A UNIMODAL, UNCERTAINTY-AWARE DEEP LEARNING APPROACH FOR ORDINAL REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Ordinal regression is an important area in machine learning, and many algorithms were proposed to approach it. In this work, we propose an ordinal regression prediction algorithm, based on deep learning machinery and inspired by the well-known Proportional Odds model. Our proposed approach has three key components: first, it is designed to guarantee unimodal output probabilities, which is a desired element in many real world applications. Second, we argue that the standard maximum likelihood is sub-optimal for ordinal regression problems and train our model using optimal transport loss, as it naturally captures the order of the classes. Third, we design a novel regularizer aiming to make the model uncertainty-aware, in the sense of making the model more confident about correct predictions, comparing to wrong predictions. In addition, we propose a novel uncertainty-awareness evaluation measure. Experimental results on eight real-world datasets demonstrate that our proposed approach consistently performs on par with and often better than several recently proposed deep learning approaches for ordinal regression, in terms of both accuracy and uncertainty-awareness, while having a guarantee on the output unimodality.

1 INTRODUCTION

Ordinal regression is an area of supervised machine learning, where the goal is to predict the value of a discrete dependent variable, whose set of (symbolic) possible values is ordered. Despite often overshadowed by more common applications like classification and regression, ordinal regression covers a wide range of important applications, such as prediction of failure times, ranking, age estimation and many more.

Many practitioners often treat ordinal regression problems as classification or regression problems (for example, this was indeed the case with many submissions to Kaggle's Diabetic Retinopathy competition[1] in 2015). While having common characteristics with both classification and regression, ordinal regression can arguably be viewed a mid-point between the two. An ordinal model is of course similar to a classification model, as both predict a discrete value ("label") out of a finite set of possible ones. However, the existence of an order on set of labels, when available, can potentially lead to an improved performance, comparing to a standard classifier, which does not assume such order. This typically occurs via distinguishing between the severity of prediction mistakes: while in classification typically "all mistakes are created equal", in ordinal regression different mistakes may be associated with different severity (for example, in the context of tumor grade prediction, predicting "3" when the ground truth value is "4" may be less severe than a "1" prediction). In regression problems, the dependent variable naturally does take values from an ordered set. However, this set is typically a continuum. Moreover, regression performance may be sensitive to monotonic transformations of the dependent variable, while such sensitivity does not take place in ordinal regression problems, as the order is invariant to monotonic transformations. Hence one may expect that typical ordinal regression algorithms have potential to outperform classification or regression approaches, when the range of the dependent variable is finite and ordered.

The arguably most fundamental ordinal regression model is the Proportional Odds Model (POM), a generalized linear model, similar in spirit to logistic regression, however where the logits are defined

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

for cumulative probabilities. POM is typically trained via maximum likelihood (as is also the case for several recently proposed deep ordinal regression approaches, which will be reviewed in section 2). We argue that likelihood is a sub-optimal measure of quality for ordinal regression setup, as it only considers the probability mass the model assigns to the true class, ignoring the remaining mass. This implicitly assumes that “all mistakes are equal”, which, as discussed above, is not the case for ordinal regression. Hence we seek for an alternative measure of quality which may be more appropriate for ordinal regression. We argue that the optimal transport divergence might be a better fit. In addition, this divergence turns out to be particularly appealing, as it obtains a simple, differentiable form in the case that one of the distributions is Dirac, which is indeed the case in ordinal regression; this will be explained in Section 3.

Another potential source of sub-optimality of POM (and of several recently-proposed approaches for deep ordinal regression) is the often-reasonable requirement that a probabilistic model for ordinal regression will output unimodal probabilities. A k-level multinomial distribution is called *unimodal* if there exists $j \in \{1, \ldots, k\}$ such that $\Pr(Y = 1) \leq \ldots \leq \Pr(Y = j) \geq \ldots \geq \Pr(Y = k)$. Although there are domains in which unimodality is not necessarily a desirable property, such as in movie rankings (where people may have either positive or negative definitive opinions about a particular movie), in many other real world domains it is a natural requirement, for example when predicting age of a person or a grade of a tumor, as it may be counter-intuitive to trust a model prediction which says that a predicted tumor grade is either "1" or "4", but not "2" or "3". However, despite often being a desired characteristic, unimodality is unfortunately not always fulfilled. While this was identified by several recent works for deep ordinal regression, unimodality is often encouraged (but not enforced) via soft targets. In Section 2, we will argue that this is a sub-optimal means to achieve unimodality.

In the current manuscript we therefore propose a novel approach for ordinal regression, based on deep learning machinery, which tackles the three issues pointed above: (i) it contains a mechanism to enforce unimodality of the output distribution, implemented via architectural design, (ii) it is trained via optimal transport loss (iii) it utilizes a novel regularization term, whose goal is to make the predictions uncertainty-aware, in the sense of making the model more confident in cases of correct predictions, comparing to cases of incorrect predictions. As a by product, we also propose a corresponding uncertainty-awareness evaluation term, which we call Entropy Ratio.

Experimentally, we provide results on eight real world image benchmark datasets which demonstrate that our proposed approach consistently performs on par with and often better than several recently proposed approaches for deep ordinal regression, in terms of both prediction accuracy and uncertainty-awareness, while having a unimodality guarantee.

2 Related Work

Being a traditional area of machine learning and statistics, there exists a large corpus of literature on ordinal regression. In this section we focus on approaches based on deep architectures. Several such approaches were proposed in the recent years. A common approach seems to be to turn the ordinal regression problem into a multi-label classification problem, for example Fu et al. (2018); Liu et al. (2017; 2018b); TV et al. (2019); Berg et al. (2020); Cheng et al. (2008); Li et al. (2021). We argue that the multi-label approach has two major problematic aspects: first, the output probabilities are not always guaranteed to be consistent, in the sense of increasing cumulative distribution (i.e., we would like to predict $\Pr(y \leq 1) \leq \Pr(y \leq 2) \leq \ldots \leq \Pr(y \leq k)$. Second, even if the output probabilities are consistent, as is the case in Liu et al. (2018a); Cao et al. (2020) for example, the predicted class probabilities are not necessarily unimodal. This is the case in several recent works, e.g., Liu et al. (2019b); Vargas et al. (2020); Pan et al. (2018); Kook et al. (2020).

Beckham & Pal (2017) proposed an elegant mechanism to obtain unimodal output probabilities, based on either the Poisson or the Binomial distributions, which are both unimodal. In both cases their model outputs a scalar (λ in case of the Poisson, p in the case of the binomial), which is then mapped to a probability mass function that uses (after normalization) as the model output probabilities. While being a convenient, architectural-based solution for the unimodality issue, their approach is inherently limited in its ability to express the level of uncertainty of the model’s prediction. To see why, note that since a single parameter determines both the location of the mode, and the decay of the probabilities, the model cannot output a highly flat or highly peaked probability vector, for example. In addition, instances of the same predicted class ought to have similar output probabilities, hence the predictions cannot reflect any level of uncertainty.

A different approach to unimodality has been to train the model with soft targets, for example Gao et al. (2017); Diaz & Marathe (2019); Liu et al. (2019a; 2020). However, we argue that utilization of soft targets suffers from two important disadvantages. First, unimodality is only encouraged, but not enforced. In section 5 we will see cases where such models trained with soft targets actually yield large amounts of non-unimodal predictions. Second, soft targets have a a pre-defined decay pattern, which is determined a-priori and hence does not reflect any level of uncertainty with respect to the prediction. Therefore, they are equivalent (in the sense of 1:1 map) to Dirac predictions (i.e., "one-hot"), and are devoid of any probabilistic insight whatsoever. As will be explained in the sequel, our approach attends both issues: we guarantee unimodal outputs, by design, and yield uncertainty-aware predictions. Other approaches we choose to mention in this context are Belharbi et al. (2019), where a constrained optimization is used to achieve unimodality on the train data. However, this does not provide any guarantee regarding the predictions on test data. Uncertainty-awareness considerations appear also in Li et al. (2021) (by mapping each input instance to a distribution in the code space) and Liu et al. (2019b) (via a Gaussian process machinery). Apart from not being unimodal, both approaches suffer from another issue, current in several works, which we explain next.

Several works use cross entropy as a training objective, while using one-hot (or binary) targets, see, for example Belharbi et al. (2019); Vargas et al. (2020); Fu et al. (2018); Beckham & Pal (2017); Berg et al. (2020); TV et al. (2019); Cao et al. (2020). As pointed out in several papers, and will also be demonstrated in section 3, in the case of one-hot targets, the cross entropy term equals the negative log of the probability assigned by the model to the true class, making it invariant to the distribution of the remaining probability mass. While a reasonable thing in a standard classification setting, this ignores the order of the classes, making it a sub-optimal choice for ordinal regression setting. To overcome this limitation of cross entropy Hou et al. (2016), followed by Beckham & Pal (2017); Liu et al. (2019a) use optimal transport loss, which is a natural way to incorporate the order of the classes into the loss term. In this sense, it is similar to the approach we take in this manuscript.

To summarize this section, we identify the following key requirements for an appropriate ordinal regression model: (i) unimodality of the model's output distribution, preferably guaranteed via architectural design (ii) the training objective function should correspond to the order induced on the label space (iii) the decay of the output probabilities should reflect the uncertainty of the model in its predictions.

These requirements naturally lead us to our proposed approach in section 4. However, before we specify it, we begin with a brief review of of the proportional odds model and optimal transport divergence.

3 Preliminaries

We begin this section with a description of the proportional odds model from a latent variable perspective. We then briefly review optimal transport as a divergence between two probability distributions.

3.1 The Proportional Odds Model

Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be random variables, having joint probability $\mathcal{P}_{XY}$, where $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{1, \ldots, k\}$, and $1, \ldots, k$ are considered as symbols. Let $\preceq$ be an order relation defined on $\mathcal{Y}$ such that $1 \preceq \ldots \preceq k$. The proportional odds model is parametrized by $\alpha \in \mathbb{R}^{k-1}, \beta \in \mathbb{R}^d$ and applies to data $\{(x_i, y_i)\}_{i=1}^n$, sampled i.i.d from $\mathcal{P}_{XY}$.

Let ϵ be a logistic random variable (thus having a sigmoid cumulative distribution function $F(x) = \frac{1}{1+\exp(-x)}$), and let Z be a random variable defined as $Z = \beta^T X + \epsilon$. The entries of α use to define the cumulative conditional probabilities via

$$\Pr(Y \preceq j | X = x) = \Pr(Z \leq \alpha_j) = F(\alpha_j - \beta^T x). \tag{1}$$

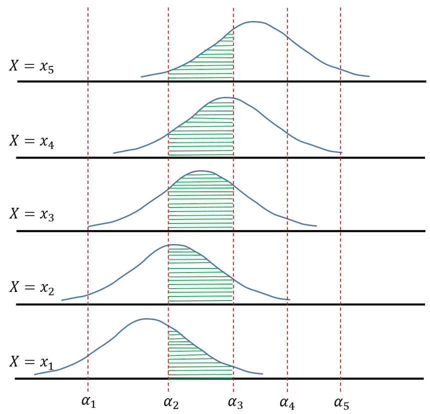

Figure 1: The proportional odds model. x_i is a realization of X. The standard logistic density is shifted by $\beta^T x_i$. The thresholds α_j define the bins which determine the probability predicted by the model to each class. For example, the green area defines the probability $\Pr(Y = 3)$.

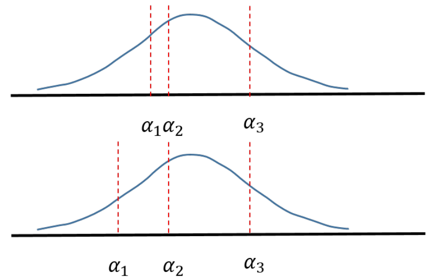

Figure 2: POM deficiencies: (a) POM does not always output unimodal probabilities: the above plot shows an example where the output probabilities are such that $\Pr(Y = 1) > \Pr(Y = 2) < \Pr(Y = 3) > \Pr(Y = 4)$. (b) The likelihood function of POM is invariant to the way the predicted probability mass of the incorrect classes is assigned: if the correct class is 3, the two instances have the same likelihood, despite the fact that in the bottom case, the probability mass assigns to neighboring class 2 is larger.

Similarly to logistic regression, this yields linear log-odds (logits), however, defined with respect to cumulative terms

$$\gamma_j \equiv \log \frac{\Pr(Y \preceq j|X = x)}{\Pr(Y \succ j|X = x)} = \alpha_j - \beta^T x.$$

It is convenient to interpret equation (1) by viewing $\beta^T x$ as a factor that shifts the standard logistic density function, while the α_j terms are thresholds, with respect to which the cumulative probabilities are defined. This is depicted in Figure 1. Let (x, y) be a realization of (X, Y). The likelihood assigned by the model to (x, y) is

$$\begin{aligned} L(\alpha, \beta; (x, y)) =& \Pr(Y = y|X = x; \alpha, \beta) \\ =& F(\alpha_y - \beta^T x) - F(\alpha_{y-1} - \beta^T x), \end{aligned} \tag{2}$$

considering $\alpha_0 = -\infty$ and $\alpha_k = \infty$. The model is typically trained in a standard fashion by maximizing the log-likelihood function on the training data.

Despite its popularity, the POM suffers from two main issues: First, the model's output probabilities are not necessarily unimodal. This is depicted in Figure 2. Second, the likelihood function in equation (2) depends only on the probability the model assigns to the correct class y, and is invariant to the way the remaining probability mass is assigned by the model. This ignores the order on the label set, and hence does not use important information that might be used to improve prediction quality, as depicted in Figure 2. The same issue occurs in the ordinal likelihood, proposed in Chu et al. (2005). In addition, it is important to mention that as the cross entropy term is essentially equivalent to model's negative log-likelihood function, this invariance to the partition of the remaining mass over the incorrect classes is common to all models trained via cross entropy minimization, as long as the target labels are one-hot. In section 4 we will show how our method overcomes these two limitations of the POM.

3.2 Optimal Transport

Let (M, d) be a finite metric space, and let p, q be probability mass functions defined on M. Optimal transport, also denoted as the 1-Wasserstein distance and the Earth Mover Distance, between p and q is

$$OT(p, q) = \inf_{\gamma \in \Gamma} \int_{M \times M} c(x, y) d\gamma(x, y), \tag{3}$$

where Γ is the set of all joint probabilities on $M \times M$, having marginals p and q, and the metric c specifies the costs of moving probability mass between every two elements of M. This amounts to the

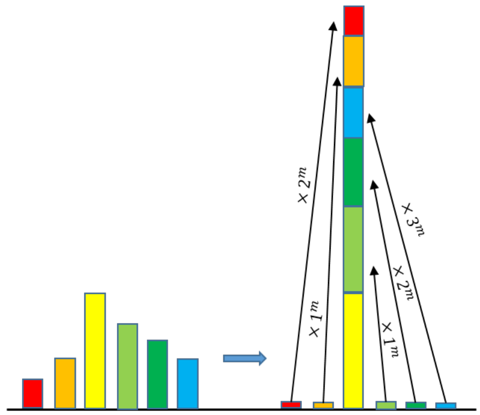

Figure 3: Optimal transport between model output probability mass function and a Dirac (one-hot) probability mass, using $c(i,j) = |i-j|^m$ cost.

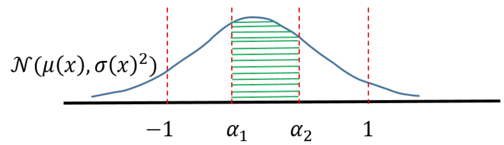

Figure 4: Generation of unimodal output probabilities for $k = 3$ classes. An input x is mapped to a (μ, σ) pair, which define a normal distribution $\mathcal{N}(\mu(x), \sigma(x)^2)$ over the real line. The output probabilities are proportional to the mass in the bins, which are of equal length. The green area equals to the un-normalized probability $\tilde{p}_2(x)$, corresponding to $\Pr(Y = 2|X = x)$.

optimal transportation of probability mass that transforms p into q and vice versa. In the general case, the distance can be found by solving a linear program, and several relaxations have been proposed to accelerate its computations while preserving its geometrical properties, see, for example, Peyré et al. (2019); Feydy et al. (2019); Cuturi (2013). However, in the case where p is a Dirac point mass (i.e., having a a one-hot probability mass function), solving equation (3) becomes trivial and becomes

$$OT(p,q) = \sum_{i=1}^{k} q_i c(i,j), \tag{4}$$

where j is the correct class, and k is the total number of classes, as is also depicted in Figure 3.

Letting q denote a model's output probabilities and p denote a one-hot target, equation (4) is of course differential with respect to the model outputs q and therefore can be used as a loss term for gradient-based optimization.

The cost metric c can incorporate domain knowledge in order to quantify the semantic distance between every two elements of M. Since in our case $M = \{1, \ldots, k\}$ is an ordered space, a natural possibility is to define $c(i,j) = |i-j|^m$, for some $m \geq 1$, i.e., ,mapping the symbolic class labels to consecutive integers, and computing powers of absolute differences. When $m = 1$, the optimal transport can also be computed as the ℓ_1 distance between the cumulative mass functions[2], $\|\text{CMF}(p) - \text{CMF}(q)\|_1$, (see Levina & Bickel (2001), for example). This is equivalent to the computation in equation (4), and also generalizes to arbitrary targets (i.e., not necessarily Dirac).

4 The Proposed Approach

In this section we describe our proposed mechanism for architectural-based generation of unimodal output probability distributions.

4.1 Rational

Achieving unimodality directly via architectural design has a major advantage, since the output probabilities are guaranteed to be unimodal for every input instance, as is also the case for the mechanism proposed by Beckham & Pal (2017). However, unlike Beckham & Pal (2017), our proposed approach employs the normal distribution, depending separately on a location parameter and a scale parameter, so that the location of the mode is detached from the decay of the probability mass. This adds flexibility to the shape if the output probability vector, comparing to Beckham &

[2]This holds when the classes are ordered.

Pal (2017) where single parameter determines both the mode and the decay. We will demonstrate in section 5 that this greater flexibility is helpful in expressing prediction uncertainty.

4.2 Unimodal Output Probabilities Generation

Inspired by the POM, we utilize thresholds to define bins, so that the total mass inside each bin is the output probability of the corresponding class. However, observe that the lack of unimodality of POM can be fixed by letting the bins be of equal length and remain fixed during training.

Therefore, instead of learning the thresholds, during training a map $x \mapsto (\mu, \sigma)$ is learned, where μ is a location parameter, and σ is a scale parameter, which define a $\mathcal{N}(\mu, \sigma^2)$ distribution, using which the output probabilities are computed.

Formally, we divide the range $[-1, 1]$ to k equal bins similarly to da Costa et al. (2008), where k is the number of classes, defined by $-1 = \alpha_0, \alpha_1, ..., \alpha_k = 1$, so that $\alpha_i - \alpha_{i-1} = \frac{2}{k}$. The (un-normalized) probabilities are given by

$$\tilde{p}_i(x) = \Phi_{\mu(x),\sigma(x)}(\alpha_i) - \Phi_{\mu(x),\sigma(x)}(\alpha_{i-1}), \tag{5}$$

where $\Phi_{\mu,\sigma}(\cdot)$ is the $\mathcal{N}(\mu, \sigma^2)$ cumulative distribution function, and we have emphasized that μ, σ are in fact functions of the input instance x. Since the bins cover [-1,1] and not the entire real line, we normalize the probabilities to obtain proper model predictions via

$$\Pr(Y = j|x) = p_i(x) \equiv \frac{\tilde{p}_i(x)}{\sum_{j=1}^{k} \tilde{p}_j(x)}. \tag{6}$$

To compensate for the fact that the probability generating mechanism depends on less parameters than POM (2 for the former, $d + k - 1$ for the latter), the map $x \mapsto (\mu, \sigma)$ is expressed via a deep network, which is therefore able to represent a complex nonlinear relation. Our proposed mechanism for generation of unimodal output probabilities is depicted in Figure 4.

The following lemma, proved in Appendix A, establishes that the model output probabilities are indeed unimodal.

Lemma 4.1. *Let $x \in \mathbb{R}^d$ be an input to the model, which is mapped to $\mu = \mu(x), \sigma = \sigma(x)$, and let $p_1, \dots p_k$ be the model output probabilities, generated via equation* (6)*. Then $p_1, \dots p_k$ define a unimodal multinomial random variable.*

We remark that all arguments made here with regard to the normal distribution also hold for other unimodal distributions, which are symmetric around μ, such as Logistic(μ, σ) and Cauchy(μ, σ), which both have slower decay patterns.

4.3 Uncertainty-Awareness Regularizer

The separation of the location and scale allows for flexibility in the representation of prediction uncertainty. Specifically, this uncertainty can be affected in two ways: (i) the standard deviation parameter (ii) the proximity of the mean parameter to the bin threshold.

To encourage uncertainty awareness, we argue that it is natural to require that the confidence in the prediction will be greater in cases of correct predictions, comparing to incorrect ones. In order to achieve this, we introduce a novel regularization term, defined for a minibatch $X = \{x_1, \dots x_m\}$ such that for instance x, whose label is y, the model, parametrized by θ outputs probability vector $p_\theta(x)$. We define the regulazier as

$$R_{\text{H}}(X; \theta) := \max\left\{0, \underset{x \in X:\text{mode}(p_\theta(x)) \neq y}{mean} H(p_\theta(x)) - \left(\underset{x \in X:\text{mode}(p_\theta(x)) = y}{mean} H(p_\theta(x)) + \epsilon_R\right)\right\} \tag{7}$$

where we write $p_\theta(x)$ to emphasize the fact that output probabilities depend on the model parameters, collectively denoted as θ, and $H(\cdot)$ denotes entropy. This hinge-like term penalizes the model if the average entropy over instances with incorrect predictions is not higher than the average entropy of instances with correct predictions by at least ϵ_R. In our experiments we used $\epsilon_R = 0.5$

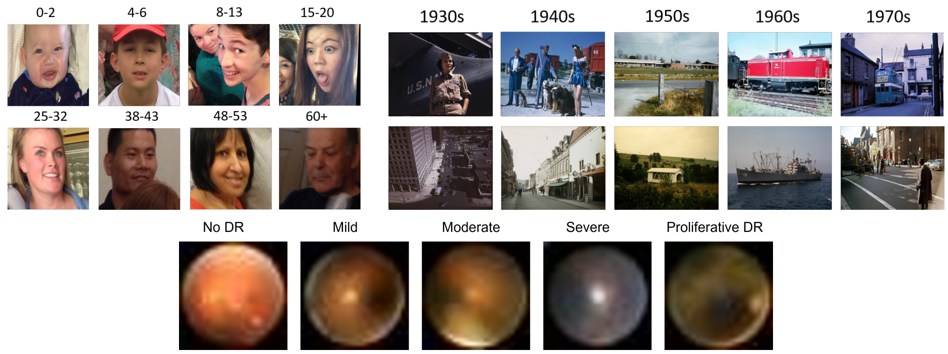

Figure 5: Top left: Examples from the Adience dataset. Age category is indicated above each image. Top right: Examples from the HCI dataset. Decades categories are indicated above. Bottom: Examples from the Retina mnist dataset. Diabetic Retinopathy classes are indicated above.

.

4.4 Training procedure

To summarize this section, given an instance x, our model outputs parameters $\mu_\theta(x_i), \sigma_\theta(x)$ of a normal distribution. $\mu_\theta(x), \sigma_\theta(x)$ determine the un-normalized categorical probabilities $\tilde{p}(x) = (p,\ i = 1, \ldots, k$ via equation(5), which are then normalized to $p_i(x),\ i = 1, \ldots, k$ using equation (6). Then given a minibatch X of instances, the objective equation used to train the model is

$$l(\theta; X) = \frac{1}{m} \sum_{x \in X} \|\text{CMF}(p_\theta(x)) - \text{CMF}(q(x))\|_1 + \gamma \cdot R_{\text{H}}(X; \theta), \tag{8}$$

where $q(x)$ is the k-length one hot vector corresponding to the true label y of x and γ is a penalty parameter.

5 Experimental Results

We report experimental results on eight real world benchmark image datasets.

5.1 Datasets

We evaluate our method on eight real world benchmark image datasets, involving various ordinal regression tasks: age-detection (Adience Eidinger et al. (2014), FG-Net Fu et al. (2014), AAF Cheng et al. (2019), AFAD-LITE Niu et al. (2016), WIKI Rothe et al. (2018)), bio-medical image classification (Retina-MNIST Yang et al. (2021)), historical image dating (HCI Palermo et al. (2012)) and image aesthetics estimation (EVA Kang et al. (2020)). A more detailed description of the datasets appears in Appendix B. Some examples from the Adience, HCI and Retina-MNIST datasets are shown in Figure 5.

5.2 Methods Compared

We compare our proposed approach to four recently proposed approaches for deep ordinal regression, with unimodal output probabilities:

- DLDL Gao et al. (2017), an approach utilizing soft labels, generated using squared exponentially decaying distributions, trained using Kullback-Leibler divergence minimization (equivalent to cross entropy minimization).
- SORD Diaz & Marathe (2019), an approach utilizing soft labels, generated using linear exponentially decaying distributions, trained using Kullback-Leibler divergence minimization.
- Beckham & Pal (2017), where architectural-based unimodal output probabilities are generated using binomial distribution (single-learned parameter), trained using optimal transport loss.

- Liu et al. (2019a), an approach utilizing soft labels, created as a mixture of Dirac, uniform and linear exponentially decaying distributions, trained using optimal transport loss.

In order to perform a fair comparison, we implemented all methods, using the same image transformations, backbone CNN and training procedures, so that the methods differ only in their output layer architectures and loss functions. We performed 5 independent trials, using the same train-validation-test splits for all methods. Additional technical details can be found in Section C. For reproducibility, the supplementary material contains code reproducing the results reported in this section.

5.3 Evaluation Metrics

We report several commonly-used evaluation metrics for ordinal regression tasks: Mean Absolute Error (MAE), One-Off Accuracy (OOA), Spearman correlation, Quadratic Weighted Kappa (QWK), as well as the percentage of test examples with unimodal predicted output probabilities.

In addition, following the discussion in section 4.3, we evaluate the level of overconfidence by comparing the ratio of the average entropy in incorrect predictions, to the average entropy in correct predictions, that is

$$\text{Entropy ratio} = \frac{mean_{x\in X:\text{mode}(p_\theta(x))\neq y}\, H(p_\theta(x))}{mean_{x\in X:\text{mode}(p_\theta(x))= y}\, H(p_\theta(x))}, \tag{9}$$

where p_θ is the model output probabilities for test instance x. A higher ratio is therefore desirable and indicates a lower level of overconfidence, and correspondingly better uncertainty awareness.

5.4 Results on Real World Datasets

Table 6 shows the test results of each method on the eight benchmark datasets. As can be seen, the proposed approach performs at least on-par and often better than the compared baselines, in a fairly consistent manner, across the various datasets and evaluation metrics. In addition, observe that only the proposed approach and the method of Beckham & Pal (2017) output unimodal probabilities, both via architectural design, while the other baselines, trained using soft targets, do not always output unimodal probabilities. Moreover, in all eight datasets, the proposed approach outperforms all baselines also in terms of overconfidence.

6 Conclusion

In this manuscript we presented an approach for deep ordinal regression, inspired by the proportional odds model, utilizing an architectural mechanism for generation of unimodal output probabilities, trained using optimal transport objective, augmented with a novel regularizer which encourages uncertainty awareness. We demonstrated that while performing on-par with or better than other recently proposed approaches for ordinal regression, our proposed method enjoys the benefits of guaranteed unimodal output probabilities, and of less overconfidence about its incorrect predictions.

In terms of limitations of the proposed approach, perhaps the most obvious one is that the cost matrix of the OT loss needs to be specified beforehand, and is sensitive to transformations. While in all our experiments we have used $m = 1$ in equation (4), other options are acceptable as well. Another limitation is that the shape of our output probability vector, as determined by the truncated Gaussian distribution, naturally has limited expressiveness. Yet, changing the Gaussian link to other distributions, we haven't noticed major differences (see appendix D).

In terms of societal impact, in this research several age estimation datasets were used as benchmarks for the proposed method. All data we used is based on the public research papers without any private information of the subjects. We do not expect any privacy violations while using these datasets. In addition, models trained for age estimation task could include biases based on ethnicity or gender, but to the best of our knowledge the datasets we used were built with fair representation of different groups of people.

Table 1: Performance of various methods on real world datasets, in a mean ± std format.

Dataset	Method	MAE ↓	OOA ↑	Spearman ↑	QWK ↑	% Unimodal ↑	Entropy ratio ↑
HCI	Beckham & Pal (2017)	.62 ± .04	.85 ± .02	.71 ± .03	.75 ± .02	**1 ± 0**	.82 ± .1
	Liu et al. (2019a)	.57 ± .05	.86 ± .02	.68 ± .04	.74 ± .03	.4 ± .04	.81 ± .1
	Gao et al. (2017)	.71 ± .04	.87 ± .02	.67 ± .03	.7 ± .03	.99 ± .02	1.28 ± .13
	Diaz & Marathe (2019)	.57 ± .04	.86 ± .02	.69 ± .03	.75 ± .02	.96 ± .01	.71 ± .06
	Proposed	**.55 ± .03**	**.88 ± .01**	**.72 ± .03**	**.77 ± .02**	**1 ± 0**	**1.42 ± 0.13**
Adience	Beckham & Pal (2017)	.53 ± .08	.94 ± .02	**.9 ± .02**	.91 ± .03	**1 ± 0**	1 ± .01
	Liu et al. (2019a)	.48 ± .06	.94 ± .02	.88 ± .03	.9 ± .03	.53 ± .05	1.08 ± .01
	Gao et al. (2017)	.5 ± .08	.94 ± .02	.88 ± .02	.9 ± .03	.6 ± .05	1. ± .01
	Diaz & Marathe (2019)	.47 ± .07	.94 ± .02	.89 ± .01	.91 ± .03	.99 ± .01	1.06 ± .03
	Proposed	**.45 ± .05**	**.95 ± .01**	**.9 ± .02**	**.92 ± .02**	**1 ± 0**	**1.19 ± .16**
Retina MNIST	Beckham & Pal (2017)	.78 ± .02	.8 ± .01	.6 ± .01	.55 ± .02	**1 ± 0**	1.02 ± 0.011
	Liu et al. (2019a)	.69 ± .02	.82 ± .01	.6 ± .02	.58 ± .02	.69 ± .04	1.05 ± .02
	Gao et al. (2017)	.76 ± .09	.8 ± .04	.59 ± .06	.54 ± .09	.94 ± .07	1. ± .01
	Diaz & Marathe (2019)	.77 ± .06	.79 ± .3	.57 ± .05	.56 ± .05	.88 ± .03	1.13 ± .01
	Proposed	**.68 ± .01**	**.83 ± .01**	**.62 ± .01**	**.6 ± .01**	**1 ± 0**	**2.08 ± .08**
FG-NET	Beckham & Pal (2017)	.46 ± .01	.94 ± .03	.75 ± .04	.8 ± .03	**1 ± 0**	1.01 ± .01
	Liu et al. (2019a)	.36 ± .05	.96 ± .01	.82 ± .06	.83 ± .05	.2 ± .04	1.27 ± .1
	Gao et al. (2017)	.46 ± .05	.94 ± .02	.75 ± .05	.77 ± .04	.09 ± .03	1. ± .1
	Diaz & Marathe (2019)	.38 ± .05	.95 ± .02	.8 ± .04	.83 ± .04	.98 ± .01	1.13 ± .04
	Proposed	**.35 ± .03**	**.98 ± .01**	**.84 ± .02**	**.87 ± .03**	**1 ± 0**	**1.58 ± 0.16**
AAF	Beckham & Pal (2017)	.44 ± .01	.97 ± .01	**.83 ± .01**	.85 ± .05	**1 ± 0**	1. ± .01
	Liu et al. (2019a)	.39 ± .01	**.98 ± .01**	.82 ± .01	.85 ± .01	.7 ± .03	1.08 ± .01
	Gao et al. (2017)	.4 ± .01	.98 ± .01	.82 ± .01	.85 ± .01	.9 ± .03	1. ± .01
	Diaz & Marathe (2019)	.39 ± .01	**.98 ± .01**	.82 ± .02	.85 ± .01	**1 ± .01**	1.05 ± .01
	Proposed	**.38 ± .01**	**.98 ± .01**	**.83 ± .01**	**.86 ± .01**	**1 ± .0**	**1.2 ± 0.01**
AFAD-LITE	Beckham & Pal (2017)	.51 ± .01	.91 ± .01	.67 ± .01	.69 ± .01	**1 ± 0**	2.18 ± .05
	Liu et al. (2019a)	.5 ± .01	.92 ± .01	.67 ± .01	.69 ± .01	.96 ± .01	1.11 ± .01
	Gao et al. (2017)	.5 ± .01	**.93 ± .01**	.67 ± .01	.69 ± .01	**1 ± .01**	1 ± .01
	Diaz & Marathe (2019)	.5 ± .01	.92 ± .01	.67 ± .01	.69 ± .01	**1 ± .01**	1.1 ± .01
	Proposed	**.49 ± .01**	**.93 ± .01**	**.68 ± .01**	**.7 ± .01**	**1 ± 0**	**2.26 ± .23**
EVA	Beckham & Pal (2017)	.63 ± .02	.92 ± .01	**.6 ± .03**	**.6 ± .02**	**1 ± 0**	1. ± .01
	Liu et al. (2019a)	.61 ± .02	.92 ± .01	.56 ± .03	.56 ± .03	.68 ± .02	1.04 ± .01
	Gao et al. (2017)	.62 ± .03	.92 ± .01	.55 ± .03	.54 ± .03	.91 ± .02	1. ± .01
	Diaz & Marathe (2019)	.59 ± .03	.93 ± .01	.57 ± .03	.55 ± .03	.99 ± .01	1.01 ± .01
	Proposed	**.59 ± .02**	**.94 ± .01**	.54 ± .03	.56 ± .03	**1 ± 0**	**1.1 ± .05**
WIKI	Beckham & Pal (2017)	.68 ± .01	.92 ± .01	.68 ± .01	.68 ± .01	**1 ± 0**	1. ± .01
	Liu et al. (2019a)	**.42 ± .01**	95. ± .01	.69 ± .01	.7 ± .01	.95 ± .01	1.05 ± .01
	Gao et al. (2017)	.44 ± .01	.94 ± .01	.68 ± .01	.7 ± .01	.97 ± .01	1. ± .01
	Diaz & Marathe (2019)	.44 ± .01	.94 ± .01	.69 ± .01	.7 ± .01	.99 ± .01	1.03 ± .01
	Proposed	.43 ± .0	**.95 ± .00**	**.72 ± .01**	**.7 ± .00**	**1 ± 0**	**1.27 ± .01**

REFERENCES

Christopher Beckham and Christopher Pal. Unimodal probability distributions for deep ordinal classification. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 411–419, 2017.

Soufiane Belharbi, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger. Non-parametric uni-modality constraints for deep ordinal classification. *arXiv*, pp. arXiv–1911, 2019.

Axel Berg, Magnus Oskarsson, and Mark O'Connor. Deep ordinal regression with label diversity. *arXiv preprint arXiv:2006.15864*, 2020.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. A neural network approach to ordinal regression. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1279–1284. IEEE, 2008.

Jingchun Cheng, Yali Li, Jilong Wang, Le Yu, and Shengjin Wang. Exploiting effective facial patches for robust gender recognition. *Tsinghua Science and Technology*, 24(3):333–345, 2019.

Wei Chu, Zoubin Ghahramani, and Christopher KI Williams. Gaussian processes for ordinal regression. *Journal of machine learning research*, 6(7), 2005.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013.

Joaquim F Pinto da Costa, Hugo Alonso, and Jaime S Cardoso. The unimodal model for the classification of ordinal data. *Neural Networks*, 21(1):78–91, 2008.

Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4738–4747, 2019.

Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690, 2019.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2002–2011, 2018.

Yanwei Fu, Timothy M Hospedales, Tao Xiang, Shaogang Gong, and Yuan Yao. Interestingness prediction by robust learning to rank. In *European conference on computer vision*, pp. 488–503. Springer, 2014.

Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*, 2016.

Chen Kang, Giuseppe Valenzise, and Frédéric Dufaux. Eva: An explainable visual aesthetics dataset. In *Joint Workshop on Aesthetic and Technical Quality Assessment of Multimedia and Media Analytics for Societal Trends*, pp. 5–13, 2020.

Lucas Kook, Lisa Herzog, Torsten Hothorn, Oliver Dürr, and Beate Sick. Deep and interpretable regression models for ordinal outcomes. *arXiv preprint arXiv:2010.08376*, 2020.

Elizaveta Levina and Peter Bickel. The earth mover's distance is the mallows distance: Some insights from statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pp. 251–256. IEEE, 2001.

Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13896–13905, 2021.

Xiaofeng Liu, Yang Zou, Yuhang Song, Chao Yang, Jane You, and BV K Vijaya Kumar. Ordinal regression with neuron stick-breaking for medical diagnosis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0, 2018a.

Xiaofeng Liu, Xu Han, Yukai Qiao, Yi Ge, Site Li, and Jun Lu. Unimodal-uniform constrained wasserstein training for medical diagnosis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019a.

Xiaofeng Liu, Fangfang Fan, Lingsheng Kong, Zhihui Diao, Wanqing Xie, Jun Lu, and Jane You. Unimodal regularized neuron stick-breaking for ordinal classification. *Neurocomputing*, 2020.

Yanzhu Liu, Adams Wai-Kin Kong, and Chi Keong Goh. Deep ordinal regression based on data relationship for small datasets. In *IJCAI*, pp. 2372–2378, 2017.

Yanzhu Liu, Adams Wai Kin Kong, and Chi Keong Goh. A constrained deep neural network for ordinal regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 831–839, 2018b.

Yanzhu Liu, Fan Wang, and Adams Wai Kin Kong. Probabilistic deep ordinal regression based on gaussian processes. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5301–5309, 2019b.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4920–4928, 2016.

Frank Palermo, James Hays, and Alexei A Efros. Dating historical color images. In *European Conference on Computer Vision*, pp. 499–512. Springer, 2012.

Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5285–5294, 2018.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4): 144–157, 2018.

Vishnu TV, Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, et al. Data-driven prognostics with predictive uncertainty estimation using ensemble of deep ordinal regression models. *arXiv preprint arXiv:1903.09795*, 2019.

Víctor Manuel Vargas, Pedro Antonio Gutiérrez, and César Hervás-Martínez. Cumulative link models for deep ordinal classification. *Neurocomputing*, 2020.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *arXiv preprint arXiv:2110.14795*, 2021.

A Proof of Lemma 4.1

Proof. Let p_i, p_{i+1} be the output probabilities of two adjacent classes, and let $-1 = \alpha_0, \ldots, \alpha_k = 1$ be the thresholds. We will show that (i) if $\mu \leq \alpha_{i-1}$ then $p_i \geq p_{i+1}$. Symmetrical argument will then imply that if $\mu \geq \alpha_{i+1}$ then $p_{i+1} \geq p_i$ (ii) if $\mu \in (\alpha_{i-1}, \alpha_i)$ then $p_i > p_{i+1}$, whenever the latter exists. Similarly, this would imply that $p_i > p_{i-1}$. Together, (i) and (ii) will imply the statement of the lemma.

Denote by f the density of the $\mathcal{N}(\mu, \sigma^2)$ distribution. To prove (i) observe that $\tilde{p}_i > \frac{2f(\alpha_i)}{k} > \tilde{p}_{i+1}$, which implies $p_i > p_{i+1}$.

To prove (ii), divide the i'th bin to two sub-bins $B_{i,1}, B_{i,2}$, of lengths $a = \mu - \alpha_{i-1}$ and $b = \alpha_i - \mu$, respectively. Similarly, divide the $i+1$'th bin to two bins $B_{i+1,1}, B_{i+1,2}$ lengths b and a. Then from (i)

$$\int_{B_{i,2}} f(x)dx > \int_{B_{i+1,1}} f(x)dx. \tag{10}$$

In addition, observe that

$$\begin{aligned}\int_{B_{i,1}} f(x)dx &= \int_0^a f(\mu + x)dx \\ &> \int_0^a f(\mu + 2b + x)dx \\ &= \int_{B_{i+1,2}} f(x)dx. \end{aligned} \tag{11}$$

Adding up equations (10) and (11), we obtain $\tilde{p}_i > \tilde{p}_{i+1}$, which gives $p_i > p_{i+1}$.

□

B Datasets

Tabel 2 contains information on the benchmark datasets used for our experiments.

HCI (Historical Color Image) dataset contains 1326 images, partitioned to 5 classes, corresponding to decades from 1930s to 1970s, and the task is to associate each images with the decade it was taken at. Random affine, radom horizontal / vertical flips and random crops of 224 transformations are applied during the training. The images are normalized in each color channel with mean and standard deviation of 0.5. The dataset was randomly split to the train/test as described in the Table 3.

Adience: During the training the images are resized to (256,256) and random crop of size 224 and random horizontal flip are applied as augmentations.

FG-Net: We partitioned the dataset to 8 classes, corresponding to decades. Augmentations were same as in the Adience experiment.

RetinaMNIST dataset has 5 classes and we apply random affine, horizontal and vertical flips as augmentations during the training. The size of the images is (28,28) as it provided by the dataset contributors. The train/test splits are proved by the contributors and were used as-is.

AFAD-LITE is a subset of the full AFAD dataset, which contains images of 22 continuous ages (for 15 to 40), in the amount of 60K. We partitioned the dataset into 4 classes and augmentations were the same as in the Adience experiment.

EVA (Explainable Visual Aesthetics) dataset containes 4070 images aesthetically ranked from 0 to 10 by multiple voters. We calculate the average score for each image and partition the data into 5 classes in accordance with the average score. Augmentations were same as in the Adience experiment.

AAF (All-Age-Faces) dataset is already pre-processed and contains 13,322 face images (mostly Asian), distributed across all ages (from 2 to 80). We partitioned the dataset into 6 classes and augmentations were the same as in the Adience experiment.

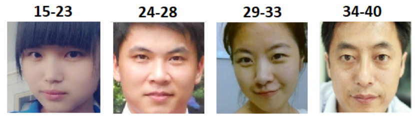

Figure 6: Examples from the AFAD-LITE dataset. Age classes are indicated above.

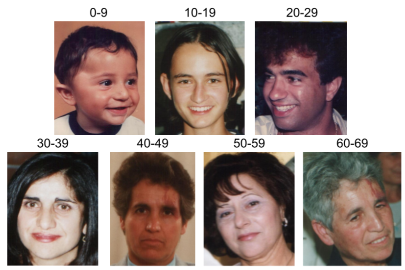

Figure 7: Examples from the FG-Net dataset. Age classes are indicated above.

WIKI dataset contains 62,328 images of celebrities from wikipedia with ages ranging form 1 to 100. Images with wrong timestamps are removed, and the dataset is partitioned into 6 classed. Augmentations were same as in the Adience experiment.

Figures 6, 7, 8, 9, 10 show examples from the AFAD-LITE, FG-Net, EVA, AAF and WIKI datasets.

Table 2: Benchmark datasets characteristics

Dataset	Task	Train images	Test Images	Classes
Adience	age estimation	pre-defined splits		8
HCI	image dating	1,276	50	5
FG-Net	age estimation	902	100	8
RetinaMNIST	DR classification	1200	400	5
AFAD-LITE	age estimation	37980	11869	4
AAF	age estimation	9058	2665	6
EVA	aesthetics estimation	2940	611	5
WIKI	age estimation	38660	12082	6

C TECHNICAL DETAILS

Table 3 shows the technical details for the experiments on the real world benchmark datasets reported in this manuscript.

Table 3: Technical details of the experiments

Dataset	Epochs	Batch size	initial LR	Decay LR after (epochs)	Weight Decay	γ	ϵ_R
Adience	100	64	10^{-4}	40	10^{-5}	0	.5
HCI	500	16	10^{-4}	100, 300	10^{-3}	.1	.5
FG-Net	100	64	10^{-4}	40	10^{-4}	0	.5
RetinaMNIST	100	16	10^{-4}	80, 90	10^{-4}	0	.5
AFAD-LITE	100	64	10^{-4}	40	10^{-5}	0	.5
AAF	100	64	10^{-4}	40	10^{-5}	0	.5
EVA	100	64	10^{-4}	40	10^{-5}	.1	.5
WIKI	100	64	10^{-4}	40	10^{-5}	.25	.5

The Adam optimizer was used in all experiments, with the default $\beta = (0.9, 0.999)$. The means and standard deviations reported in table 6 are based on 10 repetitions of each experiment, differing in

Figure 8: Examples from the EVA dataset. Aesthetics classes are indicated above.

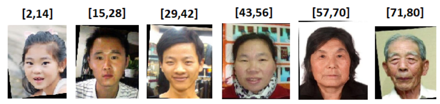

Figure 9: Examples from the AAF dataset. Age classes are indicated above.

weights initialization and random train-test splits, except for Adience, for which we repeated the experiment five times, using the same train-test splits as the creators of the dataset[3].

D EXTENDED EXPERIMENTS ON HCI DATASET WITH DIFFERENT DISTRIBUTION FUNCTIONS

Table 4: Performance of the proposed method on HCI dataset with different distribution functions, in a mean ± std format.

Distribution	MAE	OOA	Spearman	QWK	% Unimodal	Entropy ratio
Cauchy	.58 ± .04	.87 ± .02	.70 ± .03	.76 ± .02	1 ± 0	.81 ± .2
Laplace	.58 ± .03	.87 ± .01	.69 ± .02	.75 ± .02	1 ± 0	.83 ± .05
Logistic	.63 ± .09	.85 ± .03	.72 ± .06	.75 ± .04	1 ± 0	.92 ± .2
Normal	.54 ± .03	.89 ± .01	.70 ± .03	.77 ± .02	1 ± 0	.95 ± 0.13

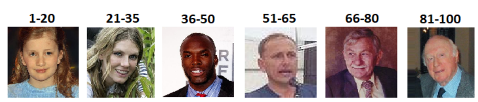

Figure 10: Examples from the WIKI dataset. Age classes are indicated above.

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