ICIC Express Letters
Part B: Applications
Volume **8**, Number **3**, March **2017**

ICIC International ©2017 ISSN 2185-2766
pp. 545–552

LOCAL DISCRIMINANT ANALYSIS ORDINAL REGRESSION

Xiaoming Wang, Zengxi Huang and Yajun Du

School of Computer and Software Engineering
Xihua University
No. 999, Jinzhou Road, Jinniu District, Chengdu 610039, P. R. China
wxmwm@aliyun.com

Received July 2016; accepted October 2016

Abstract. *Kernel discriminant learning ordinal regression (KDLOR) is a method designed to tackle the ordinal regression problems. However, it ignores the local structure characteristics of the data and might provide undesired results. In this paper, we propose a novel ordinal regression method called local discriminant analysis ordinal regression (LDAOR). The proposed method explicitly incorporates both the local structure characteristics and the discriminant information in the data space, and so achieves better generalization performance over its counterparts. In the paper, we first discuss the linear model of LDAOR and then develop its nonlinear version by using the representation theorem for reproducing kernel Hilbert spaces (RKHS). The experimental results on several synthetic and benchmark datasets validate the effectiveness of LDAOR.*
Keywords: Machine learning, Ordinal regression, Discriminant analysis, Local structure characteristic

1. **Introduction.** Machine learning often involves a type of problems in which ones want to explore an order among different categories. It is referred to as ordinal regression (OR). OR is actually a type of supervised learning problems [1-3]. In OR, the labels of the data points are discrete and ordinal. Thus, it differs from the traditional regression or classification problems. OR has been successfully applied in a wide range of machine learning applications such as information retrieval [4], medical analysis [5], facial age estimation [6], facial beauty assessment [7], image classification [8], and text classification [9].

During the past decade, many methods have been designed to deal with the OR problems. In [10], Sun et al. extended the kernel discriminant analysis (KDA) algorithm to handle the OR problems. This method is referred to as kernel discriminant learning ordinal regression (KDLOR). KDLOR shows inspiring generalization ability in solving the OR problems. However, if the data points in some categories are multimodal, then KDLOR tends to provide undesired results. The reason is that KDLOR ignores the local structure characteristic of the data. In [11], Liu et al. further extended the manifold learning idea to deal with OR and proposed a method called ordinal regression method via manifold learning (ORML). In contrast with KDLOR, this method further takes fully consideration of the local structure characteristics of the data space. However, this method does not explicitly utilize the discriminant information of the data in its objective function and so is unsupervised. As a result, ORML sometimes gives some unreasonable results.

In this paper, we propose a novel ordinal regression learning method called local discriminant analysis ordinal regression (LDAOR). In contrast with KDLOR, the model of LDAOR incorporates the local structure characteristics of the data space. And it utilizes the discriminant information in the training data but ORML does not. These are the key differences between LDAOR and its counterparts. On the other hand, LDAOR is similar to KDLOR and ORML in the form of the optimization problem, i.e., its optimization problem is also a convex QP problem with linear constraints and many techniques can

be employed to solve it. In the paper, we first formulate the linear model of LDAOR and discuss how to solve it. Then, we develop its nonlinear version. The experimental results indicate that KDLOR is effective and can achieve superior generalization performance over its counterparts.

The rest of this paper is organized as follows. In Section 2, the linear model of LDAOR is first formulated and how to solve it is then discussed. After that, the nonlinear version of LDAOR is developed in Section 3. The experimental results are reported in Section 4. Finally, conclusions are drawn in Section 5.

2. The Proposed Ordinal Regression Method.

In this paper, we will address an OR problem with C ordinal categories. The training data contains N data points and is denoted by $\{(\mathbf{x}_i, y_i)|\mathbf{x}_i \in R^d,\ y_i \in \{1, \ldots, C\},\ i = 1, \ldots, N\}$, where $\mathbf{x}_i$ is the ith input data point and y_i represents its order. Here d is the dimension of the data space.

2.1. Constructing the locality within-class scatter matrix.

In order to develop our method, we first construct a matrix to capture the local structure characteristics of the data. For the above given training data, we define a matrix as follows

$$\widehat{\mathbf{S}}^c = \sum_{\mathbf{x}_i \in \mathbf{X}^c} \sum_{\mathbf{x}_j \in \mathbf{X}^c} A_{ij}^c (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \tag{1}$$

where $\mathbf{X}^c = \{\mathbf{x}_i | y_i = c,\ i = 1, \ldots, N^c\}$, the matrix $\mathbf{A}^c$ is the affinity matrix constructed to model the neighborhood relationship between the data points in the cth category and we will detail how to construct it in Section 2.3. We refer to the above matrix $\widehat{\mathbf{S}}^c$ as locality scatter matrix of the c-category data. Further, we define the following matrix

$$\widehat{\mathbf{S}}_{\mathrm{W}} = \sum_{c=1}^{C} \widehat{\mathbf{S}}^c \tag{2}$$

We refer to the matrix as locality within-class scatter matrix. This matrix incorporates the local structure characteristics of the data since it uses the locality scatter matrix of each category. Moreover, it utilizes the discriminant information of the data and is supervised.

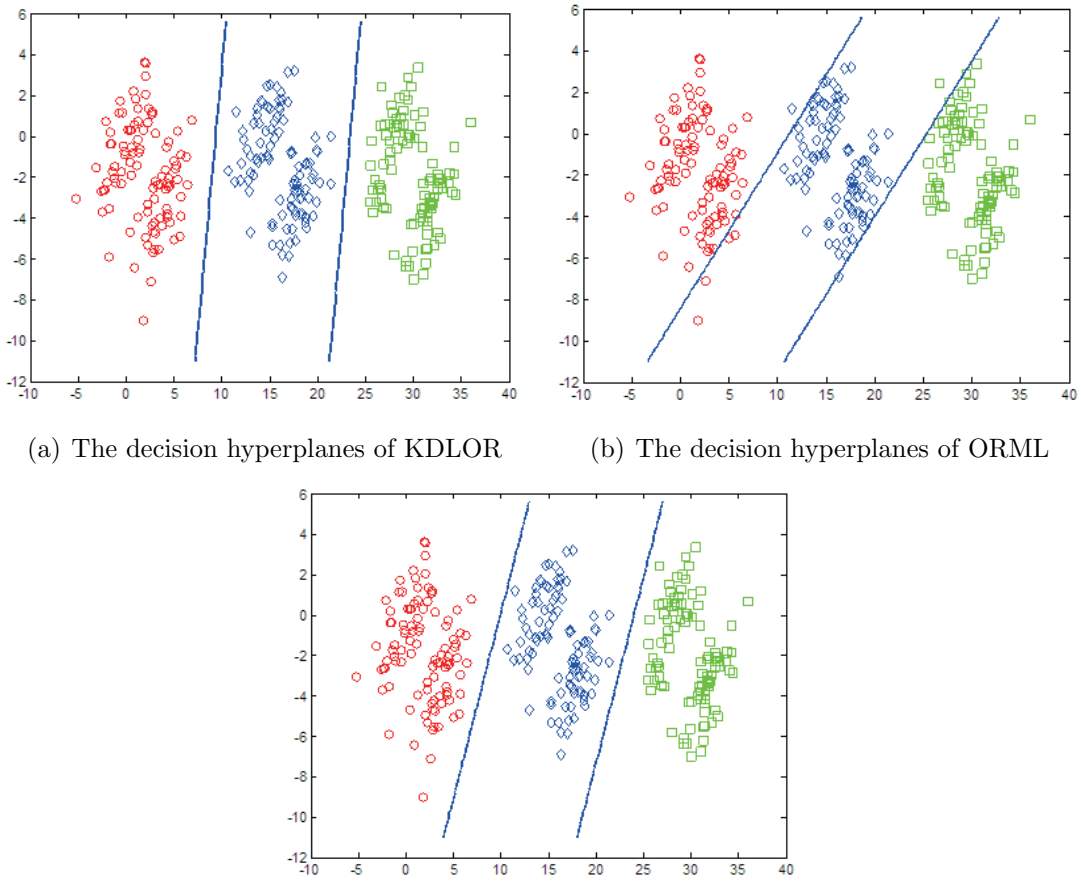
2.2. Formulation of the proposed ordinal regression method.

By following the basic idea of KDLOR and using the above defined matrix $\widehat{\mathbf{S}}_{\mathrm{W}}$ in (2), we define the primal optimization problem of LDAOR as follows

$$\begin{aligned} &\min_{\mathbf{w}} \mathbf{w}^T \widehat{\mathbf{S}}_{\mathrm{W}} \mathbf{w} - \lambda\rho \\ &\text{s.t. } \mathbf{w}^T(\mathbf{u}^{c+1} - \mathbf{u}^c) \geq \rho,\ c = 1, 2, \ldots, C-1 \end{aligned} \tag{3}$$

where λ is a penalty parameter. As KDLOR, LDAOR tries to construct $C-1$ hyperplanes $\mathbf{w}^T\mathbf{x} + b_c = 0$ which have the same direction $\mathbf{w}$ and different constraint thresholds b_c $(c = 1, \ldots, C)$. However, our method introduces the locality within-class scatter matrix $\widehat{\mathbf{S}}_{\mathrm{W}}$ in the objective function instead of the within-class scatter matrix $\mathbf{S}_{\mathrm{W}}$ in KDLOR. In this way, the local structure characteristics of the data are taken fully into consideration. LDAOR also explicitly utilizes the discriminant information contained in the training data since the used matrix $\widehat{\mathbf{S}}_{\mathrm{W}}$ in its objective function is supervised, whereas OLML does not.

Figure 1 illustrates the differences between LDAOR and the other two methods. Here, we consider a synthetic OR task with 3 ordered categories and each category of which consists of 100 data points. Figure 1(a) shows the decision hyperplanes of KDLOR on the data and Figure 1(b) illustrates ones generated by ORML. Figure 1(c) is the experimental results of LDAOR. Obviously, the hyperplanes of LDAOR are more reasonable in contrast with ones of LDAOR and ORML and they reflect the local characteristic of the data. This example clearly demonstrates the limitations of KDLOR and ORML and the advantage of LDAOR over them.



(a) The decision hyperplanes of KDLOR

(b) The decision hyperplanes of ORML

(c) The decision hyperplanes of LDAOR

FIGURE 1. The decision hyperplanes of KDLOR, ORML and LDAOR on a synthetic data

Similarly to KDLOR, through using the duality theory [12], if the matrix $\mathbf{S}_{\mathrm{W}}$ is non-singular or invertible, we can obtain the Wolf dual problem of (3) as follows

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \sum_{c=1}^{C-1} \alpha_c (\mathbf{u}^{c+1} - \mathbf{u}^c)^T \widehat{\mathbf{S}}_{\mathrm{W}}^{-1} \sum_{c=1}^{C-1} \alpha_c (\mathbf{u}^{c+1} - \mathbf{u}^c) \\ &\text{s.t. } 0 \le \alpha_c \le \lambda,\ c = 1, \ldots, C-1 \\ &\quad \sum_{c=1}^{C-1} \alpha_c = \lambda \end{aligned} \tag{4}$$

This is a convex QP problem and can be solved through the same technique as in KDLOR. Suppose $\boldsymbol{\alpha}^*$ is the solution of the above optimization problem, $\mathbf{w}$ is obtained as follows

$$\mathbf{w} = \widehat{\mathbf{S}}_{\mathrm{W}}^{-1} \sum_{c=1}^{C-1} \alpha_c^* (\mathbf{u}^{c+1} - \mathbf{u}^c) \tag{5}$$

and so the discriminant function value for an unknown data point $\mathbf{x}$ is

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} = \sum_{c=1}^{C-1} \alpha_c^* (\mathbf{u}^{c+1} - \mathbf{u}^c) \widehat{\mathbf{S}}_{\mathrm{W}}^{-1} \mathbf{x} \tag{6}$$

Thus, the predictive ordinal decision function is given by

$$\min_c \arg\{c : f(\mathbf{x}) < b_c\} \tag{7}$$

Here b_c is the threshold and we employ the same strategy as in [10] to compute it, i.e.,

$$b_c = \mathbf{w}^T(\mathbf{u}^{c+1} - \mathbf{u}^c)/2 \tag{8}$$

Note, in the practical applications, as in KDLOR, LDAOR may encounter the singularity problem of $\widehat{\mathbf{S}}_{\mathrm{W}}$ since its inverse matrix is needed when solving the problem. To deal with this problem, ones can add a constant $\eta > 0$ to the diagonal elements of $\widehat{\mathbf{S}}_{\mathrm{W}}$ as $\widehat{\mathbf{S}}_{\mathrm{W}} = \widehat{\mathbf{S}}_{\mathrm{W}} + \eta\mathbf{I}$. Here $\mathbf{I}$ is an identity matrix. An appropriate value of η is generally estimated through a cross-validation technique.

2.3. **Discussion on constructing the affinity matrix.** The affinity matrix $\mathbf{A}^c$ of the c-category data models the neighborhood and there are two main ways to construct it. In general, the matrix can be constructed as follows

$$A_{ij}^c = \begin{cases} \exp\left(- \parallel \mathbf{x}_i^c - \mathbf{x}_j^c \parallel /t\right), & \text{if } \mathbf{x}_i^c \in N_k(\mathbf{x}_j^c) \text{ or } \mathbf{x}_j^c \in N_k(\mathbf{x}_i^c); \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

where $\mathbf{x}_i^c$ is the ith data point of the c-category data, t is the heat kernel parameter, and $N_k(\mathbf{x}_i^c)$ represents the k nearest neighbors of $\mathbf{x}_i^c$ in the c-category data. This way of defining the affinity matrix is used in spectral clustering [13] and locality preserving projection (LPP) [14].

Another way to define the affinity matrix is

$$A_{ij}^c = \begin{cases} 1/N^c, & \text{if } \mathbf{x}_i^c \in N_k(\mathbf{x}_j^c) \text{ or } \mathbf{x}_j^c \in N_k(\mathbf{x}_i^c); \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

This way does not use the heat kernel and so its parameter is avoided. This way was used in local Fisher discriminant analysis (LFDA) [15].

2.4. **Connection to other methods.** First, if we defined the affinity matrix of the c-category data as

$$A_{ij}^c = \begin{cases} 1/N^c, & \text{if } y_i = y_j = c \\ 0, & \text{if } y_i \neq y_j \end{cases} \tag{11}$$

Then, the locality within-class scatter matrix $\widehat{\mathbf{S}}_{\mathrm{W}}$ is reduced to the within-class scatter matrix $\mathbf{S}_{\mathrm{W}}$ which is used in KDLOR. As a result, our method LDAOR is KDLOR. So, our method can be viewed as a generalized version of KDLOR and further takes consideration of the local structure characteristics of the data space.

On the other hand, the locality within-class scatter matrix can be formulated as

$$\widehat{\mathbf{S}}_{\mathrm{W}} = \sum_{c=1}^{C} \widehat{\mathbf{S}}^c = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \tag{12}$$

where

$$A_{ij} = \begin{cases} A_{ij}^c, & \text{if } y_i = y_j = c \\ 0, & \text{if } y_i \neq y_j \end{cases} \tag{13}$$

So, if we define the affinity matrix as

$$A_{ij} = \begin{cases} \exp\left(- \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2 /t\right), & \text{if } \mathbf{x}_i \in N_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_k(\mathbf{x}_i); \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

then our method is reduced to ORML. The main limitation of ORML is that it ignores the discriminant information in its objective function. However, LDAOR explicitly utilizes the discriminant information of the training data in its objective function.

3. **Extension to Nonlinear Case.** In the nonlinear case, ones generally use the kernelization trick [16] to map the data space into a high-dimensional feature space. In the feature space, we define the optimization problem of KDLOR as

$$\begin{aligned} &\min_{\mathbf{w}} \mathbf{w}^T \widehat{\mathbf{S}}_{\mathrm{W}}^{\phi} \mathbf{w} - \lambda\rho \\ &\text{s.t. } \mathbf{w}^T(\mathbf{u}_{c+1}^{\phi} - \mathbf{u}_c^{\phi}) \geq \rho,\ c = 1, 2, \ldots, C-1 \end{aligned} \tag{15}$$

where $\mathbf{u}_c^{\phi}$ is the mean vector in the feature space and $\widehat{\mathbf{S}}_{\mathrm{W}}^{\phi}$ is the corresponding locality within-class scatter matrix. Note, $\widehat{\mathbf{S}}_{\mathrm{W}}^{\phi}$ can be further rewritten as

$$\widehat{\mathbf{S}}_{\mathrm{W}}^{\phi} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j))(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j))^T \tag{16}$$

Here $\phi(\mathbf{x})$ denotes the data point in the feature space.

Further, according to the representation theorem [16], the vector $\mathbf{w}$ can be formulated as $\mathbf{w} = \sum_{i=1}^{N} a_i\phi(\mathbf{x}_i)$ in reproducing kernel Hilbert spaces (RKHS). Here $a_i \in \boldsymbol{R}$. Thus, the optimization problem (15) can be reformulated as

$$\begin{aligned} &\min_{\mathbf{a}} \tfrac{1}{2}\mathbf{a}^T \bar{\mathbf{S}}_{\mathrm{W}} \mathbf{a} - \lambda\rho \\ &\text{s.t. } \mathbf{a}^T(\bar{\mathbf{u}}_{c+1} - \bar{\mathbf{u}}_c) \geq \rho,\ c = 1, 2, \ldots, C-1 \end{aligned} \tag{17}$$

where $\bar{\mathbf{S}}_{\mathrm{W}}$ is formulated as $\bar{\mathbf{S}}_{\mathrm{W}} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}(\mathbf{z}_i - \mathbf{z}_j)(\mathbf{z}_i - \mathbf{z}_j)^T$, $\bar{\mathbf{u}}_c = \frac{1}{N_j}\sum_{i=1}^{N_j} \mathbf{z}_i$, and $\mathbf{a} = [a_1, \ldots, a_N]^T$. Here the vectors $\mathbf{z}_i$ are defined as $\mathbf{z}_i = [k(\mathbf{x}_i, \mathbf{x}_1), k(\mathbf{x}_i, \mathbf{x}_2), \ldots, k(\mathbf{x}_i, \mathbf{x}_N)]^T$. Here $k(\mathbf{x}_i, \mathbf{x}_j) = \varphi^T(\mathbf{x}_i)\varphi(\mathbf{x}_j)$ is a predefined kernel function. This is the final formulation of the optimization problem of the nonlinear LDAOR. It should be noted that the above optimization problem (17) actually is an optimization problem defined by linear LDAOR since $\bar{\mathbf{S}}_{\mathrm{W}}$ is the locality within-class scatter matrix of the data which consists of $\mathbf{z}_i$ $(i = 1, \ldots, N)$. Thus, according to the previous discussion about the linear LDAOR, (17) can be efficiently solved.

Suppose $\mathbf{a}$ is the solution of the above optimization problem, the discriminant function value for an unknown data point $\mathbf{x}$ is

$$f(\mathbf{x}) = \mathbf{a}^T\mathbf{k} \tag{18}$$

where $\mathbf{k} = [k(\mathbf{x}_1, \mathbf{x}), k(\mathbf{x}_2, \mathbf{x}), \ldots, k(\mathbf{x}_N, \mathbf{x})]^T$. In the nonlinear case, the thresholds b_c can be determined by

$$b_c = \mathbf{a}^T(\bar{\mathbf{u}}_{c+1} - \bar{\mathbf{u}}_c)/2 \tag{19}$$

Thus, the predictive ordinal decision function is given by using (7).

4. **Experiments.** In this section, we report the experimental results. First, we evaluate its generalization performance on several benchmark datasets by comparing it with other methods. Then, we test the proposed method on a real dataset.

4.1. **Benchmark dataset.** In order to evaluate the performance of LDAOR, we conducted the experiments on several benchmark datasets, which were selected from [3] and include Pyrimidines (74 data points with 27 attributes), Triazines (186 data points with 60 attributes), Wisconsin Breast Cancer (194 data points with 32 attributes), Machine CPU (209 data points with 6 attributes), Auto MPG (392 data points with 7 attributes), Boston Housing (506 data points with 13 attributes), Stocks Domain (950 data points with 9 attributes) and Abalone (4177 data points with 8 attributes). For each dataset, following the way in [3], we discretized the target values into five ordinal quantities using equal-frequency binning. In each experiment, each dataset was randomly selected of 70% to form the training data and the rest was used as the test data.

In the experiments, we adopted the Gaussian kernel, i.e., $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \parallel \mathbf{x} - \mathbf{y} \parallel^2)$. The kernel parameter γ and the penalty parameter λ were respectively determined by

TABLE 1. Mean-absolute-error (MAE) on the selected benchmark datasets

Datasets	KDLOR	ORML	LDAOR
Pyrimidines	0.4461±0.0382	0.4351±0.0194	**0.4232**±0.0215
Triazines	0.6874±0.0223	0.6901±0.0452	**0.6697**±0.0553
Wisconsin Breast Cancer	1.0243±0.0973	1.0114±0.0822	**1.0007**±0.0748
Machine CPU	**0.1835**±0.0318	0.1984±0.0251	0.1839±0.0219
Auto MPG	0.2522±0.0276	0.2565±0.0197	**0.2502**±0.0163
Boston Housing	**0.2553**±0.0332	0.2762±0.0231	0.2622±0.0412
Stocks Domain	0.1218±0.0167	0.2141±0.0325	**0.1032**±0.0311
Abalone	0.1935±0.0136	0.1917±0.0241	**0.1706**±0.0025

TABLE 2. Mean-zero-one-error (MZE) (%) on the selected benchmark datasets

Datasets	KDLOR	ORML	LDAOR
Pyrimidines	41.34±4.21	40.34±4.21	**38.34**±4.21
Triazines	50.11±3.25	48.73±4.75	**47.11**±3.62
Wisconsin Breast Cancer	69.71±5.37	68.42±6.18	**66.44**±5.29
Machine CPU	**16.53**±2.72	18.18±3.06	**16.53**±3.43
Auto MPG	**24.11**±3.01	25.12±2.12	24.18±3.45
Boston Housing	**22.33**±2.89	24.11±3.52	22.96±3.27
Stocks Domain	13.79±1.67	12.36±1.24	**11.54**±2.81
Abalone	18.34±4.21	17.91±2.27	**16.89**±1.38

5-fold cross validation technique. For simplicity, the adjacency matrix is constructed with (10). Generally, there are two evaluation metrics which can be used to evaluate the performance of the OR methods [3,11]. One is mean-absolute-error (MAE) and the other is mean-zero-one-error (MZE). We use these two metrics.

For each dataset, the experiment was repeated 50 times independently and the average and the standard deviation were computed. Table 1 shows the experimental results on MAE. It is easy to find that the proposed method LDAOR has lower MAE on the whole in comparison with KDLOR and ORML. These indicate that it is competitive with the other two methods in generalization ability. The reason is that not only is the local characteristic of the data explicitly considered but also the discriminant information is embodied in LDAOR. Actually, ORML also takes the local characteristic of the data into consideration. However, it ignores the discriminant information in contrast with the proposed method LDAOR. Similar phenomenon can be observed in Table 2 which reports the experimental results on MZE.

4.2. **USPS digit dataset.** To further evaluate the performance of the proposed method, we conducted an experiment on a real dataset. The used dataset is the USPS dataset [17], which comprises 11000 hand written digital character images. Each image is grayscale and normalized to 16×16. All images are divided into 10 categories and each category consists of 1100 images. As in the above experiments, the Gaussian kernel is adopted and the relevant parameters were determined by 5-fold cross validation technique. Similarly, the adjacency matrix is constructed with (10).

In this experiment, our aim is ranking the data in terms of the true digit. In each experiment, we randomly selected p $(= 10, 20, 50, 100, 200, 500)$ images in each category for training and the rest are used for testing. We repeated the experiments for 20 times and report the average results and the standard deviation. As shown in Table 3 and Table 4, ORML performs better compared with KDLOR. The reason is that ORML makes use of the underlying manifold structure in the data space and it has already been verified that

Table 3. Mean-absolute-error (MAE) on the USPS dataset

The number of the training/ test data in each category	KDLOR	ORML	LDAOR
10/1090	2.6012±0.4521	2.5034±0.4417	**2.4834**±0.4317
20/1080	2.1538±0.3251	2.0873±0.4175	**1.9711**±0.5623
50/1050	1.9697±0.5374	1.7842±0.6418	**1.7144**±0.5629
100/1000	1.8432±0.2729	1.7118±0.3706	**1.5653**±0.3243
200/900	1.6615±0.2101	1.5912±0.2412	**1.4282**±0.2345
500/600	1.5243±0.3392	1.4216±0.3823	**1.2962**±0.3275

Table 4. Mean-zero-one-error (MZE) (%) on the USPS dataset

The number of the training/ test data in each category	KDLOR	ORML	LDAOR
10/1090	18.27±4.41	16.34±4.39	**15.62**±3.58
20/1080	13.56±1.26	12.73±2.47	**11.17**±2.53
50/1050	8.48±2.62	8.64±1.98	**8.11**±1.46
100/1000	7.43±2.72	7.18±2.24	**7.03**±2.26
200/900	6.31±1.49	6.12±2.12	**5.88**±1.65
500/600	5.45±2.76	5.01±3.02	**4.24**±1.27

the USPS dataset contains underlying manifold structure [11]. However, in contrast with ORML, the proposed method LDAOR achieves lower MAE and MZE. This is because it explicitly further embodies the discriminant information contained in the data space in its objective function but ORML does not. In addition, LDAOR incorporates the local structure characteristic of the data as well as ORML.

5. **Conclusions.** In this paper, we proposed a novel ordinal regression method called LDAOR. LDAOR explicitly takes account of the local structure characteristics in the data space and the discriminant information contained in the training data. LDAOR achieves better generalization performance in contrast with its counterparts. The experimental results indicate the effectiveness of LDAOR by comparing it with ORML and KDLOR. In the future work, we will extend our method to more practical applications such as medical analysis, and facial age estimation.

Acknowledgment. This work is supported in part by the Scientific Research Project "Chunhui Plan" of Ministry of Education of China (Grant No. Z2015102), the Key Scientific Research Foundation of Sichuan Provincial Department of Education (Grant No. 11ZA004) and the National Science Foundation of China (Grant Nos. 61472329, 61532009, and 61602390).

REFERENCES

[1] P. McCullagh, Regression models for ordinal data, *Journal of the Royal Statistical Society. Series B: Methodological*, vol.42, no.2, pp.109-142, 1980.

[2] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro and C. Hervás-Martínez, Ordinal regression methods: Survey and experimental study, *IEEE Trans. Knowledge and Data Engineering*, vol.28, no.1, pp.127-146, 2016.

[3] W. Chu and Z. Ghahramani, Gaussian processes for ordinal regression, *Journal of Machine Learning Research*, vol.6, no.3, pp.1019-1041, 2005.

[4] W. Chu and S. S. Keerthi, Support vector ordinal regression, *Neural Computation*, vol.19, no.3, pp.792-815, 2007.

[5] M. Pérez-Ortiz, M. Cruz-Ramírez, M. D. Ayllón-Terán, N. Heaton, R. Ciria and C. Hervás-Martínez, An organ allocation system for liver transplantation based on ordinal regression, *Applied Soft Computing*, vol.14, no.1, pp.88-98, 2014.
[6] C. Li, Q. Liu, J. Liu and H. Lu, Ordinal distance metric learning for image ranking, *IEEE Trans. Neural Networks and Learning Systems*, vol.26, no.7, pp.1551-1559, 2015.
[7] H. Yan, Cost-sensitive ordinal regression for fully automatic facial beauty assessment, *Neurocomputing*, vol.129, pp.334-342, 2014.
[8] Q. Tian, S. Chen and X. Tan, Comparative study among three strategies of incorporating spatial structures to ordinal image regression, *Neurocomputing*, vol.136, pp.152-161, 2014.
[9] S. Baccianella, A. Esuli and F. Sebastiani, Feature selection for ordinal text classification, *Neural Computation*, vol.26, no.3, pp.557-591, 2014.
[10] B. Y. Sun, J. Li, D. D. Wu, X. M. Zhang and W. B. Li, Kernel discriminant learning for ordinal regression, *IEEE Trans. Knowledge and Data Engineering*, vol.22, no.6, pp.906-910, 2010.
[11] Y. Liu, Y. Liu and K. C. C. Chan, Ordinal regression via manifold learning, *Proc. of the 25th AAAI Conference on Artificial Intelligence*, pp.398-403, 2011.
[12] R. Fletcher, *Practical Methods of Optimization*, 2nd Edition, Wiley, New York, 1987.
[13] I. S. Dhillon, Y. Guan and B. Kulis, Kernel k-means: Spectral clustering and normalized cuts, *Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.551-556, 2004.
[14] X. He and P. Niyogi, Locality preserving projections, *Advances in Neural Information Processing Systems 16*, pp.153-160, 2003.
[15] M. Sugiyama, Local fisher discriminant analysis for supervised dimensionality reduction, *The 23rd International Conference on Machine Learning*, pp.905-912, 2006.
[16] B. Schölkopf and A. Smola, *Learning with Kernels*, MIT Press, Cambridge, 2002.
[17] J. J. Hull, A database for handwritten text recognition research, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.16, no.5, pp.550-554, 1994.