Low-Rank-Sparse Subspace Representation for Robust Regression

Yongqiang Zhang
Harbin Institute of Technology
Harbin, China
seekever@foxmail.com

Daming Shi
Harbin Institute of Technology;Shenzhen University
Harbin, China;Shenzhen, China
d.m.shi@hotmail.com

Junbin Gao
The University of Sydney
Sydney, Australia
junbin.gao@sydney.edu.au

Dansong Cheng
Harbin Institute of Technology
Harbin, China
cdsinhit@hit.edu.cn

Abstract

Learning robust regression model from high-dimensional corrupted data is an essential and difficult problem in many practical applications. The state-of-the-art methods have studied low-rank regression models that are robust against typical noises (like Gaussian noise and out-sample sparse noise) or outliers, such that a regression model can be learned from clean data lying on underlying subspaces. However, few of the existing low-rank regression methods can handle the outliers/noise lying on the sparsely corrupted disjoint subspaces. To address this issue, we propose a low-rank-sparse subspace representation for robust regression, hereafter referred to as LRS-RR in this paper. The main contribution include the following: (1) Unlike most of the existing regression methods, we propose an approach with two phases of low-rank-sparse subspace recovery and regression optimization being carried out simultaneously;(2) we also apply the linearized alternating direction method with adaptive penalty to solved the formulated LRS-RR problem and prove the convergence of the algorithm and analyze its complexity; (3) we demonstrate the efficiency of our method for the high-dimensional corrupted data on both synthetic data and two benchmark datasets against several state-of-the-art robust methods.

1. Introduction

As one of the most important machine learning technique, multivariate linear regression attempts to model the relationship between dependent variables and independent variables by fitting a linear mapping to observed samples. Generally, the Ordinary Least Squares (OLS) regression is represented as $\min_{\mathbf{T}} ||\mathbf{Y}-\mathbf{TX}||_F^2$, where $\mathbf{X}$ denotes the independent variables, $\mathbf{Y}$ the dependent variables, and $\mathbf{T}$ the mapping relationship between $\mathbf{X}$ and $\mathbf{Y}$.

In many real-world applications, such linear regression models suffer from two drawbacks: lack of robustness to outliers/noises and the curse of dimensionality. A typical solution to the former is to estimate noises under an assumed parametric distribution such as Gaussian, whereas a solution to the latter is to select appropriate features by using dimensionality reduction such as Principal Component Analysis(PCA). However, if there exist a small number of gross outliers among the samples, the estimate of model parameters would drift obviously. Moreover, the linear regression models often do not work well in processing high-dimensional data. For regression tasks on high-dimensional data like face images, we often cannot collect and label enough samples.

As a matter of fact, outliers/noise are due to three sources: the first, statistically salient data; the second, miscollected noised data; and the third, occluded multi-class data. Almost all the existing regression methods tackle the outlier/noise problem by getting rid of the first and second types under an assumption of noise distribution, whereas by "Gambling". The third type of outlier can be dealt with by using maximum likelihood. Obviously, only the first type of outlier can arguably follow a parametric distribution. The second and the third types should be treated as sparsely corrupted data. In the meantime, the third type is also caused by linearly non-separable problem, which requires a systematic solution to handle multiple subspaces/multi-classes.

Recently the robust methods have been studied widely to overcome the impact of outliers/noise, such as robust methods based on least sum of squares [32, 30] and methods using the least median of squares[3, 32] in the field of statistics. Approaches via subset selection have been studied in many computer vision applications, like Random Sample Consensus(RANSAC) [35, 8]) which randomly picks sam-

ples to construct a clean model. These methods tend to be expensive in computing when the number of samples and the dimension of sample space are large, and may consequently fail if there are limited inliers. There are also some methods that improve the robustness of Linear Discriminant Analysis (LDA)[16, 9, 22, 41]. Although these methods can remove outliers that are far away from the good samples in each dimension, however, they cannot tackle the partial sample corruptions that occur only in some of the dimensions.

Low-rank regression models can reduce noise and detect outliers partially, though they are firstly proposed to solve the curse-of-dimensionality problem. Many low-rank regression models [4, 38, 2, 6] have been studied to incorporate the correlations among different dimensions, and these models have been proved to be very effective by considering the low-rank structure in real applications. Cai *et. al.* [6] show that the low-rank regression is equivalent to the regularized regression in a learned LDA projected subspace, which can reduce normal distribution outliers/noise. However, these methods cannot deal with outliers or large noise outside of the main subspace, which is the focus of recent robust subspace recovery methods, like robust principle analysis [7]. These methods usually remove noise in independent variables in an unsupervised manner, thus lacking of correlation with dependent variables. More recently, Low-Rank Robust Regression (denoted as LR-RR here) [19] has been proposed to learning 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. An independent subspace and a set of disjoint subspaces are illustrated in Fig. 1. It can be seen that if the outliers or noise occur from non-orthogonal subspaces, the LR-RR tends to fail.

Our work is inspired by the low-rank robust regression [19], low-rank-sparse subspace clustering [36] and some earlier rank minimization methods like Robust Principal Component Analysis (RPCA) [7]. In this paper, we aim to detect intra-sample outliers within disjoint subspaces for robust regression. We propose a new robust regression method via Low-Rank-Sparse-Representation (LRSR), which not only recovers the low-rank disjoint subspaces but also performs a robust regression via sparsity optimization.

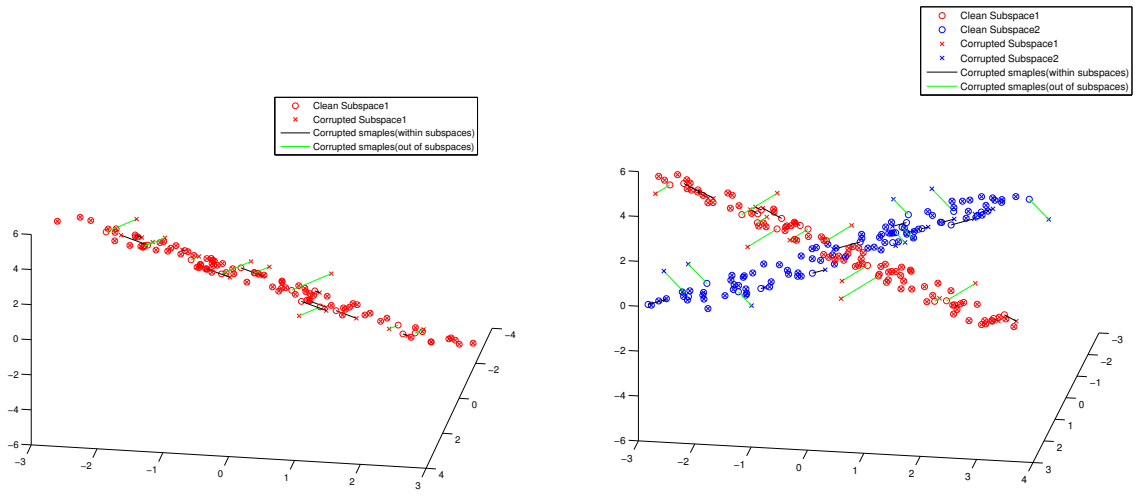
In Section 2, we review several regression approaches and give a new view from subspace learning, including Least Square Regression(LSR) methods [37, 40], Ride Regression [29], Least Absolute Shrinkage and Selection Operator(LASSO) [12], Least Angle Regression [10], rank-reduced regression methods — Low-Rank Regression [38] and Low-Rank Ridge Regression [6], subspace-learning regression — Principal Component Regression (PCR) [34], LDA [31], Support Vector Regression (SVR)[39], Relevance Vector Machine Regression (RVM)[13], Partial Least Squares (PLS) Regression [1], Canonical Correlation Analysis (CCA)[15], Robust PCA[7] and LR-RR[19]. In Section 3, we first give a further explanation about solving Low-Rank-Spares Regression in disjoint subspaces. Second, we propose a low-rank-sparse regression method via the framework of Low Rank subspace Sparse Representation (LRSR) by a supervised manner. Section 4 is dedicated to convergence analysis for the proposed algorithms. In Section 5, we introduce the evaluation metric — *Relative Absolute Error (RAE)* for regression and conduct experiments on synthetic data. Moreover, we evaluate our methods against the state-of-the-art methods RPCA+LSR [4] and LR-RR [19] on two benchmark datasets.

2. Subspace view of Regression Approaches

High-dimensional data like face images or shapes often lie in low-dimensional subspaces, with some entries of the samples are often corrupted. The noises or outliers often exist both inside and outside the main subspaces (orthogonal to main subspaces).Typically, there are two steps in subspace approaches to regression, namely, subspace recovery and regression Optimization. The former aims to extract the principal dimensions that data have spanned, while the latter aims to learning robust regression model from the recovered clean observations.

2.1. Subspace View of Typical Regression Models

To some extent the classical regression methods are related to supervised subspace learning like LSR [37, 40], Ride Regression [29], LASSO [12] and Least Angle Regression [10]. Generally, LSR learns a series of linear mappings from input variables and output variables, which can be represented as a group of hyperplanes lying on certain subspaces. Ride Regression adds an ℓ_2 norm term to regularize the regression model, which is a biased regression method. The learned hyperplane is a balance between the sample space and uniform space (spanned by all unit vectors).LASSO can select a sparse linear mapping of features from the whole sample space, which can be retreated as a linear representation within the subspace spanned by the selected feature space by LASSO. LARS also learns a sparse linear regression model within a selected feature subspace. Principal Component Regression [34] learns coefficients matrix from subspace spanned by principal components of input variables, while LDA [31] learns the most discriminant subspace of input variables that appears to contain all of the class variability. Compared to unsupervised PCR, LDA is a supervised subspace learning method, resulting in better classification or regression result. In the SVR method, an output variable is represented as the linear or non-linear combination of support vectors, which is



(a) LR-RR (b) LRS-RR

Figure 1. Corrupted Subspaces. (a) All data are within a subspace, as LR-RR assumes; (b) All data are distributed among two overlapping subspaces, as LRS-RR assumes. In case (a), the LR-RR works well; In case (b), it tends to fail because that there are some samples lying on more than one subspace, which is not considered by LR-RR.

equivalent to the low-rank representation in the subspace only spanned by support vectors. The dimension of this subspace is equal to the one of a matrix concatenated by all the support vectors. Similar to SVR, Relevance Vector Machine Regression [13] can also be treated as a regression model among the subspace represented by all the relevance vectors.

The PLS [1] finds a linear regression model by projecting the input variables and response variables onto a new distinguishing space. Because both the $\mathbf{X}$ and $\mathbf{Y}$ data are projected to the new subspace, PLS is a bilinear subspace regression model. Different from PLS, CCA projects both $\mathbf{X}$ and $\mathbf{Y}$ data onto a mutual subspace that is related to mutual information between $\mathbf{X}$ and $\mathbf{Y}$ data. Both the PLS and the CCA are supervised subspace learning methods that can be used for regression. Moreover, Lang *et. al.* [23] establish explicit connections between LS, PCR and PLS, and a finite number of other related methods, regression/prediction process of which can be referred to a general cyclic subspace regression. Several rank-reduced regression models have been proposed recently, such as Low-Rank Regression [38] and Low-Rank Ridge Regression [6]. Cai *et. al.* [6] have proven that low-rank regression is equivalent to regression in a regularized LDA subspace. There are also several nonlinear regression methods like Orthogonal Forward Regression [18, 17] and K-Nearest Neighbors ($K-NN$) regression [20]. The former type of methods takes a set of radial basis functions based on input variables as the linear subspace, the dimension of which is equal to the number of radial basis functions. Then an output vector can be represented as a linear combination of projected samples within the subspace. Instead of using a kernel function directly, $K-NN$ represents a sample as the linear combination of its neighbors. It is a local linear model and can be treated as a linear representation in a local subspace, the rank of which is not larger than the neighborhood size K.

2.2. Subspace Recovery

Given that data corrupted with gross errors and outliers are ubiquitous in modern applications, how to recover clean data is essential for robust regression. Robust subspace recovery is a basic problem, in which we assume the clean data set was sampled from several fixed subspaces while outliers/corrupted data may be spread throughout the whole ambient space. One attempt is to recover the underlying fixed subspaces from the corrupted observed data. Modeling high-dimension data in low-dimensional subspaces is the most useful paradigm in subspace recovery. PCA can be regarded as a subspace recovery technique minimizing the sum of squared errors of data points under the assumption that data are from a single fixed unknown subspace. However, its sensitivity to grossly corrupted observations often jeopardizes its robustness. That is, a single grossly corrupted entry in the data could render the estimated subspace far from the true one.

Various methods like [11, 21] have been proposed to augment the accuracy of subspaces recovery. As a breakthrough, RPCA [7] provides clear analysis of exact low-rank recovery with an unspecified rank and the certain ratio of large-scale sparse corruptions. Given a large data matrix $\mathbf{X}$, it may be decomposed as $\mathbf{X} = \mathbf{D} + \mathbf{E}$, where $\mathbf{D}$ has low rank and $\mathbf{E}$ is sparse. To estimate the two com-

ponents, RPCA minimizes a weighted combination of the nuclear norm and the ℓ_1 norm, formulated as follows:

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{E}} \ \|\mathbf{D}\|_* + \lambda\|\mathbf{E}\|_1, \\ &\text{s.t.} \ \ \mathbf{X} = \mathbf{D} + \mathbf{E}, \end{aligned} \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\mathbf{D}$ denotes the matrix of clean data lying in a low-dimensional subspace and $\mathbf{E}$ can be considered as the deviation of $\mathbf{X}$ from the intrinsic low-dimensional subspace. For example, in video analysis, $\mathbf{E}$ represents moving objects in the low-dimensional background.

The Alternating Direction Method (ADM) algorithm [24] can achieve much higher accuracy and better convergence performance than other algorithms [25, 5] when solving the optimization problem (1). However, RPCA only recovers an independent subspace in an unsupervised manner and can not remove noises or outliers inside the subspace. Therefore, RPCA tends to transform data points into a common low-dimensional subspace which may result in weakening the distances between samples of different subjects and the principal angles between subspaces in a practical application.

2.3. Regression Optimization

Robust regression methods aim at learning a robust regression model from observations corrupted by outliers/noise. An idea solution is the removal of outliers/noise by a supervised learning method as follow:

$$\begin{aligned} &\min_{\mathbf{T}} \|(\mathbf{Y} - \mathbf{T}\hat{\mathbf{D}})\|_F^2 \\ &\text{s.t. } \mathbf{X} = \mathbf{D} + \mathbf{E},\ \hat{\mathbf{D}} = \left[\mathbf{D}; \mathbf{1}^{\mathbf{T}}\right]. \end{aligned} \tag{2}$$

where $\mathbf{D}$ is an idea clean data embedded within the main low-rank subspaces, and $\mathbf{E}$ represents outliers/noise both small Gaussian noise and large scale outliers, inside or outside main subspaces. A representative method is Low-Rank Robust Regression by Huang *et al.* [19]. The LR-RR reveals a single low-rank subspace from data by seeking the low-rank representation with sparse noise as follows:

$$\begin{aligned} \min_{\mathbf{T},\mathbf{D},\mathbf{E}} \frac{\eta}{2}\|\mathbf{W}(\mathbf{Y} - \mathbf{T}\hat{\mathbf{D}})\|_F^2 + rank(\mathbf{D}) + \lambda\|\mathbf{E}\|_{\mathbf{0}} \\ \text{s.t. } \mathbf{X} = \mathbf{D} + \mathbf{E},\ \hat{\mathbf{D}} = \left[\mathbf{D}; \mathbf{1}^{\mathbf{T}}\right]. \end{aligned} \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{\mathbf{d_y}\times\mathbf{d_y}}$ is a diagonal matrix that weights the output dimensions, $\mathbf{T} \in \mathbb{R}^{\mathbf{d_y}\times(\mathbf{d_x}+\mathbf{1})}$ is the regression matrix (the extra dimension is for the regression bias term). η and λ are scalars that weight the first and third term in Eq.3 respectively. LR-RR explicitly avoids projecting the outlier matrix $\mathbf{E}$ to the output space by learning the regression $\mathbf{T}$ only from the augmented noise-free data $\hat{\mathbf{D}} = \left[\mathbf{D}; \mathbf{1}^{\mathbf{T}}\right] \in \mathbb{R}^{(d_x+1)\times n}$. The second term $\mathbf{D}$ is a low-dimensional subspace constraint, as a good prior for many computer vision applications. The third term $\mathbf{E}$ ensures noise/outliers to be sparse.

Although LR-RR cleans the data in a supervised manner, it does not work well for corrupted input data from disjoint subspaces, as mentioned above.

3. Low Rank Subspace Sparse Representation for Regression

In this section, we first propose a low-rank-sparse model for robust disjoint subspace regression, and then utilize an efficient linearized alternating direction method with adaptive penalty (LADMAP) [27] to solve the proposed model.

3.1. The LRS-RR Regression model

As analyzed above, low rank representation can capture global information critical for revealing the structure of lower dimensional subspace and removing large disturbances in the original data. LR-RR has excellent performance in regard to analyzing corrupted data drawn from independent subspaces. However, using the original data contaminated with large noises as the dictionary is by no means a good choice. Moreover, LR-RR often fails in the case of disjoint subspaces or overlapping subspaces. For example, in the synthesized data used in the experiment shown in Fig. 1, the first and the second subspaces are intersected with each other, thus we sometimes get some wrong recovery points. Although LR-RR performs very well with data from orthogonal subspaces, it is weaker than LRS-RR in recovering the global subspace structures from the corrupted data. The result of LR-RR in synthetic data in Fig. 1 shows that LR-RR cannot remove the impact of large noises within disjoint subspaces.

Based on the analyses and observations above, we propose a method to combine the low rank and sparse representations for subspace regression, especially for the cases when the subspaces are not independent and data are corrupted by large noise. For example, the corruptions caused by the uneven illumination result in that a relatively large number of within-class data points drifting to other subspaces. Therefore, methods like LR-RR may bring points from different subspaces into the same subspace with the uneven illumination corruption. Nevertheless, LR-RR is better at handling corruptions than RPCA does due to supervised learning. However, we cannot ensure the face image subspaces are orthogonal to or independent from each other. Therefore, we extend the framework by learning a clean dictionary-the basis for subspaces-that satisfies the condition of sparse noise. The model proposed is defined

as follows:

$$\begin{aligned}\hat{\mathscr{J}}_{\text{LRS-RR}} = &\min_{\mathbf{T},\hat{\mathbf{D}},\mathbf{A},\mathbf{Z},\mathbf{J},\mathbf{E}} \frac{\eta}{2}\left\|\mathbf{W}(\mathbf{Y}-\mathbf{T}\hat{\mathbf{D}})\right\|_F^2 + \|\mathbf{A}\|_* \\ &+ \|\mathbf{Z}\|_* + \lambda_2\|\mathbf{J}\|_1 + \lambda_1\|\mathbf{E}\|_1 \\ \text{s.t. } &\mathbf{X}=\mathbf{A}\mathbf{Z}+\mathbf{E},\ \hat{\mathbf{D}}=\left[\mathbf{A}\mathbf{Z};\mathbf{1}^{\mathbf{T}}\right],\ \mathbf{Z}=\mathbf{J},\mathbf{J}\geq\mathbf{0}.\end{aligned} \tag{4}$$

where $\mathbf{W}\in\mathbb{R}^{\mathbf{d_y}\times\mathbf{d_y}}$ is a diagonal matrix that weights the output dimensions, $\mathbf{T}\in\mathbb{R}^{\mathbf{d_y}\times(\mathbf{d_x}+\mathbf{1})}$ is the regression matrix (the extra dimension is for the regression bias term), $\mathbf{A}$ is a clean low-rank dictionary, the cols of which span the main subspaces, and $\mathbf{Z}$ is a coefficient matrix as the low-rank representation of clean samples by dictionary $\mathbf{A}$. η, λ_2, λ_1 are scalars that weight the first, fourth, and fifth term in Eq.4 respectively. $\mathbf{E}$ represents the sample-specific corruptions. As $\text{rank}(\mathbf{A}^*\mathbf{Z}^*)\leq\min\{\text{rank}(\mathbf{A}^*),\text{rank}(\mathbf{Z}^*)\}$, $\mathbf{A}^*\mathbf{Z}^*$ is the low rank recovery of the original data. Instead of low-rank data with sparse noise model in LR-RR, low-rank representation with sparse noise is taken as the constraint term in our model.

Benefiting from low-rank representation model [28], LRS-RR can handle outliers/noise lying in disjoint subspaces. $\mathbf{AZ}$ explicitly avoids projecting the outlier matrix $\mathbf{E}$ to the output space by learning the regression $\mathbf{T}$ only from the augmented noise-free data $\hat{\mathbf{D}}=\left[\mathbf{A}\mathbf{Z};\mathbf{1}^{\mathbf{T}}\right]\in\mathbb{R}^{(\mathbf{d_x}+\mathbf{1})\times\mathbf{n}}$. Note that there are infinite possible decompositions of $\mathbf{X}$ into $\mathbf{AZ}$ and $\mathbf{E}$. LRS-RR thus adds the second, third, fourth and fifth terms in Eq.4 to constrain the possible solutions. The second term constrains the dictionary $\mathbf{A}$ to lie in low-dimensional disjoint subspaces. The third and fourth terms constrain the representation to be low-rank-sparse. The fifth term regularizes $\mathbf{E}$ to be sparse.

3.2. LADMAP solution for LRS-RR model

The optimization problem (4) can be solved using an Augmented Lagrange Muliplier (ALM) technique. First we write its ALM form as follows,

$$\begin{aligned}&\hat{\mathscr{J}}_{\text{LRS-RR}} \\ =&\min_{\mathbf{T},\hat{\mathbf{D}},\mathbf{A},\mathbf{Z},\mathbf{J},\mathbf{E}} \frac{\eta}{2}\left\|\mathbf{W}(\mathbf{Y}-\mathbf{T}\hat{\mathbf{D}})\right\|_F^2 + \|\mathbf{A}\|_* \\ &+\|\mathbf{Z}\|_* + \lambda_2\|\mathbf{J}\|_1 + \lambda_1\|\mathbf{E}\|_1 \\ &+\langle\mathbf{Y_1},\mathbf{X}-\mathbf{A}\mathbf{Z}-\mathbf{E}\rangle + \frac{\mu_1}{2}\|\mathbf{X}-\mathbf{A}\mathbf{Z}-\mathbf{E}\|_{\mathbf{F}}^{\mathbf{2}} \\ &+\langle\mathbf{Y_2},\hat{\mathbf{D}}-\left[\mathbf{A}\mathbf{Z};\mathbf{1}^{\mathbf{T}}\right]\rangle + \frac{\mu_2}{2}\left\|\hat{\mathbf{D}}-\left[\mathbf{A}\mathbf{Z};\mathbf{1}^{\mathbf{T}}\right]\right\|_{\mathbf{F}}^{\mathbf{2}} \\ &+\langle\mathbf{Y_3},\mathbf{Z}-\mathbf{J}\rangle + \frac{\mu_3}{2}\|\mathbf{Z}-\mathbf{J}\|_{\mathbf{F}}^{\mathbf{2}},\end{aligned} \tag{5}$$

where $\mathbf{Y_1}\in\mathbb{R}^{\mathbf{d_x}\times\mathbf{n}}$, $\mathbf{Y_2}\in\mathbb{R}^{(\mathbf{d_x}+\mathbf{1})\times\mathbf{n}}$ and $\mathbf{Y_3}\in\mathbb{R}^{\mathbf{n}\times\mathbf{n}}$ are Lagrange multiplier matrices, and μ_1, μ_2 and μ_3 are the penalty parameters. According to the LADMAP method [27], Eq.5 can be rewritten as,

$$\begin{aligned}\hat{\mathscr{J}}_{\text{LRS-RR}} = &\min_{\mathbf{T},\hat{\mathbf{D}},\mathbf{A},\mathbf{Z},\mathbf{J},\mathbf{E}} \frac{\eta}{2}\left\|\mathbf{W}(\mathbf{Y}-\mathbf{T}\hat{\mathbf{D}})\right\|_F^2 + \|\mathbf{A}\|_* \\ &+\|\mathbf{Z}\|_* + \lambda_2\|\mathbf{J}\|_1 + \lambda_1\|\mathbf{E}\|_1 \\ &+\frac{\mu_1}{2}\|\mathbf{X}-\mathbf{A}\mathbf{Z}-\mathbf{E}+\mathbf{Y_1}/\mu_1\|_F^2 \\ &+\frac{\mu_2}{2}\left\|\hat{\mathbf{D}}-\left[\mathbf{A}\mathbf{Z};\mathbf{1}^{\mathbf{T}}\right]+\mathbf{Y_2}/\mu_2\right\|_F^2 \\ &+\frac{\mu_3}{2}\|\mathbf{Z}-\mathbf{J}+\mathbf{Y_3}/\mu_3\|_F^2.\end{aligned} \tag{6}$$

For each of the six matrices $\mathbf{T},\hat{\mathbf{D}},\mathbf{A},\mathbf{Z},\mathbf{J},\mathbf{E}$ to be solved in particularly Eq. 6, the cost function is convex if the remaining five matrices are kept fixed. Eq. 6 can be solved iteratively via the following subproblems:

1. Fixing $\hat{\mathbf{D}},\mathbf{A},\mathbf{Z},\mathbf{J},\mathbf{E}$, solve (6) for $\mathbf{T}$ by the following problem, denoted by LRS-RR-1,

$$\min_{\mathbf{T}} \frac{\eta}{2}\left\|\mathbf{W}(\mathbf{Y}-\mathbf{T}\hat{\mathbf{D}})\right\|_F^2 \tag{7}$$

which is an ordinary least square regression problem, whose solution is,

$$\mathbf{T}=(\hat{\mathbf{D}}(\hat{\mathbf{D}})^{\mathbf{T}}+\gamma\mathbf{I}_{\mathbf{d_x}+\mathbf{1}})^{-\mathbf{1}}\mathbf{Y}(\hat{\mathbf{D}})^{\mathbf{T}}, \tag{8}$$

where λ is a positive scalar for regularizing the solution to $\mathbf{T}$.

2. Fixing $\mathbf{T},\mathbf{A},\mathbf{Z},\mathbf{J},\mathbf{E}$, solve (4) for $\hat{\mathbf{D}}$ by the following problem, denoted by LRS-RR-2,

$$\min_{\hat{\mathbf{D}}} \frac{\eta}{2}\left\|\mathbf{W}(\mathbf{Y}-\mathbf{T}\hat{\mathbf{D}})\right\|_F^2 + \frac{\mu_2}{2}\left\|\hat{\mathbf{D}}-\left[\mathbf{A}\mathbf{Z};\mathbf{1}^{\mathbf{T}}\right]+\mathbf{Y_2}/\mu_2\right\|_F^2. \tag{9}$$

which is also an ordinary least square regression problem, to which the solution of which is,

$$\hat{\mathbf{D}}=[\eta\mathbf{T}^{\mathbf{T}}\mathbf{W}^{\mathbf{T}}\mathbf{W}\mathbf{T}+\mu_2\mathbf{I_d}]^{-1}[\eta\mathbf{T}^{\mathbf{T}}\mathbf{W}^{\mathbf{T}}\mathbf{W}\mathbf{Y}-\mathbf{Y_2}+\mu_2[\mathbf{A}\mathbf{Z};\mathbf{1}^{\mathbf{T}}]]. \tag{10}$$

3. Fixing $\mathbf{T},\hat{\mathbf{D}},\mathbf{Z},\mathbf{J}$, solve (6) for $\mathbf{A}$ and $\mathbf{E}$ by the following problem, denoted by LRS-RR-3,

$$\begin{aligned}&\min_{\mathbf{A},\mathbf{E}} \|\mathbf{A}\|_* + \lambda_1\|\mathbf{E}\|_1 \\ &+\frac{\mu_1}{2}\|\mathbf{X}-\mathbf{A}\mathbf{Z}-\mathbf{E}+\mathbf{Y_1}/\mu_1\|_F^2 \\ &+\frac{\mu_2}{2}\left\|\hat{\mathbf{D}}-\left[\mathbf{A}\mathbf{Z};\mathbf{1}^{\mathbf{T}}\right]+\mathbf{Y_2}/\mu_2\right\|_F^2.\end{aligned} \tag{11}$$

which is a slight variation of low-rank representation problem [28], and the linear ADM solution is,

$$\begin{aligned}\mathbf{A}^{\mathbf{k+1}} =&\mathcal{D}_{1/\beta_{\mathbf{A}}}(\mathbf{A}^{\mathbf{k}}-\mathbf{F_A}^{\mathbf{k}}/\beta_{\mathbf{A}}), \\ \mathbf{E}^{\mathbf{k+1}} =&\mathcal{S}_{\lambda_1/\mu_1}(\mathbf{X}-\mathbf{A}^{\mathbf{k}}\mathbf{Z}+\mathbf{Y_1}/\mu_1),\end{aligned} \tag{12}$$

where $\mathcal{D}$ is the singular value thresholding[5], $\mathcal{S}$ is the shrinkage operator [42], $\beta_{\mathbf{A}} = (\mu_1 + \mu_2)\tau_{\mathbf{A}}/2$, $\tau_{\mathbf{A}} > \rho(\mathbf{Z^T Z})$ is the proximal parameter, $\rho(\mathbf{Z^T Z})$ denotes the spectral radius of $\mathbf{Z^T Z}$, and $\mathbf{F_A}^{\mathbf{k}}$ is the derivative by $\mathbf{A^k}$ for the second and third terms in Eq. 11,

$$\begin{aligned}\mathbf{F_A}^{\mathbf{k}} = ((\mu_1 + \mu_2)\mathbf{A^k Z} - \mu_1(\mathbf{X} - \mathbf{E}) - \mathbf{Y_1} \\ - (\mu_2 \hat{\mathbf{D}} + \mathbf{Y_2})_{(\mathbf{1:d_x},\cdot)})\mathbf{Z^T}.\end{aligned} \tag{13}$$

4. Fixing $\mathbf{T}, \hat{\mathbf{D}}, \mathbf{A}, \mathbf{E}$, solve (6) for $\mathbf{Z}$ and $\mathbf{J}$ by the following problem, denoted by LRS-RR-4,

$$\begin{aligned}\min_{\mathbf{Z}} \ &\|\mathbf{A}\|_* + \|\mathbf{Z}\|_* + \lambda_2 \|\mathbf{J}\|_1 \\ &+ \frac{\mu_1}{2} \|\mathbf{X} - \mathbf{AZ} - \mathbf{E} + \mathbf{Y_1}/\mu_1\|_F^2 \\ &+ \frac{\mu_2}{2} \left\|\hat{\mathbf{D}} - \left[\mathbf{AZ}; \mathbf{1^T}\right] + \mathbf{Y_2}/\mu_2\right\|_F^2 \\ &+ \frac{\mu_3}{2} \|\mathbf{Z} - \mathbf{J} + \mathbf{Y_3}/\mu_3\|_F^2 .\end{aligned} \tag{14}$$

Similar to Eq. 11, the above optimization problem is also a variation of low-rank representation problem, to which the LADMAP solution is,

$$\begin{aligned}\mathbf{Z^{k+1}} &= \mathcal{D}_{1/\beta_{\mathbf{Z}}}(\mathbf{Z^k} - \mathbf{F_Z}^{\mathbf{k}}/\beta_{\mathbf{Z}}) \\ \mathbf{J^{k+1}} &= \max(\mathbf{Q}, \mathbf{0})\end{aligned} \tag{15}$$

where $\beta_{\mathbf{Z}} = (\mu_1 + \mu_2 + \mu_3)\tau_{\mathbf{Z}}/2$, $\tau_{\mathbf{Z}} > \rho(\mathbf{AA^T})$ is the proximal parameter, $\rho(\mathbf{AA^T})$ denotes the spectral radius of $\mathbf{AA^T}$, and $\mathbf{F_Z}^{\mathbf{k}}$ is the derivative by $\mathbf{Z^k}$ for the second, third and fourth terms in Eq. 14,

$$\begin{aligned}\mathbf{F_Z}^{\mathbf{k}} =& \mathbf{A^T}((\mu_1 + \mu_2)\mathbf{A^{k+1}Z^k} - \mu_1(\mathbf{X} - \mathbf{E}) \\ &- \mathbf{Y_1} - (\mu_2 \hat{\mathbf{D}} + \mathbf{Y_2})_{(\mathbf{1:d_x},\cdot)}) \\ &+ \mu_3(\mathbf{Z^k} - \mathbf{J^k}) + \mathbf{Y_3}\end{aligned} \tag{16}$$

and $\mathbf{Q^{k+1}}$ is a shrinkage operator, defined as

$$\mathbf{Q^{k+1}} = \mathcal{S}_{\lambda_2/\mu_3}(\mathbf{Z^{k+1}} + \mathbf{Y_3}/\mu_3) \tag{17}$$

Finally, the Lagrange multiplier matrices $\mathbf{Y_1}, \mathbf{Y_2}, \mathbf{Y_3}$ and regularization terms μ_1, μ_2, μ_3 are updated based on LADM,

$$\begin{aligned}\mathbf{Y_1^{k+1}} &= \mathbf{Y_1^k} + \mu_1^{k+1}((\mathbf{X}) - \mathbf{A^{k+1}Z^{k+1}} - \mathbf{E^{k+1}}) \\ \mathbf{Y_2^{k+1}} &= \mathbf{Y_2^k} + \mu_2^{k+1}((\mathbf{D}) - [\mathbf{A^{k+1}Z^{k+1}}; \mathbf{1^T}]) \\ \mathbf{Y_3^{k+1}} &= \mathbf{Y_3^k} + \mu_3^{k+1}((\mathbf{Z})^{\mathbf{k+1}} - \mathbf{J^{k+1}}) \\ \mu_1^{k+1} &= \min(\mu_{max}, \rho\mu_1^k) \\ \mu_2^{k+1} &= \min(\mu_{max}, \rho\mu_2^k) \\ \mu_3^{k+1} &= \min(\mu_{max}, \rho\mu_3^k)\end{aligned} \tag{18}$$

where ρ is a positive scalar.

4. Convergence Analysis and Computation Complexity Analysis

4.1. Convergence Analysis

Because LRS-RR-1 and LRS-RR-2 are ordinary least square problems, we mainly introduce two theorems regarding the convergence of the augmented Lagrangian multiplier algorithms for subproblems LRS-RR-3 and LRS-RR-4.

Subproblem LRS-RR-3 is similar to the transposed standard LR-RR model, thus the convergence analysis in [27] can be applied to this model. We present a convergence theorem below.

Theorem 1. *If $\{\mu_1\}, \{\mu_2\}$ are non-decreasing and upper bounded, $\tau_A > \rho(\mathbf{ZZ}^T)$, then the sequence $\left(\mathbf{A}^k, \mathbf{E}^k, \mathbf{Y}_1^k, \mathbf{Y}_2^k\right)$ generated by* (12)–(13) *converges to a KKT point of LRS-RR-3.*

For the LRS-RR-4 model (15), there are two blocks of primary variables. For the cases of less than three blocks of primary variables, a naive linearized version of ADM tends to converge. A slight difference is that the variable $\mathbf{Z}$ is non-negative by constraining $\mathbf{Z} = \mathbf{J}, \mathbf{J} \geq \mathbf{0}$. Following the convergence analysis in [27, 26], we immediately have the following theorem

Theorem 2. *If $\{\mu_1\}, \{\mu_2\}, \{\mu_3\}$ are non-decreasing and upper bounded, $\tau_Z > \rho(\mathbf{A}^T\mathbf{A})$, then the sequence $\left(\mathbf{Z}^k, \mathbf{J}^k, \mathbf{Y}_3^k\right)$ generated by* (15)–(17) *converges to a KKT point of LRS-RR-4.*

Please refer to the proof in [26]. Finally we have the following convergence for LRS-RR-1–LRS-RR-4.

4.2. Computation Complexity Analysis

In the iterations of LRS-RR-1 – LRS-RR-4, the computational costs are mainly matrix inversion and SVD. For the data matrix $\mathbf{X} \in \mathbb{R}^{d_x \times n}$, the computation complexity of full SVD is $\mathcal{O}(d_x n^2)(d_x > n)$. Each iteration of LRS-RR-3 mainly includes SVD — $\mathcal{O}(d_x n^2)$ and matrix multiplication of $\mathbf{A}^{k+1}\mathbf{Z}$ — $\mathcal{O}(d_x n^2)$. Then the whole computational complexity for LRS-RR-3 is $\mathcal{O}(t_1 * 2d_x n^2)$ and t_1 is the iteration number of LRS-RR-3. Similar to LRS-RR-3, since rank($\mathbf{Z}$) <= rank($\mathbf{X}$) the whole computation complexity for LRS-RR-4 is $\mathcal{O}(t_1 * 2d_x n^2)$ and t_1 is the iteration number. LRS-RR-1 and LRS-RR-2 contains iterations of matrix inverse of $\hat{\mathbf{D}}(\hat{\mathbf{D}})^T + \gamma\mathbf{I}_{\mathbf{d_x+1}}$ and $\eta\mathbf{T}^T\mathbf{W}^T\mathbf{WT} + \mu_2\mathbf{I}_{d_x}$, whose complexity is $\mathcal{O}((d_x + 1)^3)$, and $\mathcal{O}({d_x}^3)$, respectively. Hence, the total complexity for LRS-RR is $\mathcal{O}(T * (t_1 * 2d_x n^2 + t_1 * ({d_x}^3)))$.

5. Experiments

In this section, we evaluate the performance of our proposed algorithms on both synthetic data and the data from

Method	$RAE_{\mathbf{T}}$	$RAE_{\mathbf{Y}}$
LSR	0.269 ± 0.121	0.035 ± 0.012
RANSAC	0.256 ± 0.133	0.036 ± 0.013
RPCA+LSR	0.464 ± 0.030	0.051 ± 0.006
LR-RR	0.035 ± 0.015	0.015 ± 0.006
LRS-RR	$\mathbf{0.005 \pm 0.0005}$	$\mathbf{0.011 \pm 0.003}$

Table 1. RAE and its standard deviation on synthetic data (10 repetitions).

real vision tasks. All the experiments were done on a PC with the same hardware —i5 CPU(2.57 GHz), 8GB RAM and operating system —WIN-10.

We compare our LRS-RR method against the state-of-the-art approaches in four experiments for regression and classification. The approaches include: (1) standard LSR; (2) RANSAC [35]; (3) RPCA+LSR, which firstly performs RPCA [7] on the input data and then learns the regression model on the cleaned data using standard LSR; (4) LR-RR.

In the first experiment, synthetic data lying in the disjoint subspaces were used to validate and compare the our method with the popular approaches, as well as to illustrate the convergence of our method. In the second experiment, we apply LRS-RR to the problem of head pose estimation from partially corrupted images, and we also compare the CPU computing time. The third experiment illustrates the application of LRS-RR to reconstruction of corrupted faces.

5.1. Synthetic Data for Accuracy and Convergence Validation

This section illustrates the benefits of RR in a synthetic example. We generate 400 three-dimensional samples, 200 samples for one subspace and 200 for the other. The first two components are generated from a uniform distribution between [-6; 6]. The third dimension of the first subspace is generated by $\mathbf{z} = \mathbf{x} + \mathbf{y}$, and the other is generated by $\mathbf{z} = \mathbf{x} - \mathbf{y}$, as two joint subspaces. We compare our RR with five methods: (1)LSR; (2) RANSAC; (3)RPCA+LSR, and (4) LR-RR. We randomly select 200 samples for training and used the remaining 200 data points for testing. Both the training and testing sets contain half of the corrupted samples. We compute the Relative Absolute Error (RAE) between true regression matrix $\mathbf{T}^*$ and learned $\mathbf{T}$: $RAE_{\mathbf{T}} = \|\mathbf{T} - \mathbf{T}^*\|_F / \|\mathbf{T}\|_F$, and the RAE between true output $\mathbf{Y}^*$ and learned $\mathbf{Y}$: $RAE_{\mathbf{Y}} = \|\mathbf{Y} - \mathbf{Y}^*\|_Y / \|\mathbf{F}\|_F$, the result is shown in Tab. 5.1. It can be seen that our method learns the most exact model, with the smallest prediction errors among all the methods. Fig. 2 gives the plots of relative Frobenius norm errors of learned dictionary $\mathbf{D}$ and separated noise $\mathbf{E}$ varying with iteration number. It indicates that our methods can converge quickly.

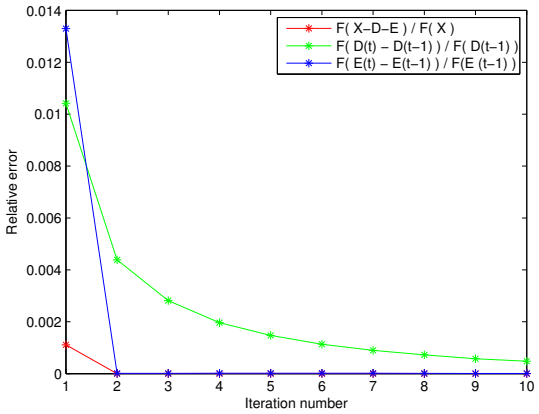


Figure 2. Convergence result of LRS-RR on Synthetic Data.

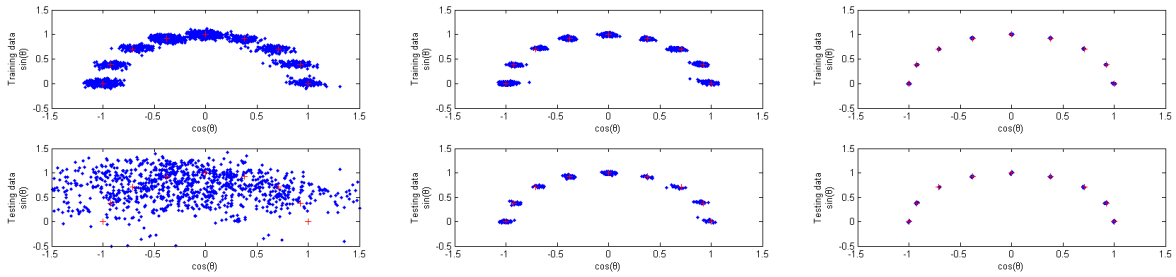
Method	Pose Angle Err	Time(s)
LSR	$27.56^o \pm 23.60^o$	0.05
RANSAC	$23.20^o \pm 20.39^o$	0.22
RPCA+LSR	$20.45^o \pm 19.51^o$	0.25
LR-RR	$1.97^o \pm 5.77^o$	3.03
LRS-RR	$\mathbf{1.03^o \pm 5.65^o}$	10.02

Table 2. Comparison of yaw angle error and standard deviation on a subset of CMU PIE.

5.2. CMU PIE Database for Pose Estimation

This section demonstrates the performance of LRS-RR in the problem of head pose estimation. A subset of the CMU PIE database[33] is used, which contains over 5000 face images from 53 subjects. The face regions are labeled carefully by hand. These faces cover 9 head poses(from -90^o to $+90^o$, step 22.5^o), each with a random lighting direction. Each image is cropped around the face region and resized to 48×48. We reshape each image into a vector in the matrix $\mathbf{X}$ and the yaw angles of the images are used as the output data $\mathbf{Y} = [\cos(\theta); \sin(\theta)]$.

Similar to the previous section, we have compared LRS-RR with above methods: (1) RANSAC, (2) RPCA+LSR, (3) RPCA+LSR, (4)LR-RR. The 53 subjects were randomly divided into 5 folds for cross-validation and selection of best parameters of these methods for fairness. Tab. 5.2 gives the the averaged angle error of different methods and also shows the time cost for testing of different methods. Though our method achieves least errors, the computing process is still needs to be sped up. A visual result of pose predilection is also illustrated in Fig. 3, indicating that our method achieves more narrow prediction than popular low-rank methods (which is not listed due to the paper's layout).



(a) RPCA+LSR (b) LR-RR (c) LRS-RR

Figure 3. Pose projection in the output space $[cos(\theta), sin(\theta)]$. The red '+' denotes the ground truth.

5.3. YaleB Database for Reconstruction of Corrupted Faces

The YaleB database [14] contains over 2,300 frontal face images of 38 subjects under different illumination changes.There are 64 near-frontal images taken for each subject. In this experiment, we use the cropped face images (196 $\times$ 128 pixels) of the first 15 subjects. First, we calculate 20-dimensional eigenfaces using training images as the input matrix $\mathbf{X}$. Then for each tested face image, 10 blocks(30 $\times$ 30 pixels) are randomly selected as synthetic corruption (set to 255, see the first column on the Fig. 4 for reference). To evaluate the reconstruction accuracy, we first calculate the real regression model with un-blocked tested images by eigenfaces, and then we relearn the regression models of different robust methods by taking eigenfaces as the input matrix $\mathbf{X}$ and blocked tested images as responses $\mathbf{Y}$. Finally we calculate the output errors between learned models and the real model, as well as the errors between prediction of the real model response (Goal) and learned models. As shown in Tab. 5.3, we compare face reconstruction accuracy of RANSAC, RPCA+LSR, LR-RR and our LSR-RR. Some examples of face reconstruction by different methods are also given in Fig. 4, which shows that our method gets the most similar visual result to the prediction goal.

Figure 4. Reconstruction of corrupted faces on YaleB.

Method	Model Err	Fitting Err
RANSAC	1.058 ± 0.040	0.185 ± 0.007
RPCA+LSR	1.075 ± 0.051	0.187 ± 0.007
LR-RR	1.069 ± 0.044	0.185 ± 0.006
LRS-RR	$\mathbf{1.045 \pm 0.049}$	$\mathbf{0.164 \pm 0.006}$

Table 3. Face Reconstruction error and standard deviation on YaleB under synthetic corruption.

6. Conclusions and Future Work

This paper addresses the problem of supervised low-rank-spare subspace representation for robust regression in high-dimensional data and presents a LADMAP solution for LRS-RR. Our method jointly learns a regression model while removing the outliers/noise that are little correlated with the regression responses. Compared to previous robust regressions under a low-rank subspace constraint, our method can deal with outliers/noise inside or outside the disjoint subspaces, and can obtain much more exact regression model in this complex situation. We illustrated the benefits of LRS-RR in several computer vision problems including head pose estimation and face reconstruction. We showed that by filtering outliers/noise via a reasonably supervised low-rank-spare subspace learning, our method can recover the clean data better and outperforms state-of-the-art approaches in both the random and low-rank ones. The framework of LRS-RR is useful to solve high-dimensional problems in many real-life applications. Moreover, our approach is extensible and can easily be integrated into other regression methods, such as cascaded regression for face alignment and tracking. However, the current LRS-RR is time-consuming, the optimization algorithm will be parallelized in our future work.

References

[1] *Partial Least-Squares Regression*. Springer New York, 2008.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[3] H. Barreto and D. Maharry. Least median of squares and regression through the origin. *Comput. Stat. Data Anal.*, 50(6):1391–1397, 2006.

[4] F. Bunea, Y. She, and M. H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *Statistics*, 39(2):1282–1309, 2010.

[5] J. F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[6] 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*, pages 1124–1132, 2013.

[7] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

[8] S. Choi, T. Kim, and W. Yu. Performance evaluation of ransac family. In *BMVC*, 2009.

[9] C. Croux and C. Dehon. Robust linear discriminant analysis using s-estimators. *Canadian Journal of Statistics*, 29(3):473–494, 2010.

[10] B. Efron and R. Tibshirani. Least angle regression. *Mathematics*, 32(2):2004, 2004.

[11] J. Gao. Robust L1 principal component analysis and its Bayesian variational inference. *Neural Computation*, 20(2):555–572, 2008.

[12] J. Gao, P. W. Kwan, and D. Shi. Sparse kernel learning with lasso and bayesian inference algorithm. *Neural Networks*, 23(2):257–264, 2010.

[13] J. B. Gao, S. R. Gunn, C. J. Harris, and M. Brown. Regression with input-dependent noise: a relevance vector machine treatment. *IEEE Trans. Neural Netw. Learn. Syst.*, 2001.

[14] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):643–660, 2001.

[15] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[16] X. He and W. K. Fung. High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis*, 72(2):151–162, 2000.

[17] X. Hong, S. Chen, J. Gao, and C. J. Harris. Nonlinear identification using orthogonal forward regression with nested optimal regularization. *Cybernetics IEEE Transactions on*, 45(12):2925–2936, 2015.

[18] X. Hong, S. Chen, Y. Guo, and J. Gao. l1-norm penalized orthogonal forward regression. *Computer Science*, 2015.

[19] D. Huang, R. Cabral, and F. D. l. Torre. Robust regression. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):363–375, 2016.

[20] F. Khelifi and J. Jiang. KNN regression to improve statistical feature extraction for texture retrieval. *IEEE Transactions on Image Processing*, 20(1):293–8, 2011.

[21] E. Kim, M. Lee, C.-H. Choi, N. Kwak, and S. Oh. Efficient-norm-based low-rank matrix approximations for large-scale problems using alternating rectified gradient method. *IEEE Trans. Neural Netw. Learn. Syst.*, 26(2):237–251, 2015.

[22] S. J. Kim, A. Magnani, and S. P. Boyd. Robust fisher discriminant analysis. *Advances in Neural Information Processing Systems*, pages 659–666, 2005.

[23] P. Lang, A. Gironella, and R. Venema. Properties of cyclic subspace regression. *Journal of Multivariate Analysis*, 98(3):625–637, 2007.

[24] Z. Lin, M. Chen, and Y. Ma. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC technical report UIULC-ENG-09-2215*, 2010.

[25] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. *Computational Advances in Multi-Sensor Adaptive Processing*, 61, 2009.

[26] Z. Lin, R. Liu, and H. Li. Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning. *Machine Learning*, 99(2):287–325, 2015.

[27] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *NIPS*, pages 612–620, 2011.

[28] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, pages 663–670, 2010.

[29] McDonald and C. Gary. Ridge regression. *Wiley Interdisciplinary Reviews Computational Statistics*, 1(1):93–100, 2009.

[30] P. Meer. Robust techniques for computer vision. *Imsc Press Multimedia*, 2004.

[31] R. H. Riffenburgh and C. W. Clunies-Ross. *Linear Discriminant Analysis*. Springer New York, 2013.

[32] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. Wiley-Interscience,, 2003.

[33] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(12):1615 – 1618, 2010.

[34] L. M. Surhone, M. T. Timpledon, S. F. Marseken, C. Correlation, T. S. O. Squares, and R. Analysis. Principal component regression. *Betascript Publishing*, pages 1954–1954, 2010.

[35] P. H. S. Torr and A. Zisserman. Mlesac: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1):138–156, 2000.

[36] J. Wang, D. Shi, D. Cheng, Y. Zhang, and J. Gao. LRSR: Low-rank-sparse representation for subspace clustering. *Neurocomputing*, 2016.

[37] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang. Discriminative least squares regression for multiclass classification and feature selection. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(11):1738–1754, 2012.

[38] S. Xiang, Y. Zhu, X. Shen, and J. Ye. Optimal exact least squares rank minimization. In *ACM SIGKDD*, pages 480–488, 2012.

[39] Q. Zhang, X. Hu, and B. Zhang. Comparison of -norm svr and sparse coding algorithms for linear regression. *IEEE Trans. Neural Netw. Learn. Syst.*, 26(8):1828–1833, 2014.

[40] X. Y. Zhang, L. Wang, S. Xiang, and C. L. Liu. Retargeted least squares regression algorithm. *IEEE Trans. Neural Netw. Learn. Syst.*, 26(9):1, 2015.

[41] Y. Zhang and D. Y. Yeung. Worst-case linear discriminant analysis. In *NIPS*, pages 2568–2576, 2010.

[42] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu. Non-negative low rank and sparse graph for semi-supervised learning. In *CVPR*, pages 2328–2335. IEEE, 2012.