.93±1.5
RR	97.62±0.65	96.78±2.42	83.29±2.16	82.05±2.40	87.27±2.25	84.13±1.65
L21LR	91.98±9.22	72.82±3.11	65.90±11.01	80.06±3.82	86.65±2.18	77.71±2.85
LR	95.86±0.62	73.44±3.01	59.50±3.75	70.19±4.43	86.27±2.44	68.91±1.58

TABLE 5: Computing time of different methods on single-view datasets (mean±std%).

Dataset	AR	Extended Yale B	Cal101-HOG	Cal101-LBP	Cal101-Sift	USPS
NNLR	14.25±0.92	13.83±2.41	1.77±0.30	2.60±0.35	13.48±1.02	0.63±0.07
LR-RR	12.85±0.97	3.42±0.25	0.47±0.06	2.16±0.21	20.78±0.44	0.14±0.02
LR-ELM	35.8±2.91	34.87±1.85	0.79±0.03	11.14±0.36	10.13±0.62	14.50±1.76
LRRR	27.21±1.29	3.11±0.16	0.59±0.03	2.96±0.17	27.35±6.79	0.14±0.02
RR	21.90±0.87	3.54±0.33	0.59±0.08	3.64±0.44	34.14±3.08	0.06±0.01
L21LR	129.55±23.46	3.61±0.32	7.72±1.01	71.24±3.78	566.2±19.32	2.50±0.24
LR	13.68±1.23	2.88±0.22	0.34±0.04	2.50±0.22	26.17±2.30	0.07±0.02

TABLE 6: Recognition rate of different methods on multi-view datasets (mean±std%).

Dataset	MSRC	Cal101-HOG&LBP	Cal101	Digits
NNLR	**96.76±1.75**	**87.76±1.31**	91.68±1.50	**98.42±0.15**
LRRR	94.19±2.52	87.08±1.16	91.24±1.46	97.61±0.23
RR	93.43±2.99	86.46±1.24	**91.74±1.00**	97.69±0.17
L21LR	94.48±2.05	77.2±8.56	86.09±5.67	98.02±0.24
LR	93.14±2.72	76.58±3.18	84.66±3.32	97.72±0.19

TABLE 7: Computing time of different methods on multi-view datasets (mean±std%).

Dataset	MSRC	Cal101-HOG&LBP	Cal101	Digits
NNLR	**0.21±0.04**	6.77±0.43	**6.63±0.54**	0.63±0.09
LRRR	8.56±0.55	15.06±0.46	282±47.02	0.87±0.08
RR	1.78±0.20	4.21±0.16	35.5±2.03	0.12±0.01
L21LR	4.08±0.23	73.85±1.51	848.98±25.18	1.73±0.23
LR	0.55±0.04	**3±0.24**	25.88±2.14	**0.05±0.01**

LR in single-view datasets Cal101 and USPS. FIGURE 3 and TABLE 5 show that all the regularized linear regression methods have better recognition rate than LR in all multi-view datasets. The reason is that those regularized method provide priori constraints on the coefficient matrix and improve the discriminant ability of methods. This indicates that regularization item plays an important role in enhancing the performance of linear regression.

2) **Comparison with other low-rank regularized regression methods.** Both LRRR and NNLR are based on ridge regression, and they explore the low rank structure of data and alleviate the rank deficiency problem. In addition, NNLR takes into account the global geometric structure which described by nuclear norm of projected data. FIGURE 2, FIGURE 3 and TABLE 4, TABLE 5 show that NNLR has better recognition rate than LRRR in all single-view datasets and all multi-view datasets. This indicates that preserve global geometric structure is useful to improve the discriminant capability of feature decreased dimension and enhance the learning model. From TABLE 4, we can see LR-RR achieves lower performance than NNLR and LRRR. The reason may be that LR-RR is mainly focus on the case that the data is

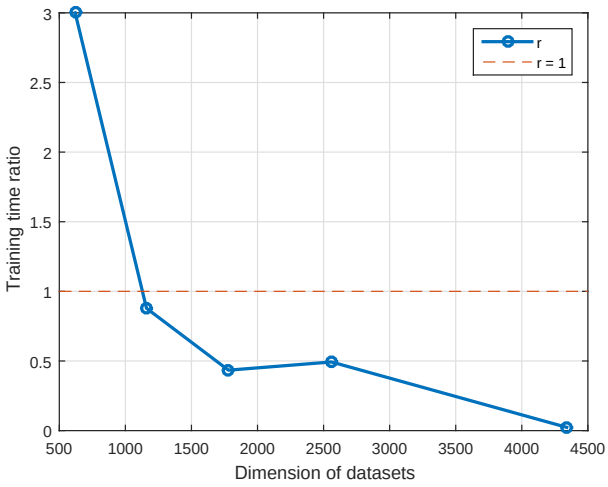

FIGURE 4: The curve of r versus dimension.

'dirty', i.e. data has much noise. When the data is clean, its performance degenerate to RR.

3) **Comparison of sigle-view methods and multi-view methods.** Multi-view regression methods use data or information from multiple channels, such as different image features, both webpage citations view and contents view. Generally, we expect that multi-view regression can produce better results by exploiting information from multiple views. TABLE 4 and TABLE 5 show that the results of all single-view methods in sigle-view datasets Cal101-HOG and Cal101-LBP is lower than the results of multi-view methods in multi-view datasets Cal101-HOG&LBP and Cal101, the results of all single-view methods in sigle-view datasets Cal101-HOG, Cal101-HOG and Cal101-sift is lower than the results of multi-view methods in multi-view dataset Cal101. The reason is that Cal101-HOG&LBP and Cal101 have much more view data and provide more useful information. The multi-view method in our experiment can be seen as an ensemble of single-view method on single-view data. As a result, it has better accuracy than single-view methods.
4) **Comparison of computing time of different methods.** The computing time of different methods are showed in TABLE 5 and 7. FIGURE 4 shows the training time ratio of NNLR to LRRR versus dimension of data. From which, we can find that when there is increasing dimension of data, training time of NNLR is less than LRRR. The reason is that when the dimension of data is big and the number of samples in dataset is approximate, the computation complexity is mainly affected by data dimension p. Ignoring other variables, the computational cost of NNLR and LRRR can be simplified as $o(p)$ and $o(p^2)$, which is consistent with FIGURE 4. There is a prerequisite for this conclusion that the number of samples is small. Therefore, when there are small samples and large dimension, NNLR gets higher recognition rate and lower uses less computing time.

Thus, our newly proposed NNLR method are efficient and practical low rank method for machine learning applications.

V. CONCLUSION

In this paper, we proposed a novel low rank regression named NNLR. NNLR uses a nuclear norm item of the product of the data matrix and the coefficient matrix to explore the low rank structure of data based on ridge regression. Since there is no closed form solution to the optimization problem, we design an iterative algorithm to solve the optimal problem. Compared to classical method and related method such as LRRR and L21LR, we perform extensive experiment on some well-known dataset, from which we can find that NNLR has the best performance. For the dataset that the dimension is large and the number is small, NNLR has less computing-time consumption than LRRR which has a closed form solution.

ACKNOWLEDGMENT

REFERENCES

[1] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics, vol. 42, no. 1, pp. 80–86, 1970.
[2] D. W. Marquardt, "Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation," Technometrics, vol. 12, no. 3, pp. 591–612, 1970.
[3] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," Journal of the Royal Statistical Society, vol. 73, no. 3, pp. 273–282, 2011.
[4] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," Annals of Statistics, vol. 32, no. 2, pp. 407–451, 2004.
[5] Z. Hui and T. Hastie, "Regularization and variable selection via the elastic net," Journal of the Royal Statistical Society, vol. 67, no. 2, pp. 301–320, 2005.
[6] T. W. Anderson, "Estimating linear restrictions on regression coefficients for multivariate normal distributions," Annals of Mathematical Statistics, vol. 22, no. 3, pp. 327–351, 1951.
[7] F. Bunea, Y. She, and M. H. Wegkamp, "Optimal selection of reduced rank estimators of high-dimensional matrices," Annals of Statistics, vol. 39, no. 2, pp. 1282–1309, 2011.
[8] S. Zheng, X. Cai, C. Ding, F. Nie, and H. Huang, "A closed form solution to multi-view low-rank regression," in Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, pp. 1973–1979.
[9] J. F. Cai, Cand, E. J. S, and Z. Shen, "A singular value thresholding algorithm for matrix completion," Siam Journal on Optimization, vol. 20, no. 4, pp. 1956–1982, 2008.
[10] G. C. Reinsel and R. P. Velu, Multivariate Reduced-Rank Regression. Springer,, 1998.
[11] T. Hochkirchen, "Modern multivariate statistical techniques: Regression, classification, and manifold learning," Journal of the Royal Statistical Society, vol. 173, no. 2, pp. 2536–2541, 2010.
[12] A. J. Izenman, "Reduced-rank regression for the multivariate linear model," Journal of Multivariate Analysis, vol. 5, no. 2, pp. 248–264, 1975.
[13] A. Mukherjee and J. Zhu, "Reduced rank ridge regression and its kernel extensions," Statistical Analysis and Data Mining, vol. 4, no. 6, p. 612, 2011.
[14] X. Cai, C. Ding, F. Nie, and H. Huang, "On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions," in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 1124–1132.
[15] G. Rabusseau and H. Kadri, "Low-rank regression with tensor responses," in Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 1867–1875.
[16] J. Qian, J. Yang, F. Zhang, and Z. Lin, "Robust low-rank regularized regression for face recognition with occlusion," in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 21–26.

[17] M. Yuan, "Dimension reduction and coefficient estimation in multivariate linear regression," Journal of the Royal Statistical Society, vol. 69, no. 3, pp. 329–346, 2007.
[18] Toh, Kim-Chuan, Yun, and Sangwoon, "An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems," Pacific Journal of Optimization, vol. 6, no. 3, pp. 615–640, 2009.
[19] D. Ruppert, "The elements of statistical learning: Data mining, inference, and prediction," Mathematical Intelligencer, vol. 99, no. 466, pp. 567–567, 2010.
[20] D. Huang, R. S. Cabral, and F. D. L. Torre, "Robust regression," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 363–375, 2016.
[21] Y. Zhang, D. Shi, J. Gao, and D. Cheng, "Low-rank-sparse subspace representation for robust regression," in IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7445–7454.
[22] Q. Li, Y. Liu, S. Wang, Q. Gao, and X. Gao, "Image classification using low-rank regularized extreme learning machine," IEEE Access, vol. 7, pp. 877–883, 2018.
[23] A. Hoerl and R. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics, vol. 12, no. 1, pp. 55–67, 2000.
[24] A. Mukherjee and J. Zhu, "Reduced rank ridge regression and its kernel extensions." Statistical Analysis and Data Mining, vol. 4, no. 6, pp. 612–622, 2011.
[25] H. Lian and S. Ma, "Reduced-rank regression in sparse multivariate varying-coefficient models with high-dimensional covariates," Statistics, 2013.
[26] D. P. Bertsekas, Constrained optimization and Lagrange multiplier methods. Academic Press, 1982.
[27] N. Liqiang, S. Xuemeng, and C. Tatseng, "Learning from multiple social networks," Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan Claypool Publishers, 2016.
[28] S. RÍźping and T. Scheffer, "Learning with multiple views," In Proc. ICML Workshop on Learning with Multiple Views, 2005.
[29] B. C. J. C. . Zhou D, "Spectral clustering and transductive learning with multiple views," Machine Learning, Twenty-fourth International Conference, Corvallis, Oregon, Usa, June., 2007.
[30] A. M. Martinez, "The ar face database," Cvc Technical Report, vol. 24, 1998.
[31] J. L. Yong and K. Grauman, "Foreground focus: Unsupervised learning from partially matching images," International Journal of Computer Vision, vol. 85, no. 2, pp. 143–166, 2009.