Multimed Tools Appl (2017) 76:17461–17477
DOI 10.1007/s11042-016-3980-3

Leverage triple relational structures via low-rank feature reduction for multi-output regression

Shichao Zhang[1] · **Lifeng Yang**[1] · **Zhenyun Deng**[1] · **Debo Cheng**[1] · **Yonggang Li**[1]

Received: 17 May 2016 / Revised: 6 September 2016 / Accepted: 14 September 2016 /
Published online: 8 October 2016

Abstract Multi-output regression aims at learning a mapping from feature variables to multiple output variables. It is significant to utilize variety of inherent relational structure information of observations to conduct multi-output regression task when learning a best mapping from high-dimensional data. In this paper, we propose a new multi-output regression method, which simultaneously takes advantage of the low-rank constraint, sample selection, and feature selection in a unified framework. We first take the effect of low-rank constraint to search the correlation of output variables and impose $\ell_{2,p}$-norm regularization on the coefficient matrix to capture the correlation between features and outputs. And then, the $\ell_{2,p}$-norm on the loss function is designed to discover the correlation between samples, so as to select those informative samples to learn the model for improving predictive capacity. Thirdly, orthogonal subspace learning is exploited to ensure multi-output variables share the same low-rank structure of data by rotating the results of feature selection. In addition, to get the optimal solution of the objective function, we propose an effective iterative optimization algorithm. Finally, we conduct sets of experimental results on real datasets, and show the proposed method outperforms the state-of-the-art methods in terms of aCC and aRMSE.

Keywords Multi-output regression · Low-rank regression · Feature selection · Orthogonal subspace learning

1 Introduction

Multi-output regression [4], also known as multi-target and multi-response regression, is a significant research topic in machine learning and statistics. It aims at simultaneously

✉ Shichao Zhang
zhangsc@mailbox.gxnu.edu.cn

[1] Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin, Guangxi 541004, China

predicting multiple real-value output variables from the same feature variables (can be regarded as the input variables). As the matter of fact, multi-output regression is encountered frequently in various fields of applications, such as ecological model requires to predict many kinds of conditions or quality of the vegetation [18]; stock analysis needs to predict many attributions of stocks, i.e., the price of a stock in the future via utilizing the related economic variables and prices happened in the past [22]; supervised analysis simultaneously estimates some different biophysical parameters from the remote sensing images [25]; and so on.

As known, the data used for multi-output regression have high-dimensional features and its quantity is large scale, due to the development of science and technology. Therefore, it comes to a big challenge of leveraging the large amount of high-dimensional data to conduct multi-output regression effectively. Many previous works have given attentions to find the relational structure of high-dimensional data for adding the effect of algorithm, for instance manifold structure [5], low-rank structure [1, 9, 16], and so on. And, most methods always effectively complete the task of dealing with high-dimensional data and removing the interference of noise by conducting feature reduction [12, 13, 19, 38]. As well as, we desire to search for all the most possible relational structures of information inherent in these data for improving the multi-output regression model.

To deal with big scale of high-dimensional data for multi-output regression effectively, we fully consider the three kinds of inhered information of the data, interpretation as Fig. 1, i.e., the relationship of output and output, the relationship of feature and output, the relationship of sample and sample, and then we propose a new method leveraging these triple relational structure information into a unified framework called Low-rank Feature Reduction for multiple-output regression (shorted for LFR). The rationale of the proposed solution for multi-output regression is to simultaneously conducting feature selection and sample selection in the low-rank regression model. Specially, we employ the low-rank constraint on the rank of the coefficient matrix to perform the searching of relation structure among multiple output variables. Meantime, the unimportant features should not participate in the learning of regression model, due to the redundant features always hold little informative to explore the relational structures and sometimes cost more computational cost. Therefore, we also conduct feature selection by imposing an $\ell_{2,p}$-norm regularization term to penalize all coefficients in the same row of the regression coefficient matrix for the removal of the noisy features. Additionally, we also employ the $\ell_{2,p}$-norm on the loss function to conduct sample selection, aiming at finding the relation structure among samples and eliminating the helpless samples, i.e., the

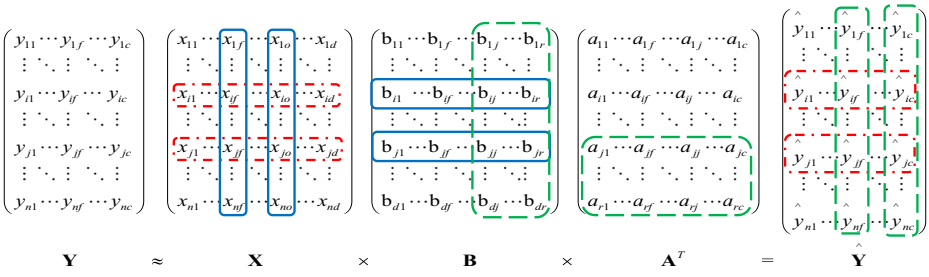


Fig. 1 An illustration of the three kinds of relational structures inhered in high-dimensional data. The *red dotted rectangles*, the *blue solid rectangles*, and the *green dash rectangles*, imply the relationship of sample-sample, feature-output, output-output, respectively

outliers. Although the derived objective function is hard to deal with as which is convex and non-smooth, we further propose a novel iterative algorithm to optimize it efficiently.

The main contributions of our LFR method are described as follow:

- This work uses an $\ell_{2,p}$-norm regularization term to make row-sparse of the coefficient matrix, so that it can effectively take advantage of the correlation between features and multiple outputs, and select useful features to improve the predictive ability of model. Moreover, it can avoid encountering the problem of "curse of dimension" in high-dimensional data.
- In our framework, we innovative combine the $\ell_{2,p}$-norm with loss function term to search the correlation among samples. Then, the proposed method can effectively remove the interference of outliers and select more informative samples to improve the capability of regression model. This idea is always neglected by existing regression methods.
- The proposed method makes use of a low-rank constraint on the rank of coefficient matrix to construct the low-rank regression model, as mentioned in previous literatures, which has the ability to find the correlation among multiple outputs effectively. Therefore, we can reduce the number of model's parameters and improve the model's predictive accuracy.

The rest of the paper is organized as follow: we will briefly summarize reviews of the previous feature selection and low-rank regression for multi-output regression task in Section 2. And we detailly introduce the proposed multi-output regression method for high-dimensional data, and present an effective algorithm to get the optimal solution of the objective function in Section 3. We conduct extensive experiments to demonstrate the effectiveness of our algorithm in Section 4. Finally, the conclusion is drawn in Section 5.

2 Related works

The traditional linear regression is an useful approach of settling the problem of single-output regression, but when confronts with the multiple-output regression, it usually obtains the solution via separately conducting the single-regression to predict each output variable. In fact, there are always some inhered correlation structures in multiple output variables [21]. However, the traditional linear regression method does not take advantage of the fact that multiple outputs are likely correlation.

Some researchers [27] wanted to break out the obstacle by adding the process of sparse learning on the traditional linear regression, which is equivalent to conduct feature selection to seek for the relationship of the regression coefficients. This kind of methods usually can obtain the faithful results, but they still do not catch the correlation among multiple outputs. To do this for improving the regression model, Aderson, et al. [1], proposed an effective way of using a low-rank constraint on the rank of coefficient matrix. After that, researchers [2] proposed imposing the trace-norm regularization on the coefficient matrix to discover the low-rank structure existing among the multiple outputs. However, the method can not explicitly select or tune the rank of the coefficient matrix. Izenwan, et al. [16], proposed the so-called reduced-rank regression (RRR) method, in which the low-rank constraint implies that the coefficient matrix can be expressed as the

product of two matrices owned a lower artificial fixed rank. And this method can effectively reduce the effective number of parameters to be estimated to improve the efficiency of estimation, and also can explicitly select the rank of the coefficient matrix. However, these high-dimensional data have large amount of features, some of them might have unrelated information for regression prediction and may result in more unnecessary computational cost.

At that point, some researchers [5] proposed some methods leveraged the additional process of sparse learning along with the low-rank regression modal to seek for the relationship between features and outputs, that is, conducting feature selection to select some features with more useful information for improving the model. But, there are usually involved many outliers and corrupted samples by noise in data, which might have a bad effect on the learning of multi-output regression model.

Our work pursues to obtain a better multi-output regression model for large amount of high-dimensional data. Therefore, we not only consider the triple relational structures in data, but also remove the interference of redundant features and outlier samples for improving the predictive capacity and efficient of model.

3 Method

In this section, we consider three kinds of inhered relational structure information in data, as described in Fig. 1. Firstly, we leverage these correlations in the data to construct multi-output regression model. Then, we apply the proposed iterative algorithm to optimize objective function for obtaining the optimal solution. Finally, we analyse the convergence of the objective function.

3.1 Notations

In this paper, we denote matrices as boldface uppercase letters, vectors as boldface lowercase letters, and scalars as normal italic letters, respectively. For a matrix $\mathbf{X} = [x_{ij}]$, its i-th row and j-th column are denoted as x^i and x_j, respectively. Also we denote the Frobenius norm, ℓ_1-norm, ℓ_2-norm, $\ell_{2,1}$-norm, and $\ell_{2,p}$-norm of a matrix X as $\|\mathbf{X}\|_F = \sqrt{\sum_i \|x^i\|_2^2} = \sqrt{\sum_j \|x_j\|_2^2}$, $\|\mathbf{X}\|_1 = \sum_i \sum_j |x_{ij}|$, $\|\mathbf{X}\|_2 = \sum_i \sum_j |x_{ij}|^2$, $\|\mathbf{X}\|_{2,1} = \sum_i \|x_i\|_2^2 = \sum_i \sqrt{\sum_j x_{ij}^2}$, and $\|\mathbf{X}\|_{2,P} = [\sum_i (\sum_j |x_{ij}|^2)^{p/2}]^{1/p}$, respectively. We further denote the transpose operator, the trace operator, and the inverse of a matrix $\mathbf{X}$ as $\mathbf{X}^T$, $tr(\mathbf{X})$, and $\mathbf{X}^{-1}$, respectively.

3.2 LFR method for multiple-output regression

Given a training dataset $\mathbf{D} = \{(x^1, y^1), \cdots, (x^n, y^n)\} = (\mathbf{X}, \mathbf{Y}) \in R^{n \times (d+c)}$, where (x^i, y^i) denotes a sample, $x^i \in R^{1 \times d} (i = 1, \cdots, n)$ denotes a input feature vector with d dimensional features, $y^i \in R^{1 \times c} (i = 1, \cdots, n)$ denotes a output variable owned c outputs, and n is the number of samples, let $\mathbf{X} = [x^1, \cdots, x^i, \cdots, x^n] \in R^{n \times d}$ and $\mathbf{Y} = [y^1, \cdots, y^i, \cdots, y^n] \in R^{n \times c}$ respectively be the feature matrix (also known as input matrix) and output matrix. The purpose of multi-output regression is to find a model that can exactly predict the $\mathbf{Y}$ via the $\mathbf{X}$ in dataset $\mathbf{D}$, that is

$$x^i \xrightarrow{\text{predict}} y^i$$

It is known that the linear regression model finds the linear relation between x^i and y^i, i.e.,

$$\mathbf{Y}_{ij} = \mathbf{X}^i\mathbf{W}_j + e_{ij}, \quad (i = 1, \cdots, n \text{ and } j = 1, \cdots, c)$$

Where $\mathbf{W}$ denotes the regression coefficient matrix and e_{ij} means the error term. As there are n samples in the dataset, therefore the model can be described in matrix notation as follow,

$$\mathbf{Y} = \mathbf{XW} + \mathbf{E}$$

where $\mathbf{E} \in R^{n \times c}$ is the error matrix. However, we require the predictive outputs $\hat{\mathbf{Y}} = \mathbf{XW}$ got closest to the ground truth outputs $\mathbf{Y}$, i.e., the $\mathbf{E}$ obtains the minimal values. This problem is often solved via the least square loss function, that is

$$\min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{XW}\|_F^2 \tag{1}$$

The solution of Eq. (1) is $\mathbf{W} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Note that, it is equivalent to transfer the multi-output regression problem to the single-output regression problem. However, there are always some underlying relational structures among large number of outputs. Therefore, imposing a low-rank constraint on the rank of $\mathbf{W}$ in Eq. (1) is used to obtain the possible correlation among multiple outputs, i.e.,

$$\min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{XW}\|_F^2, \quad s.t., \ rank(\mathbf{W}) \leq \min(d, c) \tag{2}$$

From the above equation, it is easy to know the rank of the regression coefficient matrix can be explicitly determined and the number of the effective number of parameters is reduced (for example, only r features are selected by the model), and hence the model predictive efficiency is improved. To get clearer interpretation of using the low-rank constraint $rank(\mathbf{W}) = r \leq \min(d, c)$, the $\mathbf{W}$ can be expressed as a product of two rank r matrixes, i.e.,

$$\mathbf{W} = \mathbf{BA}^T \tag{3}$$

Where $\mathbf{B} \in R^{d \times r}$ and $\mathbf{A} \in R^{c \times r}$. For the fixed r, then the Eq. (2) becomes as the following optimization problem

$$\min_{\mathbf{A},\mathbf{B}} \left\|\mathbf{Y} - \mathbf{XBA}^T\right\|_F^2 \tag{4}$$

Feature selection is often used to select some discriminating features from the high-dimensional data to reduce the dimension of data [31]. Consequently, it can lead to reduce the computational cost of various analyses for high-dimensional data, and simultaneously eliminate the noise in data. Therefore, conducting feature selection can result in a better model performance in practice. Accordingly, feature selection has been getting more and more widely adopted in many applications [11, 12, 14, 35, 37, 39] for high-dimensional data analysis.

Recently, sparse-based feature selection algorithms [6, 13, 26, 28, 32, 36] have already attracted increasing researchers' attentions. Such methods always can exploit the correlation among features and imply the fact that real high-dimensional data could be represented by sparse features. Researchers [9, 34] attempted to imposing an $\ell_{2,1}$-norm regularization on the coefficient matrix to conduct effectively feature selection. But researchers [29] showed that feature selection via utilizing an $\ell_{2,p}$-norm regularization

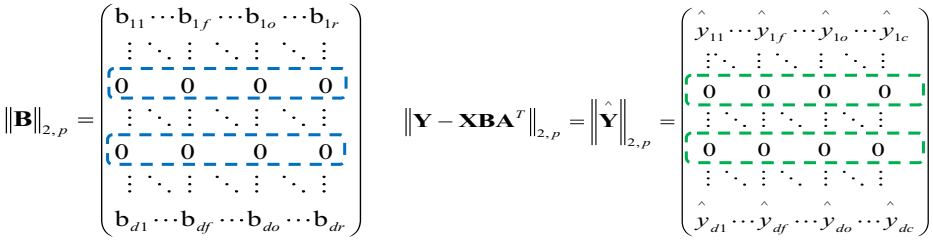


Fig. 2 An illustration to the $\ell_{2,p}$-norm regularization, especially, the *blue dash rectangles* in (**a**) imply the corresponding features can be eliminated. The *green dash rectangles* in (**b**) imply the corresponding samples can be removed

always outperforms feature selection via using the $\ell_{2,1}$-norm regularization and other popular feature selection methods. Just as the interpretation in Fig. 2a, if the feature has little information to help predict output variables, then all elements of the correspond row in matrix $\mathbf{B}$ are 0, otherwise, not all are 0.

The $\ell_{2,p}$-norm is defined as

$$\|\mathbf{W}\|_{2,P} = \left[\sum_i \left(\sum_j |w_{ij}|^2\right)^{p/2}\right]^{1/p} = \left(\sum_{i=1}^{d} \|\mathbf{W}^i\|^p\right)^{1/p}$$

Where $\mathbf{W} \in R^{d \times c}$ denotes an arbitrary matrix, $p(0 < p < 2)$ is a parameter. The $\ell_{2,p}$ -norm can effectively make some matrix's rows shrink to zero, i.e., it makes $\mathbf{W}$ row-sparse. And the nonzero row is corresponding to the selected features.

To exclude the redundant features, it is significant to perform feature selection. From the Fig. 1, note that excluding a feature corresponds to setting an entire row of the matrix $\mathbf{B}$ as zeros. Therefore, we add an $\ell_{2,p}$-norm regularization term on the matrix $\mathbf{B}$ and one orthogonal constraint to keep the low-rank structure of data by rotating the results of feature selection. Specifically, we solve

$$\min_{\mathbf{A},\mathbf{B}} \left\|\mathbf{Y}-\mathbf{X}\mathbf{B}\mathbf{A}^T\right\|_F^2 + \lambda\|\mathbf{B}\|_{2,P}, \quad s.t., \ \mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{5}$$

Where λ is the tuning parameter and the constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ is introduced for identifiability purpose, $\mathbf{I} \in R^{r \times r}$ denotes the identity matrix. Note that $p(0 < p < 2)$ is a tuning parameter which controls the degree of correlated structures among features. After that, the redundant features will be excluded in the low-rank regression. Moreover, the orthogonal rotation constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ is used to conduct a subspace learning to enhance the performance of the feature selection process.

After the above steps, we can leverage the correlation among multiple outputs via the low-rank regression and the relationship of feature-output via imposing an $\ell_{2,p}$-norm regularization on the matrix $\mathbf{B}$. However, there are many predict-helpless samples in large amount of high dimensional data, i.e., the outlier samples, which might be the obstacle of model's learning. Thus it is necessary to remove the useless samples in the dataset. To this end, an effective method [19] of taking advantage of the likely sample-sample relationship may impose an $\ell_{2,p}$-

norm on the predict-output matrix ($\mathbf{Y}-\mathbf{XBA}^T$). Therefore we consider the following final optimization problem

$$\min_{\mathbf{A},\mathbf{B}} \left\|\mathbf{Y}-\mathbf{XBA}^T\right\|_{2,p} + \lambda\|\mathbf{B}\|_{2,P}, \quad s.t.,\ \mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{6}$$

Where, parameter $p = 1$, the $\ell_{2,p}$-norm will be transferred to the standard $\ell_{2,1}$-norm. Regarding $\hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{XBA}^T$ as the predict-output matrix, then we impose the $\ell_{2,p}$-norm regularization on the matrix $\hat{\mathbf{Y}}$ to obtain its row-sparse. An interpretation of sample selection as Fig. 2b, i.e., all outputs in the same row of the matrix $\hat{\mathbf{Y}}$ are 0, which represents the corresponding sample is useless to prediction and should be removed.

According to the above analyses, we know that the Eq. (6) considers simultaneously the process of low-rank and subspace learning to improve the performance of prediction. Specifically, we use a rank constraint $rank(\mathbf{BA}^T) = r \le \min(d, c)$ to limit the rank of $\mathbf{A}$ and $\mathbf{B}$. Where, the low-rank constraint on $\mathbf{B}$ is used to select the high related features by consider the correlation among features, and then remove the redundancy or irrelevant features. Further, the sparsity of features and samples are realized by $\ell_{2,p}$-norm penalizing each row of Eq. (6) regression model. Therefore, we can select the representative features and samples by the sparse regression model. Especially the distribution of low-rank representation structures may different from each other after conducting feature selection, therefore the subspace learning is used to maintain the structure of multi-output by rotating the output matrix.

3.3 Optimization

Equation (6) is convex but non-smooth due to involving in the $\ell_{2,p}$-norm term. As a result, we propose an algorithm of iteratively optimizing with respect to $\mathbf{A}$ and $\mathbf{B}$. In detail, we iteratively conduct the following two steps until satisfy the predefined conditions:

1) Update $\mathbf{B}$ with the fixed $\mathbf{A}$

Since the constraint $\mathbf{AA}^T = \mathbf{I}$, such that there is an orthogonal matrix $(\mathbf{A}, \mathbf{A}^\perp)$, where $\mathbf{A}^\perp$ is a matrix with orthogonal column. Then the optimization problem of Eq. (6) is equivalent to the following problem:

$$\begin{aligned}\min_{\mathbf{B}} \left\|\mathbf{Y}-\mathbf{XBA}^T\right\|_{2,p} + \lambda\|\mathbf{B}\|_{2,P} &= \min_{\mathbf{B}} \left\|\left(\mathbf{Y}-\mathbf{XBA}^T\right)(\mathbf{A}, \mathbf{A}^\perp)\right\|_{2,p} + \lambda\|\mathbf{B}\|_{2,P} \\ &= \min_{\mathbf{B}} \left(\|\mathbf{YA}-\mathbf{XB}\|_{2,p} + \|\mathbf{YA}^\perp\|_{2,p}\right) + \lambda\|\mathbf{B}\|_{2,P}\end{aligned} \tag{7}$$

Therefore, we fixed $\mathbf{A}$, the above Eq. (7) can be reduced to

$$\min_{\mathbf{B}} \|\mathbf{YA}-\mathbf{XB}\|_{2,p} + \lambda\|\mathbf{B}\|_{2,P} \tag{8}$$

At the following section, we will prove the convergence of the value of Eq. (8). Note that Eq. (8) can be rewritten as follow

$$\min_{\mathbf{B}}\ tr\left[(\mathbf{YA}-\mathbf{XB})^T\mathbf{N}(\mathbf{YA}-\mathbf{XB})\right] + \lambda tr\left(\mathbf{B}^T\mathbf{QB}\right) \tag{9}$$

Where both $\mathbf{N} \in R^{n \times n}$ and $\mathbf{Q} \in R^{d \times d}$ are diagonal matrix with respectively diagonal elements $\mathbf{N}_{ii} = \frac{p}{2\|(\mathbf{YA}-\mathbf{XB})^i\|_2^{2-p}}$ $(i = 1, \cdots, n)$ and $\mathbf{Q}_{jj} = \frac{p}{2\|\mathbf{B}^j\|_2^{2-p}}$ $(j = 1, \cdots, d)$. By setting the derivative of Eq. (9) *w.r.t.* $\mathbf{B}$ to zero, then we have

$$\mathbf{B} = \left(\mathbf{X}^T\mathbf{NX} + \lambda\mathbf{Q}\right)^{-1}\mathbf{X}^T\mathbf{NYA} \tag{10}$$

2) Update A with the fixed B

As matrix $\mathbf{B}$ is fixed, thus the optimization problem in Eq. (6) can be reduced to

$$\min_{\mathbf{A}} \left\|\mathbf{Y}-\tilde{\mathbf{X}}\sim\mathbf{A}^T\right\|_{2,p}, \quad s.t., \ \mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{11}$$

Where $\tilde{\mathbf{X}}\sim = \mathbf{XB} \in R^{n \times r}$. According to literature [15], the Eq. (11) is an orthogonal Procrustes problem, such that the optimal solution of $\mathbf{A}$ is $\mathbf{UV}^T$, where $\mathbf{U} \in R^{n \times r}$ and $\mathbf{V} \in R^{r \times r}$ are obtained from the singular value decomposition of $\mathbf{Y}^T\mathbf{X} = \mathbf{UDV}^T$, where $\mathbf{D} \in R^{r \times r}$ is a diagonal matrix.

The discussion above leads to the following algorithm 1 [20, 30].

Algorithm 1 the pseudo code of solving Eq.(6)

Input: $\mathbf{X} \in R^{n \times d}$, $\mathbf{Y} \in R^{n \times c}$, λ, p, r

Output: $\mathbf{A} \in R^{c \times r}$, $\mathbf{B} \in R^{d \times r}$

1. Initialize the iterative number t=0;
2. Initialize $\mathbf{A}(0)$ as a random diagonal matrix;
3. **While** the value of objective function (6) not converged **do**;
4. Update $\mathbf{B}(t+1)$ via Eq.(10);
5. Update $\mathbf{A}(t+1)$ via Eq.(11);
6. Compute the diagonal matrix $\mathbf{N}(t+1)$ as $\mathbf{N}_{ii} = \frac{p}{2\left\|(\mathbf{YA}-\mathbf{XB})^i\right\|_2^{2-p}} (i = 1, \cdots, n)$;
7. Compute the diagonal matrix $\mathbf{Q}(t+1)$ as $\mathbf{Q}_{ii} = \frac{p}{2\left\|\mathbf{B}^j\right\|_2^{2-p}} (j = 1, \cdots, d)$;
8. Compute the value of objective function (6);
9. t=t+1;

End.

3.4 Proving of the convergence

It can be proved that the objective function value in Eq. (8) monotonically decreases in each of iteration. Note that, the objective function (8) is equivalent to

$$\min_{\mathbf{B}} tr\left[(\mathbf{YA}-\mathbf{XB})^T\mathbf{N}(\mathbf{YA}-\mathbf{XB})\right] + \lambda tr\left(\mathbf{B}^T\mathbf{QB}\right)$$

Therefore, we have

$$tr\left[\left(\mathbf{YA}_{(t+1)}-\mathbf{XB}_{(t+1)}\right)^{T}\mathbf{N}_t\left(\mathbf{YA}_{(t+1)}-\mathbf{XB}_{(t+1)}\right)\right]+\lambda tr\left({\mathbf{B}_{(t+1)}}^{T}\mathbf{Q}_t\mathbf{B}_{(t+1)}\right)$$

$$\le tr\left[\left(\mathbf{YA}_t-\mathbf{XB}_t\right)^{T}\mathbf{N}_t\left(\mathbf{YA}_t-\mathbf{XB}_t\right)\right]+\lambda tr\left({\mathbf{B}_t}^{T}\mathbf{Q}_t\mathbf{B}_t\right)$$

$$\Rightarrow\sum_{i=1}^{n}\frac{\left\|y^i\mathbf{A}_{(t+1)}-x^i\mathbf{B}_{(t+1)}\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2-p}}+\sum_{i=1}^{d}\frac{\left\|{b^i}_{(t+1)}\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|{b^i}_t\right\|_2^{2-p}}\le\sum_{i=1}^{n}\frac{\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2-p}}$$

$$+\sum_{i=1}^{d}\frac{\left\|{b^i}_t\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|{b^i}_t\right\|_2^{2-p}}\quad\Rightarrow\sum_{i=1}^{n}\left\|y^i\mathbf{A}_{(t+1)}-x^i\mathbf{B}_{(t+1)}\right\|_2^{2-p}-\sum_{i=1}^{n}\left\|y^i\mathbf{A}_{(t+1)}-x^i\mathbf{B}_{(t+1)}\right\|_2^{2-p}$$

$$+\sum_{i=1}^{n}\frac{\left\|y^i\mathbf{A}_{(t+1)}-x^i\mathbf{B}_{(t+1)}\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2-p}}\quad+\lambda\sum_{i=1}^{d}\left\|{b^i}_{(t+1)}\right\|_2^{2-p}-\lambda\sum_{i=1}^{d}\left\|{b^i}_{(t+1)}\right\|_2^{2-p}$$

$$+\lambda\sum_{i=1}^{d}\frac{\left\|{b^i}_{(t+1)}\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|{b^i}_t\right\|_2^{2-p}}\quad\le\sum_{i=1}^{n}\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2-p}-\sum_{i=1}^{n}\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2-p}$$

$$+\sum_{i=1}^{n}\frac{\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2-p}}\quad+\lambda\sum_{i=1}^{d}\left\|{b^i}_t\right\|_2^{2-p}-\lambda\sum_{i=1}^{d}\left\|{b^i}_t\right\|_2^{2-p}+\lambda\sum_{i=1}^{d}\frac{\left\|{b^i}_t\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|{b^i}_t\right\|_2^{2-p}}$$

$$\Rightarrow\sum_{i=1}^{n}\left\|y^i\mathbf{A}_{(t+1)}-x^i\mathbf{B}_{(t+1)}\right\|_2^{2-p}+\lambda\sum_{i=1}^{d}\left\|{b^i}_{(t+1)}\right\|_2^{2-p}$$

$$-\left(\sum_{i=1}^{n}\left\|y^i\mathbf{A}_{(t+1)}-x^i\mathbf{B}_{(t+1)}\right\|_2^{2-p}-\sum_{i=1}^{n}\frac{\left\|y^i\mathbf{A}_{(t+1)}-x^i\mathbf{B}_{(t+1)}\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2-p}}\right)$$

$$-\lambda\left(\sum_{i=1}^{d}\left\|{b^i}_{(t+1)}\right\|_2^{2-p}-\sum_{i=1}^{d}\frac{\left\|{b^i}_{(t+1)}\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|{b^i}_t\right\|_2^{2-p}}\right)\quad\le\sum_{i=1}^{n}\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2-p}+\lambda\sum_{i=1}^{d}\left\|{b^i}_t\right\|_2^{2-p}$$

$$-\left(\sum_{i=1}^{n}\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2-p}-\sum_{i=1}^{n}\frac{\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2-p}}\right)$$

$$-\lambda\left(\sum_{i=1}^{d}\left\|{b^i}_t\right\|_2^{2-p}-\sum_{i=1}^{d}\frac{\left\|{b^i}_t\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|{b^i}_t\right\|_2^{2-p}}\right)$$

It had been showed in [7] that for any nonzero vectors have

$$\sum_i\left\|{b^i}_{(t+1)}\right\|_2^{2-p}-\sum_i\frac{\left\|{b^i}_{(t+1)}\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|{b^i}_{(t+1)}\right\|_2^{2-p}}\le\sum_i\left\|{b^i}_t\right\|_2^{2-p}-\sum_i\frac{\left\|{b^i}_t\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|{b^i}_t\right\|_2^{2-p}}$$

$$\sum_{i=1}^{n}\left\|y^i\mathbf{A}_{(t+1)}-x^i\mathbf{B}_{(t+1)}\right\|_2^{2-p}-\sum_{i=1}^{n}\frac{\left\|y^i\mathbf{A}_{(t+1)}-x^i\mathbf{B}_{(t+1)}\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2-p}}\le\sum_{i=1}^{n}\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2-p}$$

$$-\sum_{i=1}^{n}\frac{\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2(2-p)}}{\left(2/p\right)\left\|y^i\mathbf{A}_t-x^i\mathbf{B}_t\right\|_2^{2-p}}$$

And it can be easily known that

$$\sum_{i=1}^{n} \left\| y^i \mathbf{A}_{(t+1)} - x^i \mathbf{B}_{(t+1)} \right\|_2^{2-p} + \lambda \sum_{i=1}^{d} \left\| b^i{}_{(t+1)} \right\|_2^{2-p} \leq \sum_{i=1}^{n} \left\| y^i \mathbf{A}_t - x^i \mathbf{B}_t \right\|_2^{2-p} + \lambda \sum_{i=1}^{d} \left\| b^i{}_t \right\|_2^{2-p}$$

Thus, it can be demonstrated that the value of objective function (8) decreases in each iteration. At the meantime, Eq. (8) is a convex function, which indicates [33] that it will converge to the global optimum solution of the Eq. (8).

4 Experimental results

In this section, we will compare the performance of our proposed LFR method with the state-of-the-art methods on multi-output datasets in terms of aCC and aRMSE. We firstly introduce the benchmark datasets used in experiments. And then, we summarize the comparing algorithms and the experimental setting. Finally, we analyze the results and obtain a conclusion.

4.1 Dataset descriptions

We summarize the general information of datasets used in our experiments in Table 1.

EDM [17]: dataset for the Electrical Discharge Machining aims to predict two target variables using 16 input variables representing mean values and deviations of the observed quantities of the considered machining parameters.
ATP1d and ATP7d [24]: datasets of Airline Ticket Prices, the input variables include details about the flights and the 6 target variables are the minimum prices observed over the next 7 days for 6 flight preferences.
OES97 and OES10 [24]: gathered from the annual Occupation Employment Survey compiled by the US Bureau of Labor Statistics for the years 1997 (OES97) and 2010 (OES10). The input variables are a randomly sequenced subset of employment types, and

Table 1 The general information of experiment datasets, where n, d, and c denote the number of samples, features and outputs, respectively

Dataset	Samples (n)	Features (d)	Outputs (c)
EDM	154	16	2
ATP1d	337	411	6
ATP7d	296	411	6
SF1	323	10	3
SF2	1066	10	3
WQ	1060	16	14
OES97	334	263	16
OES10	403	298	16
SCM1D	1658	280	16
SCM20D	1503	61	16

16 targets are randomly selected from the entire set of categories above the 50 % threshold.

SF1 and SF2 [3]: datasets for predicting three potential types of Solar Flare from the ten feature variables describing active regions on the sun.

WQ [10]: provided by the Hydrometeorological Institute of Slovenia for inferring chemical from biological parameters of river water quality. It includes the measured values of 16 different chemical parameters and 14 bioindicator taxa.

SCM1D and SCM20D [23]: contain 16 regression targets, each target corresponding to the next day mean price (SCM1D) or mean price for 20-days in the future (SCM20D) for each product in the simulation.

4.2 Experimental settings

For each dataset, we first randomly split the dataset into 10 parts. Then according to the standard 10-fold cross validation, we select one part for testing and use the remaining 9 parts for training, repeat the whole process 10 times to avoid the possible bias. The final results for different methods are reported. For the model selection, we set tuning parameter $\lambda \in \{10^{-5}, \cdots, 10^{5}\}$, the rank of coefficient matrix $r \in \{1, \cdots, \min(d, c)\}$ and parameter $p \in \{0.1, \cdots, 1.9\}$ in the $\ell_{2,p}$-norm, and $(c, g) \in \{10^{-5}, \cdots, 10^{5}\}$ in the LIBSVM toolbox by a 5-fold inner cross-validation.

The evaluation of different methods is based on two metrics [4]: aCC (average Correlation Coefficient) and aRMSE (average Root Mean Square Error). Specially, the aCC can effectively reflect the correlation between predicted outputs and its corresponding truth outputs, i.e., the bigger aCC is, the predicted outputs get more closed to its corresponding truth ground outputs, namely, the prediction can achieve more faithful results. And the aRMSE is often used to reflect the stability of algorithm, the smaller aRMSE is, the algorithm owns a better stability. Let y_i and $\hat{y}_i$ be the truth outputs and the predictive outputs, respectively. $\bar{y}_i$ And $\tilde{y}_i\sim$ are the means of truth outputs and the predictive outputs respectively. The definition of them as follow

$$aCC = \frac{1}{d}\sum_{i=1}^{d} \frac{\sum_{j=1}^{ntest}\left(y_i^{(j)} - \bar{y}_i\right)\left(\hat{y}_i^{(j)} - \tilde{y}_i\sim\right)}{\sqrt{\sum_{j=1}^{ntest}\left(y_i^{(j)} - \bar{y}_i\right)^2 \sum_{j=1}^{ntest}\left(\hat{y}_i^{(j)} - \tilde{y}_i\sim\right)^2}}$$

$$aRMSE = \frac{1}{d}\sum_{i=1}^{d} \sqrt{\frac{\sum_{j=1}^{ntest}\left(y_i^{(j)} - \hat{y}_i^{(j)}\right)^2}{ntest}}$$

We compare the proposed method with the following representative state-of-the-art methods:

- SMART [27]: Sparse Multi-tAsk Regression and feature selection model includes both $\ell_{2,1}$-norm and ℓ_1-norm regularizers for feature selection, however without imposing a low-rank constraint on the rank of the coefficient matrix.
- LSG21 [6]: New Graph Structured Sparsity Model has the process of sparse learning via imposing the $\ell_{2,1}$-norm on the coefficient matrix, but does not adopt the low-rank constraint.

- CSFS [8]: Convex Semi-supervised multi-label Feature Selection can conduct the feature selection via the $\ell_{2,1}$-norm regularization, but might neglects the low-rank structure of data.
- LRLR [16]: Low-Rank Linear Regression model is the original linear regression modal, but different from the traditional linear regression, because it makes a low-rank constraint on the rank of regression coefficient matrix.
- LRRR [5]: Low-Rank Ridge Regression model own the low-rank constraint, and also impose the ℓ_2-norm on the coefficient matrix.
- SLRR [5]: Sparse Low-Rank Regression model not only uses the low-rank constraint to seek for the low-rank structure in data, but also utilizes the $\ell_{2,1}$-norm on the coefficient matrix for feature selection.

The comparing methods above discussion can be parted into three groups. First, the algorithms only have the process of sparse learning without the low-rank constraint, i.e., SMART, LSG21 and CSFS algorithm. Second, the algorithms only have a low-rank regression but without sparse learning, i.e., LRLR algorithm. Finally, the algorithms own both the low-rank constraint and sparse learning, i.e., LRRR and SLRR algorithm, the difference of them is the regularizer on the coefficient matrix.

4.3 Regression results

We summarize the performances of the comparing methods on Tables 2 and 3. From the Table 2, we can know that the proposed LFR method outperformed all the comparing methods in terms of the aCC. For example, the proposed LFR method increased on average by 1.435 %, compared to the LRLR algorithm which owns the low-rank constraint on the rank of coefficient matrix but did not had the process of feature selection, and increased on average by 1.608 %, 1.230 % compared to the LRRR algorithm and SLRR algorithm respectively, which not only own the low-rank constraint for seeking the correlation among multiple outputs, but also respectively utilized the F-norm and $\ell_{2,1}$-norm to penalize the coefficient matrix for feature selection, and increased on average by 1.550, 1.456 and 1.439 % compared to the SMART algorithm, CSFS algorithm and LSG21 algorithm respectively, which could conduct feature selection but without the process of low-rank constraint on the coefficient

Table 2 The aCC results of all algorithms experimented on the multi-output datasets

Dataset	LRLR	LRRR	SLRR	SMART	CSFS	LSG21	LFR
EDM	0.8097	0.8051	0.8051	0.8105	0.8002	0.8048	0.8183
ATP1d	0.9220	0.9229	0.9233	0.9201	0.9299	0.9233	0.9442
ATP7d	0.8893	0.8899	0.8902	0.8871	0.8786	0.8907	0.9417
SF1	0.5395	0.5294	0.5521	0.5506	0.5381	0.5458	0.5657
SF2	0.5430	0.5457	0.5461	0.5453	0.5472	0.5428	0.5520
WQ	0.3806	0.3778	0.3796	0.3613	0.3818	0.3664	0.3863
OES97	0.9018	0.9020	0.9021	0.9014	0.9023	0.9028	0.9042
OES10	0.9412	0.9389	0.9499	0.9407	0.9473	0.9503	0.9567
SCM1D	0.9579	0.9572	0.9574	0.9574	0.9571	0.9575	0.9582
SCM20D	0.9389	0.9377	0.9386	0.9380	0.9393	0.9391	0.9401

Table 3 The aRMSE results of all algorithms experimented on the multi-output datasets

Dataset	LRLR	LRRR	SLRR	SMART	CSFS	LSG21	LFR
EDM	0.0453	0.0474	0.0474	0.0475	0.0478	0.0474	0.0450
ATP1d	0.0070	0.0070	0.0069	0.0071	0.0068	0.0069	0.0068
ATP7d	0.0067	0.0067	0.0066	0.0067	0.0070	0.0067	0.0065
SF1	0.0203	0.0205	0.0202	0.0200	0.0203	0.0202	0.0199
SF2	0.0108	0.0107	0.0107	0.0108	0.0107	0.0108	0.0107
WQ	0.0245	0.0245	0.0245	0.0246	0.0245	0.0247	0.0245
OES97	0.0163	0.0163	0.0162	0.0129	0.0162	0.0128	0.0128
OES10	0.0125	0.0130	0.0127	0.0126	0.0127	0.0126	0.0124
SCM1D	0.0011	0.0012	0.0012	0.0012	0.0012	0.0012	0.0011
SCM20D	0.0014	0.0013	0.0014	0.0013	0.0014	0.0014	0.0013

matrix. This validated that the proposed method can lead to the best faithful results over all the comparison methods for the tasks of multi-output regression.

Note that the LRLR method obtained a little smaller aCC result than all other comparing methods on the experimental datasets, which may attribute to the LRLR method can take advantage of the low-rank constraint on the rank of coefficient matrix to find the inherent low-rank structure in data, but without making the coefficient matrix sparsity. This implied that the method could use the relationship of output-output but could not conduct feature selection. Therefore, this may be the reason of resulting in it hard to make multiple-output regression for high-dimensional data.

The LRRR method and SLRR method could obtain better aCC results than the other comparing methods. Both of these two methods have the low-rank constraint and the process of feature selection via leveraging different regularization on the coefficient matrix, i.e., the LRRR uses the ℓ_2-norm on the coefficient matrix, but the SLRR imposes the $\ell_{2,1}$-norm on the coefficient matrix to lead the matrix row-sparse. However the proposed LFR method outperform these two methods, may attributes to the fact that LFR method owns the explicitly selection of the parameter p on the $\ell_{2,p}$-norm and conduct the sample selection of removing the useless samples for learning the model.

From Table 2, we can also know SMART method, CSFS method and LSG21 method always obtained the better aCC results than the LRLR method, but sometimes got little smaller aCC results than the LRRR method and the SLRR method, the reason for which may because these three methods all had the process of feature selection, however without low-rank constraint to take advantage of the correlation among multiple outputs for improving the regression model.

Moreover, to reflect the stability of algorithms, we summarize the aRMSE results of all the comparing methods in Table 3.

From Table 3, we can easily know that the proposed LFR method achieved the minimum aRMSE results compared to all the comparison methods. For example, the proposed LFR method reduced the aRMSE values on average by 0.049, 0.076, 0.068, 0.037, 0.076 and 0.037 %, respectively compared to LRLR method, LRRR method, SLRR method, SMART method, CSFS method and LSG21 method in terms of aRMSE. This demonstrates that the proposed LFR method has the best stability compared to over all the comparing methods.

The reason for our LFR method outperforms all comparing methods may be that, the proposed method fully considered the three kinds of inhered relational structure information in data, i.e., the relationship of output-output, feature-output, and sample-sample, by simultaneously leveraging the process of low-rank constraint on the coefficient matrix, feature selection, and sample selection for improving the multi-output regression.

5 Conclusion

In order to perform the multi-output regression task for large amount of high-dimensional data, this paper has proposed a new method, i.e., Leverage Triple Relational Structures via Low-rank Feature Reduction for Multi-output Regression, by simultaneously utilizing low-rank constraint to obtain the correlation among outputs, imposing $\ell_{2,p}$-norm regularizer to find the correlation between features-outputs, and combining $\ell_{2,p}$-norm on the loss function to search the correlation among samples. Moreover, an orthogonal subspace constraint has been exploited to ensure the multi-output variables share the same low-rank structure of data by rotating the results of feature selection. Consequence, the method's effectiveness was demonstrated theoretically. Finally, sets of results experimented on datasets showed that the proposed LFR method is effective to deal with the multi-output regression for large number of high-dimensional data.

Acknowledgments This work was supported in part by the China "1000-Plan" National Distinguished Professorship; the National Natural Science Foundation of China (Grant Nos: 61450001, 61263035, 61573270 and 61672177); the China 973 Program (Grant No: 2013CB329404); the China Key Research Program (Grant No: 2016YFB1000905); the Guangxi Natural Science Foundation (Grant Nos: 2012GXNSFGA060004 and 2015GXNSFCB139011); the Innovation Project of Guangxi Graduate Education (Grant Nos: YCSZ2016046 and YCSZ2016045); the Guangxi Higher Institutions' Program of Introducing 100 High-Level Overseas Talents; the Guangxi Collaborative Innovation Center of Multi-Source Information Integration and Intelligent Processing; and the Guangxi Bagui Scholar Teams for Innovation and Research Project.

References

1. Anderson T (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. Ann Math Stat 22(3):327–351
2. Argyriou A, Evgeniou T, Pontil M (2006) Multi-task feature learning. Adv Neural Inf Process Syst 41–48
3. Bache K, Lichman M UCI machine learning repository. http://archive.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Sciences
4. Borchani H, Varando G, Bielza C et al (2015) A survey on multi-output regression. Data Min Knowl 5(5): 216–233
5. Cai X, Ding C, Nie F (2013) On the equivalent of low-rank regressions and linear discriminant analysis based regressions. In: Proceedings of the 19th ACM SIGKDD, pp 1124–1132
6. Cai X, Nie F, Cai W et al (2013) New graph structured sparsity model for multi-label image annotations. In: ICCV, pp 801–808
7. Candes EJ, Recht B (2009) Exact matrix completion via convex optimization. Found Comput Math 9(6): 717–772
8. Chang X, Nie F, Yang Y et al (2014) A convex formulation for semi-supervised multi-label feature selection. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, pp 1171–1177
9. Chen B, Liu G, Huang Z et al (2011) Multi-task low-rank affinities pursuit for image segmentation. In: Proc. IEEE Int'l Conf. Computer Vision, pp 2439–2446

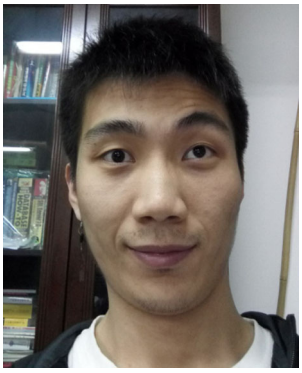
10. Dzeroski S, Demsar D, Grbovic J (2000) Predicting chemical parameters of river water quality from bioindicator data. Appl Intell 13(1):7–17
11. Feng J, Zhou L, Xu H et al (2014) Robust subspace segmentation with block-diagonal prior. In: CVPR, pp 3818–3825
12. Gao L, Song J et al (2015) Learning in high-dimensional multimedia data: the state of the art. J Multimed Syst 1–11
13. Gao L, Song J et al (2015) Optimal graph learning with partial tags and multiple features for image and video annotation. In: CVPR, pp 4371–4379
14. Gao L, Song J et al (2016) Graph-without-cut: an ideal graph learning for image segmentation. In: AAAI Conference on Artificial Intelligence, pp 1188–1194
15. Gower JC, Dijksterhuis GB (2004) Procrustes problems. Oxford Statistical Science Series, 30. Oxford, UK: Oxford University Press
16. Izenman AJ (1975) Reduced-rank regression for the multivariate linear model. J Multivar Anal 5(2):248–264
17. Karalic A, Bratko I (1997) First order regression. Mach Learn 26(2):147–176
18. Kocev D, Dzeroski S, White MD et al (2009) Using single-target and multi-target regression trees and ensembles to model a compound index of vegetation condition. Ecol Model 220(8):1159–1168
19. Nie F, Huang H, Cai X et al (2010) Efficient and robust feature selection via joint L21-norms minimization. In: Proc. NIPS, pp 1813–1821
20. Qin Y, Zhang S, Zhu X et al (2007) Semi-parametric optimization for missing data imputation. Appl Intell 27(1):79–88
21. Rai P, Kumar A, Daumé H III (2012) Simultaneously leveraging output and task structures for multiple-output regression. Adv Neural Inf Proces Syst 25:1–9
22. Rothman AJ, Levina E, Zhu J (2010) Sparse multivariate regression with covariance estimation. J Comput Graph Stat 19(4):947–962
23. Spyromitros-Xioufis E, Tsoumakas G, Groves W et al (2016) Multi-target regression via input space expansion: treating targets as inputs. Mach Learn 1–44
24. Spyromitros-Xious E, Groves W, Tsoumakas G et al (2012) Multi-label classification methods for multi-target regression. arXiv preprint arXiv:1211.6581, Cornell University Library, pp 1159–1168
25. Tuia D, Verrelst J, Alonso L et al (2011) Multioutput support vector regression for remote sensing biophysical parameter estimation. IEEE Geosci Remote Sens Lett 8(4):804–808
26. Wang S, Chang X, Li X, Sheng Q, Chen W et al (2016) Multi-task support vector machines for feature selection with shared knowledge discovery. Signal Process 120:746–753
27. Wang H, Nie F, Huang H et al (2011) Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In: Proc IEEE Int Conf Comput Vis. pp 557–562. doi:10.1109/ICCV.2011.6126288
28. Wu F, Yuan Y, Zhuang Y (2010) Heterogeneous feature selection by group lasso with logistic regression. In: ACM MM, pp 983–986
29. Zhang M, Ding C, Zhang Y et al (2014) Feature selection at the discrete limit. In: Twenty-Eighth AAAI Conference on Artificial Intelligence, pp 1355–1361
30. Zhang S, Qin Z, Ling CX et al (2005) "Missing is useful": missing values in cost-sensitive decision trees. IEEE Trans Knowl Data Eng 17(12):1689–1693
31. Zhao Y, Zhang S (2006) Generalized dimension-reduction framework for recent-biased time series analysis. IEEE Trans Knowl Data Eng 18(2):231–244
32. Zhu X, Huang Z, Cheng H et al (2013) Sparse hashing for fast multimedia search. ACM Trans Inf Syst 31(2):9.1–9.24
33. Zhu X, Huang Z, Cheng H et al (2013) Sparse hasing for fast multimedia search. ACM Trans Inf Syst 31(2)
34. Zhu X, Huang Z, Shen HT et al (2012) Dimensionality reduction by mixed kernel canonical correlation analysis. Pattern Recogn 45(8):3003–3016
35. Zhu X, Li X, Zhang S et al (2016) Robust joint graph sparse coding for unsupervised spectral feature selection. IEEE Transactions on Neural Networks and Learning Systems, PP(99), pp 1–13
36. Zhu X, Li X, Zhang S (2015) Block-row sparse multiview multilabel learning for image classification. IEEE Trans Cybern 46(2):450–461
37. Zhu X, Suk HI, Lee SW et al (2015) Subspace regularized sparse multi-task learning for multi-class neurodegenerative disease identification. IEEE Trans Biomed Eng 63(3):607–618
38. Zhu X, Zhang S, Jin Z, Zhang Z, Zuoming X (2011) Missing value estimation for mixed-attribute datasets. IEEE Trans Knowl Data Eng 23(1):110–121
39. Zhu P, Zuo W, Zhang L et al (2015) Unsupervised feature selection by regularized self-representation. Pattern Recogn 48(2):438–446

Shichao Zhang is a Distinguished Professor and the director of Institute of School of Computer Science and Information Technology at the Guangxi Normal University, Guilin, China. He holds a Ph.D. degree in Computer Science from Deakin University, Australia. His research interests include data analysis and smart pattern discovery. He has published over 50 international journal papers and over 60 international conference papers. He has won over 10 nation-class grants, such as the China NSF, China 863 Program, China 973 Program, and Australia Large ARC. He is an Editor-in-Chief for International Journal of Information Quality and Computing, and is served as an associate editor for IEEE Transactions on Knowledge and Data Engineering, Knowledge and Information Systems, and IEEE Intelligent Informatics Bulletin. Email: zhangsc@mailbox.gxnu.edu.cn

Lifeng Yang is with the Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, M. S. China. Email: 517567113@qq.com

Zhenyun Deng is with the Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, M. S. China. Email: 597277287@qq.com

Debo Cheng is with the Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, M. S. China. Email: 15676209686@163.com



Yonggang Li is with the Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, M. S. China. Email: 574717541@qq.com