On The Equivalent of Low-Rank Regressions and Linear Discriminant Analysis Based Regressions

Xiao Cai
Dept. of Comp. Sci. & Eng.
University of Texas at Arlington
Arlington, Texas, 76092
xiao.cai@mavs.uta.edu

Chris Ding
Dept. of Comp. Sci. & Eng.
University of Texas at Arlington
Arlington, Texas, 76092
chqding@uta.edu

Feiping Nie
Dept. of Comp. Sci. & Eng.
University of Texas at Arlington
Arlington, Texas, 76092
feipingnie@gmail.com

Heng Huang
Dept. of Comp. Sci. & Eng.
University of Texas at Arlington
Arlington, Texas, 76092
heng@uta.edu

ABSTRACT

The low-rank regression model has been studied and applied to capture the underlying classes/tasks correlation patterns, such that the regression/classification results can be enhanced. In this paper, we will prove that the low-rank regression model is equivalent to doing linear regression in the linear discriminant analysis (LDA) subspace. Our new theory reveals the learning mechanism of low-rank regression, and shows that the low-rank structures exacted from classes/tasks are connected to the LDA projection results. Thus, the low-rank regression efficiently works for the high-dimensional data.

Moreover, we will propose new discriminant low-rank ridge regression and sparse low-rank regression methods. Both of them are equivalent to doing regularized regression in the regularized LDA subspace. These new regularized objectives provide better data mining results than existing low-rank regression in both theoretical and empirical validations. We evaluate our discriminant low-rank regression methods by six benchmark datasets. In all empirical results, our discriminant low-rank models consistently show better results than the corresponding full-rank methods.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.2.8 [**Database Management**]: Database Applications-Data Mining

General Terms

Algorithms

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

Keywords

Low-Rank Regression, Low-Rank Ridge Regression, Sparse Low-Rank Regression, Linear Discriminant Analysis

1. INTRODUCTION

As one of most important data mining and machine learning technique, multivariate linear regression attempts to model the relationship between predictors and responses by fitting a linear equation to observed data. Such linear regression models suffer from two deficiencies when they are applied to the real-world applications. First, the linear regression models usually have low performance for analyzing the high-dimensional data. In many data mining and machine learning applications, such as gene expression, document classification, face recognition, the input data have a large number of features. To perform accurate regression or classification tasks on such data, we have to collect an enormous number of samples. However, due to the data and label collection difficulty, we often cannot obtain enough samples and suffer from the curse-of-dimensionality problem [8]. To solve this problem, the dimensionality reduction methods, such as linear discriminant analysis (LDA) [10], were often used to reduce the feature dimensionality first.

Second, the linear regression models don't emphasize the correlations among different responses. Standard least squares regression is equivalent to regressing each response on the predictors separately. To incorporate the response (*i.e.* classes or tasks) correlations into the regression model, Anderson introduced the reduced rank regression method [1], which is a multivariate regression model with a coefficient matrix with reduced rank. Later many researchers worked on the low-rank (or reduced) regression models [26, 5, 13, 1, 2, 20], in which the classes/tasks correlation patterns are explored by the low-rank structure and utilized to enhance the regression/classification results.

In this paper, we propose new and important theoretical foundations of the low-rank regression. We first present the discriminant low-rank linear regression, which reformulates the standard low-rank regression to a more interpretable objective. After that, we prove that the low-rank regression model is indeed equivalent to doing linear regression in the LDA subspace, *i.e.* the learned low-rank classes/tasks

correlation patterns are connected to the LDA projection results. Our new theorem explains the underlying computational mechanism of low-rank regression, which performs the LDA projection and the linear regression on data points simultaneously. In our special case, when the low-rank regression coefficient matrix becomes a full-rank matrix, our result is connected to Ye's work on the equivalence between the multivariate linear regression and LDA [27].

Motivated by our new theoretical analysis, we propose two new discriminant low-rank regression models, including low-rank ridge regression (LRRR) and sparse low-rank regression (SLRR). Both methods are equivalent to performing the regularized regression tasks in the regularized LDA subspace (two methods have different regularization terms). Because the regularization term avoids the rank deficiency problem in both regression and LDA, our LRRR method outperforms the low-rank regression in both theoretical analysis and experimental results. Using the structured sparsity-inducing norm based regularization term, our SLRR method can explore both classes/tasks correlations and feature structures. All our new discriminant low-rank regression models can simultaneously analyze the high-dimensional data in the discriminant subspace without any pre-processing step and incorporate the classes/tasks correlations. We evaluate the proposed methods on six benchmark data sets. In all experimental results, our discriminant low-rank models consistently outperform their corresponding full-rank counterparts.

Notations. In this paper, matrices are written as uppercase letters and vectors are written as bold lowercase letters. For matrix $W = \{w_{ij}\}$, its i-th row, j-th column are denoted as $\mathbf{w}^i$, $\mathbf{w}_j$ respectively. Tr (W) means the trace operation for matrix W and $||W||_*$ means the trace norm of matrix W.

2. LOW-RANK REGRESSION AND LDA+LR

One of the main result of this paper is to prove that the **low-rank linear regression (LRLR)** is equivalent to doing **standard linear regression in LDA subspace (we call this as "LDA+LR")**.

2.1 Low-Rank Linear Regression Revisit

Traditional Linear Regression model for classification is to solve the following problem:

$$\min_W ||Y - X^T W||_F^2, \tag{1}$$

where $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \Re^{d\times n}$ is the centered training data matrix and $Y \in \Re^{n\times k}$ is the normalized class indicator matrix, i.e. $Y_{i,j} = 1/\sqrt{n_j}$ if the i-th data point belongs to the j-th class and $Y_{i,j} = 0$ otherwise and n_j is the sample size of the j-th class. The model outputs the parameter matrix $W \in \Re^{d\times k}$, which can be used to predict any test data point $\mathbf{x} \in \Re^{d\times 1}$ by $W^T\mathbf{x}$.

When the class or task number is large, there are often underlying correlation structures between classes or tasks. To explore these hidden structures and utilize such patterns to improve the learning model, in recent work [3], researchers presented to learn a low-rank projection W in the regression model by imposing the trace norm regularization as:

$$\min_W ||Y - X^T W||_F^2 + \lambda||W||_*. \tag{2}$$

The trace norm regularization can discover the low-rank structures existing between classes or tasks. Using Eq. (2), the rank of coefficient matrix W, which is decided by the selection of parameter λ, cannot be explicitly selected and tuned.

In related research work, the low-rank regression was studied in statistics and machine learning communities [26, 5, 13, 1, 2, 20]. In the low-rank regression, the rank of W is explicitly decided by constraining the rank of W to be $s < min(n, k)$ and solving the following problem:

$$\min_W ||Y - X^T W||_F^2, \;\; s.t. \;\; rank(W) \leq s. \tag{3}$$

Because the rank of coefficient matrix can be explicitly determined, the low-rank regression in Eq. (3) is better than the trace norm based objective in Eq. (2) in practical applications. Although the general rank minimization is a non-convex and NP-hard problem, the objectives with rank constraints are solvable, *e.g.* the global solution was given in [26, 5].

2.2 Relation to LDA+LR

In this section, we will show that the low-rank linear regression (LRLR) is equivalent to perform Linear Discriminant Analysis (LDA) and linear regression simultaneously (LDA+LR). In other words, the learned low-rank structures and patterns are induced by the LDA projection (with regression). The low rank s is indeed the projection dimension of LDA.

Before introducing our main theorems, we first propose the following discriminant Low-Rank Linear Regression formulation (LRLR):

$$\min_{A,B} ||Y - X^T AB||_F^2, \tag{4}$$

where $A \in \Re^{d\times s}$, $B \in \Re^{s\times k}$, $s < min(n, k)$. Thus $W = AB$ has low-rank s. The above LRLR objective has the same solutions as Eq. (3), but it has clearer discriminant projection interpretation. Eq. (4) can be written as

$$\min_{A,B} ||Y - (A^T X)^T B||_F^2. \tag{5}$$

This shows A can be viewed as a projection. Interestingly as we show in Theorem 1, A is exactly the optimal subspace defined by the classic LDA.

THEOREM 1. *The low-rank linear regression method of Eq. (4)) is identical to doing standard linear regression in LDA subspace.*

Proof: Denoting $J_1(A, B) = ||Y - X^T AB||_F^2$ and taking its derivative w.r.t. B, we have,

$$\frac{\partial J_1(A,B)}{\partial B} = -2A^T XY + 2A^T XX^T AB. \tag{6}$$

Setting Eq. (6) to zero, we obtain,

$$B = (A^T XX^T A)^{-1} A^T XY. \tag{7}$$

Substituting Eq. (7) back into Eq. (4), we have,

$$\min_A ||Y - X^T A(A^T XX^T A)^{-1} A^T XY||_F^2, \tag{8}$$

which is equivalent to

$$\max_A \text{Tr}\ ((A^T(XX^T)A)^{-1} A^T XYY^T X^T A). \tag{9}$$

Note that

$$S_t = XX^T, \quad S_b = XYY^T X^T, \tag{10}$$

where S_t and S_b are the total-class scatter matrix and the between-class scatter matrix defined in the LDA, respectively. Therefore, the solution of Eq. (9) can be written as:

$$A^* = \arg\max_A \text{ Tr } [(A^T S_t A)^{-1} A^T S_b A], \tag{11}$$

which is exactly the problem of LDA, and the global optimal solution to Eq. (11) is the top s eigenvectors of $S_t^{-1} S_b$ corresponding to the nonzero eigenvalues (if S_t is singular, we compute the eigenvectors of $S_t^+ S_b$ corresponding to the nonzero eigenvalues, where S_t^+ denotes the pseudo-inverse of S_t). Now Eq. (5) implies that we do linear regression on the projected data $\widetilde{X} = A^T B$. Since A is the LDA projection, thus Eq. (5) implies we do regression on the LDA subspace. □

Note that in Eq. (4), the class indicator matrix Y is normalized, but not centered. However X is centered. The following Theorem 2 shows that we obtain the optimal solution whatever Y is centered or not.

THEOREM 2. *The optimal solution (A^*, B^*) for the following problem*

$$\min_{A,B} ||PY - X^T AB||_F^2 \tag{12}$$

is identical to those of Eq. (4); here $P = I - \boldsymbol{e}\boldsymbol{e}^T/n \in \Re^{n\times n}$ is the centering matrix, and $\boldsymbol{e} = (1 \cdots 1)^T$.

For this reason, the bias (intercept) term are already automatically incorporated in Eq. (4).

Proof: The key point of the proof is the fact that in the solution for both B and A of Eq. (7) and Eq. (9), Y always appears together with X as combination

$$XY = (XP)Y = XP^2Y = (XP)(PY),$$

because X is centered and $P^2 = P$. In other words, as long as X is centered, Y is automatically centered. □

This results can be easily extended to the standard linear regression. In fact we have

Remark 1. *As long as X is centered, the optimal solution W^* for the standard linear regression of Eq.(1) remains identical no matter Y is centered or not.*

Our new results provide the theoretical foundation to explain the mechanism behind the low-rank regression methods. Meanwhile, the above proof process also indicates a concise algorithm to achieve the global solution of LRLR in Eq. (4), as well as Eq. (3). The Algorithm to solve Eq. (4) is summarized in Alg. 1.

Moreover, we note that Theorem 1 also provides clarification to a long-standing puzzle in multi-class LDA, as explained below.

2.3 LDA: Trace-of-Ratio or Ratio-of-Trace?

The original Fisher LDA is on 2-class problem, where only $k - 1 = 1$ projection direction $\mathbf{a}$ is needed. The Fisher objective is

$$\max_{\mathbf{a}} \frac{\mathbf{a}^T S_b \mathbf{a}}{\mathbf{a}^T S_w \mathbf{a}}.$$

The generalization to multi-class has two natural formulations [10], either the trace-of-ratio formulation

$$\max_A \text{ Tr } \frac{A^T S_b A}{A^T S_w A} \tag{13}$$

Algorithm 1 The algorithm to solve LRLR

Input:
1. The centralized training data $X \in \Re^{d\times n}$.
2. The normalized training indicator matrix $Y \in \Re^{n\times k}$.
3. The low-rank parameter s.

Output:
1. The matrices $A \in \Re^{d\times s}$ and $B \in \Re^{s\times k}$.

Process:
1. Calculate A by Eq. (11)
2. Calculate B by Eq. (7)

where $A = (\mathbf{a}_1 \cdots \mathbf{a}_{k-1})$, or the ratio-of-trace formulation[1]

$$\max_A \frac{\text{Tr } A^T S_b A}{\text{Tr } A^T S_w A} \tag{14}$$

Our Theorem 1 lends support to the trace-of-ratio objective function because this formulation arises directly from the linear regression.

2.4 Full-Rank Linear Regression and LDA

Here we note an important connection. In the special case, the low-rank regression coefficient matrix W becomes a full-rank matrix. Without loss of generality we assume $s = k \leq n$, because the number of data points n is usually larger than the number of classes k. The matrix $B \in \Re^{k\times k}$ becomes a square matrix. Because $rank(W) = rank(AB) = k$ and $k \leq n$, $rank(A) \geq k$ and $rank(B) \geq k$. Thus, $rank(B) = k$ and B is a full rank matrix, *i.e.* the matrix B is invertible.

The Theorem 1 is still correct for the special case. Moreover, we can further conclude the equivalence between the multivariate linear regression and LDA results. We can simply prove this conclusion. Because the matrix A includes the LDA subspaces and the matrix B can be considered as an invertible rotational matrix, thus AB is also one of the infinite number global solutions of LDA [15]. Thus, in the special full-rank case, the multivariate linear regression is equivalent to the LDA result, which was shown in Ye's work [27] with the assumptions: the reduced dimension is $k - 1$ and $rank(S_b) + rank(S_w) = rank(S_t)$. Our proof is more general and doesn't need the rank assumption.

2.5 Low-Rank Ridge Regression (LRRR)

As we know, by adding a Frobenius norm based regularization on the linear regression loss, ridge regression can achieve better performance than linear regression [12]. Thus, it is important and necessary to add the ridge regularization into low-rank regression formulation. We propose the following Low-Rank Ridge Regression (LRRR) objective as,

$$\min_{A,B} ||Y - X^T AB||_F^2 + \lambda ||AB||_F^2, \tag{15}$$

where $A \in \Re^{d\times s}$, $B \in \Re^{s\times k}$, $s < min(n, k)$, λ is the regularization parameter. Similarly, we can see that the LRRR objective is equivalent to the following objective:

$$\min_W ||Y - X^T W||_F^2 + \lambda ||W||_F^2, \quad s.t. \quad rank(W) \leq s. \tag{16}$$

Compared to Eq. (16), Eq. (15) provides better chance for us to understand the learning mechanism of LRRR. We will

[1]In Eqs.(13,14), the optimal solution remains the same when S_w is replaced by S_t.

Algorithm 2 The algorithm to LRRR

Input:
1. The centralized training data $X \in \Re^{d\times n}$.
2. The normalized training indicator matrix $Y \in \Re^{n\times k}$.
3. The low-rank parameter s.
4. The regularization parameter λ.

Output:
1. The matrices $A \in \Re^{d\times s}$ and $B \in \Re^{s\times k}$.

Process:
1. Calculate A by Eq. (21)
2. Calculate B by Eq. (18)

show that our new LRRR objective is connected to the regularized discriminant analysis, which provides better projection results than the standard LDA. We will also derive the global solution of the non-convex problems in Eq. (15) and Eq. (16).

THEOREM 3. *The proposed Low-Rank Ridge Regression (LRRR) method (both Eq. (15) and Eq. (16)) is equivalent to doing the regularized regression in the regularized LDA subspace.*

Proof: Denoting $J_2(A,B) = ||Y - X^TAB||_F^2 + \lambda||AB||_F^2$, and taking its derivative w.r.t. B, we have,

$$\frac{\partial J_2(A,B)}{\partial B} = -2A^TXY + 2A^TXX^TAB + 2\lambda A^TAB. \quad (17)$$

Setting Eq. (17) to zero, we get,

$$B = (A^T(XX^T + \lambda I)A)^{-1}A^TXY, \quad (18)$$

where $I \in \Re^{d\times d}$ is the identity matrix. Substituting Eq. (18) back into Eq. (15), we have

$$\min_A ||Y - X^TA(A^TXX^TA + \lambda A^TA)^{-1}A^TXY||_F^2 + \lambda||A(A^T(XX^T + \lambda I)A)^{-1}A^TXY||_F^2, \quad (19)$$

which is equivalent to the following problem:

$$\max_A \{(A^T(XX^T + \lambda I)A)^{-1}A^TXYY^TX^TA\}. \quad (20)$$

Similarly, the solution of Eq. (20) can be written as:

$$A^* = \arg\max_A \{Tr((A^T(S_t + \lambda I)A)^{-1}A^TS_bA)\}, \quad (21)$$

which is exactly the problem in regularized LDA [9]. After we get the optimal solution A, we can re-write Eq. (15) as:

$$\min_B ||Y - (A^TX)^TB||_F^2 + \lambda||AB||_F^2, \quad (22)$$

which is the regularized regression, and the optimal solution is given by Eq. (18). Thus, the LRRR of Eq. (15) is equivalent to performing ridge regression in regularized-LDA subspace. □

Similar to Theorem 2, we can show that Y is automatically centered as long as X is centered.

Another interest point is that although our LRRR model is a non-convex problem, Theorems 1 and 3 show that they have the global optimal solutions. The Algorithm to solve LRRR of Eq. (15) is described in Alg. 2.

2.6 Full-Rank Ridge Regression and Regularized LDA

In the special case, the low-rank regression coefficient matrix W becomes a full-rank matrix. Similar to §2.4, we have the following lemma:

LEMMA 1. *The full-rank ridge regression result is equivalent to the solution of regularized LDA (S_t is replaced by the regularized form $S_t + \lambda I$).*

Similar to the proof in §2.4, we can easily prove the coefficient matrix W in full-rank ridge regression is one of the global solutions of LDA regularized by λI.

3. SPARSE LOW-RANK REGRESSION FOR FEATURE SELECTION

Besides exploring and utilizing the class/task correlations and structure information, the learning models also prefer to select and use the important features to avoid the "curse of dimensionality" problem in high-dimensional data analysis. Thus, it is important to extend our discriminant low-rank regression formulations to feature selection models.

Due to the intrinsic properties of real world data, the structured sparse learning models have shown superior feature selection results in previous research [22, 28, 21, 17, 6, 23, 25, 24, 7]. One of the most effective ways for selecting features is to impose sparsity by inducing hybrid structured $\ell_{2,1}$-norm on the coefficient matrix W as the regularization term [19, 3]. Therefore, following our LRLR and LRRR methods, we propose a new **Sparse Low-Rank Regression (SLRR)** method, which reserves the low-rank constraint and adds the mixed $\ell_{2,1}$-norm regularization term to induce both desired low-rank structure of classes/tasks correlations and structured sparsity between features. To be specific, "low-rank" means $rank(AB) = s < min(n,k)$ and "structured sparsity" means most rows of AB are zero to help feature selection. Thus, we solve:

$$\min_{A,B} ||Y - X^TAB||_F^2 + \lambda||AB||_{2,1}\,, \quad (23)$$

where $A \in \Re^{d\times s}$, $B \in \Re^{s\times k}$, $s < min(n,k)$. Similarly, we can see that the above SLRR objective is equivalent to the following objective:

$$\min_W ||Y - X^TW||_F^2 + \lambda||W||_{2,1}, \quad s.t. \quad rank(W) \le s. \quad (24)$$

Both Eq. (23) and Eq. (24) are new objectives to simultaneously learn low-rank classes correlation patterns and features structured sparsity.

3.1 Connection to Discriminant Analysis

Interestingly our new SLRR method also connects to the regularized discriminant analysis by the following theorem.

THEOREM 4. *The optimal solution of the proposed SLRR method (Eq. (23) and Eq. (24)) has the same column space of a special regularized LDA.*

Proof: Eq. (23) is equivalent to the following problem,

$$\min_{A,B} ||Y - X^TAB||_F^2 + \lambda \mathrm{Tr}\,(B^TA^TDAB), \quad (25)$$

where $D \in \Re^{d\times d}$ is a diagonal matrix and each element on the diagonal is defined as follows:

$$d_{ii} = \frac{1}{2||\mathbf{g}^i||_2}, \quad i = 1, 2, ..., d, \quad (26)$$

Algorithm 3 The algorithm to SLRR

Input:
1. The centralized training data $X \in \Re^{d\times n}$.
2. The normalized training indicator matrix $Y \in \Re^{n\times k}$.
3. The low-rank parameter s.
4. The regularization parameter λ.

Output:
1. The matrices $A \in \Re^{d\times s}$ and $B \in \Re^{s\times k}$.

Initialization:
1. Set $t = 0$
2. Initialize $D^{(t)} = I \in \Re^{d\times d}$.

Repeat:
1. Calculate $A^{(t+1)}$ by Eq. (30)
2. Calculate $B^{(t+1)}$ by Eq. (28)
3. Update the diagonal matrix $D^{(t+1)} \in \Re^{d\times d}$, where the i-th diagonal element is $\frac{1}{2||(A^{(t+1)}B^{(t+1)})^i||_2}$.
4. Update $t = t + 1$

Until Converge.

where $\mathbf{g}^i$ is the i-th row of matrix $G = A^*B^*$. Denoting $J_3(A,B) = ||Y - X^TAB||_F^2 + \lambda \text{Tr}\ (B^TA^TDAB)$ and taking its derivative w.r.t. B, we have,

$$\frac{\partial J_3(A,B)}{\partial B} = -2A^TXY + 2A^TXX^TAB + 2\lambda A^TDAB. \quad (27)$$

Setting the above equation to be zero, we can get,

$$B = (A^T(XX^T + \lambda D)A)^{-1}A^TXY, \quad (28)$$

where $D \in \Re^{d\times d}$ is the diagonal matrix defined in Eq. (26). Substituting Eq. (28) back into Eq. (25), then we need solve the following problem to get A,

$$\max_A \text{Tr}\ ((A^T(XX^T + \lambda D)A)^{-1}A^TXYY^TX^TA). \quad (29)$$

The solution of Eq. (29) is:

$$A^* = \arg\max_A \{\text{Tr}\ ((A^T(S_t + \lambda D)A)^{-1}A^TS_bA)\}, \quad (30)$$

Since the column space of $W^* = A^*B^*$ is identical to the column space of A^*, the proposed SLRR has the same column space of a special regularized LDA (S_t is replaced with $S_t + \lambda D$). □

After we get the optimal solution A, we can solve Eq. (23) through Eq. (25), which is the regularized regression problem. Again, similar to Theorm 2, we can prove that if Y is centered or not will not affect the learnt model A^* and B^*.

3.2 Algorithm to Solve SLRR

Solving SLRR objective in Eq. (23) is nontrivial, there are two variables A and B needed to be optimized, and the non-smooth regularization also makes the problem more difficult to solve. Interestingly, a concise algorithm can be derived to solve this problem based on the above proof. The detailed algorithm is described in Algorithm 3. In next subsection, we will prove that the algorithm converges. Our experimental results show that the algorithm always converges in 5-20 iterations.

3.3 Algorithm Convergence Analysis

Because Alg. 3 is an iterative algorithm, we will prove its convergence.

THEOREM 5. *Alg. 3 decreases the objective function of Eq. (23) monotonically.*

Proof: In the t-th iteration, we have

$$< A^{(t+1)}, B^{(t+1)} >= \arg\min_{A,B} ||Y - X^TAB||_F^2 + \lambda \text{Tr}\ (B^TA^TD^{(t)}AB) \quad (31)$$

In other words,

$$||Y - X^TA^{(t+1)}B^{(t+1)}||_F^2 + \lambda \text{Tr}\ ({B^{(t+1)}}^T{A^{(t+1)}}^TD^{(t)}A^{(t+1)}B^{(t+1)}) \leq ||Y - XA^{(t)}B^{(t)}||_F^2 + \lambda \text{Tr}\ ({B^{(t)}}^T{A^{(t)}}^TD^{(t)}A^{(t)}B^{(t)}) \quad (32)$$

Denote $G^{(t)} = A^{(t)}B^{(t)}$ and $G^{(t+1)} = A^{(t+1)}B^{(t+1)}$. By the definition of matrix D in the algorithm, Eq. (32) can be rewritten as,

$$||Y - X^TG^{(t+1)}||_F^2 + \lambda \sum_{i=1}^{d} \frac{||\mathbf{g}^{i(t+1)}||_2^2}{2||\mathbf{g}^{i(t)}||_2} \leq ||Y - X^TG^{(t)}||_F^2 + \lambda \sum_{i=1}^{d} \frac{||\mathbf{g}^{i(t)}||_2^2}{2||\mathbf{g}^{i(t)}||_2} \quad (33)$$

where $\mathbf{g}^{i(t)}$ and $\mathbf{g}^{i(t+1)}$ are the i-th row of the matrix $G^{(t)}$ and $G^{(t+1)}$ respectively. Since for each i, we have

$$||\mathbf{g}^{i^{(t+1)}}||_2 - \frac{||\mathbf{g}^{i^{(t+1)}}||_2^2}{2||\mathbf{g}^{i^{(t)}}||_2} \leq ||\mathbf{g}^{i^{(t)}}||_2 - \frac{||\mathbf{g}^{i^{(t)}}||_2^2}{2||\mathbf{g}^{i^{(t)}}||_2}. \quad (34)$$

Thus, summing up d inequalities and multiplying the summation with the regularization parameter λ, we obtain:

$$\lambda \sum_{i=1}^{d} \left(||\mathbf{g}^{i^{(t+1)}}||_2 - \frac{||\mathbf{g}^{i^{(t+1)}}||_2^2}{2||\mathbf{g}^{i^{(t)}}||_2} \right) \leq \lambda \sum_{i=1}^{d} \left(||\mathbf{g}^{i^{(t)}}||_2 - \frac{||\mathbf{g}^{i^{(t)}}||_2^2}{2||\mathbf{g}^{i^{(t)}}||_2} \right) \quad (35)$$

Combining Eq. (33) and Eq. (35), we get:

$$||Y - X^TG^{(t+1)}||_F^2 + \lambda \sum_{i=1}^{d} ||\mathbf{g}^{i(t+1)}||_2 \leq ||Y - X^TG^{(t)}||_F^2 + \lambda \sum_{i=1}^{d} ||\mathbf{g}^{i(t)}||_2 \quad (36)$$

Therefore, we have:

$$||Y - X^TG^{(t+1)}||_F^2 + \lambda||G^{(t+1)}||_{2,1} \leq ||Y - XG^{(t)}||_F^2 + \lambda||G^{(t)}||_{2,1} \quad (37)$$

Since A and B are updated according to gradient, Alg. 3 will monotonically decrease the objective in Eq. (23) in each iteration. □

3.4 Full-Rank Sparse Linear Regression and Regularized LDA

In the special case, the low-rank regression coefficient matrix W becomes a full-rank matrix. Similar to §2.4, we also have the following lemma:

LEMMA 2. *The optimal solution of the full-rank sparse linear regression is one of the global solutions of LDA regularized by λD.*

Similar to the proof in §2.4, we can easily prove the coefficient matrix W in full-rank sparse linear regression is one of the global solutions of LDA regularized by λD.

4. EXPERIMENTAL RESULTS

In this section, we will evaluate the performance of our proposed LRLR, LRRR, SLRR with their corresponding full-rank counterparts. We firstly introduce the six benchmark datasets used in our experiments.

4.1 Dataset Descriptions

UMIST face dataset [11] contains 20 persons and totally 575 images. All images are cropped and resized into 112×92 pixels per image.

Binary Alphadigits 36 dataset [4] contains binary digits of 0 through 9 and capital A through Z with size 20×16. There are 39 examples of each class.

Binary Alphadigits 26 dataset [4] contains binary capital A through Z with size 20×16. There are 39 examples of each class.

VOWEL dataset [18] consists of 990 vowel recognition data used for the study of recognition of the eleven steady state vowels of British English. The speakers are indexed by integers 0-89. (Actually, there are fifteen individual speakers, each saying each vowel six times.) The vowels are indexed by integers 0-10. For each utterance, there are ten floating-point input values, with array indices 0-9.

MNIST hand-written digits dataset [14] consists of $60,000$ training and $10,000$ testing digits. It has 10 classes, from digit 0 to 9. Each image is centralized (according to the center of mass of the pixel intensities) on a 28×28 grid. We randomly select 15 images for each class in our experiment.

JApanese Female Facial Expressions (JAFFE) data set [16] contains 213 photos of 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects.

We summarize the datasets that we will use in our experiments in Table 1

4.2 Experimental Setup

All the datasets in our experiments have large number of classes (at least 10 classes). For each dataset, we randomly split the data into 5 parts. According to the standard 5-fold cross validation, in each round, we use 4 parts for training and the remaining part for testing. The average classification accuracy rates for different methods are reported. In the training stage, we use different full-rank linear regression models, *i.e.* full-rank linear regression, full-rank ridge regression, sparse full-rank linear regression to learn the coefficient matrix W directly or we solve the proposed low-rank counterparts (LRLR, LRRR, SLRR) to calculate W indirectly by $W = AB$. In all experiments, we automatically tune the regularization parameters by selecting the best parameters among the values $\{10^r : r \in \{-5, -4, -3, ...3, 4, 5\}\}$ with 5-fold cross validation on the corresponding training data only. In addition, for LRLR, LRRR, SLRR, we calculate the classification results with respect to different low-rank parameters s in the range of $[k/2, k)$, where k is the number of classes. At last, in the testing stage, we utilize the following decision function to classify the coming testing data $\mathbf{x}_t \in \Re^{d \times 1}$ into one and only one out of k classes,

$$\arg\max_{1 \le j \le k} \left(W^T \mathbf{x}_t\right)_j . \tag{38}$$

Please note that all the data are centered and we consider the model without bias. The code is written in MATLAB and we terminate our iterative optimization procedure of sparse regression when the relative change in the objective function is below 10^{-5}.

Table 1: The summary of the datasets used in our experiments. k is the number of classes, d is the number of feature dimensions, n is the number of data points.

Dataset	k	d	n
UMIST	20	10304	575
BINALPHA36	36	320	1404
BINALPHA26	26	320	1014
VOWEL	11	10	990
MNIST	10	784	150
JAFFE	10	1024	213

4.3 Classification Results

Our proposed methods can find the low-rank structure of the regression models, which are equivalent to doing regression in the regularized LDA subspace. For illustration purpose, in Fig. 1 we plot the ranked singular value of the learnt coefficient matrix $W = AB$ on the left hand side and draw the absolute value of the learnt W of the 1st fold (of the 5 fold cross validation, other folds show similar result) on the right hand side for each dataset. The corresponding rank parameter is selected based on which SLRR achieves the best classification accuracy. For example, in Fig. 1(a) shows the UMIST results, we can see the number of non-zero singular value of W is 15, i.e., the rank of the learnt coefficient matrix is 15, less than its full rank value of 20. In addition, the learnt W is sparse and is effectively used for feature selection, *e.g.* selecting the important features (non-zero rows) across all the classes.

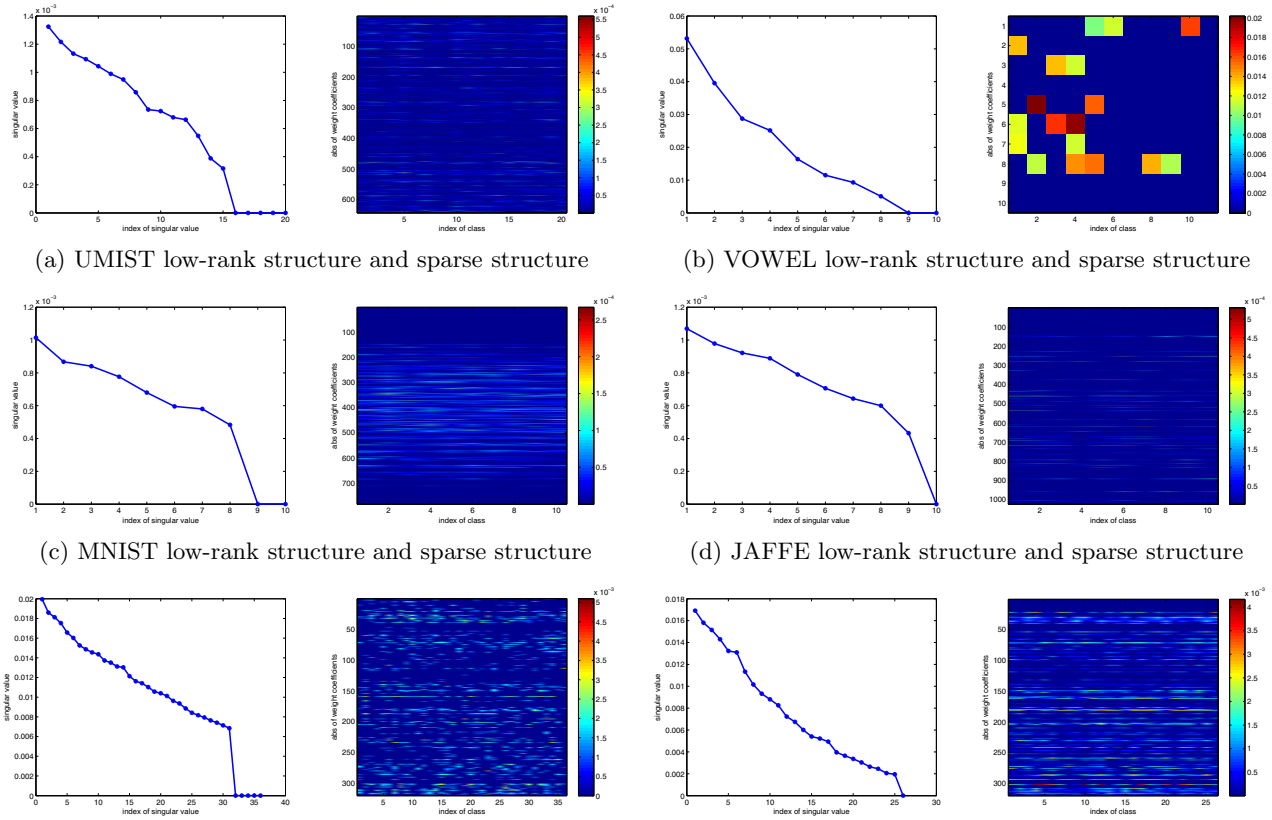
Fig. 2 shows the average classification accuracy comparisons of the above three types of full-rank regressions with the proposed low-rank counterparts with respect to different low-rank constraints. From Fig. 2, we can obviously conclude that the discriminant low-rank regressions consistently outperform their full-rank counterparts, when the specified low-rank parameter s falls in a proper range. For five out of six datasets in our experiments, the low-rank property can boost the result greatly. Only in JAFFE dataset (as shown in Fig. 2.(l)), the performance of sparse low-rank regression is competitive with that of the full-rank counterpart.

To help the researchers easily compare all methods, we also list the best classification results in terms of average accuracy and standard deviation for different regression methods in Table 2.

Our experimental results also verify our previous key point that the RLRR method is better than LRLR method. On all six datasets, the RLRR outperforms the LRLR. Surprisingly, the standard ridge regression even has better performance than the LRLR method. The LRLR is equivalent to existing low-rank regression models, and both methods may have suboptimal results due to the rank deficiency problem. In standard ridge regression or RLRR methods, because the rank constraint is imposed, both of them alleviate such matrix rank deficiency issue. Now we showed the connection between low-rank constraint and LDA projection, such that we can uncover this problem.

Table 2: The average classification accuracy using different regression methods on six datasets.

Data	Rank	Linear Regression	Ridge Regression	Sparse Linear Regression
UMIST	Full	0.6650 ± 0.1069	0.9197 ± 0.0456	0.9525 ± 0.0533
	Low	$\mathbf{0.8225 \pm 0.0937}$	$\mathbf{0.9675 \pm 0.0322}$	$\mathbf{0.9675 \pm 0.0245}$
BINALPHA36	Full	0.3488 ± 0.0241	0.6039 ± 0.0231	0.5971 ± 0.0205
	Low	$\mathbf{0.4147 \pm 0.0238}$	$\mathbf{0.6105 \pm 0.0178}$	$\mathbf{0.6069 \pm 0.0205}$
BINALPHA26	Full	0.3636 ± 0.0124	0.6732 ± 0.0258	0.6527 ± 0.0297
	Low	$\mathbf{0.4422 \pm 0.0255}$	$\mathbf{0.6771 \pm 0.0221}$	$\mathbf{0.6578 \pm 0.0281}$
VOWEL	Full	0.2960 ± 0.0405	0.3010 ± 0.0402	0.2960 ± 0.0417
	Low	$\mathbf{0.2980 \pm 0.0323}$	$\mathbf{0.3040 \pm 0.0304}$	$\mathbf{0.3020 \pm 0.0314}$
MNIST	Full	0.4067 ± 0.0830	0.4467 ± 0.1043	0.8067 ± 0.0435
	Low	$\mathbf{0.4400 \pm 0.1020}$	$\mathbf{0.7933 \pm 0.0772}$	$\mathbf{0.8267 \pm 0.0742}$
JAFFE	Full	0.6519 ± 0.1066	0.9446 ± 0.0479	0.9870 ± 0.0188
	Low	$\mathbf{0.8617 \pm 0.0813}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9951 \pm 0.0098}$



(a) UMIST low-rank structure and sparse structure

(b) VOWEL low-rank structure and sparse structure

(c) MNIST low-rank structure and sparse structure

(d) JAFFE low-rank structure and sparse structure

(e) BINALPHA36 low-rank structure and sparse structure

(f) BINALPHA26 low-rank structure and sparse structure

Figure 1: Demonstration of the low-rank structure and sparse structure found by our proposed SLRR method.

For some data with very large feature dimension ($d >> n$), like UMIST, MNIST and JAFFE, feature selection is necessary to reduce the redundancy between features and alleviate the curse of dimensionality. Our classification results both in Fig. 2 and Table 2 have shown that under such circumstances, SLRR and its full rank counterpart can achieve better classification result than RLRR and ridge regression since the $\ell_{2,1}$-norm can impose sparsity and select the features for all the classes.

Thus, our newly proposed RLRR as well as SLRR methods are more important and more practical low-rank models for machine learning applications.

5. CONCLUSION

In this paper, we provide theoretical analysis on low-rank regression models. We proved that the low-rank regression is equivalent to doing linear regression in the LDA subspace. More important, we proposed two new discriminant low-rank ridge regression and sparse low-rank regression meth-

ods. Both of them are equivalent to doing regularized regression in the regularized LDA subspace. From both theoretical and empirical views, we showed that both LRRR and SLRR methods provide better learning results than standard low-rank regression. Extensive experiments have been conducted on six benchmark datasets to demonstrate that our proposed low-rank regression methods consistently outperform their corresponding full-rank counterparts in terms of average classification accuracy.

6. ACKNOWLEDGMENTS

Corresponding Author: Heng Huang (heng@uta.edu).

This research was partially supported by NSF-IIS 1117965, NSF-CCF 0830780, NSF-DMS 0915228, NSF-CCF 0917274.

7. REFERENCES

[1] T. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22(3):327–351, 1951.

[2] T. Anderson. Asymptotic distribution of the reduced rank regression estimator under general conditions. *The Annals of Statistics*, 27(4):1141–1154, 1999.

[3] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, pages 41–48, 2006.

[4] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997.

[5] F. Bunea, Y. She, and M. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

[6] X. Cai, F. Nie, H. Huang, and C. H. Q. Ding. Multi-class $\ell_{2,1}$-norms support vector machine. In *ICDM*, pages 91–100, 2011.

[7] C. H. Q. Ding, D. Zhou, X. He, and H. Zha. R_1-pca: rotational invariant l_1-norm principal component analysis for robust subspace factorization. In *ICML*, pages 281–288, 2006.

[8] D. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32, 2000.

[9] J. Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, pages 165–175, 1989.

[10] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Pr, 1990.

[11] D. Graham and N. Allinson. Characterising virtual eigensignatures for general purpose face recognition. *NATO ASI series. Series F: computer and system sciences*, pages 446–456, 1998.

[12] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, pages 55–67, 1970.

[13] A. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] D. Luo, C. Ding, and H. Huang. Linear discriminant analysis: New formulations and overfit analysis. *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[16] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE, 1998.

[17] F. Nie, H. Huang, X. Cai, and C. H. Q. Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, pages 1813–1821, 2010.

[18] M. Niranjan and F. Fallside. Neural networks and radial basis functions in classifying static speech patterns. *Computer Speech & Language*, 4(3):275–289, 1990.

[19] G. Obozinski, B. Taskar, and M. Jordan. Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, 2006.

[20] G. Reinsel and R. Velu. *Multivariate reduced-rank regression: theory and applications*. Springer New York, 1998.

[21] L. Sun, R. Patel, J. Liu, K. Chen, T. Wu, J. Li, E. Reiman, and J. Ye. Mining brain region connectivity for alzheimer's disease study via sparse inverse covariance estimation. In *KDD*, pages 1335–1344, 2009.

[22] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[23] H. Wang, F. Nie, H. Huang, S. L. Risacher, C. Ding, A. J. Saykin, L. Shen, and ADNI. A new sparse multi-task regression and feature selection method to identify brain imaging predictors for memory performance. *IEEE Conference on Computer Vision*, pages 557–562, 2011.

[24] H. Wang, F. Nie, H. Huang, S. L. Risacher, A. J. Saykin, L. Shen, et al. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics (ISMB)*, 28(12):i127–i136, 2012.

[25] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, and L. Shen. High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction. In *NIPS*, pages 1286–1294, 2012.

[26] S. Xiang, Y. Zhu, X. Shen, and J. Ye. Optimal exact least squares rank minimization. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 480–488, 2012.

[27] J. Ye. Least squares linear discriminant analysis. In *ICML*, pages 1087–1093, 2007.

[28] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.