# CASSOR: Class-Aware Sample Selection for Ordinal Regression with Noisy Labels

Yue Yuan[1,2,3], Sheng Wan[1,2,3], Chuang Zhang[1,2,3], and Chen Gong[1,2,3{✉}]

[1] School of Computer Science and Engineering,
Nanjing University of Science and Technology, China
[2] Key Laboratory of Intelligent Perception and Systems
for High-Dimensional Information of Ministry of Education, China
[3] Jiangsu Key Laboratory of Image and Video Understanding for Social Security, China
chen.gong@njust.edu.cn

**Abstract.** Ordinal regression aims at solving the classification problem, where the categories are related in a natural order. Due to the difficulty in distinguishing between highly relevant categories, label noise is frequently present in ordinal data. Moreover, the varying degrees of relevance between categories can lead to an inconsistent distribution of misclassification loss across categories, posing a challenge to select clean data consistently from all categories for training. To overcome this limitation, we develop a sample selection method termed '**C**lass-**A**ware **S**ample **S**election for **O**rdinal **R**egression' (CASSOR). To be concrete, we devise a class-specific sample selection strategy in order to adaptively acquire sufficient clean examples for robust model training. Moreover, a label-ranking regularizer is designed to help guide the sample selection process via exploring the ordinal relationship between different examples. As a result, our proposed CASSOR is endowed with strong discrimination abilities on ordinal data. Intensive experiments have been performed on multiple real-world ordinal regression datasets, which firmly demonstrates the effectiveness of our method.

**Keywords:** Ordinal regression · Label noise · Weakly-supervised learning.

## 1 Introduction

Ordinal regression, also known as ordinal classification, aims to predict categories on an ordinal scale [5]. Unlike the nominal classification setting, ordinal regression involves labels that naturally possess a specific order [6]. To now, ordinal regression has found its applications in various fields, such as age estimation [2]. The existing methods to deal with ordinal regression tasks can be roughly divided into two types, namely regression and classification. The regression approaches aim to predict the values of the latent variable by mapping the input space to a one-dimensional real space [3] before predicting the categories of the input examples. The classification approaches, on the other hand, embed the ordinal relationship between categories into loss functions [12], labels [4,17], or architectural design [16].

The existing ordinal regression techniques are primarily designed for clean-label settings. However, the class labels observed in ordinal data may not always be correct. This is because the potential relevance between adjacent categories will make it challenging for annotators to accurately distinguish between different categories. As a result, the label noise can probably lead to performance degradation in model training. To

now, various deep learning approaches have been proposed for handling classification problems with label noise. Most of them focus on the estimation of the noise transition matrix [15] or the selection of clean examples [8,14]. The former aims to employ the transition matrix to build a risk-consistent estimator or a classifier-consistent estimator, while obtaining an accurate noise transition matrix can be challenging in practical scenarios [7]. Here, [5] is the only method designed for ordinal regression under label noise, which uses the noise transition matrix to construct the unbiased estimator of the true risk. On the other hand, the sample selection methods focus on selecting clean examples for model training and yield relatively satisfactory performance [8]. They usually predefine a loss threshold heuristically to regulate the number of clean examples, assuming that examples with loss below the threshold are probably clean [8,10].

Nevertheless, the above-mentioned sample selection methods are designed for nominal classification problems, which fail to exploit the fundamental characteristics of ordinal data. To be specific, if a category is highly relevant to its neighbors, it can be misclassified with a high probability, which leads to a large misclassification loss. Meanwhile, if a category is weakly related to its neighbors, the corresponding misclassification loss could be small. This will result in inconsistent distribution of misclassification loss across categories. Simply selecting the small-loss examples with a single threshold can lead to imbalanced sample selection across categories. As a result, highly relevant categories cannot provide sufficient information for model learning, ultimately degrading the model performance. In addition, ordinal data typically exhibit a natural label order that benefits the learning of ordinal models [6], which is, however, neglected by the nominal classification methods.

In light of the aforementioned challenges, we propose a new type of sample selection method termed **C**lass-**A**ware **S**ample **S**election for **O**rdinal **R**egression (CASSOR). Firstly, we design a class-aware sample selection strategy via calculating a class-specific score for each category. The score determines the number of examples chosen from each category, ensuring that categories with significant misclassification contribute adequate examples for model training. Considering the varying misclassification loss associated with different categories during the training phase, the class-specific score can be dynamically adjusted. This could also help prevent the model from overfitting to certain noisy examples and improve the generalization abilities. Additionally, since a biased selection of training examples is inevitable [7], we employ a dual-network architecture. As such, the potential errors caused by the biased selection can be reduced by the dual networks in a mutually beneficial manner [8]. Furthermore, to incorporate the inherent ordinal relationship between labels, we design a new type of OT loss called 'Optimal Transport regularized by label Ranking' (OTR). Unlike the traditional OT loss [1,12,16], which neglects the ordinal relationship among examples, our proposed OTR loss preserves the label order between the predicted results of the dual networks. Therefore, the inherent ordinal relationship can help guide the sample selection process and further reduce the accumulated errors caused by the biased sample selection.

## 2 Our Method

### 2.1 Preliminaries

In ordinal regression problems, the label of an example with a feature vector $\mathbf{x}$ is denoted as $y$, where $y \in \mathcal{Y} = \{1, 2, \ldots, K\}$. That is, $y$ is in a label space with $K$ dif-

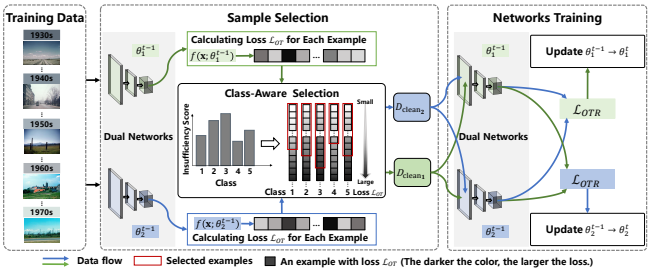


**Fig. 1.** The pipeline of our proposed method.

ferent labels, and the class labels satisfy $1 \prec 2 \prec \ldots \prec K$ with '$\prec$' representing order relation. The objective of ordinal regression is to find a classification rule or function to predict the categories of new examples given a training set of $N$ examples, namely $D = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$. Label noise refers to the situation that the observed label does not match the ground-truth label $y^*$, *i.e.*, $y \neq y^*$. To ensure the unimodality of model prediction, we adopt the architecture in the ordinal regression model [16] as the backbone of our method and the baseline methods. Let $f(\cdot;\theta)$ be the latent function for the network parameterized by $\theta$. Furthermore, the Optimal Transport (OT) loss [12,16] of the example $\mathbf{x}_i$ is employed to measure the misclassification in ordinal regression tasks:

$$\mathcal{L}_{OT}(f(\mathbf{x}_i;\theta), y_i) = \sum_{k=1}^{K} d(y_i, k)\, f_k(\mathbf{x}_i;\theta), \tag{1}$$

wherein $d(y_i, k) = |y_i - k|^m$ measures the label distance between $y_i$ and $k$ with $m \geq 1$.

## 2.2 Overall Framework

As shown in Fig. 1, the proposed method consists of two critical components which are designed for ordinal regression with label noise: (1) Class-Aware selection strategy, which adaptively selects reliable data from each category for robust modeling training (see Section 2.3); (2) Regularization with label ranking, which aims to incorporate the label order inherently contained in ordinal data for model learning (see Section 2.4).

## 2.3 Class-Aware Sample Selection

We develop a Class-Aware sample selection strategy for ordinal data in order to sufficiently learn from the categories with much misclassification. Firstly, we aim to compute an insufficiency score which can be obtained based on the distance between the average distribution of the prediction from each category and the distribution of each target category. Here, the average distribution $p_{y=k}$ of each observed category can be calculated by $p_{y=k} = \frac{1}{N_k}\sum_{i=1}^{N_k} \mathbb{1}[y_i = k] f(\mathbf{x}_i;\theta)$, where $f(x_i;\theta)$ indicates the predicted probability distribution of the example $\mathbf{x}_i$ by network parameters $\theta$, and $N_k$ is the number of examples in the $k$-th category. The $j$-th element in $p_{y=k}$ represents the probability of predicting an example of the $k$-th category to the $j$-th category. After that, we use Jensen-Shannon Divergence (JSD) [11] represented as $JS(\cdot||\cdot)$ to measure the dissimilarity between the average distribution and Dirac point mass[16] characterized by a one-hot probability mass function. We choose JSD because it is relatively simple and efficient for computation. A smaller JSD value indicates that the two distributions are more similar to each other. On this basis, we construct a matrix $\mathbf{S}$ with $\mathbf{S}_{i,j}$ denoting the JSD between the average distribution of prediction related to the $i$-th category and the one-hot distribution $Dirac(j)$ of class $j$:

$$\mathbf{S}_{i,j} = JS(p_{y=i}||Dirac(j)), \quad \forall i, j \in \{1, ..., K\}. \tag{2}$$

With Eq. (2), we can obtain the insufficiency score for the $j$-th category, which is expressed as $v_j = \frac{1}{K}\sum_{i=1}^{K}\mathbf{S}_{i,j}$. Here, the insufficiency score can be used to measure misclassification in a specific category, and a larger score often corresponds to more misclassified examples. For practical use, the insufficiency score is normalized as follows in order to eliminate the influences of different scales: $\tilde{v}_j = \frac{v_j - mean(v)}{max(v) - min(v)}$, where $\tilde{v}_j \in (-1, 1)$. The normalized insufficiency score can then be utilized to adjust the ratio of selected examples for each category. Concretely, the selection ratio of the $j$-th category is presented as $\mathbf{r}_j = 0.5 + \lambda \times \tilde{v}_j$, where $\lambda$ is a hyperparameter assigned to the insufficiency score $\tilde{v}_j$. Afterward, we select the examples with small classification loss, *i.e.*, $D_{clean}$, based on the ratio $\mathbf{r}_j$ at each epoch, so that the model can be encouraged to learn from the categories with relatively much misclassification. Note that the misclassification loss is measured by OT loss [12,16] in our method. Consequently, the model's discrimination abilities towards prone-to-misclassification categories will be enhanced.

### 2.4 Regularization with Label Ranking

We believe the ordering information of ordinal labels can enhance the performance of the model in ordinal regression tasks [6]. To this end, we have introduced an OTR loss that aims to maintain the label ranking between the predicted results of the dual networks, which consists of the traditional OT loss and a label-ranking loss $\mathcal{L}_{LR}$. Different from OT loss which focuses on the individual example, the proposed OTR loss concentrates on the relationship between each pair of examples. Here, the OTR loss is:

$$\mathcal{L}_{OTR} = \tilde{\mathcal{L}}_{OT} + \beta \times \mathcal{L}_{LR}, \tag{3}$$

where $\beta$ is a hyperparameter and $\tilde{\mathcal{L}}_{OT}$ represents the average OT loss of the selected examples. The label-ranking loss $\mathcal{L}_{LR}$ in Eq. (3) can be expressed as

$$\mathcal{L}_{LR} = \sum_{k=1}^{K-1} \frac{\sum_{d(y_i,y_j)=k} JS(f(\mathbf{x}_i;\theta_1), f(\mathbf{x}_j;\theta_2))}{\sum_{d(y_i,y_j)\geq k} JS(f(\mathbf{x}_i;\theta_1), f(\mathbf{x}_j;\theta_2))}, \tag{4}$$

where $f(\mathbf{x}_i;\theta_1)$ is the prediction of $\mathbf{x}_i$ generated from the network parameterized by $\theta_1$, and so on. In Eq. (4), $d(\cdot,\cdot)$ is the label distance function also used in Eq. (1). The objective of the Eq. (4) is to enforce a condition where given a pair of examples, the estimated distribution distance between the sample pair with a greater label distance is larger than the estimated distribution distance between the sample pair with a smaller label distance. Finally, the OTR loss of Eq. (3) is used to update the network parameters.

## 3 Experiments

### 3.1 Experimental Settings

***Datasets.*** Given the research emphasis on ordinal regression and label noise, we adhere to established practices [4,16] by using three standard datasets for assessment: Historical Color Image (HCI), Adience, and Diabetic Retinopathy (DR). HCI [13] comprises 1,325 images for a five-class ordinal task spanning the '1930s' to '1970s'. Adience [9] focuses on age estimation, where we selected and adapted the first six age groups from train-test splits[4]. DR[5] includes retinal images with diabetic retinopathy categorized into severity levels. We categorize and adapt it into 'no DR,' 'Mild,' 'Moderate,' and 'Severe DR and Proliferative DR,' and 1,680 images per class are used for evaluation.

[4] https://github.com/GilLevi/AgeGenderDeepLearning/tree/master/Folds

[5] https://www.kaggle.com/c/diabetic-retinopathy-detection/data

***Ordinal Label Noise Generation.*** Similar to [5], we hold the assumption that the probability of mislabeling decreases with the increase of label distance. By letting $\mathbf{T}$ be the noise transition matrix and $\rho$ be the total noise rate, $\mathbf{T}_{i,j}$ denotes the probability of flipping label $i$ to label $j$, $\mathbf{T}_{i,i} = 1 - \rho(i \in \{1, \ldots, K\})$, and $\mathbf{T}_{i,j}(i \neq j)$ can be calculated as $\mathbf{T}_{i,j} = \rho \frac{e^{\mathbf{T}^*_{i,j}}}{\sum_{k=1}^{K} e^{\mathbf{T}^*_{i,k}}}$. Here, $\mathbf{T}^*_{i,j}$ follows the standard normal distribution and can be formulated as $\mathbf{T}^*_{i,j} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}}$, where $\sigma$ is set to 2 empirically.

***Baseline Methods.*** We evaluate the effectiveness of our method by comparing it with multiple representative approaches, including the typical ordinal methods such as UNIORD [16], SORD [4], and CORAL[2]; label noise-robust learning algorithms such as forward correction (F-correction), backward correction (B-correction) [15], and Co-teaching [8]; and label noise-robust ordinal regression methods such as RDORLN [5]. For RDORLN [5], we use the same ordinal regression loss as our method.

**Table 1.** Experimental results of all compared methods on noisy HCI, Adience, and DR.

| Dataset | $\rho$ | | UNIORD [16] | SORD [4] | CORAL [2] | F-correction [15] | B-correction [15] | Co-teaching [8] | RDORLN [5] | **Our method** |
|---|---|---|---|---|---|---|---|---|---|---|
| HCI | 0.2 | MAE↓ | 0.790 ± 0.039 | 0.767 ± 0.047 | 0.935 ± 0.058 | 0.782 ± 0.037 | 0.862 ± 0.026 | 0.809 ± 0.049 | 0.777 ± 0.025 | **0.642 ± 0.029** |
| | | RMSE↓ | 1.224 ± 0.038 | 1.200 ± 0.058 | 1.377 ± 0.069 | 1.196 ± 0.048 | 1.301 ± 0.058 | 1.261 ± 0.061 | 1.221 ± 0.041 | **1.047 ± 0.044** |
| | 0.4 | MAE↓ | 0.990 ± 0.044 | 0.998 ± 0.069 | 1.100 ± 0.082 | 0.972 ± 0.031 | 1.106 ± 0.084 | 1.020 ± 0.075 | 1.032 ± 0.049 | **0.728 ± 0.074** |
| | | RMSE↓ | 1.403 ± 0.057 | 1.418 ± 0.074 | 1.555 ± 0.090 | 1.382 ± 0.015 | 1.573 ± 0.103 | 1.487 ± 0.088 | 1.467 ± 0.071 | **1.137 ± 0.070** |
| Adience | 0.2 | MAE↓ | 0.566 ± 0.043 | 0.566 ± 0.032 | 0.810 ± 0.081 | 0.533 ± 0.030 | 0.533 ± 0.043 | 0.423 ± 0.040 | 0.543 ± 0.039 | **0.407 ± 0.030** |
| | | RMSE↓ | 0.898 ± 0.046 | 0.886 ± 0.033 | 1.125 ± 0.075 | 0.876 ± 0.037 | 0.861 ± 0.048 | 0.737 ± 0.043 | 0.881 ± 0.039 | **0.704 ± 0.038** |
| | 0.4 | MAE↓ | 0.759 ± 0.037 | 0.797 ± 0.053 | 0.947 ± 0.072 | 0.812 ± 0.060 | 0.811 ± 0.070 | 0.492 ± 0.033 | 0.786 ± 0.022 | **0.420 ± 0.035** |
| | | RMSE↓ | 1.091 ± 0.045 | 1.150 ± 0.057 | 1.301 ± 0.065 | 1.264 ± 0.086 | 1.180 ± 0.074 | 0.797 ± 0.028 | 1.149 ± 0.031 | **0.715 ± 0.037** |
| DR | 0.2 | MAE↓ | 0.636 ± 0.015 | 0.651 ± 0.021 | 0.730 ± 0.011 | 0.635 ± 0.017 | 0.673 ± 0.016 | 0.597 ± 0.025 | 0.653 ± 0.011 | **0.577 ± 0.016** |
| | | RMSE↓ | 0.911 ± 0.014 | 0.948 ± 0.017 | 1.041 ± 0.015 | 0.917 ± 0.019 | 0.973 ± 0.019 | 0.927 ± 0.030 | 0.936 ± 0.012 | **0.862 ± 0.020** |
| | 0.4 | MAE↓ | 0.769 ± 0.012 | 0.775 ± 0.017 | 0.848 ± 0.017 | 0.777 ± 0.016 | 0.792 ± 0.021 | 0.617 ± 0.020 | 0.762 ± 0.019 | **0.609 ± 0.012** |
| | | RMSE↓ | 1.029 ± 0.011 | 1.057 ± 0.025 | 1.155 ± 0.024 | 1.048 ± 0.020 | 1.093 ± 0.029 | 0.943 ± 0.030 | 1.029 ± 0.013 | **0.902 ± 0.021** |

## 3.2 Experimental Results

For HCI, we evaluate the proposed method with synthetic label noise, where the noise rate $\rho$ is chosen from $\{0.2, 0.4\}$. We run five individual trials for all compared methods under each noise level and report the mean MAE, RMSE, and standard deviation in Table 1. Note that the performance of the ordinal regression methods consistently decreases as the noise level increases. Particularly, RDORLN [5] achieves poor results due to its reliance on the assumption that the noise transition matrix accurately reflects the true-noisy label relationship, which may not hold for the HCI dataset. In contrast, our method consistently achieves good results, showcasing its effectiveness across all noise rates. For Adience, the ordinal regression methods, such as UNIORD, SORD, and CORAL, exhibit unsatisfactory performance as a result of their inability to address label noise. Similarly, the label noise-robust methods, such as B-correction and Co-teaching, also yield poor results due to the inadequate consideration of ordinal information. We also performed well on DR, especially in RMSE.

**Table 2.** Experimental results of the proposed method with different key components.

| Dataset | $\rho$ | MAE↓ | | | | RMSE↓ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | A | B | C | D |
| DR | 0.2 | 0.632 ± 0.020 | 0.593 ± 0.008 | 0.579 ± 0.011 | 0.577 ± 0.016 | 0.964 ± 0.015 | 0.888 ± 0.015 | 0.879 ± 0.017 | 0.862 ± 0.020 |
| | 0.4 | 0.671 ± 0.006 | 0.632 ± 0.010 | 0.621 ± 0.013 | 0.609 ± 0.012 | 0.983 ± 0.007 | 0.922 ± 0.008 | 0.908 ± 0.006 | 0.902 ± 0.021 |

## 3.3 Ablation Study

Our method includes three crucial elements, namely the class-aware sample selection, the dual-network architecture, and the label-ranking regularization. We incrementally add these key components from **A** to **D**. **A**: A naïve baseline method, where $50\%$ of small-loss examples over all the training data are selected for training. **B**: Incorporating the class-aware selection strategy. **C**: Incorporating the dual-network architecture.

**D**: Incorporating the regularization with label ranking equipped with dual-network. The experimental results are shown in Table 2. As expected, the performance of the model can be improved when each component is applied.

## 4 Conclusion

In this paper, we introduce CASSOR, a novel sample selection approach for handling label noise in ordinal regression. CASSOR aims to mitigate the negative effects of inconsistent misclassification loss in the sample selection of ordinal data. Furthermore, a label-ranking regularizer is devised to guide the sample selection process with ordinal relations. As a result, our proposed method demonstrates strong performance on various real-world ordinal datasets. Future work will focus on developing a robust quantitative framework for measuring the essential differences between ordinal class labels.

## Acknowledgement

This research is supported by NSF of Jiangsu Province (Nos: BZ2021013, BK20220080).

## References

1. Beckham, C., Pal, C.: Unimodal probability distributions for deep ordinal classification. In: ICML. pp. 411–419 (2017)
2. Cao, W., Mirjalili, V., Raschka, S.: Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters **140**, 325–331 (2020)
3. Chu, W., Ghahramani, Z., Williams, C.K.: Gaussian processes for ordinal regression. Journal of Machine Learning Research **6**(7) (2005)
4. Diaz, R., Marathe, A.: Soft labels for ordinal regression. In: CVPR. pp. 4738–4747 (2019)
5. Garg, B., Manwani, N.: Robust deep ordinal regression under label noise. In: ACML. pp. 782–796 (2020)
6. Gutiérrez, P.A., P. Ortiz, M., S. Monedero, J., F. Navarro, F., H. Martinez, C.: Ordinal regression methods: survey and experimental study. TKDE **28**(1), 127–146 (2015)
7. Han, B., Yao, Q., Liu, T., Niu, G., Tsang, I.W., Kwok, J.T., Sugiyama, M.: A survey of label-noise representation learning: Past, present and future. arXiv:2011.04406 (2020)
8. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. NeurIPS **31** (2018)
9. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: CVPRW. pp. 34–42 (2015)
10. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. In: ICLR (2019)
11. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory **37**(1), 145–151 (1991)
12. Liu, X., Han, X., Qiao, Y., Ge, Y., Li, S., Lu, J.: Unimodal-uniform constrained wasserstein training for medical diagnosis. In: ICCVW. pp. 0–0 (2019)
13. Palermo, F., Hays, J., Efros, A.A.: Dating historical color images. In: ECML. pp. 499–512 (2012)
14. Patel, D., Sastry, P.: Adaptive sample selection for robust learning under label noise. In: WACV. pp. 3932–3942 (2023)
15. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: CVPR. pp. 1944–1952 (2017)
16. Shaham, U., Svirsky, J.: Deep ordinal regression using optimal transport loss and unimodal output probabilities. arXiv:2011.07607 (2020)
17. Víctor, Manuel, V., Pedro, Antonio, G., Javier, B.G., César, H.M.: Soft labelling based on triangular distributions for ordinal classification. Information Fusion **93**, 258–267 (2023)