

# Constructing and Combining Orthogonal Projection Vectors for Ordinal Regression

Bing-Yu Sun · Hai-Lei Wang · Wen-Bo Li · Hui-Jing Wang ·  
Jiuyong Li · Zhi-Qiang Du

Published online: 19 January 2014  
© Springer Science+Business Media New York 2014

**Abstract** Ordinal regression is to predict categories of ordinal scale and it has wide applications in many domains where the human evaluation plays a major role. So far several algorithms have been proposed to tackle ordinal regression problems from a machine learning perspective. However, most of these algorithms only seek one direction where the projected samples are well ranked. So a common shortcoming of these algorithms is that only one dimension in the sample space is used, which would definitely lose some useful information in its orthogonal subspaces. In this paper, we propose a novel ordinal regression strategy which consists of two stages: firstly orthogonal feature vectors are extracted and then these projector vectors are combined to learn an ordinal regression rule. Compared with previous ordinal regression methods, the proposed strategy can extract multiple features from the original data space. So the performance of ordinal regression could be improved because more information of the data is used. The experimental results on both benchmark and real datasets proves the performance of the proposed method.

---

B.-Y. Sun · H.-L. Wang · W.-B. Li (✉)  
Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui,  
People's Republic of China  
e-mail: bysun@ustc.edu

H.-L. Wang  
The Department of Automation, University of Science and Technology of China, Hefei,  
People's Republic of China

H.-J. Wang  
The Computer College, Shenzhen Institute of Information Technology, Shenzhen, Guangdong,  
People's Republic of China

J. Li  
The School of Computer and information Science, University of South Australia, Adelaide, Australia

Z.-Q. Du  
The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,  
Wuhan University, Wuhan, People's Republic of China

**Keywords** Ordinal regression · Linear discriminant analysis · Kernel discriminant analysis · Multiple feature combination

## 1 Introduction

Ordinal regression is to predict categories of ordinal scale and it shows resemblance to both regression and classification because labels are discrete and ordinal. In practice, ordinal labels typically correspond to linguistic terms such as ‘very bad’, ‘bad’, ‘average’, ‘good’, ‘very good’, expressing a difference in correctness, quality, beauty or any other characteristics of the analyzed objects. So far ordinal regression has been widely applied in information retrieval, collaborative filtering, medicine, psychology and other domains where human-generated data play an important role.

Some algorithms have been proposed to tackle ordinal regression problems from a machine learning perspective. Kramer et al. [1] investigated the use of a regression tree learner by mapping the ordinal scale to numeric values. However, a problem with this approach is that there might be no principled way of devising an appropriate mapping function since the true metric distances between the ordinal scales are unknown in most cases. Herbrich et al. [2] made a theoretical study on ordinal regression and applied the principle of structural risk minimization to ordinal regression. Crammer and Singer [3] generalized the online perceptron algorithm with multiple thresholds to seek the direction and thresholds for ordinal regression. Shashua and Levin [4] proposed two large margin principles, namely, fixed margin principle and sum of margins principle, to handle the direction and multiple thresholds. Chu and Keerthi [5] improved the methods of [4] and proposed two support vector ordinal regression methods by optimizing multiple thresholds to define parallel discriminant hyperplanes: one by adding the ordering of the thresholds as a constraint to the original optimization problem and the other by considering the training samples from all the ranks to determine each threshold. Lin and Lin [6] presented a reduction framework from ordinal regression to binary classification based on extended examples. They also showed that their framework provides a unified view for several existing ordinal regression algorithms. Cardoso and Costa [7] transformed the ordinal regression problem to a standard two-class problem by using the data replication method. They also instantiated this method in two important machine learning algorithms: support vector machines and neural networks. Liu et al. proposed to apply manifold learning on ordinal regression to uncover the embedded nonlinear structure of the dataset. They also introduced the multilinear extension of the proposed algorithm to support the ordinal regression of high order data like images [8]. To track feature selection problems needed in ordinal regression, Baccianella et al. presented four feature selection metrics specifically devised for ordinal regression. The proposed feature selection methods were tested two datasets of product review data [9].

It should be noticed that the above algorithms only seek one projection where the samples are projected to a line. This can be viewed as a feature extraction step which only one feature is selected. However, one feature is often insufficient for achieving the best performance. This is because some useful features fail to be extracted. So for some complex ordinal regression problems, extracting only one feature may result in a loss of useful discriminant information because there are still useful information remaining in the orthogonal subspace of the extracted feature. On the other hand, although several supervised feature extraction methods, such as linear discriminant analysis (LDA), maximum margin criterion (MMC) [10–14], can extract multiple features, these methods can be only used for solving classification problems and fail to extract features for ordinal regression problems. In Xia et al.

[15], have proposed an algorithm framework for ordinal regression that recursively extracts features from the decreasing subspace and learned a ranking rule from the examples represented by the new features. However, they used regression methods on ordinal regression and the main research of this paper is to find a feature extractor rather than a solution to ordinal regression. So how to obtain multiple projection vectors and correspondingly extract multiple features is still open.

In our previous work, we have proposed a novel Kernel discriminant algorithm for ordinal regression (KDLOR) and its performance has been proved by experimental results [16]. This algorithm, however, can only find one projection and thus only one feature can be obtained. Based on the principle of KDLOR, we propose a novel ordinal regression method by constructing and combining orthogonal projection vectors for ordinal regression in this paper. The proposed algorithm framework consist of two phases: recursively constructing projection vectors from orthogonal subspaces and combining these vectors to form a ranking rule. In the first phase, the projection vectors will be obtained recursively, step by step. At each step when a new projection vector is obtained, the next optimal projection vector are searched from the orthogonal subspaces of the obtained projection vectors. In the second phase, by the use of different combination strategy, the decision rule of each projection vector is combined to form the final decision. In comparison with other ordinal regression algorithms, the proposed algorithm can extract multiple features from the original training data. So the performance of the ordinal regression could be improved because more information of the training data is used. The experiments on both benchmark datasets and real datasets show the efficient and efficiency of the proposed method.

The idea of combination of orthogonal directions for large margin classifiers such as support vector machine have been proposed for several years and its performance have been proved by several papers [15, 17, 18]. This idea is based the fact that one single direction of maximum margin would not suffice for all classification problems. Therefore, it is desirable to eliminate this constraint completely if possible such that classifies can make full use of the multidimensional maximum margin. It is for the motivation of both of dimensionality reduction and accuracy improvement that we wish to suggest a recursive procedure for extracting multilevel margin features, recursive ordinal regression, which constitutes the main contribution of this letter. Our main idea is recursively deriving new maximum margin features by discarding all the information represented by the old maximum margin features. With the proposed method, a completely orthogonal basis of feature subspace spanned by the training samples can be derived, which is different with previous proposed ordinal regression methods.

The rest paper is organized as follows. Section 1 introduces the basics of LDA and KDLOR. Section 2 elaborates our proposed method. Section 3 presents experiments and results, and Section 4 concludes the paper. In this section we will introduce the basics of linear discriminant analysis algorithm and Kernel discriminant algorithm for ordinal regression.

### 1.1 Kernel Discriminant Analysis

Let  $(\mathbf{x}_i, y_i) \in R^l \times R, i = 1, \dots, N$  be a set of training samples where  $\mathbf{x}_i \in R^l$  denotes inputs,  $y_i \in \{1, 2, \dots, K\}$  denotes the corresponding class labels,  $N$  is the sample size and  $K$  is the number of classes. The samples from the  $i$ -th class is denoted as  $X_i$ . The objective of LDA is to find a linear projection from which different classes can be separately well. For the convenience of discussion, we define respectively a between-class scatter matrix and a within-class scatter matrix as follows:

$$S_w = \frac{1}{N} \sum_{k=1}^K \sum_{\mathbf{x} \in X_k} (\mathbf{x} - \mathbf{m}_k)(\mathbf{x} - \mathbf{m}_k)^T \tag{1}$$

$$S_b = \frac{1}{N} \sum_{k=1}^K N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T \tag{2}$$

where  $\mathbf{m}_k = \frac{1}{N_k} \sum_{\mathbf{x} \in X_k} \mathbf{x}$  denotes the mean vector of samples of the  $k$ -th class,  $N_k$  is the sample size of  $k$ -th class and  $\mathbf{m} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$  stands for the mean vector of all samples. The corresponding separability can be measured by two criteria, distances between projected means of classes (the larger, the better) and variances of data objects in every class on the projected direction (the smaller, the better). The objectives can be achieved by maximizing the following Rayleigh coefficient:

$$J(\mathbf{w}) = \arg \min_{\mathbf{w}} \frac{\mathbf{w}^T S_w \mathbf{w}}{\mathbf{w}^T S_b \mathbf{w}}. \tag{3}$$

The optimal solution to the optimization problem in (3) can be obtained by solving a generalized eigenvalue problem. In order to solve nonlinear discrimination problems, the kernel based idea, originally applied in support vector machine (SVM) [19], Kernel principal component analysis (KPCA) [20] and other kernel based algorithms can be adopted to extend LDA to its kernel version. For details, please refer to [10,21].

### 1.2 Kernel Discriminant Analysis for Ordinal Regression

Now we consider an ordinal regression problem with  $K$  ordered classes and denote these classes as consecutive integers  $Y = \{1, 2, \dots, K\}$  to keep the known rank information. The basic task here can be informally described as finding a projection where the ordinal information of classes can be preserved. To solve this problem, the formulation of the proposed method can be written as:

$$\begin{aligned} \min J(\mathbf{w}, \rho) &= \mathbf{w}^T \cdot S_w \cdot \mathbf{w} - C \cdot \rho \\ \text{s.t. } \mathbf{w}^T \cdot (\mathbf{m}_{k+1} - \mathbf{m}_k) &\geq \rho, \quad k = 1, 2, \dots, K - 1 \end{aligned} \tag{4}$$

where  $C$  is a penalty coefficient. This model tries to minimize the variances of the data for the same classes while simultaneously extending the difference between the projected means between two neighboring classes. If  $\rho > 0$ , then the projected means of all classes can be sorted in accord with their rank.

For nonlinear cases, we can make the following assumption:

$$\mathbf{w} = \sum_{i=1}^N \beta_i \phi(\mathbf{x}_i), \quad \beta_i \in \mathbb{R}, \tag{5}$$

and the original optimization problem (4) can be turned into the following problem:

$$\begin{aligned} \min J(\beta, \rho) &= \beta^T \cdot H \cdot \beta - C \cdot \rho \\ \text{s.t. } \beta^T \cdot (M_{k+1} - M_k) &\geq \rho, \quad k = 1, \dots, K - 1 \end{aligned} \tag{6}$$

where  $(M_k)_j = \frac{1}{N_k} \sum_{\mathbf{x} \in X_k} \phi(\mathbf{x}_j) \cdot \phi(\mathbf{x})$ ,  $H = \sum_{k=1}^K P_k(I - \mathbf{1}_{N_k})P_k^T$ ,  $P_k$  is a  $N \times N_k$  matrix with  $(P_k)_{i,j} = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$ ,  $\mathbf{x}_j \in X_k$ ,  $I$  is the identity matrix and  $\mathbf{1}_{N_k}$  is the matrix with all entries  $1/N_k$ . Then by using Mercer kernels with a set of functions of  $s(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$ , the problem (6) can be solved analogously to the problem in the linear cases.

## 2 Re-Weighting Orthogonal Discriminant Analysis for Ordinal Regression

In this section, we first describe an algorithm for extracting orthogonal discriminant vectors for improved ordinal regression performance. Then two efficient methods are proposed to combine these obtained discriminant vectors.

### 2.1 Orthogonal Discriminant Analysis for Ordinal Regression

In pattern classification, the orthogonality of discriminant vectors is a favorable property and this property has been used to extent classical Fisher discriminant analysis [17,18]. With these methods, a set of orthogonal discriminant vectors is computed based on a generalized optimization criterion. In the following we will describe how to extract orthogonal vectors based on Fisher Discriminant ordinal regression methods.

The proposed algorithm searches the orthogonal discriminant vectors based KDLOR criterion, i.e., the projection vector should not only preserve the ordinal information of classes, but also maximize the minimal distance between mean vectors of different classes. The first discriminant vector  $\mathbf{W}_1$  can be obtained through (4) directly. The next discriminant vector must minimize the KDLOR criterion, and simultaneously satisfies the orthogonal property. In general, the  $p + 1$  discrimination vector can be obtained through the following optimization problem:

$$\begin{aligned} \min J(\mathbf{w}_{p+1}, \rho_{p+1}) &= \mathbf{w}_{p+1}^T \cdot S_w \cdot \mathbf{w}_{p+1} - C \cdot \rho_{p+1} \\ \text{s.t. } \mathbf{w}_{p+1}^T \cdot (\mathbf{m}_{k+1} - \mathbf{m}_k) &\geq \rho_{p+1}, \quad k = 1, 2, \dots, K - 1 \\ \mathbf{w}_{p+1} \cdot \mathbf{w}_j &= 0, \quad j = 1, 2, \dots, p \end{aligned} \tag{7}$$

In nonlinear cases, we have the following assumption:

$$\mathbf{w}_p = \sum_{i=1}^N \beta_i^p \phi(\mathbf{x}_i), \quad \beta_i \in R, \tag{8}$$

and the original optimization problem (7) can be turned into the following problem:

$$\begin{aligned} \min J(\beta_{p+1}, \rho_{p+1}) &= \beta_{p+1}^T \cdot H \cdot \beta_{p+1} - C \cdot \rho_{p+1} \\ \text{s.t. } \beta_{p+1}^T \cdot (M_{k+1} - M_k) &\geq \rho_{p+1}, \quad k = 1, \dots, K - 1 \\ \beta_{p+1}^T \cdot Q \cdot \beta_j &= 0, \quad j = 1, 2, \dots, p \end{aligned} \tag{9}$$

where  $Q$  is a matrix with  $Q_{i,j} = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$ .

To solve (9), we can define the following Lagrangian equation:

$$\begin{aligned} J(\beta_{p+1}, \rho_{p+1}, \alpha_{p+1}, \alpha'_{p+1}) &= \beta_{p+1}^T \cdot H \cdot \beta_{p+1} - C \cdot \rho_{p+1} \\ &\quad - \sum_{k=1}^{K-1} \alpha_{p+1}^k \left\{ \beta_{p+1}^T \cdot (M_{k+1} - M_k) - \rho_{p+1} \right\} \\ &\quad - \sum_{j=1}^p \alpha'_{p+1}{}^j \cdot \beta_{p+1} \cdot Q \cdot \beta_j \end{aligned} \tag{10}$$

with Lagrange multipliers  $\alpha_{p+1}^k \geq 0$  and  $\alpha'_{p+1} \in R$ . To derive the Lagrange multipliers from (10), we can do the following differentiation:

$$\frac{\partial L}{\partial \beta_{p+1}} = 0 \Rightarrow \beta_{p+1} = \frac{1}{2} H^{-1} \left\{ \sum_{k=1}^{K-1} \alpha_{p+1}^k \cdot (M_{k+1} - M_k) \sum_{j=1}^p \alpha'_{p+1}{}^j \cdot Q^T \beta_j \right\}; \quad (11)$$

$$\frac{\partial L}{\partial \rho_{p+1}} = 0 \Rightarrow \sum_{k=1}^{K-1} \alpha_{p+1}^k = C. \quad (12)$$

Using (11) and (12), the corresponding optimization problem can be turned into

$$\begin{aligned} \min f(\alpha_{p+1}, \alpha'_{p+1}) &= \left\{ \sum_{k=1}^{K-1} \alpha_{p+1}^k \cdot (M_{k+1} - M_k)^T + \sum_{j=1}^p \alpha'_{p+1}{}^j \cdot \beta_j^T Q \right\} \cdot H^{-1} \\ &\times \left\{ \sum_{k=1}^{K-1} \alpha_{p+1}^k \cdot (M_{k+1} - M_k) + \sum_{j=1}^p \alpha'_{p+1}{}^j \cdot Q^T \beta_j \right\} \\ &s.t. \quad \alpha_{p+1}^k \geq 0, \quad k = 1, \dots, K - 1; \\ &\quad \sum_{k=1}^{K-1} \alpha_{p+1}^k = C. \end{aligned} \quad (13)$$

The above optimization problem is a convex quadratic programming (QP) one with linear constraints. To solve this problem, a variety of methods, such as interior point, active set and conjugate gradient methods can be used. The procedure of the whole algorithm is as outlined in Algorithm 1 below.

**Algorithm 1:**

- 1: Input:
- 2: Given training dataset  $\{X, Y\}$ ; the number of projection vector  $P$
- 3: For  $i = 1$  to  $P$ , do:
- 4: if  $i = 1$ , get  $w_1, \rho_1$  through (6), else, get  $w_i, \rho_i$  through (13);
- 5: end for;
- 7: Outputs:  $w_1, w_2, \dots, w_P$ .

Although it is still open to determine the value of  $P$ , We noticed that the solution can be reached quickly and generally the value of  $P$  needed not to be very large. In our experiments, we set  $P = 10$  and it caused little deviation to results. So compared to ensemble learning methods, such as Bagging, which need to train dozens of base regression models, the computational cost of the proposed method is lower.

2.2 Solving the Small Sample Size Problems

In fact, generally, the dimensionality of data is larger than the sample size, which is the case for many high-dimensional and low sample size data. In these cases  $S_w$  is singular and it is generally known as small sample size (SSS) problem. Recently, the SSS problem has been extensively addressed in classical LDA, and many solutions have been proposed. A simple method to address the SSS problem is by performing PCA projection to reduce the dimension of the feature space and make  $S_w$  nonsingular. However, there are two problems for this solution:

1. The accuracy depends very much on the dimension of the reduced PCA subspaces and how to determine the optimal dimension of this subspace remains largely an open problem;
2. Some useful information for LDA may be compromised in the intermediate PCA stage.

A more efficient method to solve this problem is to adopt a regularization method. This is to add a constant  $u > 0$  to the diagonal elements of  $S_w$  as  $S_w = S_w + uI$ , where  $I$  is an identity matrix [22,23]. The optimum value of  $u$  can be estimated through a cross validation method.

### 2.3 Decision Combination for Orthogonal Discriminant Ordinal Regression

After obtaining  $\alpha_p, \alpha'_p$ , the optimal direction  $\mathbf{w}_p$  can be calculated by substituting  $\alpha_p, \alpha'_p$  into (8) and (11). When only one direction  $\mathbf{w}$  is obtained, the rank of an unseen input vector  $\mathbf{x}$  can be predicted by the following decision rule:

$$f(x) = \min_{k \in \{1, \dots, K\}} \{k : \mathbf{w} \cdot \mathbf{x} - b_k < 0\} \tag{14}$$

where  $b_k$  is defined as  $b_k = \mathbf{w}(\mathbf{m}_{i+1} + \mathbf{m}_i)/2$  or  $b_k = \mathbf{w}(N_{k+1}\mathbf{m}_{k+1} + N_k\mathbf{m}_k)/(N_{k+1} + N_k)$ . However, when there are several direction available, we should first calculate  $\mathbf{b}^j = \{b_1^j, b_2^j, \dots, b_k^j\}^T$  for each  $\mathbf{w}_j$  using (14) and then combine them to form the final decision. In the following we will provide two combination methods.

#### 2.3.1 Majority Voting Method

Majority voting is the simplest method for combining directions. Suppose we have  $p$  directions  $\mathbf{w}_j, j = 1, \dots, p$  and corresponding  $\mathbf{b}^j$ . For each direction, we can calculate  $f_j(\mathbf{x})$  of a sample  $\mathbf{X}$  using (14). Let  $n_r = \#\{f_j(\mathbf{x}) = r\}$ , i.e., the number of directions whose decision are known to be  $r$ . Then the final rank of  $\mathbf{x}$  can be determined by

$$f(\mathbf{x}) = \arg \max_r n_r \tag{15}$$

#### 2.3.2 Weighted Average Method

Majority voting assumes all directions having the same importance in its combined decision making. For a specific ordinal regression problem, however, this assumption is not always valid. In practice, different directions may have different efficiencies and an important issue in the decision combination is how to derive a weighted scheme to balance the relative importance of different directions. This can be achieved by a weighted method, i.e., associating larger weights with more efficient projection vectors and smaller weights with less important ones. In the following we will describe how to obtain the weights of different directions. The weighted average methods is to associate weight  $\omega_j$  to  $\mathbf{w}_j$  and the rank of  $\mathbf{x}$  can be predicted as the following:

$$f(\mathbf{x}) = \min_{k \in \{1, \dots, K\}} \left\{ k : \sum_{j=1}^p \omega_j (\mathbf{w}_j \cdot \mathbf{x} - b_k^j) < 0 \right\} \tag{16}$$

where  $\omega_j, j = 1, \dots, p$  satisfy both (17) and (18)

$$0 \leq \omega_j \leq 1 \tag{17}$$

$$\sum_{j=1}^p \omega_j = 1 \tag{18}$$

Now the key issue is how to derive the value of  $\omega_j$ . From (16) it can be seen that different weights may result in different performance of the combined directions. The selection of the weights should provide the combined combined directions with better performance on infinite testing data, i.e., good generalization ability. To realize the above, we can build the following problem:

$$\begin{aligned} & \min \left\{ \sum_{j=1}^p \omega_j^2 \cdot \mathbf{w}_j^2 + \lambda \left( \sum_{i=1}^N (\xi_i^2 + \xi_i'^2) \right) \right\} \tag{19} \\ & \text{s.t. } \sum_{j=1}^p \omega_j b_{k-1}^j - \xi_i \leq \sum_{j=1}^p \omega_j \mathbf{w}_j \cdot \mathbf{x}_i \leq \sum_{j=1}^p \omega_j b_k^j + \xi_i', \\ & \quad \forall \mathbf{x}_i, \text{ if } y_i = k, i = 1, \dots, n;', \\ & \quad 0 \leq \omega_j \leq 1', \\ & \quad \sum_{j=1}^p \omega_j = 1 \end{aligned}$$

where  $\lambda > 0$  is a constant to control the tradeoff between the generalization  $\sum_{j=1}^p \omega_j^2 \cdot \mathbf{W}_j^2$  and the error  $\sum_{i=1}^N (\xi_i^2 + \xi_i'^2)$ . In nonlinear cases, by using (8), we can turn (19) as the following:

$$\begin{aligned} & \min \left\{ \sum_{j=1}^p \omega_j^2 \cdot \alpha_j' \cdot \mathbf{Q} \cdot \alpha_j + \lambda \left( \sum_{i=1}^N (\xi_i^2 + \xi_i'^2) \right) \right\} \tag{20} \\ & \text{s.t. } \sum_{j=1}^p \omega_j b_{k-1}^j - \xi_i \leq \sum_{j=1}^p \omega_j \alpha_j' \cdot \mathbf{Q} \cdot \mathbf{i} \leq \sum_{j=1}^p \omega_j b_k^j + \xi_i', \\ & \quad \forall \mathbf{x}_i, \text{ if } y_i = k, i = 1, \dots, n;', \\ & \quad 0 \leq \omega_j \leq 1', \\ & \quad \sum_{j=1}^p \omega_j = 1 \end{aligned}$$

where  $\mathbf{Q} \cdot \mathbf{i}$  is the  $i$ -th column of  $\mathbf{Q}$ . Both problem (19) and (20) are quadratic programming problems and their global optimal value of  $\omega$  can be computed easily. The value of  $\lambda$  can be set via a cross-validation approach.

Now we present an analysis of the computational complexity above combination approach. As mentioned in Sect. 2.1, the proposed combination method involves solving a QP problem. The complexity of the proposed combination method is determined primarily by the QP problem whose complexity is related mainly to the size of the Hessian matrix. However, the complexity of the proposed combination method can be made smaller because the size of the Hessian matrix, as explained below, can be reduced.



In practice, the accuracy of each direction is not bad. Thus, many of the training data can be predicted correctly by the individual directions; for all the  $p$  directions which can predict the rank of the training data correctly, the following relation should hold:

$$b_{k-1}^j \leq \mathbf{w}_j \cdot \mathbf{x}_i \leq b_k^j, \quad \forall \mathbf{x}_i, \text{ if } y_i = k, i = 1, \dots, n; \tag{21}$$

From (17) and (18), it can be seen that for all the training data satisfying (21), the following inequality should always hold:

$$\sum_{j=1}^p \omega_j b_{k-1}^j \leq \sum_{j=1}^p \omega_j \mathbf{w}_j \cdot \mathbf{x}_i \leq \sum_{j=1}^p \omega_j b_k^j, \tag{22}$$

$$\forall \mathbf{x}_i, \text{ if } y_i = k, i = 1, \dots, n;$$

So while applying the combination method, we can ignore all the training data satisfying condition (21) in solving the QP problem (20). In this way, many of the training data can be excluded; so the size of the Hessian matrix of the QP problem can be reduced considerably. Thus, the complexity of our combination method could be made smaller significantly.

### 3 Experimental Results

To evaluate the performance of the proposed method, we performed a set of experiments with both benchmark datasets and real ordinal regression datasets. We compared the proposed methods, namely, majority voting ordinal regression (MJ-OR) and weighted average ordinal regression (WA-OR) against kernel discriminant learning for ordinal regression (KDLOR), on these benchmark datasets. The following Gaussian kernel was used in our experiments:

$$s(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma^2}\right). \tag{23}$$

A Bagging method [24], which trains a number of base KDLOR from a different bootstrap sample, was also used and its performance was compared. In this paper we randomly select about half samples from full training dataset for 30 times and the final ordinal regression decision is reached based on the vote of the obtained 30 component regression models. The ten-fold cross validation was used to determine parameters  $\mu, \sigma, C$  and  $\lambda$ . Two evaluation metrics are considered to quantify the accuracy of predicted ordinal scales  $\{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N\}$  with respect to true targets  $\{y_1, y_2, \dots, y_N\}$ . The tolerance error  $\varepsilon$  of algorithm 1 is set to 0.01 and the maximal number of iteration  $\mathbf{T}$  is set to 5.

1. Mean absolute error (MAE): the average deviation of the prediction from the true rank which is treated as consecutive integers, i.e.  $\frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|$ ;
2. Mean zero-one error (MZE): the fraction of incorrect predictions.

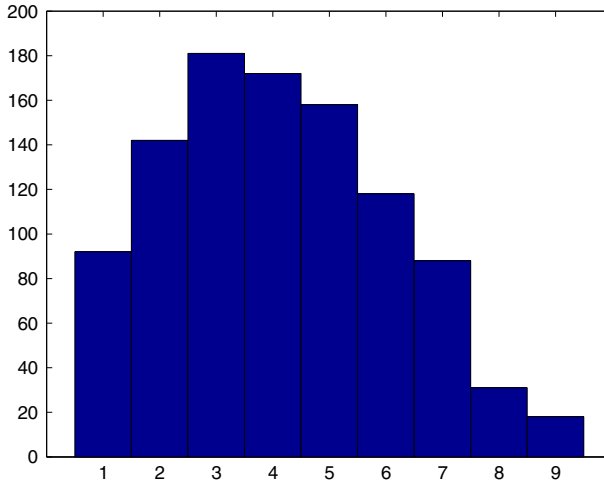
#### 3.1 Benchmark datasets

For the comparison purpose, we used the same datasets as in [5,6,16]. These five benchmark datasets were generated by quantizing some metric regression data sets with  $K = 10$ . The same training/test ratio was used and we averaged the results over 20 trials to fairly compare our results with those of other algorithms. The test result is listed in Table 1. It can be seen that generally the combination of several projections can achieve higher performance in

**Table 1** Test results of three algorithms on five datasets

Datasets	Mean zero one error			Mean absolute error				
	KDLOR	MJ-OR	WA-OR	Bagging	KDLOR	MJ-OR	WA-OR	Bagging
Pyrimidines	0.739 ± 0.05	0.73 ± 0.025	<b>0.65 ± 0.045</b>	0.73 ± 0.03	1.1 ± 0.1	1.2 ± 0.04	<b>0.97 ± 0.05</b>	1.1 ± 0.01
MachineCPU	0.48 ± 0.01	0.4 ± 0.04	<b>0.38 ± 0.02</b>	0.47 ± 0.02	0.69 ± 0.015	0.78 ± 0.2	<b>0.64 ± 0.09</b>	0.68 ± 0.02
Boston	0.56 ± 0.02	0.59 ± 0.03	<b>0.52 ± 0.08</b>	0.56 ± 0.01	0.7 ± 0.035	0.88 ± 0.03	<b>0.69 ± 0.09</b>	0.7 ± 0.02
Abalone	0.74 ± 0.02	0.74 ± 0.05	0.74 ± 0.05	0.74 ± 0.01	1.4 ± 0.05	1.6 ± 0.02	<b>1.3 ± 0.015</b>	1.38 ± 0.03
Bank	<b>0.74 ± 0.001</b>	0.745 ± 0.0025	0.75 ± 0.02	0.742 ± 0.001	1.45 ± 0.02	1.56 ± 0.02	1.56 ± 0.02	<b>1.42 ± 0.03</b>

The best performance of each dataset for different methods is in bold



**Fig. 1** Class distribution for the 488 examples of the ESL data set

comparison with the performance achievable by a single projection. Moreover, the table also shows that the weighted average method outperforms the majority voting method generally. For few datasets, such as Bank, the single projection vector achieves the best performance. The reason may be that sometime the use of several projection vectors may results in over-fitting problems. So our future work will find an efficient method to determine the optimal number of projection vectors.

### 3.2 Real Datasets

In this section, we continue the experimental study by applying the algorithms to three real ordinal regression dataset, namely, ESL dataset, wine dataset and DLBCL dataset. The first two datasets are available at the WEKA website<sup>1</sup> and UCI Machine Learning Repository.<sup>2</sup> The last dataset can be found in [25].

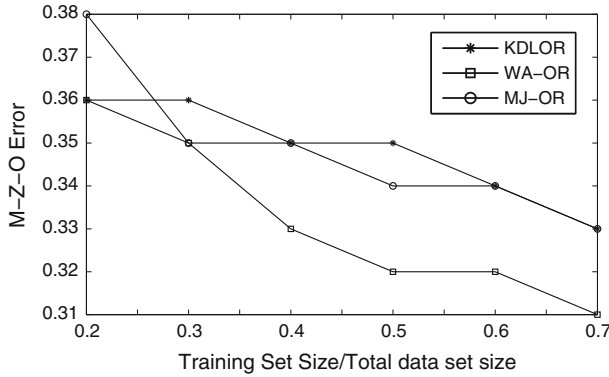
#### 3.2.1 ESL Dataset

The ESL data set contains 488 profiles of applicants for certain industrial jobs. Expert psychologists of a recruiting company, based upon psychometric test results and interviews with the candidates, determined the values of the input attributes. The output is the an overall score corresponding to the degree of fitness of the candidate to this type of job. The number of instances is 488 and the number of features is 4. The class distribution is shown as Fig. 1.

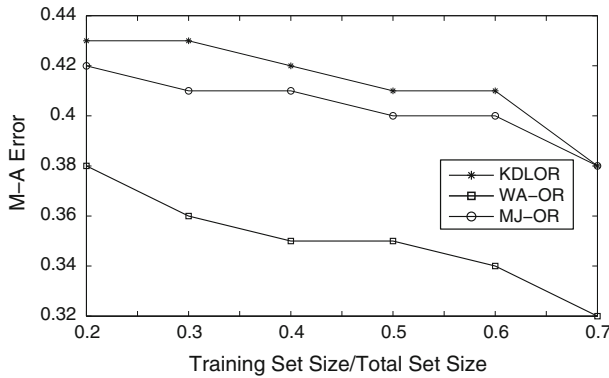
To investigate the influence of the number of training instances on different algorithms relative performance, we repeated the experiments using different training set size. Figures 2 and 3 shows the mean zero-one error rate and mean absolute error rate of three algorithms respectively when different number of data are used. The overall results suggest that the use of multiple projections can improve the ordinal regression performance.

<sup>1</sup> <http://www.cs.waikato.ac.nz/ml/index.html>.

<sup>2</sup> <http://www.ics.uci.edu/~mllearn/MLRepository.html>.



**Fig. 2** M-Z-O Error rates of different algorithms for ESL dataset when different training set sizes are used set

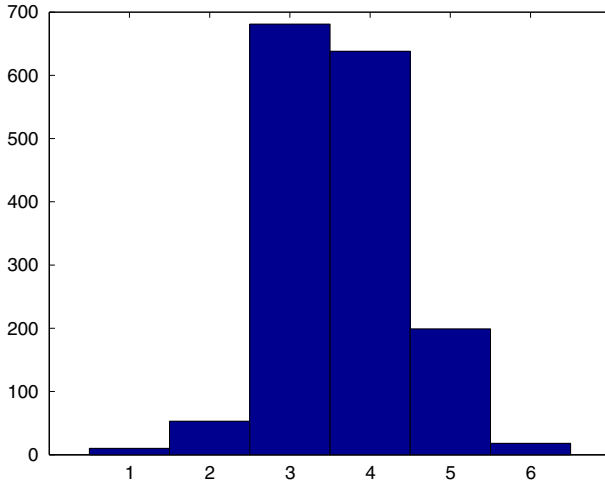


**Fig. 3** M-A Error rates of different algorithms for ESL dataset when different training set sizes are used set

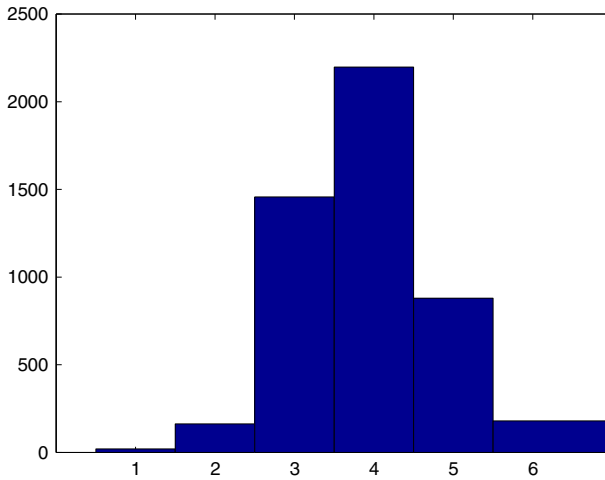
### 3.2.2 Wine Dataset

The Wine datasets are related to predict human wine taste preferences on red and white variants of the Portuguese "Vinho Verde" wine. Since the red and white tastes are quite different, the analysis is performed separately and two datasets were built with 1,599 red and 4,898 white examples. Each sample was characterized from 11 attribute, include fixed acidity, volatile acidity and so on. The preferences of the samples were evaluated by a minimum of three sensory assessors (using blind tastes), which graded the wine in a scale that ranges from 0 (very bad) to 10 (excellent). The final sensory score is given by the median of these evaluations. Figures 4 and 5 shows the class distribution of the examples of the red wine and white wine respectively. It can be seen that the distribution of the differences is unbalanced, which increase the difficulty of the ordinal regression problems.

We also repeated the experiments using different training set size on these two wine datasets. Figures 6, 7, 8 and 9 show the experimental results of different algorithms when different training set sizes are used for these two datasets respectively. Compared to the performance of a single projection vector, the use of multiple projection vectors can decrease both M-Z-O Error and M-A Error, which proved the efficient and efficiency of the proposed method.



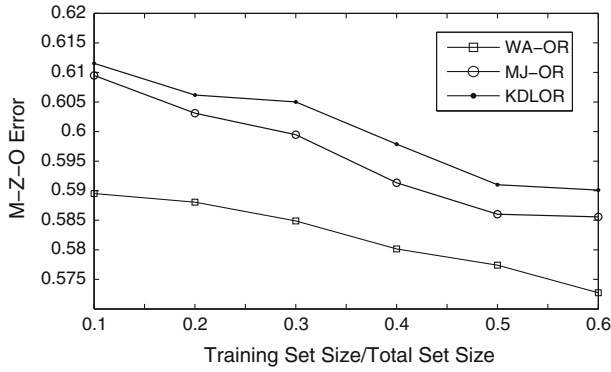
**Fig. 4** Class distribution for the 1,599 examples of the red wine data set



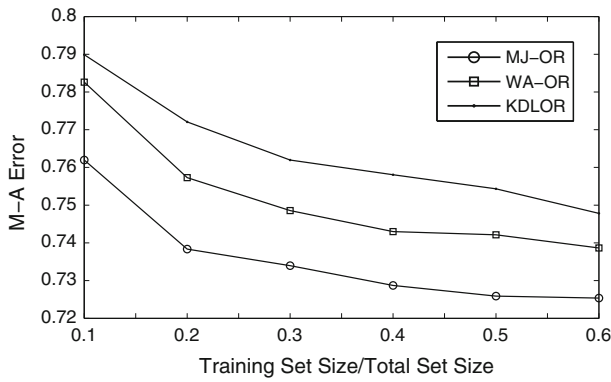
**Fig. 5** Class distribution for the 4,898 examples of the white wine data set

### 3.2.3 DLBCL Dataset

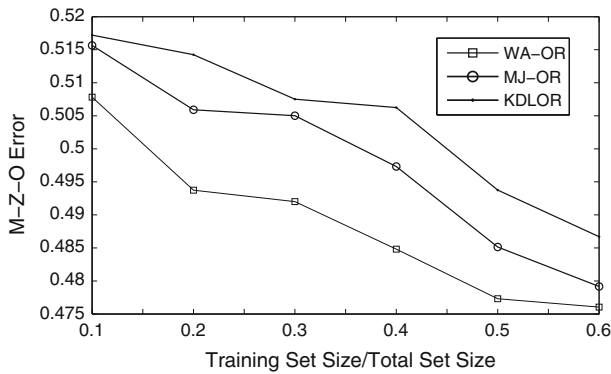
The DLBCL data set consists of measurements of 7,399 genes from 240 patients. A survival time was recorded for each patient, which ranges between 0 and 21.8 years. Among them, 138 patients were used for their exact survival time is known while others are censored. We quantized the survival time into 4 scales by 4. Figure 10 shows the class distribution of the examples. Table 2 shows the test result when different methods are used. It can be seen that the strategy proposed in this paper can also improve the ordinal regression performance for this high dimensional dataset. Compared with the Bagging method, although the MA error of the proposed WA-OR algorithm is higher, its M-Z-O error is lower. So the combination of different projection vectors can improve the ordinal regression performance in most cases.



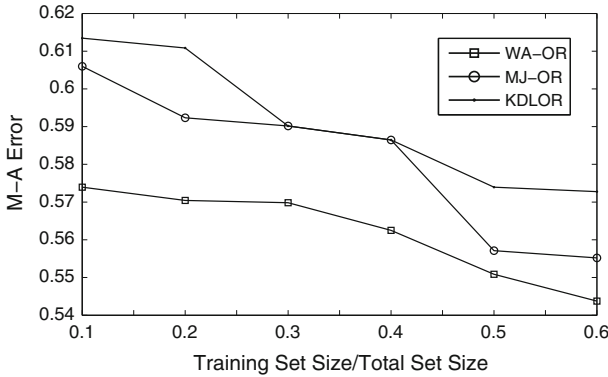
**Fig. 6** M-Z-O error rates of different algorithms for white wine dataset when different training set sizes are used



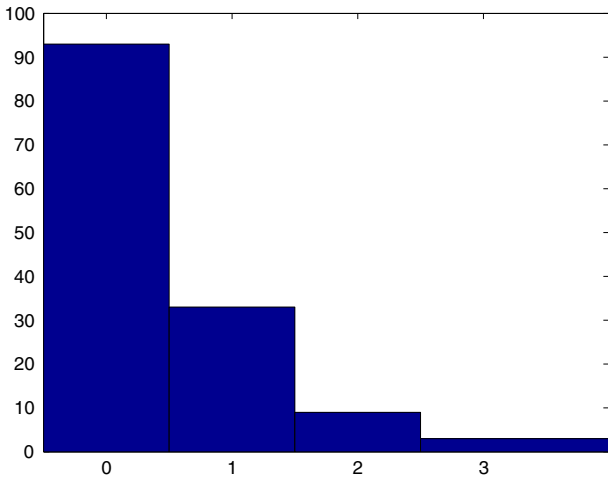
**Fig. 7** M-A error rates of different algorithms for white wine dataset when different training set sizes are used



**Fig. 8** M-Z-O error rates of different algorithms for red wine dataset when different training set sizes are used



**Fig. 9** M-A error rates of different algorithms for red wine dataset when different training set sizes are used



**Fig. 10** Class distribution for the 138 examples of the DLBCL data set

**Table 2** Experimental results on DLBCL dataset for different methods

Method	M-Z-O error	M-A error
KDLOR	0.49	0.58
MJ-OR	0.46	0.55
WA-OR	0.44	0.54
Bagging	0.46	0.52

### 4 Conclusions

In this paper, we proposed a novel ordinal regression strategy by extracting and combining orthogonal projection vectors. Traditional ordinal regression methods usually seek only one projection vector; so their performance may be unsatisfactory in case of complex problems because some useful information is lost in the sample space. On the other hand, although several methods, such as LDA, MMC, can obtain multiple projection vectors and several

features can be extracted, they do not handle ordinal classes. To address the above problems, we generalized the KDLOR algorithm of our previous work to extract multiple projection vectors, thus more than one feature can be obtained. The proposed algorithm is to first extract a projection vector and then the next optimal vector is searched within the orthogonal space of the obtained projection vector. We also developed two efficient strategy to combine multiple projection vectors for improving the performance. In comparison with existing ordinal regression methods, the proposed method can extract several projection vectors from the original samples spaces and more information of the data could be used. Experimental results on both benchmark datasets and real-world benchmark data sets demonstrate that the proposed method could improve the performance of the ordinal regression. Designing a more efficient algorithm for large and high dimensional datasets and how to determine the optimal number of the extracted features remain our future study.

**Acknowledgments** The authors sincerely thank anonymous reviewers' constructive comments. The work of this paper has been supported by the Natural Science Foundation of China (Nos: 41101516 and 61203373), Guangdong Natural Science Foundation (No. S2011010006120) and the Shenzhen Science and Technology R & D funding Basic Research Program (No. JC201105190821A).

## References

1. Kramer S, Widmer G, Pfahringer B, DeGroeve M (2001) Prediction of ordinal classes using regression trees. *Fundamenta Informaticae* 47(1–2):1–13
2. Herbrich R, Graepel T, Obermayer K (2000) Large margin rank boundaries for ordinal regression. In: Smola AJ, Bartlett PL, Schölkopf B, Schuurmans D (eds) *Advances in large margin classifiers*. MIT Press, Cambridge, pp 115–132
3. Crammer K, Singer Y (2002) Pranking with ranking. In: Dietterich TG, Becker S, Ghahramani Z (eds) *Advances in neural information processing systems*. MIT Press, Cambridge, pp 641–647
4. Shashua A, Levin A (2003) Ranking with large margin principle: two approaches. In: Becker S, Thrun S, Obermayer K (eds) *Advances in neural information processing systems*. MIT Press, Cambridge, pp 961–968
5. Chu W, Keerthi SS (2005) New approaches to support vector ordinal regression. In: *Proceedings of the 22nd international conference on machine learning (ICML 2005)*. Omnipress, pp 145–152
6. Lin L, Lin H-T (2007) Ordinal regression by extended binary classification. In: *Advances in neural information processing systems 19: proceedings of the 2006 Conference (NIPS 2006)*. MIT Press, pp 865–872
7. Cardoso JS, Pinto da Costa JF (2007) Learning to classify ordinal data: the data replication method. *J Mach Learn Res* 8:1393–1429
8. Liu Y, Liu Y, Chan KCC (2011) Ordinal regression via manifold learning. In: *Proceedings of 25th AAAI conference on artificial Intelligence (AAAI11)*, pp 398–403
9. Baccianella S, Esuli A, Sebastiani F (2010) Feature selection for ordinal regression. In: *Proceedings of the 2010 ACM symposium on applied computing (SAC '10)*. ACM, New York, pp 1748–1754
10. Bishop CM (2006) *Pattern recognition and machine learning*. Springer, Heidelberg
11. Duda RO, Hart PE, Stork D (2000) *Pattern classification*. Wiley, Chichester
12. Li H, Jiang T, Zhang K (2006) Efficient and robust feature extraction by maximum margin criterion. *IEEE Trans Neural Netw* 17(1):157–165
13. Min W, Lu K, He X (2004) Locality pursuit embedding. *Pattern Recognit* 37(4):781–788
14. Zhang T, Huang K, Li X, Yang J, Tao D (2010) Generalized discriminant analysis: a matrix exponential approach. *IEEE Trans Syst Man Cybern B* 40(1):253–263
15. Xia F, Tao Q, Wang J, Zhang W (2007) Recursive feature extraction for ordinal regression. In: *International joint conference on neural networks (IJCNN'07)*, pp 78–83
16. Sun B-Y, Li J, Wu DD, Zhang X-M, Li W-B (2010) Kernel discriminant learning for ordinal regression. *IEEE Trans Knowl Data Eng* 22(6):906–910
17. Ye J (2005) Characterization of a family of algorithms for generalized discriminant analysis on under-sampled problems. *J Mach Learn Res* 6:4831502



18. Ji S, Ye J (2008) Generalized linear discriminant analysis: a unified framework and efficient model selection. *IEEE Trans Neural Netw* 19(10):1768–1782
19. Vapnik V (1998) *The nature of statistical learning theory*. Wiley, New York
20. Muller K-R, Mika S, Ratsch G, Tsuda K, Scholkopf B (2001) An introduction to Kernel-based learning algorithms. *IEEE Trans Neural Netw* 12(2):181–201
21. Mika S (2002) *Kernel fisher discriminants*. PhD thesis, University of Technology, Berlin
22. Guo Y, Hastie T, Tibshirani R (2007) Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* 8(1):86–100
23. Kim H, Drake B, Park H (2006) Adaptive nonlinear discriminant analysis by regularized minimum squared errors. *IEEE Trans Knowl Data Eng* 18(5):603–612
24. Breiman L (1996) Bagging predictors. *Mach Learn* 24(2):123–140
25. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Staudt LM (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med* 346(25):1937–1947