Robust ordinal regression induced by lp -centroid

DOI:
10.1016/j.neucom.2018.06.041

Document Version
Accepted author manuscript

Robust Ordinal Regression Induced by l_p-Centroid

Qing Tian[1,2,3], Wenqiang Zhang[1,2], Liping Wang[4*], Songcan Chen[5], Hujun Yin[3]

1 *School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China*
2 *Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science and Technology, Nanjing 210044, China*
3 *School of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, UK*
4 *Department of Mathematics, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China*
5 *College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China*

Abstract

Ordinal regression (OR) is an important research topic in machine learning and has attracted extensive attention due to its wide applications. So far, a variety of methods have been proposed to perform OR, in which the class-center-induced threshold methods (like KDLOR and MOR) have received more attention, for their simplicity and promising performance. The class-center-induced ORs typically calculate the ordinal thresholds with class centers, which are typically derived from the l_2-norm. Unfortunately, in such a way, the class means may be biased when the data is corrupted with outliers (i.e., non-i.i.d. noises) such that the resulting OR accuracy will be deteriorated. Motivated by the success of l_p-norm in applications against noises, in this paper we propose a novel type of class centroid derived from the l_p-norm (coined as l_p-centroid) to overcome the drawbacks above, and provide an optimization algorithm and corresponding convergence analysis for computing the l_p-centroid. To evaluate the effectiveness of l_p-centroid in OR context against noises, we then combine the l_p-centroid with two representative class-center-induced ORs, namely discriminant learning based and manifold learning based ORs. Finally, extensive OR experiments on synthetic and real-world datasets demonstrate the effectiveness and superiority of the proposed methods to related existing methods.

Keywords: Ordinal regression (OR), class-center-induced threshold OR, l_p-centroid, discriminant learning, manifold learning.

1. Introduction

Ordinal regression (OR) is an interesting machine learning paradigm which aims at learning a prediction function on a set of categories so that them can be predicted with ordinal (i.e., ordered) labels, such as the grade sequence: *poor*, *average*, *good*, *very good* and *excellent*. It can be seen that compared with normal regression with continuous regression values, the regression values of OR are discrete and finite. On the other hand, OR is also different from the nominal classification learning as the latter does not care about the order of class labels, that is, the class labels are disordered. Therefore, OR is a new learning paradigm sharing the properties of both traditional classification and regression. In the past decades, OR has attracted increasing researches due to its wide applications in recommender systems [1], page ranking [44], image retrieval [40], medical image diagnosis [32], human age estimation [16], [39], etc.

To implement OR, so far a variety of models have been derived or generated. According to their modeling strategy, they generally fall in three groups. The first type of OR models are native methods. That is, they treat OR as standard classification or regression. Along this line, off-the-self classifiers like multi-SVMs [19] and neural networks [7] or regressors such as support vector regressor [15] and regression trees [22] have been adopted to perform OR. However, a main problem with these methods is that they routinely ignore the ordering information of the data labels. The second OR modeling manner is achieved by binary decompositions. That is, the original OR problem is decomposed into a set of binary problems, and then the outputs of the binary problems is integrated as the OR result. In such a setting, Frank [14] and Waegeman [41] decomposed the ordinal class labels into a sequence of binary-valued

Corresponding author: wlpmath@nuaa.edu.cn (L. Wang)

labels, and then combined the binary classification results as the final OR output. To decrease the number of binary problems, Cardoso [3] achieved the goal of OR learning by means of designing an augmented binary classifier with data duplication. Lin et al. [23] then proposed a unified modeling method by incorporating cost matrix in the objective function. Although the second type of OR approaches attempt to take into account the ordering information of labels by either coding or cost-sensitivity learning, the ordinal relationship among the labels still cannot be preserved well. The third category of OR methods assumes that after ordinal projection learning the data classes can be separated with a sequence of orderly-distributed thresholds along the projection direction. To this end, POM [28] is the first work along this line by combining a sequence of ordinal odds models. Then, to cope with more complex data sets, it was extended to nonlinear counterparts [27], [30]. Besides, the perception learning was also reconstructed for online OR by imposing a series of ordinal thresholds on the decision-making direction [10]. Latter, motivated by the success of SVM in classification, Chu et al. [8] extended it to its ordinal counterparts, i.e., SVOR-EXC and SVOR-IMC, by introducing ordinal thresholds respectively in explicit and implicit manners. Then, the discriminant statistics were adopted for modeling OR and the well-known KDLOR approach was developed [34]. Motivated by the favorable performance of KDLOR, more efficient optimization approach [31] and prior knowledge embedding [38] were successively introduced to improve it. Following the KDLOR, other ordinal versions of discriminant analysis model were also presented [4]. Besides the discriminant models, the manifold learning was also adopted for modelling OR. Along this line, the so-called MOR approach was developed [25]. To further improve its performance, several variants of MOR with multiple OR projections [26], [24], [35] and class-sample-mixed thresholds [36] were developed successively. A common characteristic of the KDLOR and MOR and their variants is that the ordinal thresholds are induced by their class centers.

Although the threshold-based, especially the KDLOR and MOR like class-center-induced threshold ORs usually yield more accurate OR performance than the other methods, their performance will be deteriorated dramatically when the training data are corrupted with outliers, since these outliers usually bias the calculation of the class centers which are typically computed with the Euclidean metric (i.e., squared l_2-norm). Therefore, to perform robust class-center-induced OR against those outliers, outlier-insensitive types of class centroid is desired to be constructed. Fortunately, related research [17] shown that the l_p-norm has promising robustness to data noises and has been successfully applied in data clustering [18], filtering [21] and face recognition [43]. Motivated by these researches, in this paper, we propose to construct the class centroid with the l_p-norm (coined as l_p-centroid) instead of the existing l_2-norm. To the best of our knowledge, although the l_p-norm concept has been researched in previous literature, it has not been specially adopted to construct class centroid, let alone incorporating in the class-center-induced OR. Therefore, this paper may be the first work in attempting to perform outlier-insensitive (i.e., robust) OR [1] associated with the centroid derived from the unified l_p-norm. Moreover, note that generating the l_p-centroid is not trivial because it involves iterative optimization algorithm together with theoretical convergence guarantee. In addition, the proposed l_p-centroid is a unified framework covering some existing algorithms, because when $p = 2$ then l_p-centroid degenerates to the class mean mostly used in existing class-center-induced ORs, and when $p = 1$ then l_p-centroid reduces to the class median [5], to name just a few. The main contributions of this work are four-fold as follows:

- A unified framework algorithm for l_p-centroid is developed.
- The optimization and corresponding convergence analyses of l_p-centroid algorithm are provided.
- Variants of representative class-center-induced ORs, i.e., KDLOR and MOR, are derived with the proposed l_p-centroid.
- Extensive experimental validations of the proposed methods in the presence of data outliers are conducted.

The remaining sections of this paper are organized below. Section 2 briefly reviews related works. Section 3 introduces the proposed algorithms and developed OR methods. Section 4 reports and analyzes the experimental results to demonstrate the superiority of the proposed algorithm. Finally, Section 5 concludes the paper.

[1] Please distinguish the so-called *robust ordinal regression* in [9], [20] from what we are studying in this paper, because the former concept was defined to assist selecting their preference set in the field of fuzzy-decision-learning, but not to perform the referred outlier-insensitive OR of this work.

2. Related work

In this section, we first review two representative class-center-induced OR methods, namely KDLOR and MOR, two mostly related methods to our work. For the sake of clarification, assume we are given N training samples from totally K ordinal classes and N_k training samples from the k-th class, which are represented as X_k. The class-center-induced ORs typically seek for a projection direction along which the ordinal classes are separated orderly by separation thresholds.

For KDLOR [34], it aims to seek for an optimal OR projection direction w along which the classes are distributed orderly w.r.t. their labels while the within-class scatters are minimized, which is formulated as follows:

$$\begin{aligned} &\min_{w} \ w^T S_w w - C\rho \\ &s.t. \quad w^T(m_{k+1} - m_k) \geqslant \rho, \quad k = 1, 2, ..., K-1, \end{aligned} \tag{1}$$

where $S_w = \frac{1}{N}\sum_{k=1}^{N}\sum_{x\in X_k}(x-m_k)(x-m_k)^T$ indicates the entire within-class scatter matrix with $m_k = \frac{1}{N_k}\sum_{x\in X_k} x$ being the mean vector of the k-th class, ρ stands for the margin separating two neighboring classes, and C is a nonnegative tradeoff parameter.

For MOR [25], it is intended to preserve the data manifold in the process of OR learning. To this end, its objective function is formulated below:

$$\begin{aligned} &\min_{w} \ w^T XLX^T w - C\rho \\ &s.t. \quad w^T(m_{k+1} - m_k) \geqslant \rho, \quad k = 1, 2, ..., K-1, \end{aligned} \tag{2}$$

in which X stands for the set of N training samples, L is the Laplacian matrix [2], the other notations are defined as in KDLOR above.

For both KDLOR and MOR, the optimal projection direction w can be obtained through optimizing their objective functions in the same manner, respectively. With the obtained w, the label $f(x)$ of a test instance x can be predicted by

$$f(x) = \min_{k\in\{1,...,K\}}\{k : w^T x - b_k < 0\}, \tag{3}$$

where $\{b_k\}_{k=1}^{K}$ are the OR thresholds, defined as

$$b_k = \frac{w^T(N_{k+1}m_{k+1} + N_k m_k)}{N_{k+1} + N_k}. \tag{4}$$

It can be seen from Eq. (4) that the OR thresholds are dominated by the mean vectors of the classes. In other words, the prediction accuracy of the class-center-induced ORs like KDLOR and MOR is seriously affected by the calculation of class means. Thus, if the the data set contains many unknown noisy outliers, then the obtained class means will be unreliable and distribution-biased, resulting in low OR accuracy. Therefore, other type of robust class centroid is required to alleviate the effect of noisy outliers to OR.

3. Proposed methodology

As shown in Eqs. (3) and (4), the prediction accuracy of class-center-induced OR is essentially dominated by the class mean vectors $\{m_k\}_{k=1}^{K}$, as defined in [34], [25]. In essence, the class mean vector m_k [2] is obtained by optimizing

[2]For the sake of clarification, we take the k-th class mean vector m_k as an example, the mean vectors of other classes can be calculated similarly.

the below objective function:

$$\begin{aligned}(m_k)^* &= arg\min_{m_k}\sum_{i=1}^{N_k}\|m_k - x_i\|_2^2 \\ &= arg\min_{m_k}\sum_{i=1}^{N_k}(m_k - x_i)^T(m_k - x_i) \\ &= \frac{1}{N_k}\sum_{i=1}^{N_k}x_i,\end{aligned} \tag{5}$$

which is convex and can be easily solved with analytical solution.

As shown in Eq. (5), the computation of the squared l_2-norm class mean m_k is contributed with the same weight by all training samples from the k-th class. Unfortunately, if the samples are mixed with outliers, especially far from the normal class distribution center, the obtained class mean will be biased greatly and the subsequent OR accuracy will be dramatically reduced. To alleviate the drawbacks of l_2-norm against noises (outliers), recent researches [42], [29] constructed various l_1-norm-based algorithms to handle the so-called Laplacian noise. They shown that the undesirable influence of noise can be alleviated by substituting the l_2-norm with l_1-norm, since the latter is less sensitive to Laplacian-type noise. However, in real scenarios, the type of noises (outliers) is usually not known and their distributions do not necessarily satisfy the Laplacian distribution, resulting in the performance of l_1-norm based methods is not always superiority to other algorithms.

3.1. l_p-centroid algorithm

To deal with distribution-unknown outliers/noises in OR, motivated by the success of l_p-norm in face recognition [43], filtering [21] and data clustering [18], we propose to construct a unified type of class centroid, coined as l_p-centroid, which is derived from the l_p-norm and automatically-robust to the noise/outliers in OR. In brief, the l_p-centroid can be obtained by substituting the l_2-norm in Eq. (5) with the l_p-norm. Formally, the l_p-centroid, denoted as $m_k^{l_p}$, can be derived from the objective function below:

$$(m_k^{l_p})^* = arg\min_{m_k^{l_p}}\sum_{i=1}^{N_k}\|m_k^{l_p} - x_i\|_p^p,\ \ p \in (0, 2], \tag{6}$$

in which $\|\cdot\|_p^p$ is called l_p-norm [3]. To address various types of outliers/noises with unknown distributions in subsequent OR process, the hyper-parameter p is assigned in the range $(0, 2]$. From Eq. (6), we can find that the normal class mean (as shown in Eq. (5)) and the l_1-norm-based class centroid are special cases of the proposed l_p-centroid with $p = 2$ and $p = 1$, respectively. Therefore, the l_p-centroid is a unified framework for calculating class centroid.

Unlike the l_2-norm wich has analytical solution (see Eq. (5)), the l_p-centroid in Eq. (6) does not have closed-form solution except $p = 2$. However, through algebraic transformations, Eq. (6) can be converted as

$$\begin{aligned}(m_k^{l_p})^* &= arg\min_{m_k^{l_p}}\sum_{i=1}^{N_k}\|m_k^{l_p} - x_i\|_p^p \\ &= arg\min_{m_k^{l_p}}\sum_{i=1}^{N_k}tr\left(\left(m_k^{l_p} - x_i\right)^T D_i\left(m_k^{l_p} - x_i\right)\right) \\ &= arg\min_{m_k^{l_p}}\sum_{i=1}^{N_k}\left(m_k^{l_p} - x_i\right)^T D_i\left(m_k^{l_p} - x_i\right) \\ &= (\sum_{i=1}^{N_k}(D_i)^{(t)})^{-1}(\sum_{i=1}^{N_k}(D_i)^{(t)}x_i),\end{aligned} \tag{7}$$

[3] To show the calculation of l_p-norm, we take one d-dimensional vector x as an example, then its l_p-norm is defined as $\|x\|_p^p = \sum_{i=1}^{d}|x_i|^p$.

where $tr(\cdot)$ denotes the trace operator on a matrix and $D_i = diag\{|m_k^{lp} - x_i|^{p-2}\}$ is a diagonal matrix. Seemingly, Eq. (7) also has closed-form solution, but D_i is coupled with the variable m_k^{lp} to be solved. Fortunately, we can adopt an iterative optimization strategy to overcome it by calculating $D_i^{(t+1)} = diag\{|(m_k^{lp})^{(t)} - x_i|^{p-2}\}$, with $(m_k^{lp})^{(t)}$ denoting the solution of Eq. (7) after t iterations. That is, m_k^{lp} of D_i is substituted with the latest iteration solution. Then, we repeat the procedure until convergence. The complete algorithm for l_p-centroid is summarized in Table 1, where $f(t)$ indicates the t-th iteration objective value of Eq. (7), and $\{(m_k^{lp})^{(t)}\}_{k=1}^K$ denote the generated centroid of all classes.

Table 1: Iterative Optimization Algorithm for l_p-centroid.

Input:	Training set $\{X_k\}_{k=1}^K \subseteq \mathbb{R}^d$, hyper-parameter p, and convergence threshold ϵ.
Output:	l_p-centroid $\{m_k^{lp}\}_{k=1}^K$.
1. **for** $k = 1, ..., K$ **do**	
2.	t=1;
3.	$(m_k^{lp})^{(t)} = \frac{1}{N_k} \sum_{i=1}^{N_k} x_i$;
4.	$f(t-1) = 1e^{10}$, $f(t) = \frac{1}{N_k} \sum_{i=1}^{N_k} \|(m_k^{lp})^{(t)} - x_i\|_p^p$;
5.	**while** $\frac{f(t-1)-f(t)}{f(t-1)} > \epsilon$ **do**
6.	$f(t-1) = f(t)$;
7.	**for** $i = 1, ..., N_k$ **do**
8.	$(D_i)^{(t)} = diag\{|(m_k^{lp})^{(t)} - x_i|^{p-2}\}$;
9.	**end for**
10.	Calculate $(m_k^{lp})^{(t+1)}$ through Eq. (7);
11.	$f(t+1) = \frac{1}{N_k} \sum_{i=1}^{N_k} \|(m_k^{lp})^{(t+1)} - x_i\|_p^p$;
12.	$t = t + 1$.
13.	**end while**
14. **end for**	

3.2. *Convergence guarantee for the l_p-centroid algorithm*

The l_p-centroid of the K classes can be respectively obtained via the algorithm in Table 1 in finite iterations of optimization according to the following Theorem 1.

Theorem 1. *Let $\{(m_k^{lp})^{(t)}\}_{t=1}^T$ denote the generated centroid sequence of the k-th class using the l_p-centroid algorithm summarized in Table 1, then we have $f(t+1) \leqslant f(t), t = 1, ..., T$. Once $f(t+1) = f(t)$ happens, then $(m_k^{lp})^* = (m_k^{lp})^{(t)}$ is a stable minimizer [4] of Eq. (6). (The proof is given in the Appendix.)*

3.3. *Advantages of tuning the p of l_p-centroid through cross-validation instead of optimization*

We can see from the algorithm in Table 1 that the p in l_p-centroid is not optimized with the class centroid. Instead, we choose to tune its value through cross-validation due to two-fold considerations: 1) the objective function of the l_p-centroid (see Eq.(6)) is not convex when $0 < p < 1$, implying it is likely to be trapped by local optimums with sub-optimal or even ill-posed centroid solutions in an optimizing manner; however, if we tune p through cross-validation with proper grid search scale, nearly or even globally optimal solutions can be obtained with high probabilities; and more importantly, 2) by means of tuning the p in cross-validation manner, we can straightforward explore the effort rules of p on the performance of the l_p-centroid algorithm by tuning its value from 0 to 2, which will be analyzed in the experiment section.

[4] When $p \in (0, 1)$, then Eq. (6) is non-convex and $(m_k^{lp})^*$ is its local minimizer; when $p \in [1, 2]$, then Eq. (6) is convex and $(m_k^{lp})^*$ is thus its global minimizer.

3.4. Ordinal regression with l_p-centroid

After the l_p-centroid of the data classes is obtained through the optimization algorithm in Table 1, we can then substitute the class means involved in the class-center-induced OR approaches with the l_p-centroid. To evaluate the effectiveness of the proposed l_p-centroid, we respectively substitute the class mean vectors m_k of Eqs. (1) and (2) with l_p-centroid m_k^{lp}, $k = 1, ..., K$. Consequently, reconstructed KDLOR and MOR with l_p-centroid can be formulated as

$$\begin{aligned} &\min_{w} \ w^T S_w w - C\rho \\ &s.t. \quad w^T(m_{k+1}^{lp} - m_k^{lp}) \geqslant \rho, \quad k = 1, 2, ..., K-1, \end{aligned} \tag{8}$$

and

$$\begin{aligned} &\min_{w} \ w^T X L X^T w - C\rho \\ &s.t. \quad w^T(m_{k+1}^{lp} - m_k^{lp}) \geqslant \rho, \quad k = 1, 2, ..., K-1, \end{aligned} \tag{9}$$

respectively. For the optimization, Eq. (8) and Eq. (9) can be solved similarly using the implementations for Eq. (1) and Eq. (2), respectively.

3.5. Comparison between the remodeled KDLOR (MOR) with l_p-centroid and the standard with l_2-centroid

By comparing Eqs. (1) ((2)) with (8) ((9)), we can easily find that the computational cost difference between the remodeled KDLOR (MOR) with l_p-centroid and the standard ones with l_2-centroid essentially lies in the difference in computing the centroid of their classes. That is, it lies in the computational cost difference between the l_p-centroid and the l_2-centroid, which are respectively formulated in Eq. (7) and Eq. (5). To be specific, the l_2-centroid can be obtained directly with analytical solution (as shown in Eq. (5)); in contrast, although the proposed l_p-centroid is being calculated in an iterative manner (as the Algorithm shown in Table 1), it in practice converges efficiently within about 4 iterations (as shown in Figure 13), and more importantly, it is more robust to data outliers and superior in ordinal regression accuracy compared with the l_2-centroid (as shown in Figure 12). So, in summary, for real applications especially these corrupted with outliers and implemented in high-performance platforms (*in real world, most of the tasks are in such a case*), the l_p-centroid is very preferable to the l_2-centroid; while for simple tasks with limited computing resources, we refer the researchers to the l_2-centroid.

4. Experiment

To evaluate the effectiveness and superiority of the proposed l_p-centroid to class-center-induced ORs, we conduct experiments on a toy data, eight benchmark datasets and a large real-world face dataset.

4.1. Toy data

To intuitively demonstrate the robustness of the proposed l_p-centroid to outliers, we first perform comparative experiment on a 2D toy data. The obtained class centers derived from l_p-centroid with varying p values are shown in Figure 1. We can see from Figure 1 that with the decreasing p from 2 (*equal to the widely-used l_2-mean as in [34] and [25]*), to 1.5 and to 1 (*equal to the l_1-median as in [5]*), the class centroid derived with l_p-centroid tends to approach the dense distributions of corresponding class samples. That is, assigning p with relatively small values can reduce the undesirable influence of outliers. For the specific value of p, we should tune it according to the data distributions no matter outliers are involved or not.

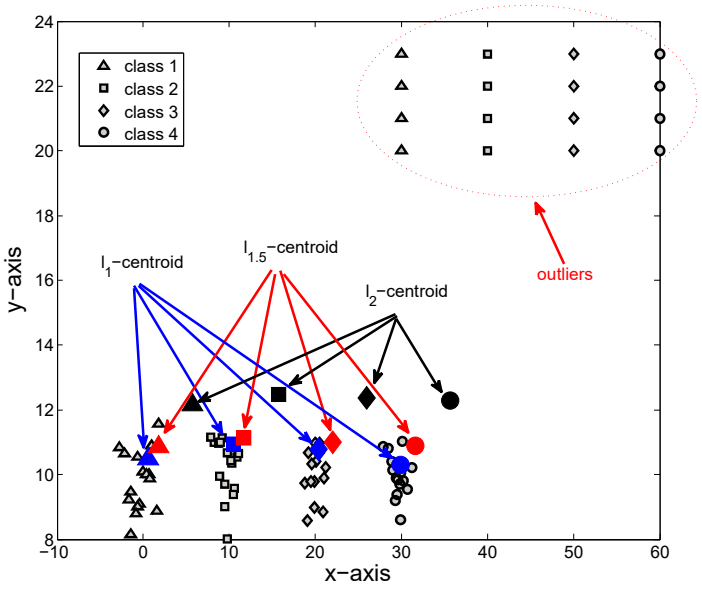


Figure 1: An intuitive comparison between class centers derived from l_p-centroid with varying p values. Totally, four class samples (indicated respectively with triangle, square, diamond and circle) are densely distributed in the bottom-left area of the figure, while some fewer outliers are distributed in the top-right area.

4.2. Benchmark datasets

To extensively evaluate the performance of the proposed algorithm to class-center-induced OR, we conduct experiments for KDLOR and MOR, in which the involved class centers are derived from the l_p-centroid as formulated in Eqs. (8) and (9). To be specific, we adopt the eight benchmark datasets used in [8] and [34] for experiment. We respectively chose 300×5, 300×5, 30×5, 200×5, 200×5, 300×5, 30×5 and 60×5 samples from the Abalone, Bank, Boston, California, Census, Computer, MachineCPU and Pyrimidines datasets for model training and the remaining as test set. All the hyper-parameters involved were tuned through 5-fold cross-validation. To practically evaluate the proposed l_p-centroid algorithm in OR, we tune p in the range of $\{0.1, 0.2, ..., 2\}$. we uniformly take the *Mean Absolute Error* (MAE) (MAE $= \frac{1}{N}\sum_{i=1}^{N}|\widehat{l_i} - l_i|$ with l_i and $\widehat{l_i}$ denoting the ground-true and regressed values, respectively) as the OR performance criterion. And, we repeated the experiments 5 times on each dataset with random data splitting and show the results from Figure 2 to Figure 9.

We can find from the comparative results from Figures 2 to 9 that, with increasing noise (i.e., outliers) ratio from 0% to 20%, the OR MAEs (*the lower the better*) of either MOR or KDLOR are increasing. It shows that outliers deteriorate the OR accuracy, which coincides with the our knowledge. Another finding is that no matter how high is the outliers ratio, MOR (KDLOR) yields the lower MAEs with the proposed l_p-centroid. It shows the superiority of the proposed method in OR accuracy and its robustness to outliers. It should be noted that the proposed l_p-centroid covers the l_2-mean (when p=2) and l_1-median (when p=1) as its special cases.

4.3. Cross-Age Celebrity Dataset

To evaluate the efficiency of the proposed method to outliers in real world applications, we also make evaluations on the Cross-Age Celebrity Dataset (CACD) [6]. Taking the CACD database, which consists of more than 160,000 face images of 2,000 celebrities drawn from 2004 to 2013, aged 16 to 62, for evaluation is due to that it is the largest cross-age face data set widely adopted for ordinal age estimation tasks. More importantly, the age labels of CACD are imprecise since they were annotated by anonymous picture uploaders. In other words, the CACD is naturally corrupted with outliers. Face examples of the CACD databse are demonstrated in Figure 10. We randomly take 5% to 20% percentage of samples by extracting HOG features [11] from CACD for training with the rest for testing.

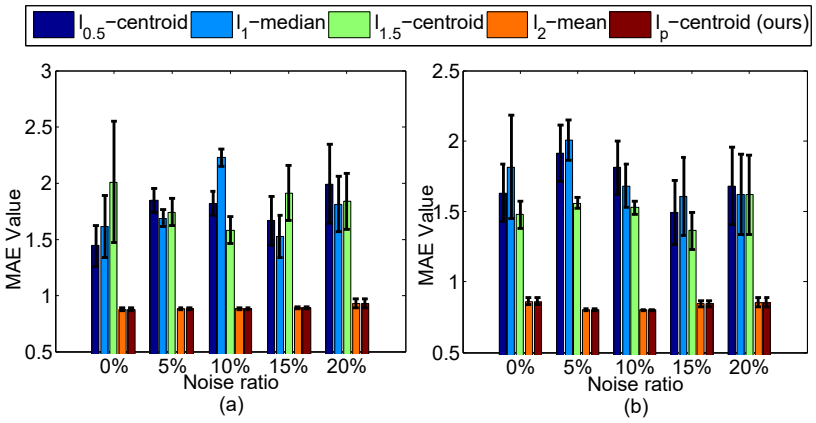


Figure 2: Comparison of OR with varying class centroid in MOR (a) and KDLOR (b) on Abalone.

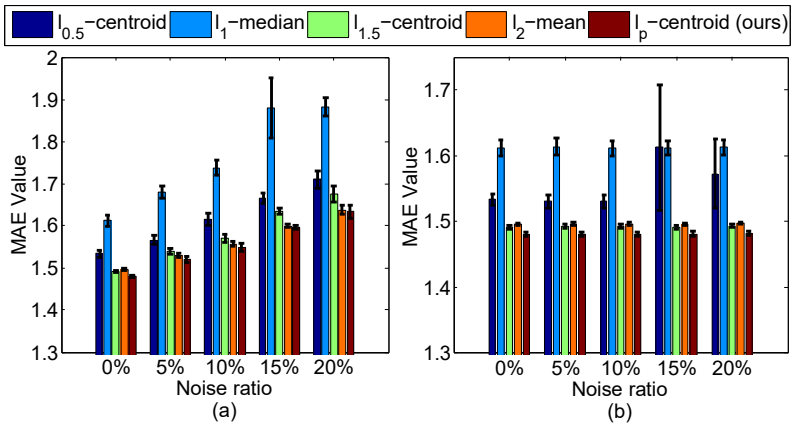


Figure 3: Comparison of OR with varying class centroid in MOR (a) and KDLOR (b) on Bank.

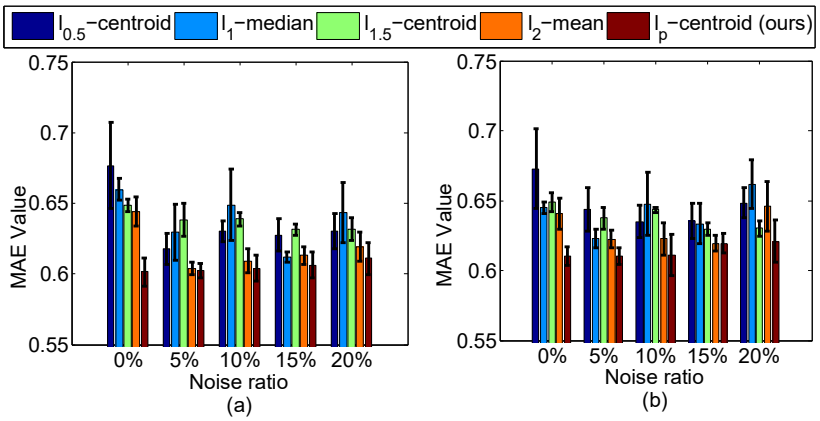


Figure 4: Comparison of OR with varying class centroid in MOR (a) and KDLOR (b) on Boston.

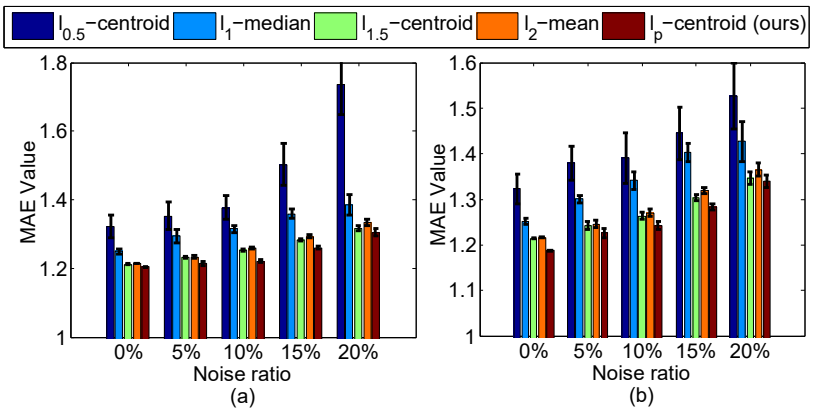


Figure 5: Comparison of OR with varying class centroid in MOR (a) and KDLOR (b) on California.

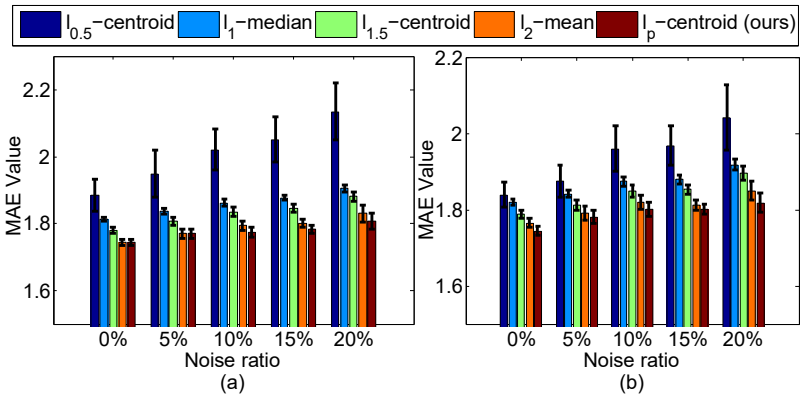


Figure 6: Comparison of OR with varying class centroid in MOR (a) and KDLOR (b) on Census.

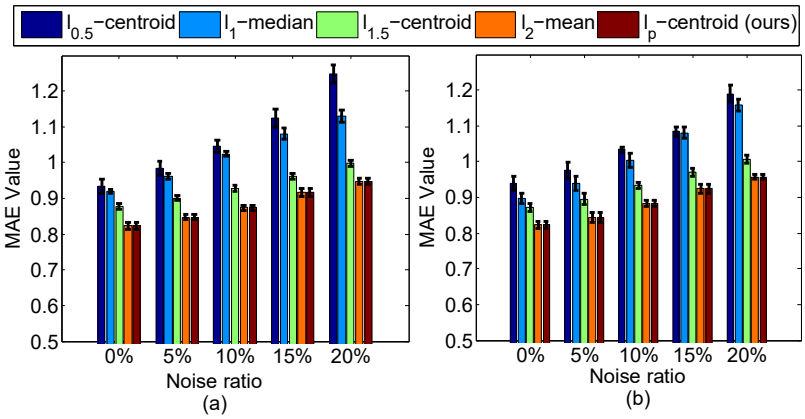


Figure 7: Comparison of OR with varying class centroid in MOR (a) and KDLOR (b) on Computer.

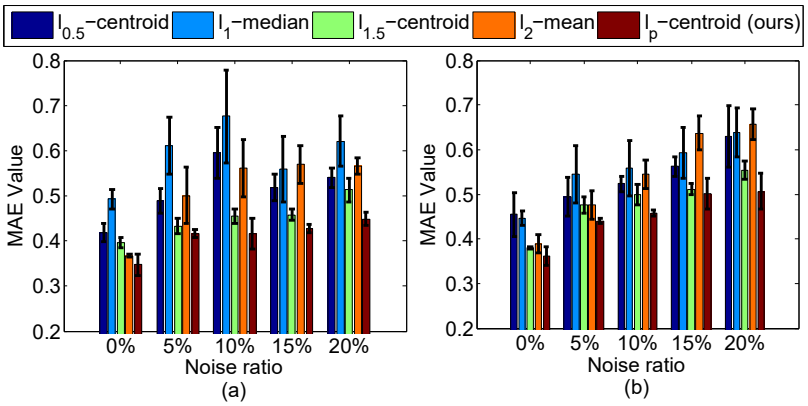


Figure 8: Comparison of OR with varying class centroid in MOR (a) and KDLOR (b) on MachineCPU.

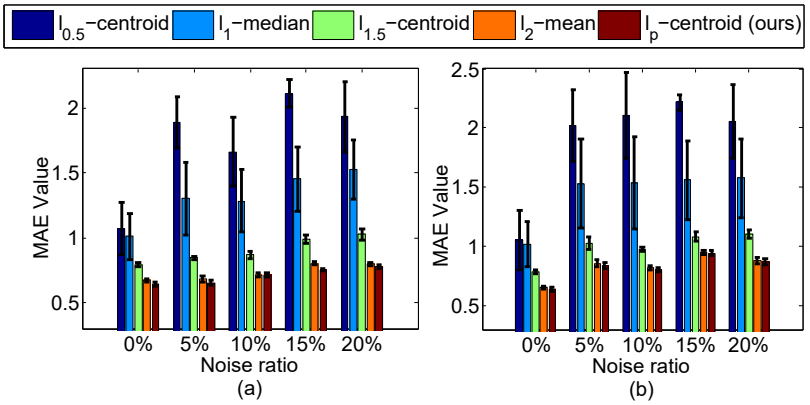


Figure 9: Comparison of OR with varying class centroid in MOR (a) and KDLOR (b) on Pyrimidines.

Figure 10: Face examples of the CACD database.

Moreover, we take the same experimental settings as in the previous sections for experiment. The experimental results averaged over ten random data partition are presented in Tables 3 and 2. From them, we can find that: 1) with increasing percentage of training samples, the MAEs yielded via both KDLOR and MOR are generally decreasing, validating the benefit of increasing training samples to improving the generalization ability of estimators; 2) the age estimation MAEs evaluated with l_1-centroid and $l_{1.5}$-centroid are generally much lower than those either by $l_{0.5}$-centroid or l_2-centroid. It is due to that although $l_{0.5}$-centroid itself can significantly remove the effort of distribution-outliers to centroid calculation, it tends to warp the natural distributions of clean data, while the l_2-centroid tends to be sensitive to the outliers; 3) the MAEs of age estimation with the l_p-centroid are significantly the lowest, showing the effectiveness and superiority of the proposed l_p-centroid method. Moreover, by comparing the results between on the eight benchmark datasets and on the CACD database, we can find that the proposed l_p-centroid seemingly does not always show significant performance superiority to other norms induced centroid including the l_2-centroid. This is partly due to the randomness of selected data for model training and testing, as well as the limited scale of the eight benchmark datasets. Because the extended experimental results on the large-scale CACD database show significant performance superiority of the proposed l_p-centroid to other types of centroid. In addition, we also explore the impact

Table 2: Evaluation results (MAE±STD, in years) yielded by MOR on CACD. The **bold** results indicate the best in each row.

Training samples percentage from each class	$l_{0.5}-centroid$	$l_1-centroid$	$l_{1.5}-centroid$	$l_2-centroid$	$l_p-centroid$
5%	16.17±1.34	17.69±1.52	21.81±2.47	21.96±1.91	**14.95±1.39**
10%	16.42±1.26	18.21±1.42	19.72±1.34	21.47±1.18	**14.83±1.24**
15%	16.85±1.18	17.90±1.38	19.08±1.42	21.89±1.38	**14.62±1.20**
20%	16.78±1.26	19.57±1.29	19.45±1.20	20.29±1.21	**14.59±1.15**

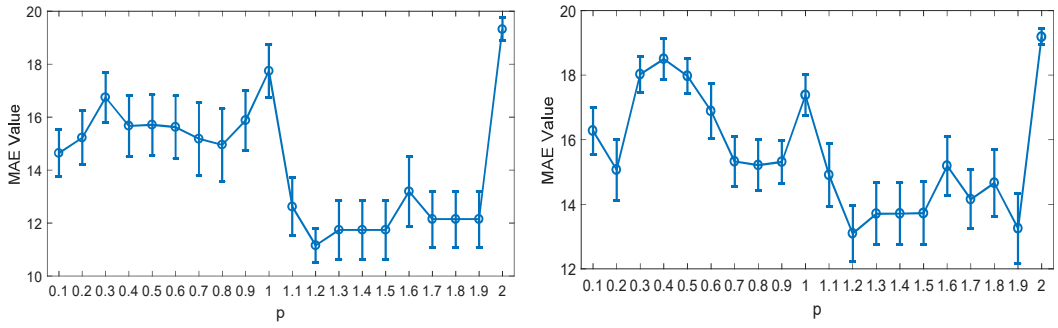
Table 3: Evaluation results (MAE±STD, in years) yielded by KDLOR on CACD. The **bold** results indicate the best in each row.

Training samples percentage from each class	$l_{0.5}-centroid$	$l_1-centroid$	$l_{1.5}-centroid$	$l_2-centroid$	$l_p-centroid$
5%	15.20±1.53	17.76±1.34	23.80±2.33	19.91±1.78	**14.51±1.83**
10%	15.97±1.26	17.82±1.55	19.72±1.86	20.09±2.07	**15.12±1.52**
15%	16.88±1.18	17.98±1.81	19.08±1.69	19.98±1.88	**14.86±1.47**
20%	16.32±1.31	17.72±1.68	19.02±1.26	19.81±1.59	**13.56±1.28**

of p in l_p-centroid on its performance. Without loss of generality, we randomly take 150 samples from each class of the CACD database for training and rest for testing, and display the results in Figure 11. It can be found that the performance variation is large when $0 < p < 1$, and is severely worse than that when $1 < p < 2$. On one hand, this is partly resulted from the non-convexity of l_p-centroid ($0 < p < 1$), likely to result in non-optimal solutions. On the other hand, when $1 < p < 2$, the ordinal regressors, KDLOR and MOR, with l_p-centroid achieve stable, lower MAEs. This implies the robust characteristic and desirable performance of l_p-centroid ($1 < p < 2$) in handing real large-scale ordinal problems involved with outliers.

4.4. Performance significance of the proposed l_p-centroid algorithm

Through the experimental comparisons in Section 4.2 and 4.3, we have shown the superiority of the l_p-centroid algorithm in performance. To evaluate the performance statistical significance of the proposed algorithm, we perform non-parametric statistical tests, i.e., the *Friedman* and *Nemenyi* tests [12], for the results in Figures 2 to 9 and Tables 3 and 2. The performance ranks and non-parametric rank statistical tests of different methods are shown in Figure 12. From them, we can see that the $l_{0.5}$-centroid method on both MOR and KDLOR ranks the worst among all the methods, it is resulted from that the nonconvex objective function of $l_{0.5}$-centroid tends to induce suboptimal centroid



(a) KDLOR (b) MOR

Figure 11: Performance rule of KDLOR (a) and MOR (b) with varying p of l_p-centroid on the CACD database.

solution. Moreover, $l_{1.5}$-centroid is in performance averagely better than l_1-centroid and more robust to outliers with comparable rank to the l_2-centroid. Most interestingly, the proposed l_p-centroid ranks the best with the smallest rank variance (see Figure 12 (a-b)) and significantly the best rank position (see Figure 12 (c-d)). In other words, the l_p-centroid is not only the most robust to outliers but also significantly superior in performance to the other types of centroid.

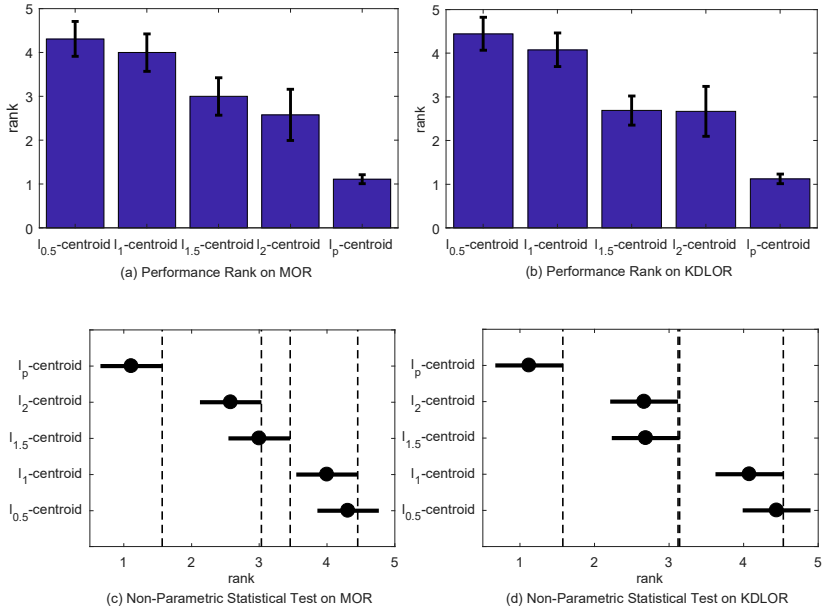


Figure 12: Performance ranks (a-b) and non-parametric statistical tests (c-d) of different centroid on MOR and KDLOR methods. For the non-parametric statistical tests, the *Friedman* and *Nemenyi* tests were adopted. In (c-d), the horizontal axis represents the values of mean rank of different methods, while the vertical axis represents the methods. For each method, • represents its mean rank value, while the line segment represents the critical difference domain. Two methods have a significant difference, if their critical domains are not overlapped; not so, otherwise.

4.5. Convergence efficiency of the proposed l_p-centroid algorithm

Besides the theoretical convergence guarantee for the proposed l_p-centroid algorithm in Theorem 1, we also evaluate its convergence efficiency experimentally. Without loss of generality, we follow the experimental settings in section 4.2, and conduct convergence experiment on the eight benchmark datasets (each corrupted with 10% outliers) and the large-scale CACD database, with results shown in Figure 13. We can see that the l_p-centroid algorithm

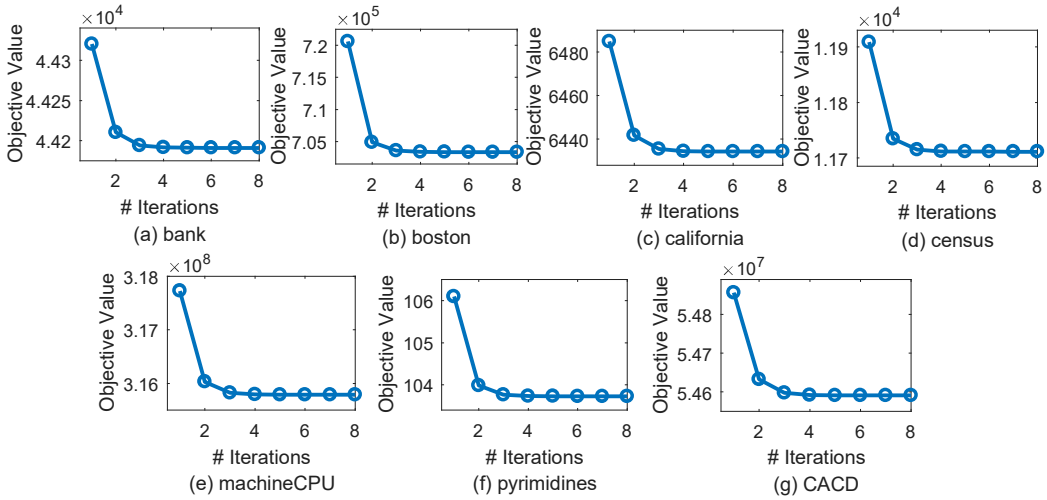


Figure 13: Objective function value convergence rule of the proposed l_p-centroid algorithm. Note that since the objective function value convergence rules of the l_p-centroid algorithm on classes of the datasets are quite similar, so their average rule is demonstrated here, and that the l_p-centroid on all classes of both *Abalone* and *Computer* degenerated to l_2-centroid with closed-form solution, so their convergence rules are not shown.

converges stably and efficiently within about 4 iterations.

5. Conclusion

In this paper, we presented a unified framework algorithm to calculate preferable class centers against outliers/noises, coined as l_p-centroid which covers the traditional l_2-norm centroid (which is typically adopted in existing class-center-induced OR) and the l_1-norm centroid (which is frequently used to handle Laplacian-type noises) as special cases. Then, we specially designed an iterative optimization algorithm to generate the l_p-centroid, and provided theoretical analyses about the convergence of the l_p-centroid algorithm. To evaluate the proposed l_p-centroid, we substituted it into two representative class-center-induced OR approaches KDLOR and MOR. Finally, extensive experiments on toy data, benchmark datasets and large-scale real-world database demonstrated the effectiveness and superiority of the proposed methods in OR accuracy and robustness to outliers. Actually, besides the ordinal estimation tasks, the proposed method can also be applied in machine learning scenarios such as clustering [13], k-nearest classification [33]. To regularize the space distributions of the samples, in the future we will consider to extend the proposed method to cross-data scenarios [37].

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China under grants 61702273, 61661136001 and 61472186, the Natural Science Foundation of Jiangsu Province under grant BK20170956, the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under grant 17KJB520022, a Project Funded by the Priority Academic Program Development of Jiangsu Higer Education Institutions, and the Startup Foundation for Talents of Nanjing University of Information Science and Technology.

Appendix:

Lemma 1. *Let $\varphi(\lambda) = \lambda - a\lambda^a$ with $a \in (0,1)$. Then for any $\lambda > 0$, we have $\varphi(\lambda) \leqslant 1 - a$ and $\lambda = 1$ is the unique maximizer.*

Proof: By taking the derivative of $\varphi(\lambda)$ with respective to λ and set it to zero, we get

$$\varphi^{'}(\lambda) = 1 - t^{\frac{1}{a}-1} = 0, \tag{10}$$

which has the unique solution $\lambda = 1$, given that $a \in (0,1)$ and $\lambda > 0$. □

Lemma 2. *Let $\{x_i\}_{i=1}^{N_k}$ denote the N_k samples of the k-th class, $\{(m_k^{lp})^{(t)}\}_{t=1}^{T}$ indicates the generated l_p-centroid sequence of the class using the l_p-centroid algorithm in Table 1, and $f(t)$ is the objective function value of Eq.* (6) *after the t-th optimization iteration. Then, we have*

$$N_k f(t+1) - \frac{p}{2}\sum_{i=1}^{N_k}((m_k^{lp})^{(t+1)} - x_i)^T (D_i)^{(t)}((m_k^{lp})^{(t+1)} - x_i) \leqslant (1 - \frac{p}{2})N_k f(t). \tag{11}$$

Proof: Let $a = \frac{p}{2} \in (0,1)$ with $p \in (0,2)$, and $\lambda = \frac{|\left((m_k^{lp})^{(t+1)} - x_i\right)_j|^p}{|\left((m_k^{lp})^{(t)} - x_i\right)_j|^p}$, j=1, ..., d, with $(\cdot)_j$ denoting the j-th element of a d-dimensional vector. Then, we have

$$\frac{|\left((m_k^{lp})^{(t+1)} - x_i\right)_j|^p}{|\left((m_k^{lp})^{(t)} - x_i\right)_j|^p} - \frac{p}{2}\frac{|\left((m_k^{lp})^{(t+1)} - x_i\right)_j|^2}{|\left((m_k^{lp})^{(t)} - x_i\right)_j|^2} \leqslant 1 - \frac{p}{2}, \tag{12}$$

which is equivalent to

$$|\left((m_k^{lp})^{(t+1)} - x_i\right)_j|^p - \frac{p}{2}\frac{|\left((m_k^{lp})^{(t+1)} - x_i\right)_j|^2}{|\left((m_k^{lp})^{(t)} - x_i\right)_j|^{2-p}} \leqslant (1 - \frac{p}{2})|\left((m_k^{lp})^{(t)} - x_i\right)_j|^p. \tag{13}$$

Summing (13) up for j=1, ..., d and i=1, ..., N_k, we have

$$\sum_{i=1}^{N_k}\|((m_k^{lp})^{(t+1)} - x_i)\|_p^p - \frac{p}{2}\sum_{i=1}^{N_k}((m_k^{lp})^{(t+1)} - x_i)^T (D_i)^{(t)}((m_k^{lp})^{(t+1)} - x_i) \leqslant (1 - \frac{p}{2})\sum_{i=1}^{N_k}\|((m_k^{lp})^{(t)} - x_i)\|_p^p. \tag{14}$$

That is,

$$N_k f(t+1) - \frac{p}{2}\sum_{i=1}^{N_k}((m_k^{lp})^{(t+1)} - x_i)^T (D_i)^{(t)}((m_k^{lp})^{(t+1)} - x_i) \leqslant (1 - \frac{p}{2})N_k f(t). \tag{15}$$

Moreover, according to Lemma 1, the equation part in (15) holds if and only if $\lambda = \frac{|\left((m_k^{lp})^{(t+1)} - x_i\right)_j|^p}{|\left((m_k^{lp})^{(t)} - x_i\right)_j|^p} = 1$, j=1, ..., d, implying $(m_k^{lp})^{(t+1)} = (m_k^{lp})^{(t)}$. □

Proof of Theorem 1:

Since $(m_k^{lp})^{(t+1)}$ is the minimizer of the t-th iteration of the l_p-centroid algorithm, i.e., $(m_k^{lp})^{(t+1)} = arg\min_{m_k^{lp}}\sum_{i=1}^{N_k}((m_k^{lp} - x_i)^T (D_i)^{(t)}(m_k^{lp} - x_i))$, then we have

$$\sum_{i=1}^{N_k}((m_k^{lp})^{(t+1)} - x_i)^T (D_i)^{(t)}((m_k^{lp})^{(t+1)} - x_i) \leqslant \sum_{i=1}^{N_k}((m_k^{lp})^{(t)} - x_i)^T (D_i)^{(t)}((m_k^{lp})^{(t)} - x_i) = N_k f(t). \tag{16}$$

By combining the above Lemma 2 and formulation (16), we have $f(t+1) \leqslant f(t)$, meaning that the objective function value of Eq. (6) is being reduced with the increasing iterations of the l_p-centroid algorithm. Furthermore, when $f(t+1) = f(t)$ happens, then $(m_k^{lp})^{(t+1)} = (m_k^{lp})^{(t)}$, which implies that $(m_k^{lp})^{(t)} = arg\min_{m_k^{lp}}\sum_{i=1}^{N_k}((m_k^{lp} - x_i)^T (D_i)^{(t)}(m_k^{lp} - x_i))$. That means that the minimizer $(m_k^{lp})^* = (m_k^{lp})^{(t)}$ is obtained. □

References

[1] Adomavicius, G., Tuzhilin, A., 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering 17, 734–749.

[2] Belkin, M., Niyogi, P., 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Advances in neural information processing systems, pp. 585–591.

[3] Cardoso, J.S., Costa, J.F., 2007. Learning to classify ordinal data: The data replication method. Journal of Machine Learning Research 8, 1393–1429.

[4] Cardoso, J.S., Sousa, R., Domingues, I., 2012. Ordinal data classification using kernel discriminant analysis: A comparison of three approaches, in: International Conference on Machine Learning and Applications, pp. 473–477.

[5] Chaudhury, K.N., Singer, A., 2012. Non-local euclidean medians. IEEE Signal Processing Letters 19, 745–748.

[6] Chen, B.C., Chen, C.S., Hsu, W.H., 2015. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. IEEE Transactions on Multimedia 17, 804–815.

[7] Cheng, J., Wang, Z., Pollastri, G., 2008. A neural network approach to ordinal regression, in: Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, IEEE. pp. 1279–1284.

[8] Chu, W., Keerthi, S.S., 2005. New approaches to support vector ordinal regression, in: Proceedings of the 22nd international conference on Machine learning, ACM. pp. 145–152.

[9] Corrente, S., Greco, S., Kadzi, Ski, M., owi, Ski, R., 2013. Robust ordinal regression in preference learning and ranking. Machine Learning 93, 381–422.

[10] Crammer, K., Singer, Y., 2002. Pranking with ranking, in: Advances in neural information processing systems, pp. 641–647.

[11] Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886–893.

[12] Demsar, J., 2006. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30.

[13] Elhamifar, E., Vidal, R., 2013. Sparse subspace clustering: Algorithm, theory, and applications. IEEE transactions on pattern analysis and machine intelligence 35, 2765–2781.

[14] Frank, E., Hall, M., 2001. A simple approach to ordinal classification. Machine Learning: ECML 2001 , 145–156.

[15] Fu, Y., Huang, T.S., 2008. Human age estimation with regression on discriminative aging manifold. IEEE Transactions on Multimedia 10, 578–584.

[16] Geng, X., Yin, C., Zhou, Z.H., 2013. Facial age estimation by learning from label distributions. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 2401–2412.

[17] Gentile, C., 2003. The robustness of the p-norm algorithms. Machine Learning 53, 265–299.

[18] Hathaway, R.J., Bezdek, J.C., Hu, Y., 2000. Generalized fuzzy c-means clustering strategies using lp norm distances. IEEE Transactions on Fuzzy Systems 8, 576–582.

[19] Hsu, C.W., Lin, C.J., 2002. A comparison of methods for multiclass support vector machines. IEEE transactions on Neural Networks 13, 415–425.

[20] Kadziski, M., Greco, S., S?owiski, R., 2014. Robust ordinal regression for dominance-based rough set approach to multiple criteria sorting. Information Sciences 283, 211–228.

[21] Kivinen, J., Warmuth, M.K., Hassibi, B., 2006. The p-norm generalization of the lms algorithm for adaptive filtering. IEEE Transactions on Signal Processing 54, 1782–1793.

[22] Kramer, S., Widmer, G., Pfahringer, B., De Groeve, M., 2001. Prediction of ordinal classes using regression trees. Fundamenta Informaticae 47, 1–13.

[23] Lin, H.T., Li, L., 2012. Reduction from cost-sensitive ordinal ranking to weighted binary classification. Neural Computation 24, 1329–1367.

[24] Liu, Y., Liu, Y., Chan, K.C., Zhang, J., 2012. Neighborhood preserving ordinal regression, in: Proceedings of the 4th International Conference on Internet Multimedia Computing and Service, ACM. pp. 119–122.

[25] Liu, Y., Liu, Y., Chan, K.C.C., 2011a. Ordinal regression via manifold learning, in: AAAI Conference on Artificial Intelligence, pp. 398–403.

[26] Liu, Y., Liu, Y., Zhong, S., Chan, K.C.C., 2011b. Semi-supervised manifold ordinal regression for image ranking, in: International Conference on Multimedea 2011, Scottsdale, Az, Usa, November 28 - December, pp. 1393–1396.

[27] Mathieson, M.J., 1996. Ordinal models for neural networks. Neural networks in financial engineering , 523–536.

[28] McCullagh, P., 1980. Regression models for ordinal data. Journal of the royal statistical society. Series B (Methodological) , 109–142.

[29] Peng, X., Lu, C., Yi, Z., Tang, H., 2016. Connections between nuclear-norm and frobenius-norm-based representations. IEEE Transactions on Neural Networks and Learning Systems 99, 1–7.

[30] Pérez-Ortiz, M., Gutiérrez, P.A., Cruz-Ramírez, M., Sánchez-Monedero, J., Hervás-Martínez, C., 2013. Kernelizing the proportional odds model through the empirical kernel mapping, in: International Work-Conference on Artificial Neural Networks, Springer. pp. 270–279.

[31] Pérez-Ortiz, M., Gutiérrez, P.A., Hervás-Martínez, C., 2014. Log-gamma distribution optimisation via maximum likelihood for ordered probability estimates, in: International Conference on Hybrid Artificial Intelligence Systems, Springer. pp. 454–465.

[32] Ribeiro, M.X., Traina, A.J., Traina Jr, C., Azevedo-Marques, P.M., 2008. An association rule-based method to support medical image diagnosis with efficiency. IEEE transactions on multimedia 10, 277–285.

[33] Samanthula, B.K., Elmehdwi, Y., Jiang, W., 2015. K-nearest neighbor classification over semantically secure encrypted relational data. IEEE transactions on Knowledge and data engineering 27, 1261–1273.

[34] Sun, B.Y., Li, J., Wu, D.D., Zhang, X.M., Li, W.B., 2010. Kernel discriminant learning for ordinal regression. IEEE Transactions on Knowledge and Data Engineering 22, 906–910.

[35] Sun, B.Y., Wang, H.L., Li, W.B., Wang, H.J., Li, J., Du, Z.Q., 2015. Constructing and combining orthogonal projection vectors for ordinal regression. Neural Processing Letters 41, 139–155.

[36] Tian, Q., Chen, S., 2015. A novel ordinal learning strategy: Ordinal nearest-centroid projection. Knowledge-Based Systems 88, 144–153.

[37] Tian, Q., Chen, S., 2017. Cross-heterogeneous-database age estimation through correlation representation learning. Neurocomputing 238, 286–295.
[38] Tian, Q., Chen, S., Tan, X., 2014. Comparative study among three strategies of incorporating spatial structures to ordinal image regression. Neurocomputing 136, 152–161.
[39] Tian, Q., Xue, H., Qiao, L., 2016. Human age estimation by considering both the ordinality and similarity of ages. Neural Processing Letters 43, 505–521.
[40] Tong, S., Chang, E., 2001. Support vector machine active learning for image retrieval, in: Proceedings of the ninth ACM international conference on Multimedia, ACM. pp. 107–118.
[41] Waegeman, W., Boullart, L., 2009. An ensemble of weighted support vector machines for ordinal regression. International Journal of Computer Systems Science and Engineering 3, 47–51.
[42] Wang, D., Lu, H., Yang, M.H., 2013. Least soft-threshold squares tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2371–2378.
[43] Wang, L., Chen, S., 2017. Joint representation classification for collective face recognition. Pattern Recognition 63, 182–192.
[44] Zhou, D., Weston, J., Gretton, A., Bousquet, O., Schölkopf, B., 2004. Ranking on data manifolds, in: Advances in neural information processing systems, pp. 169–176.

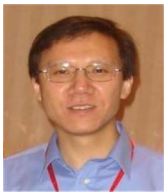
*Biography of the author(s)
Click here to download Biography of the author(s): biography.pdf

Qing Tian received the Ph.D. degree in computer science from Nanjing University of Aeronautics and Astronautics in 2016, China. He is currently an assistant professor in the School of Computer and Software, Nanjing University of Information Science and Technology, China and is currently visiting, as an academic visitor, at the University of Manchester, UK. He is the recipient of the *ICPR Best Scientific Paper Award* in 2016, the *Excellent Doctoral Dissertation Award of Jiangsu Province of China* in 2017, etc. His research interests include machine learning and pattern recognition, especially in the areas of ordinal regression and metric learning and its applications.

Wenqiang Zhang is currently a bachelor student in the School of Computer and Software, Nanjing University of Information Science and Technology, China. He has authored or co-authored several papers in the field of machine learning and pattern recognition. His research interests include machine learning and pattern recognition, especially in the areas of ordinal learning and its applications.

Liping Wang received the B.Sc. degree in mathematics from Qufu Normal University in 1998. In 2001, she received the M.Sc. degree in mathematics from Nanjing University. And, she also received her Ph.D. degree from Institute of Computational Mathematics and Scientific/Engineering Computing of the Chinese Academy of Sciences in 2004. Now she works as an associate professor in Department of Mathematics, Nanjing University of Aeronautics and Astronautics. Her research interests include optimization theory and pattern recognition.

Songcan Chen received the B.S. degree from Hangzhou University (now merged into Zhejiang University), the M.S. degree from Shanghai Jiao Tong University and the Ph.D. degree from Nanjing University of Aeronautics and Astronautics (NUAA) in 1983, 1985, and 1997, respectively. He joined in NUAA in 1986, and since 1998, he has been a full-time Professor with the Department of Computer Science and Engineering. He has authored/co-authored over 170 scientific peer-reviewed papers and ever obtained Honorable Mentions of 2006, 2007 and 2010 Best Paper Awards of Pattern Recognition Journal respectively. His current research interests include pattern recognition, machine learning, and neural computing.



Hujun Yin received the Ph.D. degree in neural networks from the University of York, York, UK, and the B.Eng. degree in electronic engineering and the M.Sc. degree in signal processing from Southeast University, Nanjing, China. He is currently a senior lecturer (associate professor) with the School of Electrical and Electronic Engineering, the University of Manchester, UK. He has published over 150 peer-reviewed articles in a wide range of topics from density modeling, image processing, face recognition, text mining and knowledge management, gene expression analysis, to novelty detection. He served or is serving as an Associate Editor for *the IEEE Transactions on Neural Networks* (2006-2010), the *International Journal of Neural Systems* (since 2005), and *the IEEE Transactions on Cybernetics* (since 2015). He has also served as the General Co-Chair for *IDEAL*

(since 2005) and Program Committee Co-Chair for *the International Symposium on Neural Networks*. His current research interests include neural networks, self-organizing learning, deep learning and pattern recognition. He is a Member of the EPSRC Peer Review College (since 2006) and a Senior Member of the IEEE (since 2003).