

Copyright © 2022 by SIAM; Unauthorized reproduction of this article is prohibited.

DOI:

<https://doi.org/10.1137/1.9781611977172.4>

Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

Continuously Generalized Ordinal Regression for Linear and Deep Models

Fred Lu^{*†}

Francis Ferraro[†]

Edward Raff^{*†}

Abstract

Ordinal regression is a classification task where classes have an order and prediction error increases the further the predicted class is from the true class. The standard approach for modeling ordinal data involves fitting parallel separating hyperplanes that optimize a certain loss function. This assumption offers sample efficient learning via inductive bias, but is often too restrictive in real-world datasets where features may have varying effects across different categories. Allowing class-specific hyperplane slopes creates generalized logistic ordinal regression, increasing the flexibility of the model at a cost to sample efficiency. We explore an extension of the generalized model to the all-thresholds logistic loss and propose a regularization approach that interpolates between these two extremes. Our method, which we term *continuously generalized ordinal logistic*, significantly outperforms the standard ordinal logistic model over a thorough set of ordinal regression benchmark datasets. We further extend this method to deep learning and show that it achieves competitive or lower prediction error compared to previous models over a range of datasets and modalities. Furthermore, two primary alternative models for deep learning ordinal regression are shown to be special cases of our framework.

Keywords: ordinal regression, ranking

1 Introduction

In ordinal regression problems, the prediction task is to choose the target y from a set of labels with an ordered relation, e.g. $1 < 2 < \dots < k$. Unlike in classification, where accuracy is paramount, in ordinal regression the loss generally increases as the model predicts classes further away from the true label. Consider predicting medication dosage, where adjacent dosage amounts may still be safe, but large differences in dosage can be fatal. Ordinal regression models may give more accurate answers and learn more efficiently than classifiers in these situations because their inductive bias is better suited to the problem.

Threshold-based models are one of the most widely used ordinal regression approaches. In such models, a continuous prediction is learned as a linear mapping from the features along with a set of thresholds which partitions the prediction into classes [28, 14]. While such models are simple and efficient to learn, the restriction often imposes unreasonably strict requirements on the nature of the ordinal relationship of the data. To address this restriction, some ordinal models introduce generalized coefficients, learning a separate linear mapping for each class [41, 5]. Such models can be viewed

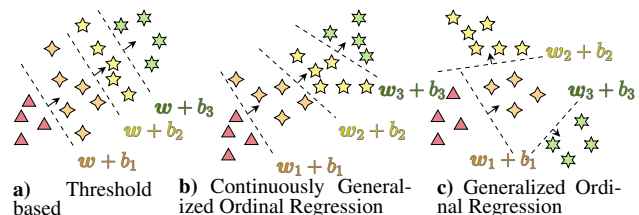


Figure 1: Figure (a) shows the standard threshold based approach on left which assumes all hyper-planes are equal, and (c) on the right shows the generalized ordinal regression which makes no assumption on the relationship between hyper-planes. Our approach coGOL allows for a continuous relaxation on the relationship between hyper-planes, allowing it to model cases that have variable relationship between hyper-planes, and captures both (a) and (c) as special cases of infinite and zero regularization respectively.

as training a separate binary classifier for each cut-point in the class ordering [21]. While this adds much-needed flexibility, it can exhibit poor stability on many datasets especially in high-noise regimes. Models of this type must also combine separate binary predictions into a consistent ordinal prediction [19].

To address these issues, we propose a continuous interpolation between the threshold and the generalized ordinal models, called *continuously generalized ordinal logit* (coGOL). By introducing generalized coefficients to the *all-thresholds* ordinal loss function with a new regularization approach, coGOL combines the flexibility and stability of the two extreme prior approaches as shown in Figure 1. By being able to adjust model flexibility, coGOL obtains better linear performance on 17 established benchmark datasets, with competitive results in deep models as well. Our extensive experiments show that coGOL achieves a statistically significant improvement over the standard ordinal logistic benchmark. This makes coGOL an ordinal approach appropriate for a wider variety of scenarios than prior methods.

Our method connects previous frameworks for threshold-based ordinal regression. We provide an overview and discuss previous approaches in section 2. The formulation of our method, with linear, kernelized, and deep versions, is given in section 3, and we detail our experimental evaluations in section 4. In section 5 we find that linear coGOL significantly

^{*}Booz Allen Hamilton. Emails: lu_fred, raff_edward@bah.com

[†]Univ. of Maryland, Baltimore County

outperforms the standard ordinal logistic baseline over 17 benchmark datasets, while deep coGOL is competitive with or better than other deep learning approaches on a set of large-scale, image-based, and sequence-based data. We also implement kernelized versions of our models and show that linear coGOL has advantages even compared to the kernel setting. Finally we conclude in section 6.

2 Related Work

2.1 Ordinal regression methods Traditional ordinal regression approaches can be grouped into three categories as suggested by [21]. In the first category, one simply fits a classification model, or a regression model whose output is then discretized. While straight-forward, this basic approach does not properly account for the ordinal regression problem. Since ordinal regression falls somewhere between classification and regression, the problem benefits from more specialized approaches. Cost-sensitive classification is a multi-class extension that specifies a cost matrix with larger loss as classes are further apart [24]. While the cost structure is ostensibly similar to the models we study, the underlying model is based on softmax classification and does not leverage the ordinal association between feature and target [25].

A second category decomposes the ordinal categories into binary classification problems, whose outputs can be combined into an ordinal prediction. Approaches to modeling and combining the outputs may involve support vector machines (SVM) [19, 37, 38] or neural networks [12, 29].

The final category contains threshold-based approaches, which simultaneously learn an output mapping and appropriate thresholds that partition the output to make ordinal predictions. For example, linear threshold models generally find the parallel hyperplanes that best separate the ordinal classes by minimizing an objective function, which include variants of SVM hinge loss and logistic loss [14, 34, 13, 25].

A well-known class of linear threshold models, known as *cumulative logit* or *proportional odds*, comes from the statistics literature and is imbued with probabilistic interpretation as a latent continuous target variable under censoring [28, 2]. Mathematically, the ordinal target y is modeled as $\mathbb{P}(y \leq j | \mathbf{x}) = \frac{\exp(\theta_j - \mathbf{w}^\top \mathbf{x})}{1 + \exp(\theta_j - \mathbf{w}^\top \mathbf{x})}$. Among many variants of this model, one generalizes the weights to be class-dependent, replacing \mathbf{w} above with \mathbf{w}_j [5], and is known as the *generalized ordered logit* model [41]. While this is an important starting point for our work, the underlying proportional odds assumption of this class of models does not generalize well to many datasets.

A unified framework for linear threshold losses is developed in [34, 33] and interprets the objective function as a convex surrogate for a natural loss function. Two such losses are the mean absolute error and the 0-1 loss and their corresponding surrogates were named the *all-thresholds* and

immediate-threshold loss. Intuitively, the difference between the two versions is that the all-thresholds loss on an (x, y) pair contains penalty terms from all ordinal classes, regardless of the value of y , while the immediate-threshold loss only contains terms from the thresholds corresponding to y . As the mean absolute error better fits the principle that the larger the misprediction, the more *wrong* the model is, this lends theoretical support for using the all-thresholds loss. This is supported by experimental evidence in [34]. Interestingly, the log-likelihood structure of the cumulative logit model is similar to the *immediate-thresholds* loss [33]. We clarify these details in our Appendix.

Our work first introduces generalized coefficients to the threshold model using the all-thresholds loss. In preliminary experiments, we found that the generalized model using immediate-threshold loss had unstable results, specifically with *higher* training and validation error than the standard ordinal logit, despite being a larger model class. While this step alone improves results on many datasets, we also add a regularization approach that permits control over the flexibility of the model. This is a novel addition which allows a bias-variance tradeoff in the model, thus combatting under- and over-fitting. Our study interpolates between the standard and generalized ordinal models using the all-thresholds loss in a novel manner that addresses shortcomings of the originals, by rewarding models which smoothly vary their hyperplane directions. This matches an inductive prior for many real-world datasets where features relate monotonically with the ordinal target, but vary to a greater extent over certain subsets of ordinal classes.

Prior works demonstrate a reduction from ordinal all-threshold models to binary classification with a specific class cost matrix [23, 25]. In fact, a generalized coefficient threshold model would reduce to classification with separate weights for each ordinal class divide. Thus our approach forms a bridge between the second (binary decomposition) and third (threshold) approach, combining flexibility and stability.

2.2 Deep ordinal regression Deep learning approaches generally also fall into the three previously described categories. Some methods handle ordinal regression as a classification problem [22, 35], while an early proposed neural network for ordinal regression adopted a threshold approach [27]. Similar to the cost-sensitive approaches in the linear case, deep models have found success with using weighted softmax labels to formulate a classification problem [17].

Following the binary decomposition approach of [23, 12], a deep learning model with multiple outputs was developed by [29] and showed strong performance on face age estimation datasets. The model (named OR-CNN) shares weights through all intermediate layers and ends with separate binary classification layers for each threshold θ_j , where each layer

predicts whether the sample's rank is higher than θ_j .

This framework was modified in [10] with additional weight sharing in the classification layers, so that only the biases differ. They then proved their model (named CORAL) to be order-consistent and to improve performance on age-estimation datasets. This model falls in the threshold model category.

In our work, we attach the continuously generalized ordinal model with the all-thresholds loss at the end of a neural network. Our model structure is in fact similar to OR-CNN, but our model further permits interpolation between the generalized and standard ordinal model. In fact, we will later show that both OR-CNN and CORAL are special cases of our framework. OR-CNN is equivalent to our generalized ordinal logit model with unconstrained weights. In contrast, CORAL is equivalent to the *all-thresholds* loss with non-generalized (parallel) thresholds. We will note that OR-CNN outperforms CORAL on all datasets, and performs more competitively with our coGOL model.

Another approach for deep ordinal regression [26] implements ordinal regression as a weighted combination of an unconstrained logistic regression and a monotonic threshold constraint based on the Hinge loss, which requires balancing the mismatched gradient scales between the two different loss functions. Where available, we include their results as a baseline.

A recent concurrent work in deep ordinal learning has also proposed regularization on the flexibility of generalized coefficients [20]. However, their regularization terms are perhaps overly complex, depending on pairwise cosine similarities between weight vectors, variance of weight norms, and constraints on bias terms. In order to balance these objectives, at least 5 separate hyperparameters are required. In contrast, our method requires one new hyperparameter and provides a much more intuitive generalization of the ordinal model. We implement their method as a baseline.

3 Continuously Generalized Ordinal Logistic

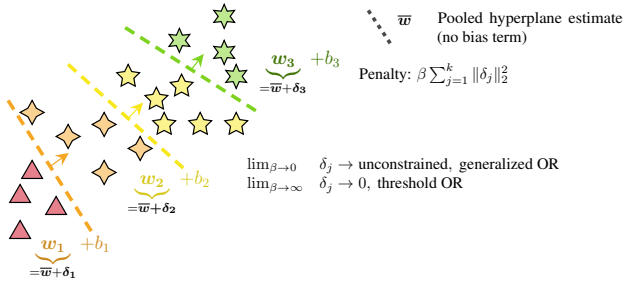


Figure 2: The essence of our approach—coGOL—is to represent the solution as a single “pooled” weight vector \bar{w} with each class's weight vector w_j being an adjustment $w_j = \bar{w} + \delta_j$ for each class. A penalty β on $\|\delta_j\|$ controls how much each class's solution may deviate from the pooled estimate \bar{w} , and an unregularized sequence of biases shift the class hyper-planes.

Suppose we have features $x \in \mathbb{R}^p$ and an ordered target variable $y \in \{1, 2, \dots, k\}$ as defined previously. Our goal is to learn a decision function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ that minimizes the risk $\mathcal{L}(f) = \mathbb{E}[\ell(y, f(x))]$ under some loss function ℓ .

Linear ordinal regression models restrict f to the set of parallel linear models $\{g(x; w, \theta)\}$ where $g_j(x) = \theta_j - w^\top x$ with bias terms $\theta \in \mathbb{R}^{k-1}$ such that $\theta_1 \leq \dots \leq \theta_{k-1}$. That is, the model is a set of $k - 1$ thresholds that partition the linear output $w^\top x$ into the ordinal classes. Prediction is then determined by the number of thresholds crossed:

$$(3.1) \quad f(x) = 1 + \sum_{j=1}^{k-1} \mathbb{I}[g_j(x) < 0]$$

where the brackets $\mathbb{I}[\cdot]$ denote the indicator function. A natural choice for the loss is mean absolute error: $\ell(y, x) = |y - x|$. Under the linear model class, the loss can be rewritten as

$$\ell(y, f(x)) = \sum_{j=1}^{y-1} \mathbb{I}[g_j(x) \geq 0] + \sum_{j=y}^{k-1} \mathbb{I}[g_j(x) < 0]$$

as shown in [33]. Next, we step through two core aspects of coGOL: a continuous loss and generalized coefficients.

A Continuous Loss Since the loss is discontinuous, we replace the indicator functions with a convex, continuous surrogate for optimization. In this work we focus on the logistic loss: $\varphi(x) = \log(1 + e^{-x})$. The loss then becomes

$$(3.2) \quad \ell(y, f(x)) = \sum_{j=1}^{y-1} \varphi(-g_j(x)) + \sum_{j=y}^{k-1} \varphi(g_j(x))$$

also known as the *all-thresholds* loss. Optimization of this loss function under monotonic θ results in the standard all-thresholds *ordinal logit* (OL) model. The monotonicity is enforced by formulating it as a convex optimization problem with the constraint on θ .

Generalized Coefficients We introduce generalized coefficients by extending the model class to $\tilde{g}_j(x) = \theta_j - w_j^\top x$, so that the threshold hyperplanes need not be parallel anymore. This yields the all-thresholds version of the *generalized ordinal logit* (GOL) model.

Finally, we wish to continuously interpolate between the parallel and generalized models in a manner that rewards smoothness between neighboring classes. Allowing $\delta_j = w_j - w_{j-1}$, we penalize the magnitude of δ_j . As in standard OL models, we also permit L2 regularization on w_j . Altogether, this results in the *continuously generalized*

ordinal logit (coGOL) loss function:

$$(3.3) \quad \begin{aligned} \ell(y, f(\mathbf{x})) = & \sum_{j=1}^{y-1} \varphi(-\tilde{g}_j(\mathbf{x})) + \sum_{j=y}^{k-1} \varphi(\tilde{g}_j(\mathbf{x})) \\ & + \underbrace{\alpha \sum_{j=1}^{k-1} \|\mathbf{w}_j\|_2^2}_{\text{standard L2 penalty}} + \underbrace{\beta \sum_{j=2}^{k-1} \|\delta_j\|_2^2}_{\text{our deviation penalty}}. \end{aligned}$$

As the penalty $\beta \rightarrow \infty$, we force $\delta_j \rightarrow 0$, recovering the threshold model. At $\beta = 0$ we recover the GOL model, and all β in between give us a continuous spectrum interpolating between these two models. We construct toy datasets to demonstrate how Equation 3.3 can make significant alterations to the learned solution, capturing the sample efficiency of OL and the expressiveness of GOL (Appendix 1).

3.1 Deep learning formulation We extend the continuously generalized all-thresholds loss to deep learning by replacing \mathbf{x} with the output of a base neural network $F(\mathbf{x})$. We can view $F(\mathbf{x})$ as a feature extractor trained alongside the ordinal regression head. We implement the ordinal regression as a final linear layer with $k-1$ independent weight vectors and $k-1$ biases and train the model using the loss of Equation 3.3. For convenience, we re-express the loss using familiar deep learning notation, with σ as the sigmoid function:

$$(3.4) \quad \begin{aligned} \ell(y, f(\mathbf{x})) = & - \sum_{j=1}^{y-1} \log \sigma(-\tilde{g}_j(F(\mathbf{x}))) \\ & - \sum_{j=y}^{k-1} \log \sigma(\tilde{g}_j(F(\mathbf{x}))) + \alpha \sum_{j=1}^{k-1} \|\mathbf{w}_j\|_2^2 + \beta \sum_{j=2}^{k-1} \|\delta_j\|_2^2. \end{aligned}$$

We note that setting $\alpha > 0$ has an important additional effect of constraining the magnitude of the weights \mathbf{w}_j besides standard weight decay. Otherwise for any value of $\beta > 0$, the feature extractor $F(\mathbf{x})$ can rescale the output to counteract the constraint on δ_j .

The linear coGOL was implemented with *cvxpy* in Python 3.6 [16, 1], while the deep learning models were implemented in Pytorch 1.6. We benchmarked our linear OL model with the *mord* package to ensure they gave identical results [33].

3.2 Connection to OR-CNN and CORAL While both OR-CNN [29] and CORAL [10] were developed as CNNs for age estimation, we adapt their models and losses for general neural networks as comparison methods. In this section we rewrite their loss functions to demonstrate their equivalence to the extreme cases of the coGOL model. Both works borrow

the convention of [23], converting the ordinal label y into an extended binary label $\tilde{y} \in \mathbb{R}^{k-1}$, such that $\tilde{y}_j = \mathbb{I}[y > j]$.

In CORAL, the output of the base network can be written as $h_j(\mathbf{x}) = \theta_j + \mathbf{w} \cdot F(\mathbf{x})$. Under uniform sample weights, their loss function $\ell(y, f(\mathbf{x}))$ is then

$$\begin{aligned} & = - \sum_{j=1}^{k-1} \log \sigma(h_j(\mathbf{x})) \tilde{y}_j + \log(1 - \sigma(h_j(\mathbf{x}))(1 - \tilde{y}_j)) \\ & = - \sum_{j=1}^{k-1} \log \sigma(h_j(\mathbf{x})) \tilde{y}_j + \log(\sigma(-h_j(\mathbf{x}))(1 - \tilde{y}_j)) \\ & = - \sum_{j=1}^{y-1} \log(\sigma(h_j(\mathbf{x}))) - \sum_{j=y}^{k-1} \log(\sigma(-h_j(\mathbf{x}))) \end{aligned}$$

which is equivalent to the all-thresholds loss by setting $g_j(\mathbf{x}) = -h_j(\mathbf{x})$. In particular, because the coefficient \mathbf{w} is not generalized, this is the standard OL model (or coGOL with $\beta \rightarrow \infty$).

In OR-CNN, the penultimate layer connects to $K-1$ separate binary cross-entropy outputs. Each such output o_j is thus the dot product of a weight vector with the penultimate layer outputs plus bias, so equivalently $\tilde{h}_j(\mathbf{x}) = \theta_j + \mathbf{w}_j \cdot F(\mathbf{x})$. Under uniform sample and task weights, their loss $\ell(y, f(\mathbf{x}))$ is then

$$\begin{aligned} & = - \sum_{j=1}^{k-1} \left(\mathbb{I}[o_j = \tilde{y}_j] \log p(o_j|\mathbf{x}, F) + \right. \\ & \quad \left. (1 - \mathbb{I}[o_j = \tilde{y}_j]) \log(1 - p(o_j|\mathbf{x}, F)) \right) \\ & = - \sum_{j=1}^{y-1} \log p(o_j|\mathbf{x}, F) - \sum_{j=y}^{k-1} \log(1 - p(o_j|\mathbf{x}, F)) \\ & = - \sum_{j=1}^{y-1} \log \sigma(\tilde{h}_j(\mathbf{x})) - \sum_{j=y}^{k-1} \log(\sigma(-\tilde{h}_j(\mathbf{x}))). \end{aligned}$$

Setting $\tilde{g}_j(\mathbf{x}) = -\tilde{h}_j(\mathbf{x})$, we get the generalized coefficient loss of Equation 3.4 but without our deviation penalty. Therefore, OR-CNN is equivalent to coGOL with $\beta = 0$.

4 Experiments

4.1 Linear model benchmarks We evaluate our method using a standard set of 17 ordinal datasets as defined in [21]. Characteristics of each dataset are in Table 1. We replicate our models using the same 30 train-test splits. We use stratified 3-fold cross-validation within each training set to tune regularization parameters β and α using 30 iterations of Optuna [3]. α and β were searched in the range $[1e-6, 10]$.

We compare the MAE, MSE, and accuracy, averaged over the 30 replications, of coGOL with OL, using the Wilcoxon signed rank test [40] to statistically compare the

Table 1: Characteristics of the 17 benchmark datasets for ordinal regression, which originate from [6, 32]. These datasets have real ordinal targets and are standard benchmarks based on prior work [21].

Dataset	Samples	Features	Classes
ERA	1000	4	9
ESL	488	4	9
LEV	1000	4	5
SWD	1000	10	4
automobile	205	71	6
balance-scale	625	4	3
bondrate	57	37	5
car	1728	21	4
contact-lenses	24	6	3
eucalyptus	736	91	5
newthyroid	215	5	3
pasture	36	25	3
squash-stored	52	51	3
squash-unstored	52	52	3
tae	151	54	3
toy	300	2	5
winequality-red	1599	11	6

models. Multiple previous studies on evaluation and model comparison have demonstrated the Wilcoxon test over multiple datasets to be more reliable for establishing improvement [7, 15], in comparison to standard error/confidence intervals in repeated trials [36, 8, 9, 4, 18]. Due to the large number of datasets and large number of trials to ensure a statistically valid conclusion, we limit our experiments to logistic loss with $\varphi(x) = \log(1 + e^{-x})$. Our framework is fully compatible with other choices such as the Hinge loss (i.e., SVM), but comparing different link functions $\varphi(\cdot)$ is beyond the scope of our work. Practitioners may choose the link function $\varphi(\cdot)$ based on what is appropriate for their problem.

4.2 Kernel model benchmarks An alternative hypothesis may be that a non-linear model would alleviate the need for coGOL, by allowing the linearly restrictive parallel hyperplanes of the standard OL approach to become more flexible with a non-linear transformation, while retaining its powerful inductive bias. We assess this hypothesis by implementing kernelized variants of all three methods: OL, GOL, and coGOL, and testing on the same datasets.

We use the Radial Basis Function (RBF) kernel. As before, hyperparameter search was done with stratified 3-fold cross-validation and 40 iterations of Optuna. We sample γ from a log-uniform distribution over the range $[0.01/(2\tau_0^2), 100/(2\tau_0^2)]$ where τ_0 is defined by the $1/K$ quantile of the pairwise Euclidean distances over the training set, as suggested in [11]. The linear and kernel models were

trained on a mix of 200 Intel Xeon Silver 4214 and Core i7-7820X CPUs.

4.3 Deep learning experiments In a second set of experiments, we consider multiple datasets suitable for a deep learning approach, due to large size or non-tabular data structure. We compare deep coGOL with the standard deep OL (that is, $\beta \rightarrow \infty$), OR-CNN [29], CORAL [10], the order regularization (Reg-Ord) model from [20], as well as a cross-entropy classifier (CE). To reasonably tune the 5 hyperparameters of Reg-Ord, we try the three most common settings from their paper and choose the best performance. Where available we also include results from Liu et al. [26]. In all experiments we tune the models using the validation set and select the best checkpoint over all epochs by validation MAE performance. We replicate each experiment 3 times and report average metrics, with the exception of the Historical Color Images Dataset (see below). Each model was trained on an NVIDIA RTX 2080 or NVIDIA RTX 6000 in under 1.5 hours. Dataset and training details follow.

BRFSS: The Behavioral Risk Factor Surveillance System (BRFSS) is an annual health survey administered by the Centers for Disease Control and Prevention. The survey collects comprehensive health and demographic information from residents across a large part of the United States. Similar to [39], we preprocess the 2016 BRFSS dataset to 80 features. Our prediction target is the BMI category of each respondent, which falls in 4 classes: underweight, normal weight, overweight, & obese.

We use a 80-10-10 split for train, validation, and test sets. As the base model, we employ a 3-layer feedforward network with 100 hidden units each and ELU activation. We train the model with Adam optimizer with learning rate $1e-3$ and batch size 10000 over 50 epochs. α, β were selected as 0.001, 0.01 respectively.

AFAD: Following previous work [29, 10], we also consider AFAD, a large dataset that has been used for age prediction from cropped and centered facial images. We used the same subset from [10], consisting of 165,501 face images with age range 15-40. The same test set was used, but we found that the original validation set was unrepresentative, containing a small subset of the set of all ages in the full dataset (i.e., most training set ages do not appear in validation). We thus reshuffled the training and validation sets with a new 85-15 split. We generally followed the training of their paper with modifications to improve performance. We adopted a ResNet34 as the base model, initialized using pretrained ImageNet weights. The models were trained with Adam optimizer at learning rate $5e-4$ for 30 epochs and batch size 256. α, β were selected as 0.001, 0.01 respectively.

Radio Frequency: The Radio Machine Learning dataset is a large collection of radio signals, both real and simulated, from 24 different channels [30]. Significant subsets of the

channels exhibit an ordinal relationship. For example, classes QPSK, 16QAM, 32QAM, 64QAM, 128QAM, 256QAM have the same encoding approach but differ only in the modulation of amplitude to change how many bits are encoded at a time (i.e., 4, 16, 32, 64, etc. bits). In our experiment, we subset the data to the above categories to assess whether an ordinal modeling approach outperforms classification in prediction error. This subset contains 106,496 samples from each channel for a total of 638,976 samples. Each signal is a length-1024 1-dimensional sequence with 2 channels. Adapting the method of [30], our base model is a modified ResNet with 6 stacks, followed by 3 FC layers (details in Appendix). We train the model with Adam optimizer with learning rate $5e-4$, batch size 2048, trained for 50 epochs. We use an 80-20 train-test split with further 10% of training used for the validation set. α, β were selected as 0.01, 0.05 respectively.

Historical Color Images: The Historical Color Image Dataset (HCID) is a collection of 1,325 historical color photographs from the 1930s to the 1970s [31]. Because of the changing technologies behind photography, the target of this task is predicting the decade in which each photograph was taken. The dataset spans 5 decades with 275 images each. Following recent work in [26], we adopt a 210-5-50 stratified split. We include their results as a benchmark. Because the original splits were not available and due to the small dataset size, we replicate our experiment over 5 random train-valid-test splits for fair comparison. We kept our training method similar to [26] but with modifications due to different model characteristics. The base model was a VGG13 with batch normalization, pretrained on ImageNet. We trained with Adam optimizer with learning rate $5e-5$, batch size 128, for 20 epochs. α, β were selected as 0.01, 0.05 respectively.

5 Results

5.1 Linear models Recalling that coGOL enables a continuous interpolation between the all-thresholds ordinal logit (OL) and the generalized ordinal logit (GOL) models, we compare coGOL to these two models. The OL model is a standard approach as described in [34, 33] so we view it as the baseline. Because we introduced generalized coefficients to the all-thresholds OL model, the GOL results serve as an intermediate result. We calculated the mean absolute error (MAE), mean squared error (MSE), and zero-one error (accuracy) over 30 test set splits.

Evaluation metrics are shown in Table 2. coGOL is competitive with or better than both comparison models, achieving lowest MAE in 11 out of 17 datasets, lowest MSE in 10, and highest accuracy in 11. In contrast, OL has the lowest MAE in 3, lowest MSE in 4, and highest accuracy in 3. Using the Wilcoxon signed rank test to compare coGOL with OL, we find that the difference in MAE is highly significant, with $p=0.0004$. The difference in accuracy is also highly

significant, with $p=0.0001$, while MSE is significant at the $\alpha = 0.1$ level at $p=0.066$.

Consider a data generation process that obeys the parallel threshold assumption. Here, OL is sample-efficient and learns the optimal model. While GOL can learn the same model, it may overfit if data variance is high, giving worse generalization. The regularization parameter on coGOL can help to control the model flexibility in this case by restricting the deviation between the different weights. On the other hand, when the data does not follow the parallel assumption, the OL model will be too restrictive. While GOL may be the best choice in low-noise data distributions, coGOL with weak regularization may still be better in the presence of sufficient data noise. Experimentally, we observe that in 5 out of the 6 largest datasets, coGOL outperforms OL. coGOL also outperforms both GOL in *bondrate*, *squash-stored*, and *squash-unstored*, 3 small datasets with high feature-to-sample ratio. On some datasets such as *balance-scale* all models perform the same, indicating that the OL assumptions are satisfactory for the data distribution.

5.2 Kernel models One may hypothesize that the non-linearity introduced by a kernelization would allow the OL model to attain the expressiveness of coGOL. However, a kernel model introduces new hyper-parameters and variance increases due to the additional degrees of freedom in the model. Because ordinal models are most valuable in low-sample size problems where the ordinal hypothesis serves as a strong inductive bias, this increases the difficulty of the task in a non-trivial manner. Indeed, Table 2 shows that the linear coGOL model performs better than any kernelized variant in 12 datasets. Of the remaining 5, only two (*balance-scale* and *toy*) show significant decreased error, indicating that the classes were not linearly separable. The last three datasets only show modest improvement over the linear models. This suggests that kernelization is not sufficient to overcome the constraints of the OL model when factoring in the increased model complexity and variance in results. This highlights that the significant performance advantage that coGOL provides in the linear case is meaningful.

5.3 Deep models Following the experimental procedure in section 4, we evaluated the MAE and root mean square error (RMSE) of coGOL compared with multiple benchmarks. Results are shown in Table 3. Across all datasets, we find that generalized coefficient models (coGOL, OR-CNN, Reg-ORD) uniformly improve performance compared to the threshold approach (OL, CORAL). coGOL consistently has lower MAE and either lowest or second-lowest RMSE. In particular there is a dramatic improvement of MAE and RMSE on HCID, indicating the standout effect of neighboring-class regularization in smaller deep learning datasets. coGOL also uniformly outperforms naive classification (CE), further high-

Table 2: Model results on 17 ordinal regression benchmark datasets using Linear and RBF kernel models. Best linear results shown in **bold**. Datasets that are underlined indicate that our Linear coGOL performs better than any kernelized model in MAE, showing that kernelization is not sufficient to tackle the learning constraints posed by OL. Our linear coGOL model shows a statistically significant improvement over the ordinal logistic (OL) baseline. The five datasets where RBF kernel performs better than linear in MAE are denoted with “*” and are the only cases where we highlight **best** and *second best* RBF results. Our RBF kernel coGOL is best or second best in these cases.

Dataset	Model	Linear			RBF			Dataset	Linear			RBF		
		MAE	MSE	Acc	MAE	MSE	Acc		MAE	MSE	Acc	MAE	MSE	Acc
<u>ERA</u>	OL	1.210	2.533	0.256	1.212	2.512	0.252	<u>eucalyptus</u>	0.396	0.457	0.634	0.516	0.712	0.560
	GOL	1.192	2.509	0.264	1.224	2.788	0.240		0.404	0.502	0.636	0.500	0.723	0.582
	coGOL	1.192	2.514	0.264	1.216	2.584	0.264		0.386	0.456	0.645	0.462	0.603	0.598
<u>ESL</u>	OL	0.310	0.345	0.705	0.303	0.328	0.713	<u>newthyroid</u>	0.030	0.030	0.970	0.037	0.037	0.963
	GOL	0.309	0.340	0.706	0.320	0.344	0.689		0.033	0.033	0.967	0.037	0.037	0.963
	coGOL	0.293	0.324	0.722	0.311	0.336	0.697		0.030	0.030	0.970	0.037	0.037	0.963
LEV*	OL	0.416	0.490	0.618	0.414	0.496	0.620	<u>pasture</u>	0.333	0.333	0.667	0.444	0.444	0.556
	GOL	0.434	0.509	0.601	<i>0.410</i>	<i>0.492</i>	0.630		0.283	0.283	0.717	0.444	0.444	0.556
	coGOL	0.421	0.494	0.613	0.406	0.482	<i>0.628</i>		0.328	0.328	0.672	0.444	0.444	0.556
<u>SWD</u>	OL	0.448	0.484	0.570	0.436	0.474	0.582	<u>squash-stored</u>	0.416	0.416	0.437	0.462	0.462	0.538
	GOL	0.435	0.464	0.579	0.438	0.472	0.580		0.381	0.430	0.643	0.462	0.538	0.615
	coGOL	0.433	0.464	0.582	0.438	0.470	0.580		0.378	0.413	0.640	0.385	0.423	0.615
<u>automobile</u>	OL	0.454	0.601	0.607	0.596	0.808	0.500	<u>squash-unstored</u>	0.289	0.289	0.711	0.308	0.308	0.692
	GOL	0.369	0.513	0.696	0.404	0.558	0.673		0.282	0.297	0.725	0.308	0.308	0.692
	coGOL	0.379	0.590	0.714	0.404	0.615	0.673		0.275	0.289	0.733	0.308	0.308	0.692
balance-scale*	OL	0.107	0.130	0.905	0.013	0.013	0.987	tae*	0.614	0.752	0.455	0.553	0.724	0.500
	GOL	0.107	0.131	0.905	0.019	0.019	0.981		0.575	0.761	0.518	0.579	<i>0.776</i>	<i>0.513</i>
	coGOL	0.107	0.131	0.905	<i>0.016</i>	<i>0.016</i>	<i>0.984</i>		0.580	0.741	0.500	<i>0.566</i>	<i>0.776</i>	0.526
<u>bondrate</u>	OL	0.549	0.767	0.556	0.600	1.133	0.600	toy*	0.953	1.464	0.287	0.020	0.020	0.980
	GOL	0.564	0.818	0.556	0.667	1.200	0.533		0.899	1.325	0.294	0.053	0.053	0.947
	coGOL	0.536	0.744	0.562	0.633	1.133	0.567		0.898	1.324	0.295	<i>0.040</i>	<i>0.040</i>	<i>0.960</i>
<u>car</u>	OL	0.082	0.083	0.919	0.412	0.587	0.676	winequality-red*	0.443	0.515	0.592	<i>0.435</i>	<i>0.540</i>	0.611
	GOL	0.073	0.078	0.930	0.339	0.426	0.701		0.436	0.511	0.600	0.446	0.556	0.605
	coGOL	0.073	0.078	0.930	0.325	0.411	0.712		0.438	0.516	0.598	0.433	0.524	<i>0.608</i>
<u>contact-lenses</u>	OL	0.428	0.601	0.659	0.667	0.917	0.500							
	GOL	0.384	0.572	0.710	0.583	0.833	0.500							
	coGOL	0.406	0.623	0.703	0.667	0.917	0.500							

lighting the benefit of using specialized ordinal regression approaches.

Small values for β often work best in the deep learning setting. For this reason, other generalized models (OR-CNN, Reg-Ord) can be competitive with coGOL. Compared to the linear benchmarks, the challenge of the deep learning datasets involve learning meaningful representations rather than dealing with data noise. This tends to favor flexible models. Furthermore, with increasing dataset complexity, it becomes less likely that the relation between features extracted by the base neural network and the ordinal class satisfies the parallel assumption. For example, we would expect for HCID that changes in camera technology were complex and nonlinear over time, so the same features that separate historical images from one decade may not useful in the next. coGOL’s flexibility lends it a better inductive prior to learn these complex decision boundaries, while enforcing neighbor-class smoothness encourages the embedding space to remain organized. In contrast, the more rigid OL models

would likely need to learn a further transformation to align the embeddings linearly. This is further supported by training analysis in Figure 3 showing validation error as a function of epoch number. We observe that the generalized models reach optimal validation loss in under 10 epochs, while the OL-based models improve much more slowly.

We also note that OR-CNN and Reg-Ord were only formulated for deep learning. In particular the softmax loss function formulation of OR-CNN and the large hyperparameter space of Reg-Ord cause significant training difficulty for the linear model benchmark. Therefore our coGOL model improves on other GOL-based models in deep learning and is novel to the linear case.

Finally, the kernel and deep results indicate that a non-linear mapping from the original feature space is not sufficient to overcome the restrictions of the OL threshold model. In other words, it does not appear efficient to learn a mapping of the data points to a space that can be separated by parallel hyperplanes.

Table 3: The performance of our model coGOL, compared to benchmarks on four deep learning datasets. The results highlight that OR-CNN and CORAL are special cases of our approach. Best results in **bold**, second best in *italics*.

	BRFSS		AFAD		RF		HCID	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
coGOL	0.580	0.825	3.209	<i>4.454</i>	0.615	<i>1.338</i>	0.674	1.033
OL	0.592	0.839	3.310	4.560	0.953	1.384	1.215	1.690
OR-CNN	<i>0.581</i>	0.825	3.229	4.483	<i>0.616</i>	1.367	0.707	1.087
CORAL	0.591	0.836	3.300	4.519	0.945	1.400	1.195	1.667
Reg-Ord	<i>0.581</i>	<i>0.828</i>	<i>3.217</i>	4.440	0.619	1.247	<i>0.689</i>	<i>1.066</i>
CE	0.618	0.918	3.439	4.871	0.654	1.500	0.750	1.193
Liu et al.	—	—	—	—	—	—	0.82	—

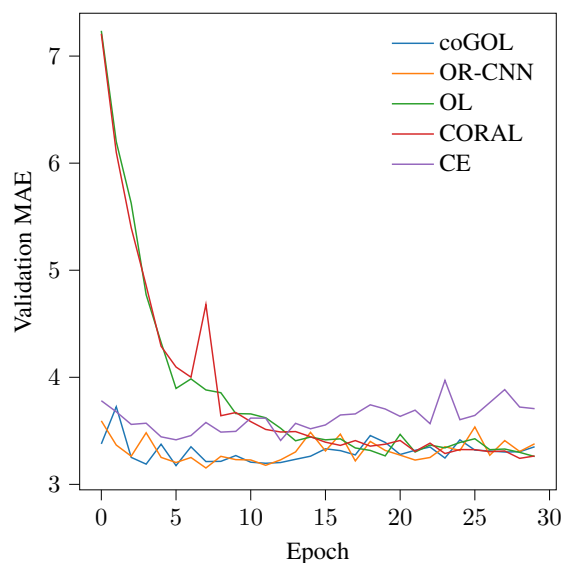


Figure 3: MAE by number of epochs for each model on the AFAD dataset.

6 Conclusion

We have developed the continuously generalized ordinal logistic model coGOL. The most common linear and deep ordinal models can be derived as special cases of our approach, which introduces a single hyper parameter β to balance model flexibility vs. helpful inductive bias. Our approach unifies the most common linear and deep models into a single framework, while also introducing a spectrum of ordinal models not previously developed. Our coGOL approach obtains comparable or better results across linear and deep learning based models, showing a statistically significant improvement.

References

- [1] A. AGRAWAL, R. VERSCHUEREN, S. DIAMOND, AND S. BOYD, *A rewriting system for convex optimization problems*, Journal of Control and Decision, 5 (2018), pp. 42–60.
- [2] A. AGRESTI, *Modeling ordinal categorical data*, University of Florida, Department of Statistics, (2013).
- [3] T. AKIBA, S. SANO, T. YANASE, T. OHTA, AND M. KOYAMA, *Optuna: A next-generation hyperparameter optimization framework*, in Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.
- [4] E. ALPAYDIN, *Combined 5×2 cv F Test for Comparing Supervised Classification Learning Algorithms*, Neural Comput., 11 (1999), pp. 1885–1892, <https://doi.org/10.1162/089976699300016007>, <http://dx.doi.org/10.1162/089976699300016007>.
- [5] C. V. ANANTH AND D. G. KLEINBAUM, *Regression models for ordinal responses: a review of methods and applications.*, International journal of epidemiology, 26 (1997), pp. 1323–1333.
- [6] A. ASUNCION AND D. NEWMAN, *Uci machine learning repository*, 2007.
- [7] A. BENAVALI, G. CORANI, AND F. MANGILI, *Should We Really Use Post-Hoc Tests Based on Mean-Ranks?*, Journal of Machine Learning Research, 17 (2016), pp. 1–10, <http://jmlr.org/papers/v17/benavoli16a.html>.
- [8] Y. BENGIO AND Y. GRANDVALET, *No Unbiased Estimator of the Variance of K-Fold Cross-Validation*, Journal of Machine Learning Research, 5 (2004), pp. 1089–1105, <http://dl.acm.org/citation.cfm?id=1005332.1044695>.
- [9] A. BLUM, A. KALAI, AND J. LANGFORD, *Beating the Hold-out: Bounds for K-fold and Progressive Cross-validation*, in Proceedings of the Twelfth Annual Conference on Computational Learning Theory, COLT '99, New York, NY, USA, 1999, ACM, pp. 203–208, <https://doi.org/10.1145/307400.307439>, <http://doi.acm.org/10.1145/307400.307439>.
- [10] W. CAO, V. MIRJALILI, AND S. RASCHKA, *Rank-consistent ordinal regression for neural networks*, arXiv preprint arXiv:1901.07884, (2019).
- [11] O. CHAPPELLE AND A. ZIEN, *Semi-supervised classification by low density separation.*, in AISTATS, vol. 2005, Citeseer, 2005, pp. 57–64.
- [12] J. CHENG, Z. WANG, AND G. POLLASTRI, *A neural network approach to ordinal regression*, in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE, 2008, pp. 1279–1284.
- [13] W. CHU, Z. GHARAMANI, AND C. K. WILLIAMS, *Gaussian processes for ordinal regression.*, Journal of machine learning research, 6 (2005).
- [14] W. CHU AND S. S. KEERTHI, *Support vector ordinal regression*, Neural computation, 19 (2007), pp. 792–815.
- [15] J. DEMŠAR, *Statistical Comparisons of Classifiers over Multiple Data Sets*, Journal of Machine Learning Research, 7 (2006), pp. 1–30, <http://dl.acm.org/citation.cfm?id=1248547.1248548>.
- [16] S. DIAMOND AND S. BOYD, *CVXPY: A Python-embedded modeling language for convex optimization*, Journal of Machine Learning Research, (2016), http://stanford.edu/~boyd/papers/pdf/cvxpy_paper.pdf. To appear.
- [17] R. DIAZ AND A. MARATHE, *Soft labels for ordinal regression*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4738–4747.

- [18] T. G. DIETTERICH, *Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms*, Neural Comput., 10 (1998), pp. 1895–1923, <https://doi.org/10.1162/089976698300017197>, <http://dx.doi.org/10.1162/089976698300017197>.
- [19] E. FRANK AND M. HALL, *A simple approach to ordinal classification*, in European conference on machine learning, Springer, 2001, pp. 145–156.
- [20] T. GUO, H. ZHANG, B. YOO, Y. LIU, Y. KWAK, AND J.-J. HAN, *Order regularization on ordinal loss for head pose, age and gaze estimation*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 1496–1504.
- [21] P. A. GUTIÉRREZ, M. PEREZ-ORTIZ, J. SANCHEZ-MONEDERO, F. FERNANDEZ-NAVARRO, AND C. HERVAS-MARTINEZ, *Ordinal regression methods: survey and experimental study*, IEEE Transactions on Knowledge and Data Engineering, 28 (2015), pp. 127–146.
- [22] G. LEVI AND T. HASSNER, *Age and gender classification using convolutional neural networks*, in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2015, pp. 34–42.
- [23] L. LI AND H.-T. LIN, *Ordinal regression by extended binary classification*, Advances in neural information processing systems, 19 (2006), pp. 865–872.
- [24] H.-T. LIN, *From ordinal ranking to binary classification*, California Institute of Technology, 2008.
- [25] H.-T. LIN AND L. LI, *Reduction from cost-sensitive ordinal ranking to weighted binary classification*, Neural Computation, 24 (2012), pp. 1329–1367.
- [26] Y. LIU, A. WAI KIN KONG, AND C. KEONG GOH, *A constrained deep neural network for ordinal regression*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 831–839.
- [27] M. J. MATHIESON, *Ordinal models for neural networks*, in Proc. 3rd Int. Conf. Neural Netw. Capital Markets, 1996, pp. 523–536.
- [28] P. McCULLAGH, *Regression models for ordinal data*, Journal of the Royal Statistical Society: Series B (Methodological), 42 (1980), pp. 109–127.
- [29] Z. NIU, M. ZHOU, L. WANG, X. GAO, AND G. HUA, *Ordinal regression with multiple output cnn for age estimation*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4920–4928.
- [30] T. J. O’SHEA, T. ROY, AND T. C. CLANCY, *Over-the-air deep learning based radio signal classification*, IEEE Journal of Selected Topics in Signal Processing, 12 (2018), pp. 168–179.
- [31] F. PALERMO, J. HAYS, AND A. A. EFROS, *Dating historical color images*, in European Conference on Computer Vision, Springer, 2012, pp. 499–512.
- [32] PASCAL, *Pascal (pattern analysis, statistical modelling and computational learning) machine learning benchmarks repository*. <http://mldata.org>, 2011.
- [33] F. PEDREGOSA, F. BACH, AND A. GRAMFORT, *On the consistency of ordinal regression methods*, The Journal of Machine Learning Research, 18 (2017), pp. 1769–1803.
- [34] J. D. RENNIE AND N. SREBRO, *Loss functions for preference levels: Regression with discrete ordered labels*, in Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling, vol. 1, Kluwer Norwell, MA, 2005.
- [35] R. ROTHE, R. TIMOFTE, AND L. VAN GOOL, *Dex: Deep expectation of apparent age from a single image*, in Proceedings of the IEEE international conference on computer vision workshops, 2015, pp. 10–15.
- [36] S. VARMA AND R. SIMON, *Bias in error estimation when using cross-validation for model selection*, 2006, <https://doi.org/10.1186/1471-2105-7-91>.
- [37] W. WAEGEMAN, L. BOULLART, ET AL., *An ensemble of weighted support vector machines for ordinal regression*, International Journal of Computer Systems Science and Engineering, 3 (2009), pp. 47–51.
- [38] H. WANG, Y. SHI, L. NIU, AND Y. TIAN, *Nonparallel support vector ordinal regression*, IEEE transactions on cybernetics, 47 (2017), pp. 3306–3317.
- [39] L. WANG AND D. ZHU, *Tackling multiple ordinal regression problems: Sparse and deep multi-task learning approaches*, arXiv preprint arXiv:1907.12508, (2019).
- [40] F. WILCOXON, *Individual Comparisons by Ranking Methods*, Biometrics Bulletin, 1 (1945), p. 80, <https://doi.org/10.2307/3001968>, <http://www.jstor.org/stable/10.2307/3001968?origin=crossref>.
- [41] R. WILLIAMS, *Understanding and interpreting generalized ordered logit models*, The Journal of Mathematical Sociology, 40 (2016), pp. 7–20.