# Unimodal Likelihood Models for Ordinal Data

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Ordinal regression (OR) is the classification of ordinal data, in which the underlying target variable is categorical and considered to have a natural ordinal relation for the explanatory variables. In this study, we suppose the unimodality of conditional probability distributions as a natural ordinal relation of the ordinal data. Under this supposition, unimodal likelihood models are considered to be promising for improving the generalization performance in OR tasks. Demonstrating that previous unimodal likelihood models have a weak representation ability, we thus develop more representable unimodal models, including the most representable one. OR experiments in this study showed that the developed more representable unimodal models could yield better generalization performance for real-world ordinal data compared with previous unimodal models and popular statistical OR models having no unimodality guarantee.

## 1 Introduction

Ordinal regression (OR, also called ordinal classification) is the classification of ordinal data, in which the underlying target variable is categorical and considered to have a natural ordinal relation for the explanatory variables; see Section 2 for a detailed formulation. Typical examples of the target label set of ordinal data are sets of grouped continuous variables like age groups {'0 to 9 years old', '10 to 19 years old', . . . , '90 to 99 years old', 'over 100 years old'} and sets of assessed ordered categorical variables like human rating {'strongly agree', 'agree', 'neutral', 'disagree', 'strongly disagree'} (Anderson, 1984), and various practical tasks have been tackled within the OR framework: for example, face-age estimation (Niu et al., 2016; Cao et al., 2019; Anonymous, 2022), information retrieval (Liu, 2011), credit or movie rating (Kim & Ahn, 2012; Yu et al., 2006), and questionnaire survey in social research (Chen et al., 1995; Bürkner & Vuorre, 2019).

Consider an example of the questionnaire survey about support for a certain idea that requires subjects to respond from {'strongly agree', 'agree', . . . }. Here, it would seem possible that subjects, who have features specific to those who typically respond 'agree', respond 'neutral', but unlikely that they respond 'disagree'. Such a phenomenon can be rephrased as the unimodality of the conditional probability distribution (CPD) of the underlying events. The hypothesis "many statisticians or practitioners often judge that the data have a natural ordinal relation and decide to treat them within the OR framework, with unconsciously expecting their unimodality" may be convincing, as we will experimentally confirm in Section 6.2 that many ordinal data, treated in previous OR studies that do not consider the unimodality, tend to have a unimodal CPD.

Commonly, the generalization performance of a classifier (OR method) based on statistical modeling depends on the underlying data distribution and the representation ability of the model. Recall that the generalization performance can be roughly decomposed into bias- and variance-dependent terms (well-known bias-variance decomposition); a model in which the representation ability is too strongly restricted will result in a large bias-dependent term if it cannot represent the underlying data distribution, and a model that is unnecessarily flexible to represent the underlying data distribution will result in a large variance-dependent term especially for a small-size training sample, and such models at both extremes may degrade their generalization performance. Therefore, assuming the unimodality of the CPD as a natural ordinal relation of ordinal data, unimodal likelihood models, which can adequately represent such unimodal data and are compact with respect to the representation ability than unconstrained statistical models such as multinomial logistic regression model, are considered to be promising for improving the generalization performance in OR tasks.

Existing studies (da Costa et al., 2008; Iannario & Piccolo, 2011; Beckham & Pal, 2017) were inspired by the shape of the probability mass function (PMF) of elementary categorical probability distributions such as binomial, Poisson, and uniform distributions and developed unimodal likelihood models. In this paper, we introduce several notions that characterize the shape of PMFs in Section 2.1, and, on the basis of them, show that their models have a weak representation ability in Section 3. We thus propose more representable unimodal models in Sections 4 and 5. In particular, a model described in Section 4.2 is the most strongly representable among the class of unimodal likelihood models.

We took experimental comparisons between 2 previous unimodal models, 2 popular statistical OR models without the unimodality guarantee, and 8 proposed unimodal models. As we show the experimental results and considerations in Section 6.3 and Appendix B, we confirmed that the proposed more representable unimodal models were effective in improving the generalization performances for the conditional probability estimation and OR tasks for many data that have been treated in previous OR studies as ordinal data.

## 2 Preliminaries

### 2.1 Ordinal Regression Tasks and Ordinal Data

The OR task is a classification task. Denoting explanatory and categorical target variables underlying the data as $\boldsymbol{X} \in \mathbb{R}^d$ and $Y \in [K] \coloneqq \{1, \ldots, K\}$, we formulate the OR task as searching for a classifier $f : \mathbb{R}^d \to [K]$ that is good in the sense that the task risk $\mathbb{E}[\ell(f(\boldsymbol{X}), Y)]$ becomes small for a specified task loss $\ell : [K]^2 \to [0, +\infty)$, where the expectation $\mathbb{E}[\cdot]$ is taken for all included random variables (here $\boldsymbol{X}$ and $Y$). Popular task losses in OR tasks include not only the zero-one loss $\ell_{\mathrm{zo}}(j,k) \coloneqq \mathbb{1}\{j \neq k\}$, where $\mathbb{1}\{c\}$ values 1 if a condition $c$ is true and 0 otherwise, but also V-shaped losses reflecting one's preference of smaller prediction errors over larger ones such as the absolute loss $\ell_{\mathrm{abs}}(j,k) \coloneqq |j-k|$ and the squared loss $\ell_{\mathrm{sq}}(j,k) \coloneqq (j-k)^2$.

In the OR framework, it is supposed that the underlying categorical target variable $Y$ of the data is equipped with an ordinal relation naturally interpretable in the relationship with the underlying explanatory variables $\boldsymbol{X}$, like examples described in the head of Section 1. We here assume that the target labels are encoded to $1, \ldots, K$ in an order-preserving manner, like from 'strongly agree', 'agree', 'neutral', 'disagree', 'strongly disagree' to $1, \ldots, 5$ or $5, \ldots, 1$. The OR framework considers the classification of such ordinal data. Note that, like most previous OR studies have discussed the OR without formal common understanding of what constitutes ordinal data and their natural ordinal relation, it is difficult to define ordinal data any more rigorously, and we in this paper refer to the data discussed in previous OR studies as ordinal data.

As we declared in Section 1, this study basically assumes, as a natural ordinal relation of ordinal data, the unimodality in the theoretical discussion or the almost-unimodality to real-world ordinal data, precisely defined in the following:

**Definition 1.** *For a vector* $\boldsymbol{p} = (p_k)_{k \in [K]} \in \mathbb{R}^K$*, we define* $M(\boldsymbol{p}) \coloneqq \min(\arg\max_k (p_k)_{k \in [K]})^1$*, and say that* $\boldsymbol{p}$ *is unimodal if it satisfies*

$$p_1 \leq \cdots \leq p_{M(\boldsymbol{p})} \text{ and } p_{M(\boldsymbol{p})} \geq \cdots \geq p_K. \tag{1}$$

*Also, we call* $M(\boldsymbol{p})$ *the mode of* $\boldsymbol{p}$ *if* $\boldsymbol{p}$ *is a PMF satisfying* $\boldsymbol{p} \in \Delta_{K-1}$*, where* $\Delta_{K-1}$ *is the* $(K-1)$*-dimensional probability simplex* $\{(p_k)_{k \in [K]} \mid \sum_{k=1}^K p_k = 1, p_k \in [0,1] \text{ for } k = 1, \ldots, K\}$*. Moreover, if the CPD* $(\Pr(Y = y | \boldsymbol{X} = \boldsymbol{x}))_{y \in [K]}$ *is unimodal at any* $\boldsymbol{x}$ *in whole the domain* $\mathbb{R}^d$ *or in its sub-domain* $\mathcal{X} \subseteq \mathbb{R}^d$ *with a large probability* $\Pr(\boldsymbol{X} \in \mathcal{X})$*, we say that the data is unimodal or almost-unimodal.*

This paper presents a qualitative discussion on the representation ability of various statistical OR models: It aims to clarify whether each model is suitable for representing data that follow a distribution with various structural properties that we consider to be expectable in unimodal ordinal data. As a preparation for that discussion, we further introduce below the notions of decay rates, scale, homoscedasticity, heteroscedasticity, and skewness that are related to structural properties of the data distribution:

**Definition 2.** *For a PMF* $\boldsymbol{p} = (p_k)_{k \in [K]} \in \Delta_{K-1}$ *having a mode* $m \in [K]$*, we call* $\frac{p_k}{p_{k+1}}$ *for* $k = 1, \ldots, m-1$ *(if* $m \neq 1$*) and* $\frac{p_k}{p_{k-1}}$ *for* $k = m+1, \ldots, K$ *(if* $m \neq K$*) the decay rates (DRs) of* $\boldsymbol{p}$*.*

---

[1] arg min and arg max may return a set, and min is applied to convert it to a point statistic and simplify the discussion.

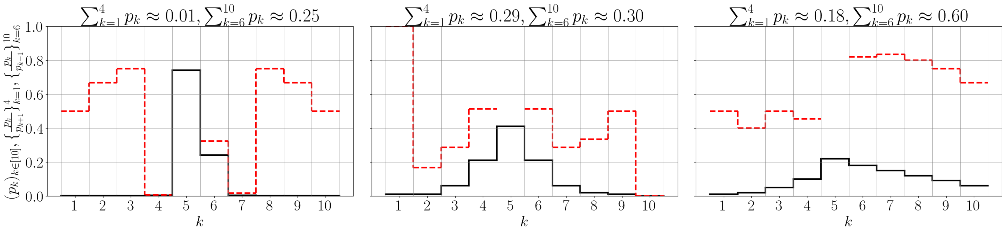


Figure 1: Instances of the unimodal 10-dimensional PMFs $(p_k)_{k\in[10]}$, in which the mode is 5, in black solid line, and their DRs $\{\frac{p_k}{p_{k+1}}\}_{k=1}^{4}$ and $\{\frac{p_k}{p_{k-1}}\}_{k=6}^{10}$ in red dotted line.

**Definition 3.** *For a PMF $\boldsymbol{p} = (p_k)_{k\in[K]} \in \Delta_{K-1}$ having a mode $m \in [K]$, we call $\sum_{k=1}^{m-1} p_k + \sum_{k=m+1}^{K} p_k = 1 - p_m$ the scale of $\boldsymbol{p}$, where we define the summation such that $\sum_{k=i}^{j} f(k)$ is zero irrelevant to the function $f$ as far as $i > j$. If the scale of the CPD $\boldsymbol{p}(\boldsymbol{x}) = (\Pr(Y = y|\boldsymbol{X} = \boldsymbol{x}))_{y\in[K]}$ is similar or dissimilar over whole the domain $\mathbb{R}^d$ of the explanatory variables $\boldsymbol{X}$, we say that the data is homoscedastic or heteroscedastic. In particular, we refer to heteroscedastic data as mode-wise heteroscedastic or overall heteroscedastic data, if the scale of their $\boldsymbol{p}(\boldsymbol{x})$ is similar or can be dissimilar in a domain where the conditional mode $\boldsymbol{M}(\boldsymbol{p}(\boldsymbol{x}))$ is the same.*

**Definition 4.** *For a PMF $\boldsymbol{p} = (p_k)_{k\in[K]} \in \Delta_{K-1}$ having a mode $m \in [K]$, we call $\sum_{k=1}^{m-1} p_k - \sum_{k=m+1}^{K} p_k$ the skewness of $\boldsymbol{p}$, and we say that $\boldsymbol{p}$ is skew or less skew when the absolute value of its skewness (absolute skewness) is large or small. If the absolute skewness of the CPD $\boldsymbol{p}(\boldsymbol{x}) = (\Pr(Y = y|\boldsymbol{X} = \boldsymbol{x}))_{y\in[K]}$ is large or small over whole the domain $\mathbb{R}^d$ of the explanatory variables $\boldsymbol{X}$, we say that the data is skew or less skew. In particular, we refer to skew data as mode-wise skew or overall skew data, if the skewness of their $\boldsymbol{p}(\boldsymbol{x})$ is similar or can be dissimilar in a domain where the conditional mode $\boldsymbol{M}(\boldsymbol{p}(\boldsymbol{x}))$ is the same.*

For a unimodal PMF, its DRs are smaller than or equal to 1. The scale and skewness of PMFs are typically treated as qualitative notions, but we measure them numerically for the sake of clarity (see $\sum_{k=1}^{m-1} p_k + \sum_{k=m+1}^{K} p_k$ and $\sum_{k=1}^{m-1} p_k - \sum_{k=m+1}^{K} p_k$ in Definitions 3 and 4). The numerical measure in each definition is intended just to further clarify our qualitative discussion, and we do not see any particular importance in that choice of the measure (other measures can be used instead). The readers should understand our treatment of them by referring to the description of Definitions 2, 3, and 4 and Figure 1. The figure displays 3 instances of the unimodal PMFs and their DRs: we say that the right PMF has a larger scale than the left PMF and that the center PMF is less skew and the right PMF is skew.

### 2.2 Likelihood Models and Ordinal Regression Methods

Suppose that one has a set of observations (ordinal data) $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, each of which is drawn independently from an identical distribution of $(\boldsymbol{X}, Y)$. Every OR method covered in this paper assumes, as a model of the conditional probability $\Pr(Y = y|\boldsymbol{X} = \boldsymbol{x})$, a certain likelihood model $\hat{\Pr}(Y = y|\boldsymbol{X} = \boldsymbol{x}) = P(y; \boldsymbol{g}(\boldsymbol{x}))$ with a fixed part $P$ (say, the softmax function in multinomial logistic regression model) and learnable part $\boldsymbol{g} \in \mathcal{G}$ (say, a certain neural network model), where we call $P$ the link function and $\boldsymbol{g}$ the learner model in a model class $\mathcal{G}$. Note that the distinction between $P$ and $\boldsymbol{g}$ is to aid in understanding relation between multiple models, and this paper does not emphasize a strict mathematical distinction (note that, for example, $(P, \mathcal{G})$ and $(2P, \{\boldsymbol{g}/2 \mid \boldsymbol{g} \in \mathcal{G}\})$ yield an equivalent likelihood model).

The key notion in the discussion of this paper is the representation ability of the likelihood model. Strictly speaking, the representation ability we discuss is defined by the following relationship:

**Definition 5.** *We say that likelihood models based on $(P_1, \mathcal{G}_1)$ have a stronger representation ability (or are more representable) than those based on $(P_2, \mathcal{G}_2)$, if there exists $\boldsymbol{g}_1 \in \mathcal{G}_1$ such that $P_1(y; \boldsymbol{g}_1(\boldsymbol{x})) = P_2(y; \boldsymbol{g}_2(\boldsymbol{x}))$ for any $\boldsymbol{x} \in \mathbb{R}^d$ and $y \in [K]$, for any $\boldsymbol{g}_2 \in \mathcal{G}_2$.*

The representation abilities of two different likelihood models may not be strictly comparable according to Definition 5, but even in such cases we in this paper discuss the relationship between the representation abilities from a qualitative understanding of those likelihood models as much as possible. Additionally, we introduce the functional degree of freedom:

**Definition 6.** *The functional degree of freedom (FDF) of a vector-valued function $\boldsymbol{g}$ defined on $\mathbb{R}^d$ refers to the maximum value over all $\boldsymbol{x} \in \mathbb{R}^d$ of the dimension at the point $\boldsymbol{g}(\boldsymbol{x})$ of the manifold $\{\boldsymbol{g}(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^d\}$.*

The FDF of the likelihood model $(\hat{\Pr}(Y = y|\boldsymbol{X} = \boldsymbol{x}))_{y\in[K]}$ can be roughly understood as the minimum number of $\boldsymbol{x}$-dependent real-valued functions needed to describe the model that is a $\boldsymbol{x}$-dependent vector-valued function. For example, general CPDs $(\Pr(Y = y|\boldsymbol{X} = \cdot))_{y\in[K]}$ including those underlying unimodal data have up to $(K-1)$-FDF (or called full-FDF) (it is not $K$ since $\sum_{y=1}^{K} \Pr(Y = y|\boldsymbol{X} = \cdot) = 1$), and models in Figure 2 are 1-FDF. We use the FDF as one simple indicator of the representation ability of statistical likelihood models; we consider that the larger its FDF, the stronger the representation ability of the statistical model tends.

In this paper, we set up an OR method as learning a model $\boldsymbol{g}$ from the class $\mathcal{G}$ through the maximum likelihood estimation $\max_{\boldsymbol{g}\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \log P(y_i; \boldsymbol{g}(\boldsymbol{x}_i))$ (which we call a conditional probability estimation task), and then constructing a classifier under the task with the task loss $\ell$ as $f(\boldsymbol{x}) = f_\ell((P(y; \hat{\boldsymbol{g}}(\boldsymbol{x})))_{y\in[K]})$ with $f_\ell((p_k)_{k\in[K]}) := \min(\arg\min_j(\sum_{k=1}^{K} p_k \ell(j,k))_{j\in[K]})$ and an obtained model $\hat{\boldsymbol{g}}$.

Therefore, the only difference of OR methods treated in this paper appears in the link function $P$ and model class $\mathcal{G}$ of the likelihood model. Under these settings, generic principles based on the bias-variance tradeoff suggest that a compact likelihood model that can adequately represent the data is promising for an OR method with good generalization performance. On the other hand, we believe that subsequent discussions hold as well even if changing the loss function used in the parameter fitting procedure and the decision function: As options for the parameter fitting other than the maximum likelihood estimation, we can apply robust alternatives (Bianco & Yohai, 1996; Croux et al., 2013) to every likelihood model treated. Anonymous (2022) has developed a decision function that allows misspecification of the likelihood model. As another decision function, it would also be possible to assume a certain prior on the model $\boldsymbol{g}$, take the Bayesian discussion further, and use a decision function based on the posterior predictive probability distribution.

## 3 Existing Unimodal Likelihood Models

### 3.1 Binomial Models

The (shifted) binomial distribution $(P_{\mathrm{b}}(k;p))_{k\in[K]}$ is unimodal at any $p \in [0,1]$, where

$$P_{\mathrm{b}}(k;p) := \binom{K-1}{k-1} p^{k-1}(1-p)^{K-k} \text{ for } k \in [K],\ p \in [0,1], \tag{2}$$

and where $\binom{k}{l} := \frac{k!}{l!(k-l)!}$ is the binomial coefficient. Inspired by the shape of the PMF of the binomial distribution, da Costa et al. (2008) considered unimodal likelihood models based on the link function

$$P_{\mathrm{bin}}(y;u) := \binom{K-1}{y-1} \left(\frac{1}{1+e^{-u}}\right)^{y-1} \left(\frac{1}{1+e^{u}}\right)^{K-y} \text{ for } y \in [K],\ u \in \mathbb{R} \tag{3}$$

and an $\mathbb{R}$-valued learner model $g(\boldsymbol{x})$ (applied to $u$ in (3)). Thereafter, Beckham & Pal (2017) introduced the scaling factor: in other words, they proposed the link function

$$P_{\mathrm{bin}}(y;u,s) := \frac{e^{\log(P_{\mathrm{bin}}(y;u))/s}}{\sum_{k=1}^{K} e^{\log(P_{\mathrm{bin}}(k;u))/s}} \text{ for } y \in [K],\ u \in \mathbb{R},\ s \in \mathbb{R}_+, \tag{4}$$

where $\mathbb{R}_+ := \{v \in \mathbb{R} \mid 0 < v < \infty\}$. They applied a $\boldsymbol{x}$-dependent model $g(\boldsymbol{x})$ to $u$ and a $\boldsymbol{x}$-independent positive parameter to $s$, and learned them; we call models consisting of the link function $P_{\mathrm{bin}}$ and learner models in $\mathcal{G} = \{(g(\cdot), s) \mid g : \mathbb{R}^d \to \mathbb{R}, s \in \mathbb{R}_+\}$ the Binomial (BIN) models.

The BIN model $(P_{\mathrm{bin}}(y; g(\boldsymbol{x}), s))_{y\in[K]}$ is parametrically constrained with the 1-FDF. One would know that the mode of the binomial distribution is

$$M((P_{\mathrm{b}}(k;p))_{k\in[K]}) = \min(\{\lceil Kp \rceil, \lfloor Kp \rfloor + 1\} \cap [K]), \tag{5}$$

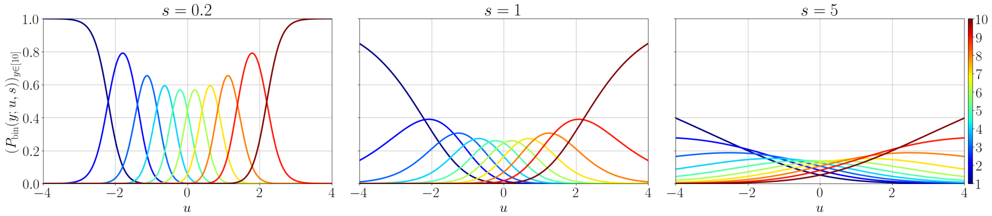


Figure 2: Instances of the BIN models (3) with $K = 10$.[2]

where $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ are the ceiling and floor functions, and hence $\max\{1, Kp\} \leq M((P_{\mathrm{b}}(k;p))_{k\in[K]}) \leq \min\{Kp + 1, K\}$. These known properties and simple calculations can show the mode, unimodality, and DRs of the BIN model:

**Theorem 1.** *It holds that*

(i) $\frac{P_{\mathrm{bin}}(y+1;u,s)}{P_{\mathrm{bin}}(y;u,s)} = \left(\frac{y(1-p)}{(K-y)p}\right)^{1/s}$ *with* $p = \frac{1}{1+e^{-u}}$ *for all* $y \in [K-1]$*,* $u \in \mathbb{R}$*,* $s \in \mathbb{R}_+$*.*

*Then, for any* $u \in \mathbb{R}$ *and* $s \in \mathbb{R}_+$*, it holds that, for* $m = M((P_{\mathrm{bin}}(y;u,s))_{y\in[K]})$ *that is* (5) *with* $p = \frac{1}{1+e^{-u}}$*,*

(ii) $P_{\mathrm{bin}}(1;u,s) \leq \cdots \leq P_{\mathrm{bin}}(m;u,s)$ *if* $m \neq 1$*, and* $P_{\mathrm{bin}}(m;u,s) \geq \cdots \geq P_{\mathrm{bin}}(K;u,s)$ *if* $m \neq K$*,*

(iii) $\frac{P_{\mathrm{bin}}(1;u,s)}{P_{\mathrm{bin}}(2;u,s)} \leq \cdots \leq \frac{P_{\mathrm{bin}}(m-1;u,s)}{P_{\mathrm{bin}}(m;u,s)}$ $(\leq 1)$ *if* $m \neq 1$*, and* $(1 \geq)$ $\frac{P_{\mathrm{bin}}(m+1;u,s)}{P_{\mathrm{bin}}(m;u,s)} \geq \cdots \geq \frac{P_{\mathrm{bin}}(K;u,s)}{P_{\mathrm{bin}}(K-1;u,s)}$ *if* $m \neq K$*.*

We call the constraint on the sequence of DRs like Theorem 1, (iii) as the DRs' unimodality (constraint), considering that $(\frac{p_1}{p_2}, \ldots, \frac{p_{m-1}}{p_m}, \frac{p_{m+1}}{p_m}, \ldots, \frac{p_K}{p_{K-1}})$ is unimodal if a PMF $(p_k)_{k\in[K]}$ having a mode $m$ satisfies that constraint. One can find that, owing to the DRs' unimodality constraint, the BIN models cannot exactly represent unimodal MPFs in Figure 1. In addition to the DRs' unimodality constraint, the BIN model $(P_{\mathrm{bin}}(y; g(\boldsymbol{x}), s))_{y\in[K]}$ is always less skew especially when its mode is close to the labels' intermediate value $\frac{1+K}{2}$ (recall that the mode (5) and median, $\lceil(K-1)p\rceil + 1$ or $\lfloor(K-1)p\rfloor + 1$, of the binomial distribution $(P_{\mathrm{b}}(k;p))_{k\in[K]}$ are close (Kaas & Buhrman, 1980)), and tends homoscedastic (see Figure 2; there, the scale $1 - \max_y(P_{\mathrm{bin}}(y;u,s))_{y\in[10]}$ for each $s$ is similar for many $u$ except at both ends). We will relax in Section 5.2 the restriction of the representation ability regarding the scale of BIN models, but they cannot avoid the restriction regarding the DRs and skewness.

### 3.2 Poisson Models

The (shifted) Poisson distribution $(P_{\mathrm{p}}(k;\lambda))_{k\in\mathbb{N}}$ is unimodal at any $\lambda > 0$, where

$$P_{\mathrm{p}}(k;\lambda) := \frac{\lambda^{k-1}e^{-\lambda}}{(k-1)!} \text{ for } k \in \mathbb{N},\ \lambda > \mathbb{R}_+. \tag{6}$$

da Costa et al. (2008) truncated (6) within $[K]$ and normalize it to develop likelihood models: $\hat{\Pr}(Y = y|\boldsymbol{X} = \boldsymbol{x}) = P_{\mathrm{p}}(y; g(\boldsymbol{x}))/\sum_{k=1}^{K} P_{\mathrm{p}}(k; g(\boldsymbol{x}))$ with an $\mathbb{R}_+$-valued learner model $g(\boldsymbol{x})$. As in the case for BIN models, Beckham & Pal (2017) developed a scaled link function

$$P_{\mathrm{poi}}(y;u,s) := \frac{e^{v_y/s}}{\sum_{k=1}^{K} e^{v_k/s}} \text{ for } y \in [K],\ u, s \in \mathbb{R}_+, \text{ with } v_k = (k-1)\log(u) - \log((k-1)!) \text{ for } k \in [K], \tag{7}$$

which is a generalization of $P_{\mathrm{poi}}(y;u,1)$ of (da Costa et al., 2008).[3] They applied a $\boldsymbol{x}$-dependent $\mathbb{R}_+$-valued learner model $g(\boldsymbol{x})$ to $u$ and a $\boldsymbol{x}$-independent positive parameter to $s$ (i.e., $\mathcal{G} = \{(g(\cdot), s) \mid g : \mathbb{R}^d \to \mathbb{R}_+, s \in \mathbb{R}_+\}$); we call these models the Poisson (POI) Models.

---

[2] The display style in Figure 1 is like showing a PMF created by cutting the curves in Figure 2 out at each $u$.

[3] In (Beckham & Pal, 2017), $v_k$ in (7) is defined with minus $u$, but which is meaningless after the softmax transformation.

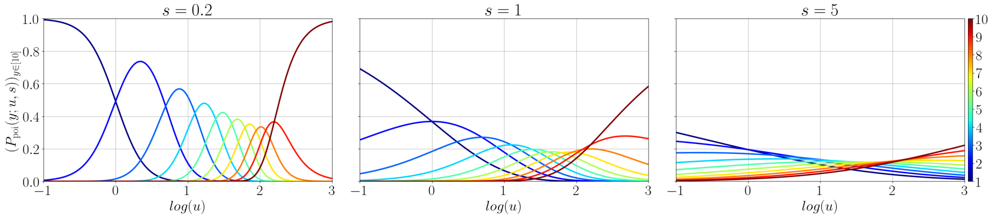


Figure 3: Instances of the POI models (7) with $K = 10$.

The POI model $(P_{\mathrm{poi}}(y; g(\boldsymbol{x}), s))_{y\in[K]}$ is also parametrically constrained with the 1-FDF. The POI model $(P_{\mathrm{poi}}(y; g(\boldsymbol{x}), s))_{y\in[K]}$ has a similar DRs' unimodality constraint to the BIN model, and always has mode-wise heteroscedasticity especially when $s$ is small and $K$ is large: its scale tends small and large respectively when $M((P_{\mathrm{poi}}(y; g(\boldsymbol{x}), s))_{y\in[K]})$ is close to 1, $K$, or smaller side of $\{2, \ldots, K-1\}$ and larger side; see Figure 3 with $s = 0.2$. They, therefore, are not suitable for representing homoscedastic or overall heteroscedastic data. Furthermore, we find that the POI models are special instances of PO-ACL (and PO-ORD-ACL) models discussed later and have a weaker representation ability; see also Sections 4.1, 5.1, and 5.2 and Theorem 4.

## 4 Novel Unimodal Likelihood Models

### 4.1 Ordered Adjacent Categories Logit Models

The BIN and POI models are unimodal, but have just 1-FDF and may be too weakly representable for some data. We thus developed more strongly representable unimodal likelihood models that have up to full-FDF, i.e., $(K-1)$-FDF. This section describes ordered adjacent categories logit (ORD-ACL) models developed by modifying existing ACL models that do not have the unimodality guarantee.

The (naïve) ACL models (see (Simon, 1974; Andrich, 1978; Goodman, 1979; Masters, 1982), and (Agresti, 2010, Section 4.1)) are designed to model the relationship between the underlying conditional probabilities of the adjacent categories, $\frac{\Pr(Y=y|\boldsymbol{X}=\boldsymbol{x})}{\Pr(Y\in\{y,y+1\}|\boldsymbol{X}=\boldsymbol{x})}$, as

$$\frac{\hat{\Pr}(Y = y|\boldsymbol{X} = \boldsymbol{x})}{\hat{\Pr}(Y \in \{y, y+1\}|\boldsymbol{X} = \boldsymbol{x})} = \frac{1}{1 + e^{-g_y(\boldsymbol{x})}} \text{ for } y \in [K-1] \tag{8}$$

with an $\mathbb{R}^{(K-1)}$-valued learner model $\boldsymbol{g}$. Considering the normalization condition $\sum_{y=1}^{K} \hat{\Pr}(Y = y|\boldsymbol{X} = \boldsymbol{x}) = 1$, it can be found that ACL models depend on the ACL link function

$$P_{\mathrm{acl}}(y; \boldsymbol{u}) := \frac{\prod_{l=1}^{y-1} e^{-u_l}}{\sum_{k=1}^{K} \prod_{l=1}^{k-1} e^{-u_l}} = \frac{e^{-\sum_{l=1}^{y-1} u_l}}{\sum_{k=1}^{K} e^{-\sum_{l=1}^{k-1} u_l}} \text{ for } y \in [K],\ \boldsymbol{u} \in \mathbb{R}^{K-1}, \tag{9}$$

together with the model class $\mathcal{G} = \{\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^{K-1}\}$, where we define the product such that $\prod_{k=i}^{j} f(k)$ is one irrelevant to the function $f$ as far as $i > j$.

We have analyzed properties of the ACL link function and then found a simple condition so that the ACL models turn to get the unimodality guarantee:

**Theorem 2.** *It holds that*

(i) $\frac{P_{\mathrm{acl}}(y+1;\boldsymbol{u})}{P_{\mathrm{acl}}(y;\boldsymbol{u})} = e^{-u_y}$ *for all* $y \in [K-1]$, $\boldsymbol{u} \in \mathbb{R}^{K-1}$.

*If* $\boldsymbol{u} \in \mathbb{R}^{K-1}$ *satisfies* $u_0(:= -\infty) \leq \cdots \leq u_{m-1} \leq 0 \leq u_m \leq \cdots \leq u_K(:= +\infty)$ *for some* $m \in [K]$, *it holds that*

(ii) $P_{\mathrm{acl}}(1; \boldsymbol{u}) \leq \cdots \leq P_{\mathrm{acl}}(m; \boldsymbol{u})$ *if* $m \neq 1$, *and* $P_{\mathrm{acl}}(m; \boldsymbol{u}) \geq \cdots \geq P_{\mathrm{acl}}(K; \boldsymbol{u})$ *if* $m \neq K$,

(iii) $\frac{P_{\mathrm{acl}}(1;\boldsymbol{u})}{P_{\mathrm{acl}}(2;\boldsymbol{u})} \leq \cdots \leq \frac{P_{\mathrm{acl}}(m-1;\boldsymbol{u})}{P_{\mathrm{acl}}(m;\boldsymbol{u})}$ $(\leq 1)$ *if* $m \neq 1$, *and* $(1 \geq)$ $\frac{P_{\mathrm{acl}}(m+1;\boldsymbol{u})}{P_{\mathrm{acl}}(m;\boldsymbol{u})} \geq \cdots \geq \frac{P_{\mathrm{acl}}(K;\boldsymbol{u})}{P_{\mathrm{acl}}(K-1;\boldsymbol{u})}$ *if* $m \neq K$.

On the basis of the unimodality guarantee stated in Theorem 2, (ii), we propose ORD-ACL models $(P_{\mathrm{acl}}(y; \acute{\boldsymbol{g}}(\boldsymbol{x})))_{y\in[K]}$ applying the ACL link function (9) and ordered learner model $\acute{\boldsymbol{g}}(\boldsymbol{x})$ satisfying that $\acute{g}_1(\boldsymbol{x}) \le \cdots \le \acute{g}_1(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^d$. Here, the ordered learner model $\acute{\boldsymbol{g}}$ can be implemented as

$$\acute{g}_k(\boldsymbol{x}) = \begin{cases} g_1(\boldsymbol{x}), & \text{for } k = 1, \\ \acute{g}_{k-1}(\boldsymbol{x}) + \rho(g_k(\boldsymbol{x})), & \text{for } k = 2, \ldots, K-1, \end{cases} \tag{10}$$

with another model $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^{K-1}$ and a fixed continuous non-negative function $\rho : \mathbb{R} \to [0, +\infty)$ that satisfies $\rho(u) = 0$ for some $u \in \mathbb{R}$ and $\lim_{u\to-\infty} \rho(u) = +\infty$ or $\lim_{u\to+\infty} \rho(u) = +\infty$ such as $\rho_{\exp}(u) := e^u$, $\rho_{\mathrm{sq}}(u) := u^2$ (we denote the procedure (10) as $\acute{\boldsymbol{g}} = \rho[\boldsymbol{g}]$; i.e., the model class $\mathcal{G} = \{\rho[\boldsymbol{g}] \mid \boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^{K-1}\}$).

It is a good aspect in the modeling of the ordinal data assumed to be unimodal that the likelihood models are guaranteed to be unimodal. In contrast, the ACL models also incidentally impose a constraint on the DRs (see Theorem 2, (i) and (iii)). Nevertheless, it would be important that the ORD-ACL models can represent arbitrary data with a unimodal CPD that has DRs' unimodality constraint, and hence that the ORD-ACL models are so representable that they fully encompass the BIN and POI models.

### 4.2 V-Shaped Stereotype Logit Models

We next describe V-shaped stereotype logit (VS-SL) models, which are unimodal, full-FDF, more flexible than ORD-ACL models, and developed by modifying SL models (or multinomial logistic regression models).

The SL models (refer to (Anderson, 1984) and (Agresti, 2010, Section 4.3)) attempt to model the conditional probabilities of multiple categories paired with a certain fixed (stereotype) category, $\frac{\Pr(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\Pr(Y\in\{1,y\}|\boldsymbol{X}=\boldsymbol{x})}$, as

$$\frac{\hat{\Pr}(Y = 1|\boldsymbol{X} = \boldsymbol{x})}{\hat{\Pr}(Y \in \{1, y\}|\boldsymbol{X} = \boldsymbol{x})} = \frac{1}{1 + e^{-g_y(\boldsymbol{x})}} \text{ for } y \in [K] \tag{11}$$

with an $\mathbb{R}^K$-valued learner model $\boldsymbol{g}$ such that $g_1(\boldsymbol{x}) = 0$ for any $\boldsymbol{x} \in \mathbb{R}^d$. Considering the normalization condition, we introduce the SL link function as

$$P_{\mathrm{sl}}(y; \boldsymbol{u}) := \frac{e^{-u_y}}{\sum_{k=1}^K e^{-u_k}} \text{ for } y \in [K],\ \boldsymbol{u} \in \mathbb{R}^K. \tag{12}$$

The SL models are based on the SL link function $P_{\mathrm{sl}}$ and model class $\mathcal{G} = \{\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^K \mid g_1(\cdot) = 0\}$.

The ACL and SL link functions have only minor differences in parameterization, but looking at the SL link function will reveal another simple way to create a unimodal likelihood model as follow:

**Theorem 3.** *If $\boldsymbol{u} \in \mathbb{R}^K$ satisfies that $u_1 \ge \cdots \ge u_m$ if $m \ne 1$ and $u_m \le \cdots \le u_K$ for some $m \in [K]$ if $m \ne K$, it holds that $P_{\mathrm{sl}}(1; \boldsymbol{u}) \le \cdots \le P_{\mathrm{sl}}(m; \boldsymbol{u})$ if $m \ne 1$ and $P_{\mathrm{sl}}(m; \boldsymbol{u}) \ge \cdots \ge P_{\mathrm{sl}}(K; \boldsymbol{u})$ if $m \ne K$.*

On the ground of this theorem, we proposed VS-SL model $(P_{\mathrm{sl}}(y; \check{\boldsymbol{g}}(\boldsymbol{x})))_{y\in[K]}$ based on the SL link function $P_{\mathrm{sl}}$ and a V-shaped learner model $\check{\boldsymbol{g}} = \tau[\acute{\boldsymbol{g}}]$ (described below) with $\acute{\boldsymbol{g}} = \rho[\boldsymbol{g}]$ and another $\mathbb{R}^K$-valued learner model $\boldsymbol{g}$ (i.e., the model class $\mathcal{G} = \{\tau[\rho[\boldsymbol{g}]] \mid \boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^K\}$ for the link function $P_{\mathrm{sl}}$). First, $\rho$ transforms an arbitrary model $\boldsymbol{g}$ to an ordered model $\acute{\boldsymbol{g}}$ (so $\acute{g}_1(\boldsymbol{x}) \le \cdots \le \acute{g}_K(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^d$). Next, $\tau$ transforms an ordered model $\acute{\boldsymbol{g}}$ to a V-shaped model $\tau[\acute{\boldsymbol{g}}]$, where the notation $\tau[\boldsymbol{u}]$ for an $\mathbb{R}^K$-valued object $\boldsymbol{u}$ implies the element-wise application $(\tau \circ u_k)_{k\in[K]}$. The function $\tau(u)$ for $u \in \mathbb{R}$ is supposed to be continuous, non-negative, non-increasing in $u < 0$, and non-decreasing in $u > 0$, and satisfy $\tau(0) = 0$ and $\lim_{u\to\pm\infty} \tau(u) = +\infty$ such as $\tau_{\mathrm{abs}}(u) := |u|$ and $\tau_{\mathrm{sq}}(u) := u^2$. It holds that $\check{g}_1(\boldsymbol{x}) \ge \cdots \ge \check{g}_m(\boldsymbol{x})$ and $\check{g}_m(\boldsymbol{x}) \le \cdots \le \check{g}_K(\boldsymbol{x})$ with $m = \min(\arg\min_k (|\acute{g}_k(\boldsymbol{x})|)_{k\in[K]})$ if $\tau = \tau_{\mathrm{abs}}, \tau_{\mathrm{sq}}$, as required in Theorem 3. Finally, the SL link function transforms $\check{\boldsymbol{g}}$ to a unimodal likelihood model that is normalized and remains non-negative.

Most importantly, the VS-SL model $(P_{\mathrm{sl}}(y; \check{\boldsymbol{g}}(\boldsymbol{x})))_{y\in[K]}$ is ensured to be unimodal and further can represent arbitrary unimodal data if ignoring the restriction owing to the structure of an implemented learner model $\boldsymbol{g}$ in $\tau[\rho[\boldsymbol{g}]]$ (or equivalently, if using an extremely flexible $\boldsymbol{g}$). Therefore, the VS-SL models can be viewed the most representable ones among the class of the unimodal likelihood models.

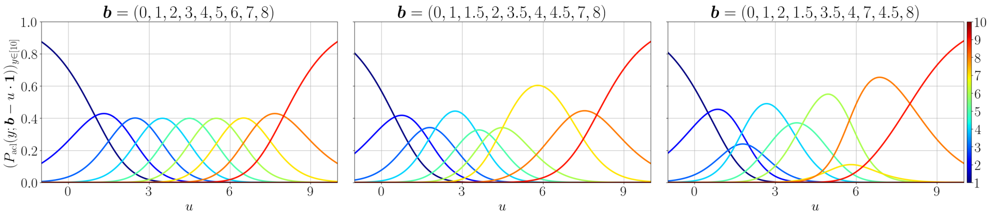


Figure 4: Instances of the PO-ACL models with $K = 10$.

## 5 Variant Models

### 5.1 Proportional-Odds Models

The ACL model $(P_{\mathrm{acl}}(y; \boldsymbol{g}(\boldsymbol{x})))_{y \in [K]}$, ORD-ACL model $(P_{\mathrm{acl}}(y; \rho[\boldsymbol{g}](\boldsymbol{x})))_{y \in [K]}$, SL model $(P_{\mathrm{sl}}(y; \boldsymbol{g}(\boldsymbol{x})))_{y \in [K]}$, and VS-SL model $(P_{\mathrm{sl}}(y; \tau[\rho[\boldsymbol{g}]](\boldsymbol{x})))_{y \in [K]}$ have full-FDF and may be too flexible and difficult to learn for some data especially when the sample size is not large enough. In the statistical research (e.g., see (McCullagh, 1980)), the proportional-odds (PO) constraint that constrains the FDF of the likelihood model to 1 is often applied to avoid such a trouble. We call $(P_{\mathrm{acl}}(y; \boldsymbol{b} - g(\boldsymbol{x}) \cdot \mathbf{1}))_{y \in [K]}$ the PO-ACL model, $(P_{\mathrm{acl}}(y; \rho[\boldsymbol{b}] - g(\boldsymbol{x}) \cdot \mathbf{1}))_{y \in [K]}$ the PO-ORD-ACL model, and $(P_{\mathrm{sl}}(y; \tau[\rho[\boldsymbol{b}] - g(\boldsymbol{x}) \cdot \mathbf{1}]))_{y \in [K]}$ the PO-VS-SL model, where $\boldsymbol{b}$ and $\mathbf{1}$ are learnable parameter and all-1 vector with appropriate dimension ($(K - 1)$ for ACL and ORD-ACL models, or $K$ for VS-SL models) and $g$ is an $\mathbb{R}$-valued learner model.[4]

The PO-ORD-ACL and PO-VS-SL models are novel. The significance of these models we expect lies in the theoretical guarantee of the unimodality, not in the improvement of probability prediction and classification performances (e.g., in comparison between PO-ORD-ACL and ORD-ACL models): For example, the PO-ACL model $(P_{\mathrm{acl}}(y; \boldsymbol{b} - g(\boldsymbol{x}) \cdot \mathbf{1}))_{y \in [K]}$ may result in ordered parameter $\boldsymbol{b}$ depending on the overall tendency of the training data even if not explicitly imposing the ordering constraint during the parameter fitting (but not always).

Additionally, recalling the POI model reviewed in Section 3.2, it would be clear that the POI model is a special instance of the PO-ACL models:

**Theorem 4.** *For any $y \in [K]$, $u \in \mathbb{R}$, $s \in \mathbb{R}_+$, it holds that $P_{\mathrm{poi}}(y; u, s) = P_{\mathrm{acl}}(y; \boldsymbol{b} - v \cdot \mathbf{1})$ with $\boldsymbol{b} = (b_k)_{k \in [K-1]}$ satisfying $b_k = \log(k)/s$ for $k = 1, \ldots, K - 1$ and $v = \log(u)/s$.*

The PO-ACL model can further represent homoscedastic data and more various mode-wise heteroscedastic data because it can adjust its parameter $\boldsymbol{b}$ according to the model fit to the data, compared with the POI model; compare Figure 3 for the POI models and Figure 4 for the PO-ACL models. On the other hand, we note that the statement "the PO-ORD-ACL model generalizes the BIN model" is incorrect in general.

### 5.2 Overall Heteroscedastic Models

The BIN, POI, and PO models have just 1-FDF. The strong constraint of these models can be interpreted as incidentally assuming the homoscedasticity or mode-wise heteroscedasticity. We in this section describes their overall homoscedastic (OH) extension, OH models.

McCullagh (1980) studied an OH extension of a certain PO model. According to his idea, we propose the OH-ACL model $(P_{\mathrm{acl}}(y; \{\boldsymbol{b} - g(\boldsymbol{x}) \cdot \mathbf{1}\}/s(\boldsymbol{x})))_{y \in [K]}$, OH-ORD-ACL model $(P_{\mathrm{acl}}(y; \{\rho[\boldsymbol{b}] - g(\boldsymbol{x}) \cdot \mathbf{1}\}/s(\boldsymbol{x})))_{y \in [K]}$, and OH-VS-SL model $(P_{\mathrm{sl}}(y; \tau[\rho[\boldsymbol{b}] - g(\boldsymbol{x}) \cdot \mathbf{1}]/s(\boldsymbol{x})))_{y \in [K]}$, where $\boldsymbol{b}$ and $\mathbf{1}$ are learnable parameter and all-1 vector with appropriate dimension, and $g$ and $s$ are $\mathbb{R}$- and $\mathbb{R}_+$-valued learner models. The $\mathbb{R}_+$-valued scale

---

[4] The PO-SL model $P_{\mathrm{sl}}(y; \boldsymbol{b} - g(\boldsymbol{x}) \cdot \mathbf{1}) = e^{-b_y} / \sum_{k=1}^{K} e^{-b_k}$ is constant regarding $\boldsymbol{x}$ and may be less representable. Therefore, Anderson (1984) considered to use a learner model $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{a} - g(\boldsymbol{x}) \cdot \boldsymbol{c}$ with learnable parameters $\boldsymbol{a}, \boldsymbol{c}$ ($\in \mathbb{R}^K$ here) s.t. $a_1 = c_1 = 0$ and $\mathbb{R}$-valued learner model $g : \mathbb{R}^d \to \mathbb{R}$ (we call $\boldsymbol{g}$ a rotatable-odds (RO) learner model) for the SL link function; we call this model the RO-SL model. We can also consider novel RO-ACL models: they are more representable than the PO-ACL (including PO-ORD-ACL) models, but have no unimodality guarantee since the RO learner model $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{a} - g(\boldsymbol{x}) \cdot \boldsymbol{c}$ does not necessarily satisfy the ordering condition $g_1(\boldsymbol{x}) \leq \cdots \leq g_{K-1}(\boldsymbol{x})$ (depending on $\boldsymbol{c}$).

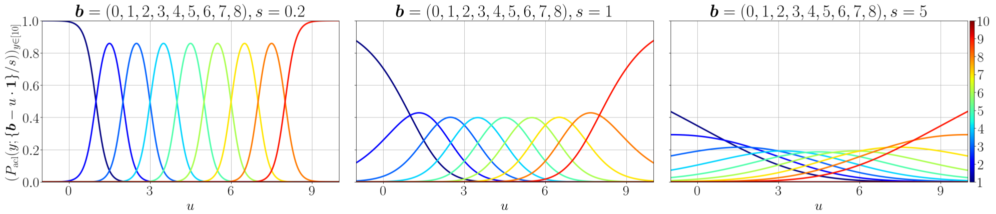


Figure 5: Instances of the OH-ACL models with $K = 10$.

model $s$ can be implemented as $s(\boldsymbol{x}) = r + \rho(t(\boldsymbol{x}))$ with a constant $r > 0$ for avoiding zero-division, fixed non-negative function $\rho$, and another $\mathbb{R}$-valued learner $t$.

The significance of OH models is that their scaling factor can vary depending on $\boldsymbol{x}$. Extending previous BIN and POI models by (Beckham & Pal, 2017), we also introduce the OH-BIN model $(P_{\rm bin}(y; g(\boldsymbol{x}), s(\boldsymbol{x})))_{y\in[K]}$ and OH-POI model $(P_{\rm poi}(y; g(\boldsymbol{x}), s(\boldsymbol{x})))_{y\in[K]}$. For the implementation $s(\boldsymbol{x}) = r + \rho(t(\boldsymbol{x}))$ of the scaling factor for these models, the use of a small $r > 0$ enables to represent a wider class distribution. Note that the OH-POI model is a special instance of the OH-ORD-ACL models, as Theorem 4 shows.

These OH models can represent overall heteroscedastic data, but their representation ability is parametrically constrained with the 2-FDF; compare Figures 2, 3, 4, and 5.

### 5.3 Combination-with-Uniform Models

When a likelihood model $(\hat{\Pr}(Y = y|\boldsymbol{X} = \boldsymbol{x}))_{y\in[K]}$ is unimodal, the combination-with-uniform (CU) likelihood model, $q(\boldsymbol{x}) \cdot (\frac{1}{K})_{k\in[K]} + \{1 - q(\boldsymbol{x})\} \cdot (\hat{\Pr}(Y = y|\boldsymbol{X} = \boldsymbol{x}))_{y\in[K]}$ with e.g., $q(\boldsymbol{x}) = \frac{1}{1+e^{-t(\boldsymbol{x})}}$ for an $\mathbb{R}$-valued model $t$, is also unimodal. For example, Piccolo (2003); Iannario & Piccolo (2011) studied a combination of uniform and binomial model. The CU extension of low-FDF (like PO and OH) models can increase their FDF by 1.

## 6 Numerical Experiments

### 6.1 Experimental Purposes

From the bias-variance tradeoff, it can be expected that compact likelihood models that can adequately represent the data will yield better generalization performance in the conditional probability estimation task, and accordingly that OR methods based on such likelihood models will yield better generalization performance in the OR task. We took numerical experiments in order to verify whether the proposed likelihood models, developed based on this working hypothesis, provide better performances than previous unimodal likelihood models with weak representation ability and popular statistical OR models with no unimodality guarantee.

### 6.2 Experimental Settings

We selected 21 real-world datasets of those used in experiments by the previous OR study (Gutierrez et al., 2015) with the total sample size $n_{\rm tot}$ that is 1000 or more, and used them for our numerical experiments.[5] AB5, …, CE5' (resp. AB10, …, CE10') are datasets generated by discretizing a real-valued target of datasets, which are often used to benchmark regression methods, by 5 (resp. 10) different bins with equal proportions. SW, …, CA originally have a categorical target, and the authors of (Gutierrez et al., 2015) judged that their targets have a natural ordinal relation.

Table 1 shows the dataset name, dataset properties, $n_{\rm tot}$, $d$, and $K$, and mean and standard deviation (STD) of 100 test MUs and test DRs' MUs: The mean unimodality (MU) is a numerical criterion to evaluate the unimodality of the conditional probability distribution of the data, and it is defined, for a likelihood

---

[5] One can get the datasets from a researchers' site (`http://www.uco.es/grupos/ayrna/orreview`) of (Gutierrez et al., 2015), or our GitHub repository (`https://github.com/Anonymous`) together with our used program codes.

Table 1: It shows the dataset name, dataset properties $n_{\text{tot}}$, $d$, and $K$, and mean and STD (as 'mean ± STD') of test MUs and test DRs' MUs.

| dataset name | $n_{\text{tot}}$ | $d$ | $K$ | MU | DRs' MU |
|---|---|---|---|---|---|
| AB5 (abalon5) | 4177 | 10 | 5 | $.8915 \pm .0662$ | $.4042 \pm .0743$ |
| BA5 (bank5) | 8192 | 8 | 5 | $.9944 \pm .0387$ | $.4293 \pm .1389$ |
| BA5' (bank5') | 8192 | 32 | 5 | $.9887 \pm .0167$ | $.7041 \pm .1210$ |
| CO5 (computer5) | 8192 | 12 | 5 | $.9956 \pm .0137$ | $.5623 \pm .1041$ |
| CO5' (computer5') | 8192 | 21 | 5 | $1.0000 \pm .0001$ | $.5851 \pm .1170$ |
| CH5 (cal.housing5) | 20640 | 8 | 5 | $.9065 \pm .1027$ | $.3993 \pm .1206$ |
| CE5 (census5) | 22784 | 8 | 5 | $.7887 \pm .0929$ | $.3221 \pm .1050$ |
| CE5' (census5') | 22784 | 16 | 5 | $.8332 \pm .0736$ | $.3827 \pm .0978$ |
| AB10 (abalon10) | 4177 | 10 | 10 | $.3218 \pm .1315$ | $.0076 \pm .0164$ |
| BA10 (bank10) | 8192 | 8 | 10 | $.8923 \pm .1463$ | $.0225 \pm .0432$ |
| BA10' (bank10') | 8192 | 32 | 10 | $.5101 \pm .2261$ | $.0019 \pm .0054$ |
| CO10 (computer10) | 8192 | 12 | 10 | $.8006 \pm .1431$ | $.0935 \pm .0883$ |
| CO10' (computer10') | 8192 | 21 | 10 | $.8189 \pm .1617$ | $.0517 \pm .0737$ |
| CH10 (cal.housing10) | 20640 | 8 | 10 | $.3311 \pm .1713$ | $.0051 \pm .0095$ |
| CE10 (census10) | 22784 | 8 | 10 | $.2042 \pm .1017$ | $.0008 \pm .0018$ |
| CE10' (census10') | 22784 | 16 | 10 | $.3151 \pm .1153$ | $.0027 \pm .0053$ |
| SW (SWD) | 1000 | 10 | 4 | $.9993 \pm .0027$ | $.9853 \pm .0204$ |
| LE (LEV) | 1000 | 4 | 5 | $.9547 \pm .0666$ | $.2208 \pm .1555$ |
| ER (ERA) | 1000 | 4 | 9 | $.7909 \pm .1045$ | $.2126 \pm .1403$ |
| WR (winequality-red) | 1599 | 11 | 6 | $.9894 \pm .0457$ | $.1415 \pm .1259$ |
| CA (car) | 1728 | 21 | 4 | $.8735 \pm .2037$ | $.0276 \pm .0223$ |

Table 2: It shows a ratio that $10^6$ samples of the PMF uniformly and randomly drawn from $\Delta^{K-1}$ satisfied the unimodality (resp. the DRs' unimodality) in the row "MU" (resp. in the row "DRs' MU").

| $K$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| MU | .6666 | .3330 | .1329 | .0444 | .0125 | .0032 | .0007 | .0001 |
| DRs' MU | .6666 | .3330 | .1329 | .0276 | .0055 | .0009 | .0001 | .0000 |

model $\hat{\Pr}(Y = \cdot|\boldsymbol{X} = \cdot)$ and $n$ used data points, as $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{(\hat{\Pr}(Y = y|\boldsymbol{X} = \boldsymbol{x}_i))_{y\in[K]}$ is unimodal.$\}$. As that likelihood model, we used the test SL model under the task-P (described below) that can represent any conditional probability distribution. We trained a likelihood model with a training sample of size $n_{\text{tra}} = 800$, and evaluated the MU with an obtained likelihood model and a remaining test sample of size $n_{\text{tes}} = n_{\text{tot}} - n_{\text{tra}}$. We repeated this procedure 100 trials with a randomly-set different sample setting and initial parameters of the likelihood model to obtain 100 test MUs. We also give a relative reference value in Table 2: It shows a ratio that $10^6$ samples of the PMF uniformly and randomly drawn from the probability simplex in $\mathbb{R}^K$ was unimodal for $K = 3, \ldots, 10$. Comparing the test MU in Table 1 and the relative reference value of the same $K$ in Table 2, one can find that many real-world data treated as ordinal data by existing OR research tend strongly unimodal. Additionally, we introduced the DRs' MU $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{(\hat{\Pr}(Y = y|\boldsymbol{X} = \boldsymbol{x}_i))_{y\in[K]}$ satisfies the DRs' unimodality.$\}$, and took a similar procedure to that for the MU; see Tables 1 and 2. The DRs' MU for ordinal data were larger than those for uniform random PMF, but their difference appears to be relatively smaller than that for the MU.[6]

We considered 4 tasks: a conditional probability estimation task (task-P), and 3 OR tasks with the zero-one task loss $\ell_{\text{zo}}$ (task-Z), absolute task loss $\ell_{\text{abs}}$ (task-A), and squared task loss $\ell_{\text{sq}}$ (task-S). Results for the task-P, -Z, -A, and -S are respectively evaluated based on the negative log likelihood (NLL), mean zero-one error (MZE), mean absolute error (MAE), and mean squared error (MSE). Note that, for a likelihood model $\hat{\Pr}(Y = \cdot|\boldsymbol{X} = \cdot)$ and $n$ used data points, the NLL is defined as $-\frac{1}{n}\sum_{i=1}^{n} \log \hat{\Pr}(Y = y_i|\boldsymbol{X} = \boldsymbol{x}_i)$, and the MZE, MAE, and MSE are respectively defined as $\frac{1}{n}\sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i)$ with $f(\boldsymbol{x}) = f_\ell((\hat{\Pr}(Y = y|\boldsymbol{X} = \boldsymbol{x}))_{y\in[K]})$ for $\ell = \ell_{\text{zo}}, \ell_{\text{abs}}, \ell_{\text{sq}}$. A method that yields a smaller criterion value is better for the corresponding task.

We tried 12 likelihood models (see Figure 6 for the relationship of their representation abilities): previous 1-FDF BIN and POI models; proposed 1-FDF PO-ORD-ACL and PO-VS-SL models; proposed 2-FDF OH-BIN, OH-POI, OH-ORD-ACL, and OH-VS-SL models; proposed full-FDF ORD-ACL and VS-SL models; previous full-FDF ACL and SL models with no unimodality guarantee. Here, we used link functions $\rho = \rho_{\text{exp}}$ to ensure the non-negativity and for a positive learner model for POI and OH-POI models and $\tau = \tau_{\text{sq}}$ to ensure the V-shape, and a positive constant $r = 0.01$ in the scaling model $s(\boldsymbol{x}) = r + \rho(t(\boldsymbol{x}))$ for OH models. We implemented all learner models with a 4-layer fully-connected neural network model that shares weights in except for the final layer and has 100 nodes activated with the sigmoid function in addition to bias nodes in every hidden layer. Note that, for the OH models, we implemented their real-valued learner and scaling models ($g$ and $t$ in the notation in Section 5.2) with two isolated weight-shared networks (each of which is described above), because their performance was significantly degraded when the real-valued learner and scaling models were implemented with a single weight-shared network. We trained a model with

---

[6] For notions related to the scale and skewness, we introduced numerical measures in Definitions 3 and 4 just for the sake of clarity of the qualitative discussion. However, we did not introduce metrical meaning into these measures (and it is quite difficult): for example, a gap between scales 0.1 and 0.3 and a gap between scales 0.7 and 0.9 would not have the same meaning. Therefore, we could not evaluate the heteroscedasticity or mode-wise skewness without misleading.

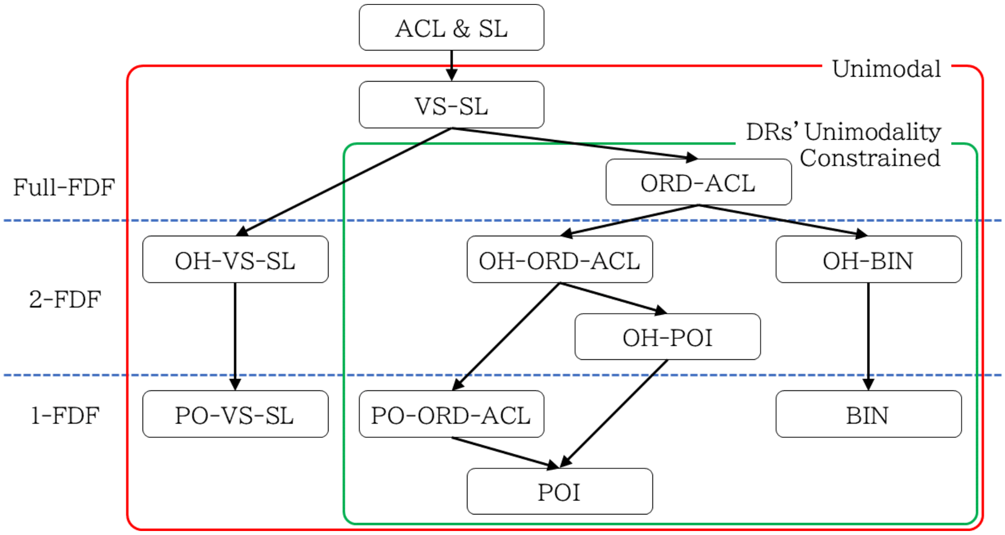


Figure 6: It shows the relationship between the representation abilities of the 12 methods studied in the numerical experiments as the directed tree graph. The directed tree graph implies that models in a root edge are more representable than models in a leaf edge in the sense of Definition 5.

a training sample and Adam optimization for 1000 epochs according to the maximum likelihood estimation, and evaluated the NLL, MZE, MAE, and MSE with a remaining test sample at the end of each epoch. Then, for the task-P, -Z, -A, or -S, we adopted a model at the timing when the test NLL, MZE, MAE, or MSE got minimum as the test model under the coresponding task.

We experimented with 6 training sample size settings $n_{\rm tra} = 25, 50, 100, 200, 400, 800$, to see the dependence of behaviors of each method on the training sample size.

For all combinations of 21 datasets, 4 tasks, 12 likelihood models, and 6 training sample size settings, we repeated above-described procedure 100 trials with a randomly-set different training sample and initial parameters, to obtain 100 test errors under the corresponding task.

### 6.3 Experimental Results

**Outline of Results** Table 3 shows the results of a simultaneous comparison of all methods. In summary, POI, PO-VS-SL, OH, and ORD-ACL models were especially good when $n_{\rm tra}$ was small, ORD-ACL and VS-SL models were especially good when $n_{\rm tra}$ was large, for every task. These results imply that methods based on likelihood models with weak (resp. strong) representation ability are more effective when the training sample size is small (resp. large), and that unimodal models yielded better performances than ACL and SL models, which are more representable but have no unimodality guarantee, under our implementation of the models (neural network architecture) and within the sample sizes that we considered ($n_{\rm tra} \leq 800$). These clarified the significance of the proposed more representable unimodal likelihood models.

**Detail of Results** For a better understanding of the obtained results, we further provide experimental considerations about comparisons within each set of methods that have interesting differences while maintaining commonality; see Table 4 and following paragraphs. We believe that these considerations will be helpful for future development.

**POI *v.s.* PO-ORD-ACL, and OH-POI *v.s.* OH-ORD-ACL** The POI model is a special instance of the PO-ORD-ACL models with fixed mode-wise heteroscedasticity, and has a weaker representation ability. This relation might make the POI model work better when $n_{\rm tra}$ was small and the PO-ORD-ACL model work better when $n_{\rm tra}$ was large, in every task. The situation for OH-POI *v.s.* OH-ORD-ACL is similar to that for POI *v.s.* PO-ORD-ACL.

Table 3: A cell for a method and training sample size $n_{\text{tra}}$ in a sub-table for an error shows the total (over the 21 datasets) numbers of times that the method was top 1, 2, and 3 in the rank regarding the mean of the test errors within those for 12 methods. In each block for $n_{\text{tra}}$ in a sub-table, the 1st, 2nd, and 3rd best results are respectively highlighted in the red, green, and blue colors.

NLL

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| BIN | 0,4,5 | 0,3,5 | 0,1,5 | 0,3,5 | 2,2,2 | 0,0,0 |
| POI | 4,9,11 | 5,7,10 | 3,5,7 | 2,2,3 | 1,1,1 | 0,0,0 |
| PO-ORD-ACL | 0,0,2 | 0,1,1 | 1,1,1 | 1,1,1 | 0,1,2 | 0,0,0 |
| PO-VS-SL | 3,3,4 | 1,3,4 | 0,1,2 | 0,0,1 | 0,1,2 | 0,2,2 |
| OH-BIN | 0,6,8 | 1,7,10 | 5,7,9 | 2,3,4 | 1,1,2 | 1,1,3 |
| OH-POI | 6,9,12 | 6,8,11 | 3,7,9 | 3,4,7 | 0,2,4 | 0,1,1 |
| OH-ORD-ACL | 0,0,1 | 0,0,0 | 1,1,3 | 1,2,3 | 0,0,1 | 0,0,3 |
| OH-VS-SL | 0,0,3 | 1,1,2 | 0,2,3 | 0,1,2 | 3,4,5 | 5,6,8 |
| ORD-ACL | 4,6,9 | 4,7,10 | 6,10,10 | 7,10,11 | 0,9,12 | 0,7,11 |
| VS-SL | 0,0,0 | 0,0,2 | 0,0,1 | 0,3,8 | 6,7,11 | 9,12,13 |
| ACL | 4,5,8 | 3,5,8 | 2,6,11 | 4,10,15 | 6,10,15 | 4,9,16 |
| SL | 0,0,0 | 0,0,0 | 0,1,2 | 1,3,3 | 2,4,6 | 2,4,6 |

MAE

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| BIN | 2,4,6 | 1,6,9 | 1,4,7 | 1,2,6 | 1,4,5 | 0,0,0 |
| POI | 1,10,12 | 1,4,6 | 0,2,2 | 0,0,1 | 0,0,0 | 0,1,1 |
| PO-ORD-ACL | 0,1,2 | 1,3,4 | 1,3,6 | 0,4,4 | 0,0,2 | 0,0,0 |
| PO-VS-SL | 4,5,9 | 5,5,8 | 5,6,11 | 2,4,7 | 1,2,2 | 0,2,4 |
| OH-BIN | 2,4,7 | 2,3,10 | 1,5,6 | 4,4,6 | 0,1,3 | 1,1,2 |
| OH-POI | 7,10,12 | 3,9,9 | 4,5,8 | 2,4,4 | 1,4,4 | 0,2,2 |
| OH-ORD-ACL | 0,1,1 | 0,0,1 | 0,0,0 | 0,1,1 | 0,0,1 | 0,2,3 |
| OH-VS-SL | 1,1,2 | 1,2,2 | 0,2,2 | 0,2,2 | 2,3,5 | 2,2,2 |
| ORD-ACL | 4,5,10 | 7,9,10 | 8,10,13 | 7,11,13 | 7,10,13 | 5,10,15 |
| VS-SL | 0,1,1 | 0,1,3 | 1,4,6 | 3,6,10 | 6,10,14 | 6,10,12 |
| ACL | 0,0,1 | 0,0,1 | 0,1,2 | 2,4,8 | 2,7,12 | 5,9,15 |
| SL | 0,0,0 | 0,0,0 | 0,0,0 | 0,0,1 | 1,1,2 | 2,3,7 |

MZE

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| BIN | 2,2,5 | 0,2,3 | 1,1,2 | 0,1,2 | 0,0,0 | 0,0,0 |
| POI | 1,7,8 | 1,3,5 | 0,0,1 | 0,0,0 | 0,1,1 | 0,0,1 |
| PO-ORD-ACL | 0,1,6 | 1,2,3 | 0,1,1 | 0,0,0 | 0,0,1 | 0,0,1 |
| PO-VS-SL | 1,4,7 | 2,2,4 | 1,3,3 | 1,2,3 | 0,0,0 | 0,1,1 |
| OH-BIN | 4,6,10 | 3,9,11 | 5,11,13 | 9,13,13 | 8,12,14 | 5,7,7 |
| OH-POI | 9,11,12 | 5,8,12 | 3,5,11 | 2,4,7 | 0,3,5 | 0,3,5 |
| OH-ORD-ACL | 1,1,1 | 0,1,2 | 0,1,4 | 0,2,3 | 0,0,1 | 1,3,5 |
| OH-VS-SL | 1,2,3 | 2,3,4 | 1,4,5 | 1,3,4 | 3,6,10 | 3,3,8 |
| ORD-ACL | 2,6,6 | 6,9,13 | 9,11,12 | 5,8,10 | 2,7,8 | 2,6,12 |
| VS-SL | 0,1,4 | 1,2,4 | 1,4,8 | 1,4,14 | 3,5,12 | 5,9,10 |
| ACL | 0,1,1 | 0,1,2 | 0,1,3 | 1,4,4 | 4,7,8 | 5,7,9 |
| SL | 0,0,0 | 0,0,0 | 0,0,0 | 1,1,3 | 1,1,3 | 0,3,4 |

MSE

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| BIN | 1,2,7 | 3,5,8 | 2,7,11 | 1,5,9 | 2,6,9 | 1,3,6 |
| POI | 4,9,12 | 4,6,12 | 1,3,6 | 0,1,5 | 0,1,3 | 0,1,4 |
| PO-ORD-ACL | 1,1,1 | 0,1,2 | 0,2,4 | 0,2,3 | 1,1,3 | 0,3,6 |
| PO-VS-SL | 2,3,5 | 2,2,4 | 4,4,5 | 3,6,6 | 0,5,5 | 2,5,8 |
| OH-BIN | 2,6,10 | 2,5,11 | 2,4,5 | 0,1,4 | 0,0,1 | 0,1,1 |
| OH-POI | 8,10,13 | 4,9,11 | 2,7,10 | 2,2,5 | 0,2,3 | 1,1,2 |
| OH-ORD-ACL | 0,1,1 | 0,0,0 | 0,0,0 | 0,0,1 | 0,1,3 | 0,2,2 |
| OH-VS-SL | 0,1,2 | 0,1,1 | 0,1,1 | 1,1,1 | 1,2,3 | 2,2,3 |
| ORD-ACL | 1,6,7 | 4,11,11 | 8,10,14 | 8,11,12 | 7,9,12 | 7,9,11 |
| VS-SL | 1,2,3 | 1,1,2 | 2,4,6 | 5,7,9 | 6,8,11 | 6,7,7 |
| ACL | 1,1,2 | 1,1,1 | 0,0,1 | 1,4,6 | 3,5,6 | 1,5,8 |
| SL | 0,0,0 | 0,0,0 | 0,0,0 | 0,2,2 | 1,2,4 | 1,3,5 |

**BIN *v.s.* POI *v.s.* PO-ORD-ACL *v.s.* PO-VS-SL (1-FDF)** The PO-ORD-ACL and PO-VS-SL models can adjust the mode-wise heteroscedasticity, while the BIN and POI models cannot. In this respect, the PO-ORD-ACL and PO-VS-SL models tend more representable than the BIN and POI models. For this reason, the BIN and POI models worked better when $n_{\text{tra}}$ was small, and the PO-ORD-ACL and PO-VS-SL models worked better when $n_{\text{tra}}$ was large.

**BIN *v.s.* OH-BIN, and POI *v.s.* OH-POI** The OH-BIN and OH-POI models are respectively OH-generalizations of the BIN and POI models, and have a stronger representation ability. Thus, OH models were better regarding the NLL that was directly optimized. However, OH models were bad for some tasks (especially, task-A and -S) when $n_{\text{tra}}$ was large, which is a counter-intuitive result and requires further analysis but may be because the difference of the restricted structures of the two likelihood models makes a difference in the compatibility between the model and task.

**OH-BIN *v.s.* OH-POI *v.s.* OH-ORD-ACL *v.s.* OH-VS-SL (2-FDF)** The situation is similar to that for the 1-FDF models, BIN *v.s.* POI *v.s.* PO-ORD-ACL *v.s.* PO-VS-SL.

**PO-ORD-ACL *v.s.* OH-ORD-ACL *v.s.* ORD-ACL, and PO-VS-SL *v.s.* OH-VS-SL *v.s.* VS-SL** The VS-SL, OH-VS-SL, and PO-VS-SL models (and the ORD-ACL, OH-ORD-ACL, and PO-ORD-ACL models) are more representable in that order. Presumably for this reason, the PO- and OH-VS-SL models worked better with small $n_{\text{tra}}$, and the VS-SL model worked better with large $n_{\text{tra}}$.

**ORD-ACL *v.s.* ACL** The ACL model can represent any data, and the ORD-ACL model can represent unimodal data with the DRs' unimodality constraint. The unimodality of the ORD-ACL model might improve the generalization performance when $n_{\text{tra}}$ was small, but the ORD-ACL model was not good when $n_{\text{tra}}$ was large perhaps because its DRs' unimodality constraint restricted its representation ability too much for our tried data; see again DRs' MU in Tables 1 and 2.

**VS-SL *v.s.* SL** The SL model can represent any data, and the VS-SL model can represent any unimodal data. Therefore, the VS-SL model has more restricted representation ability that the SL model. Unlike the comparison between ACL and ORD-ACL models, the VS-SL model gave better results than the SL model, probably thanks to the validity of the unimodal hypothesis.

Table 4: A cell for an error and training sample size $n_{\text{tra}}$ in a sub-table for a group of specified methods shows the total (over the 21 datasets) number of times that each method won in the Mann-Whitney U-test with $p$-value 0.05 regarding the error for each pair of two methods within the group. In each cell in a sub-table, methods that won the most times are highlighted in the corresponding color.

POI *v.s.* PO-ORD-ACL

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| NLL | 12,5 | 12,5 | 13,4 | 14,5 | 11,8 | 7,7 |
| MZE | 6,5 | 5,7 | 3,11 | 2,10 | 3,11 | 1,11 |
| MAE | 5,1 | 4,2 | 1,2 | 1,4 | 2,6 | 2,5 |
| MSE | 7,1 | 6,1 | 3,2 | 1,2 | 2,2 | 1,4 |

BIN *v.s.* POI *v.s.* PO-ORD-ACL *v.s.* PO-VS-SL

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| NLL | 37,36,17,20 | 37,36,14,23 | 33,34,13,20 | 32,33,13,20 | 29,30,16,20 | 23,24,16,25 |
| MZE | 16,20,6,10 | 16,14,9,16 | 23,7,15,20 | 25,6,13,23 | 24,7,18,26 | 22,6,19,26 |
| MAE | 11,13,2,5 | 7,8,4,7 | 6,3,5,11 | 12,2,5,12 | 13,4,9,11 | 15,5,9,10 |
| MSE | 15,15,2,4 | 8,13,2,5 | 8,8,4,6 | 10,2,4,5 | 5,4,5,5 | 5,4,9,9 |

BIN *v.s.* OH-BIN

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| NLL | 3,13 | 4,14 | 4,11 | 4,7 | 7,9 | 6,11 |
| MZE | 0,6 | 0,6 | 0,11 | 1,15 | 1,17 | 0,16 |
| MAE | 1,1 | 2,5 | 3,4 | 5,5 | 6,6 | 7,3 |
| MSE | 0,1 | 1,2 | 2,3 | 8,2 | 14,1 | 15,0 |

POI *v.s.* OH-POI

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| NLL | 2,7 | 4,10 | 5,10 | 5,11 | 6,11 | 4,14 |
| MZE | 1,8 | 0,12 | 0,16 | 0,18 | 1,17 | 1,18 |
| MAE | 1,6 | 0,7 | 4,9 | 5,8 | 5,8 | 8,6 |
| MSE | 0,3 | 2,6 | 4,6 | 6,3 | 9,3 | 13,1 |

OH-POI *v.s.* OH-ORD-ACL

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| NLL | 13,1 | 12,2 | 10,4 | 8,8 | 6,11 | 7,11 |
| MZE | 10,1 | 9,3 | 10,4 | 8,4 | 8,6 | 5,10 |
| MAE | 8,0 | 10,0 | 11,1 | 8,2 | 5,5 | 6,7 |
| MSE | 13,0 | 9,0 | 9,0 | 5,3 | 4,5 | 1,10 |

OH-BIN *v.s.* OH-POI *v.s.* OH-ORD-ACL *v.s.* OH-VS-SL

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| NLL | 28,37,8,10 | 32,35,9,13 | 23,28,13,13 | 14,21,17,19 | 15,15,24,30 | 16,14,23,36 |
| MZE | 22,30,4,4 | 23,24,5,9 | 29,27,6,16 | 36,18,6,16 | 34,12,12,24 | 26,12,15,25 |
| MAE | 19,22,3,2 | 22,26,3,2 | 19,26,4,4 | 25,20,4,6 | 19,12,12,11 | 17,14,17,13 |
| MSE | 23,29,2,1 | 20,23,3,2 | 17,22,2,2 | 13,14,7,7 | 7,12,13,11 | 9,10,24,18 |

PO-ORD-ACL *v.s.* OH-ORD-ACL *v.s.* ORD-ACL

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| NLL | 7,15,28 | 6,13,28 | 1,14,26 | 3,16,22 | 4,14,30 | 1,14,33 |
| MZE | 6,2,17 | 6,2,22 | 3,7,28 | 2,10,34 | 3,13,27 | 3,16,27 |
| MAE | 7,2,16 | 9,1,22 | 8,2,26 | 9,1,27 | 7,5,22 | 8,7,26 |
| MSE | 2,0,13 | 5,0,15 | 3,1,23 | 9,1,24 | 8,4,18 | 12,4,19 |

PO-VS-SL *v.s.* OH-VS-SL *v.s.* VS-SL

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| NLL | 11,22,17 | 12,24,19 | 11,26,16 | 3,20,21 | 5,15,29 | 4,19,29 |
| MZE | 9,5,10 | 8,5,10 | 4,9,17 | 2,14,25 | 0,18,28 | 1,21,24 |
| MAE | 19,1,9 | 21,0,11 | 14,1,17 | 11,3,22 | 7,5,26 | 10,5,25 |
| MSE | 13,1,10 | 13,0,12 | 11,3,13 | 13,3,15 | 8,3,17 | 12,5,16 |

ORD-ACL *v.s.* ACL

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| NLL | 9,5 | 8,3 | 8,5 | 5,9 | 2,7 | 3,7 |
| MZE | 16,0 | 17,0 | 11,0 | 8,2 | 5,3 | 5,6 |
| MAE | 15,0 | 14,0 | 12,1 | 8,1 | 8,1 | 6,2 |
| MSE | 14,0 | 14,0 | 10,0 | 7,1 | 8,2 | 5,2 |

VS-SL *v.s.* SL

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| NLL | 14,2 | 15,3 | 13,5 | 13,5 | 12,3 | 12,4 |
| MZE | 21,0 | 19,0 | 18,0 | 17,2 | 14,2 | 10,2 |
| MAE | 21,0 | 20,0 | 19,0 | 17,0 | 15,1 | 12,2 |
| MSE | 20,0 | 20,0 | 19,0 | 15,0 | 15,1 | 13,1 |

ORD-ACL *v.s.* VS-SL *v.s.* ACL *v.s.* SL

| | $n_{\text{tra}}=25$ | $n_{\text{tra}}=50$ | $n_{\text{tra}}=100$ | $n_{\text{tra}}=200$ | $n_{\text{tra}}=400$ | $n_{\text{tra}}=800$ |
|---|---|---|---|---|---|---|
| NLL | 44,16,40,4 | 46,16,36,3 | 41,14,38,8 | 24,16,36,9 | 14,26,30,8 | 14,31,26,10 |
| MZE | 43,37,21,0 | 44,33,20,0 | 40,32,18,0 | 30,29,18,4 | 22,27,20,4 | 21,22,21,5 |
| MAE | 42,29,20,0 | 42,29,19,0 | 37,29,19,0 | 29,28,19,0 | 26,26,15,2 | 25,22,18,3 |
| MSE | 40,28,19,0 | 36,30,19,0 | 37,28,14,0 | 29,21,18,0 | 27,25,14,2 | 24,24,14,3 |

**ORD-ACL *v.s.* VS-SL *v.s.* ACL *v.s.* SL (Full-FDF)** The ORD-ACL model was the best in many cases, and the VS-SL model was better when $n_{\text{tra}}$ was large. This is thought due to their unimodality and the DRs' unimodality constraint of the ORD-ACL model.

For further details of the experimental results, refer to Appendix B.

## 7 Conclusion

In this paper, we pointed out that previous unimodal BIN and POI models have a weak representation ability from the perspective of the DRs' unimodality constraint, scale, skewness, and FDF, and then developed more representable unimodal ORD-ACL and VS-SL models as well as their PO-constrained version and OH-extension. In our experiments, 1-FDF or 2-FDF OH models worked better when the training sample size was small, and full-FDF ORD-ACL and VS-SL models worked better than low-FDF unimodal models when the training sample size was large. Also, full-FDF ORD-ACL and VS-SL models were better than full-FDF models with no unimodality guarantee, due to their unimodality guarantee reasonable for many real-world data that are almost unimodal.

## References

Alan Agresti. *Analysis of Ordinal Categorical Data*, volume 656. John Wiley & Sons, 2010.

John A Anderson. Regression and ordered categorical variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(1):1–22, 1984.

David Andrich. A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573, 1978.

Anonymous. Modified threshold method for ordinal regression. *Submitted to Transactions on Machine Learning Research*, 2022. URL `https://openreview.net/forum?id=PInXz6Gasv`. Under review.

Christopher Beckham and Christopher Pal. Unimodal probability distributions for deep ordinal classification. In *International Conference on Machine Learning*, pp. 411–419, 2017.

Ana M Bianco and Victor J Yohai. Robust estimation in the logistic regression model. In *Robust Statistics, Data Analysis, and Computer Intensive Methods*, pp. 17–34. 1996.

Paul-Christian Bürkner and Matti Vuorre. Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101, 2019.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Consistent rank logits for ordinal regression with convolutional neural networks. *arXiv preprint arXiv:1901.07884*, 6, 2019.

Chuansheng Chen, Shin-Ying Lee, and Harold W Stevenson. Response style and cross-cultural comparisons of rating scales among east asian and north american students. *Psychological Science*, 6(3):170–175, 1995.

Christophe Croux, Gentiane Haesbroeck, and Christel Ruwet. Robust estimation for ordinal regression. *Journal of Statistical Planning and Inference*, 143(9):1486–1499, 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.

Leo A Goodman. Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74(367):537–552, 1979.

Pedro Antonio Gutierrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2015.

Maria Iannario and Domenico Piccolo. Cub models: Statistical methods and empirical evidence. *Modern Analysis of Customer Surveys: with Applications using R*, pp. 231–258, 2011.

Rob Kaas and Jan M Buhrman. Mean, median and mode in binomial distributions. *Statistica Neerlandica*, 34(1):13–18, 1980.

Kyoung-Jae Kim and Hyunchul Ahn. A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. *Computers & Operations Research*, 39(8):1800–1811, 2012.

Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer Science & Business Media, 2011.

Geoff N Masters. A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, 1982.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

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.

Domenico Piccolo. On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, 5(1):85–104, 2003.

Gary Simon. Alternative analyses for the singly-ordered contingency table. *Journal of the American Statistical Association*, 69(348):971–976, 1974.

Shipeng Yu, Kai Yu, Volker Tresp, and Hans-Peter Kriegel. Collaborative ordinal regression. In *Proceedings of the International Conference on Machine Learning*, pp. 1089–1096, 2006.

## A Proof of Theorems

*Proof of Theorem 1.* First, we prove Theorem 1, (i): Letting $p = \frac{1}{1+e^{-u}}$, one has that

$$\begin{aligned}\frac{P_{\mathrm{bin}}(y+1;u,s)}{P_{\mathrm{bin}}(y;u,s)} &= \frac{e^{\log(P_{\mathrm{bin}}(y+1;u))/s}}{\sum_{k=1}^{K} e^{\log(P_{\mathrm{bin}}(k;u))/s}} \bigg/ \frac{e^{\log(P_{\mathrm{bin}}(y;u))/s}}{\sum_{k=1}^{K} e^{\log(P_{\mathrm{bin}}(k;u))/s}} = \frac{e^{\log(P_{\mathrm{bin}}(y+1;u))/s}}{e^{\log(P_{\mathrm{bin}}(y;u))/s}} \\ &= e^{\{\log(P_{\mathrm{bin}}(y+1;u))-\log(P_{\mathrm{bin}}(y;u))\}/s} = e^{\{-\log(K-y)+\log(1-p)+\log(y)-\log(p)\}/s} \\ &= e^{\frac{1}{s}\log\left(\frac{y(1-p)}{(K-y)p}\right)} = e^{\log\left(\frac{y(1-p)}{(K-y)p}\right)^{1/s}} = \left(\frac{y(1-p)}{(K-y)p}\right)^{1/s}.\end{aligned} \tag{13}$$

Next, we prove Theorem 1, (ii) and (iii): Let $r(y,p) := \frac{y(1-p)}{(K-y)p}$ so that $\frac{P_{\mathrm{bin}}(y+1;u,s)}{P_{\mathrm{bin}}(y;u,s)} = \{r(y,p)\}^{1/s}$. (5) shows $\max\{1, Kp\} \leq M((P_{\mathrm{b}}(k;p))_{k\in[K]}) \leq \min\{Kp+1, K\}$ with $p = \frac{1}{1+e^{-u}}$. Since $(P_{\mathrm{b}}(k;p))_{k\in[K]} = (P_{\mathrm{bin}}(y;u,1))_{y\in[K]}$ and the scaling factor does not change the mode, one has that $\max\{1, Kp\} \leq m = M((P_{\mathrm{bin}}(y;u,s))_{y\in[K]}) \leq \min\{Kp+1, K\}$. This implies $Kp \leq m \leq Kp+1$ (or equivalently $(m-1)/K \leq p \leq m/K$). Since $r(y,p)$ is increasing in $y \in [K-1]$ and decreasing in $p \in [0,1]$, one has that

$$r(y,p) = \frac{y(1-p)}{(K-y)p} \leq \left.\frac{y(1-p)}{(K-y)p}\right|_{y=m-1, p=(m-1)/K} = 1 \text{ for } y = 1, \ldots, m-1, \tag{14}$$

$$r(y,p) = \frac{y(1-p)}{(K-y)p} \geq \left.\frac{y(1-p)}{(K-y)p}\right|_{y=m, p=m/K} = 1 \text{ for } y = m, \ldots, K-1, \tag{15}$$

which show Theorem 1, (ii). Also, these results and the fact that $r(y,p)$ is increasing in $y \in [K-1]$ prove that

$$r(1,p) \leq r(2,p) \leq \cdots \leq r(m-1,p) \leq 1, \text{ and } 1 \geq r(m,p)^{-1} \geq r(m+1,p)^{-1} \geq \cdots \geq r(K-1,p)^{-1}, \tag{16}$$

which implies Theorem 1, (iii). □

*Proof of Theorem 2.* First, we prove Theorem 2, (i):

$$\frac{P_{\mathrm{acl}}(y+1;\boldsymbol{u})}{P_{\mathrm{acl}}(y;\boldsymbol{u})} = \frac{\prod_{l=1}^{y} e^{-u_l}}{\sum_{k=1}^{K}\prod_{l=1}^{k-1} e^{-u_l}} \bigg/ \frac{\prod_{l=1}^{y-1} e^{-u_l}}{\sum_{k=1}^{K}\prod_{l=1}^{k-1} e^{-u_l}} = \frac{\prod_{l=1}^{y} e^{-u_l}}{\prod_{l=1}^{y-1} e^{-u_l}} = e^{-u_y}. \tag{17}$$

Next, we prove Theorem 2, (ii) and (iii): Under the assumption that $u_0(:= -\infty) \leq \cdots \leq u_{m-1} \leq 0 \leq u_m \leq \cdots \leq u_K(:= +\infty)$ for some $m \in [K]$, one has that

$$\frac{P_{\mathrm{acl}}(1;\boldsymbol{u})}{P_{\mathrm{acl}}(2;\boldsymbol{u})} = e^{u_1} \leq \frac{P_{\mathrm{acl}}(2;\boldsymbol{u})}{P_{\mathrm{acl}}(3;\boldsymbol{u})} = e^{u_2} \leq \cdots \leq \frac{P_{\mathrm{acl}}(m-1;\boldsymbol{u})}{P_{\mathrm{acl}}(m;\boldsymbol{u})} = e^{u_{m-1}} \leq 1, \tag{18}$$

$$1 \geq \frac{P_{\mathrm{acl}}(m+1;\boldsymbol{u})}{P_{\mathrm{acl}}(m;\boldsymbol{u})} = e^{-u_m} \geq \frac{P_{\mathrm{acl}}(m+2;\boldsymbol{u})}{P_{\mathrm{acl}}(m+1;\boldsymbol{u})} = e^{-u_{m+1}} \geq \cdots \geq \frac{P_{\mathrm{acl}}(K;\boldsymbol{u})}{P_{\mathrm{acl}}(K-1;\boldsymbol{u})} = e^{-u_{K-1}}. \tag{19}$$

These results show Theorem 2, (ii) and (iii). □

*Proof of Theorem 3.* Theorem 3 would be trivial. We omit the proof. □

*Proof of Theorem 4.* For $\boldsymbol{b} = (b_k)_{k\in[K-1]}$ satisfying $b_k = \log(k)/s$ for $k = 1, \ldots, K-1$, $v = \log(u)/s$, and $\boldsymbol{w} = (w_k)_{k\in[K-1]} = \boldsymbol{b} - v \cdot \boldsymbol{1}$, one has that

$$P_{\mathrm{acl}}(y;\boldsymbol{w}) = \frac{\prod_{l=1}^{y-1} e^{-w_l}}{\sum_{k=1}^{K}\prod_{l=1}^{k-1} e^{-w_l}} = \frac{e^{-\sum_{l=1}^{y-1}\{\log(l)-\log(u)\}/s}}{\sum_{k=1}^{K} e^{-\sum_{l=1}^{k-1}\{\log(l)-\log(u)\}/s}} = \frac{e^{\{(y-1)\log(u)-\log((y-1)!)\}/s}}{\sum_{k=1}^{K} e^{\{(k-1)\log(u)-\log((k-1)!)\}/s}}, \tag{20}$$

which is equal to $P_{\mathrm{poi}}(y;u,s)$ in (7). This concludes the proof. □

## B Supplemental Experimental Results

We here describe details of the experimental results. Figures 7, 8, 9, and 10 respectively show errorbar-plots of mean and STD of the test NLLs, MZEs, MAEs, MSEs of all methods for each dataset.

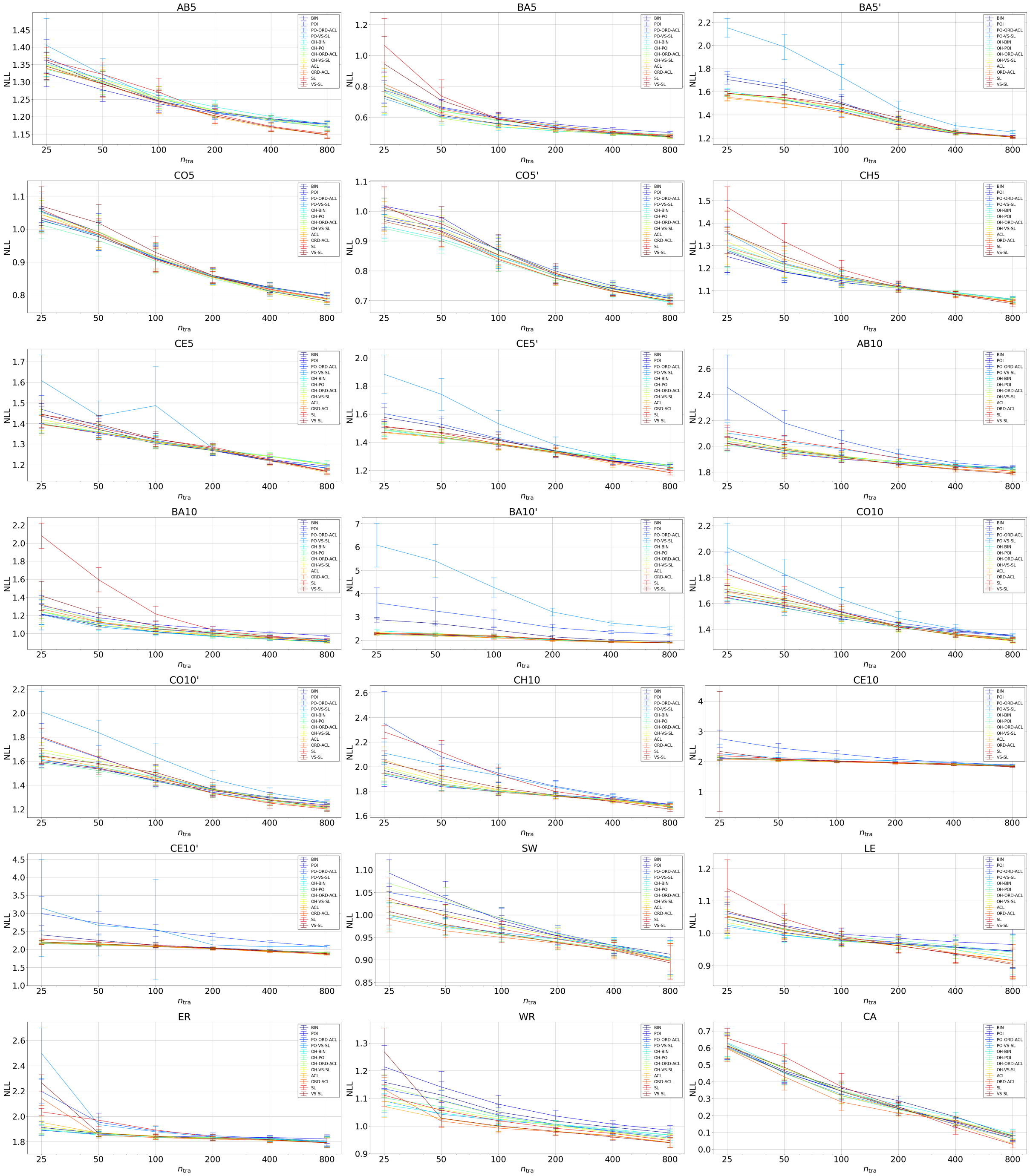
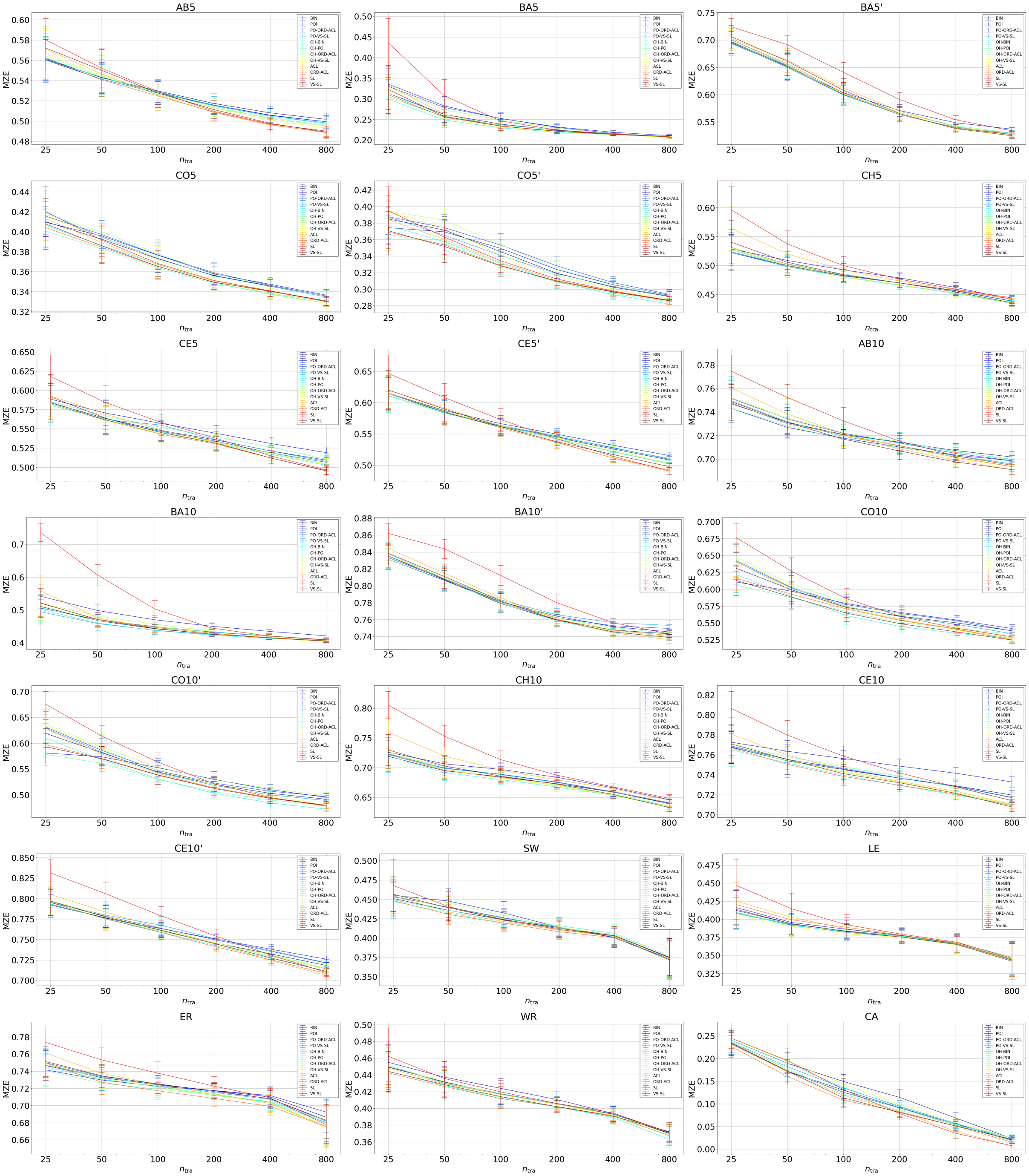


Figure 7: It shows errorbar-plots of mean and STD of the test NLLs of all methods for each dataset.



AB5
BA5
BA5'
CO5
CO5'
CH5
CE5
CE5'
AB10
BA10
BA10'
CO10
CO10'
CH10
CE10
CE10'
SW
LE
ER
WR
CA

Figure 8: It shows errorbar-plots of mean and STD of the test MZEs of all methods for each dataset.

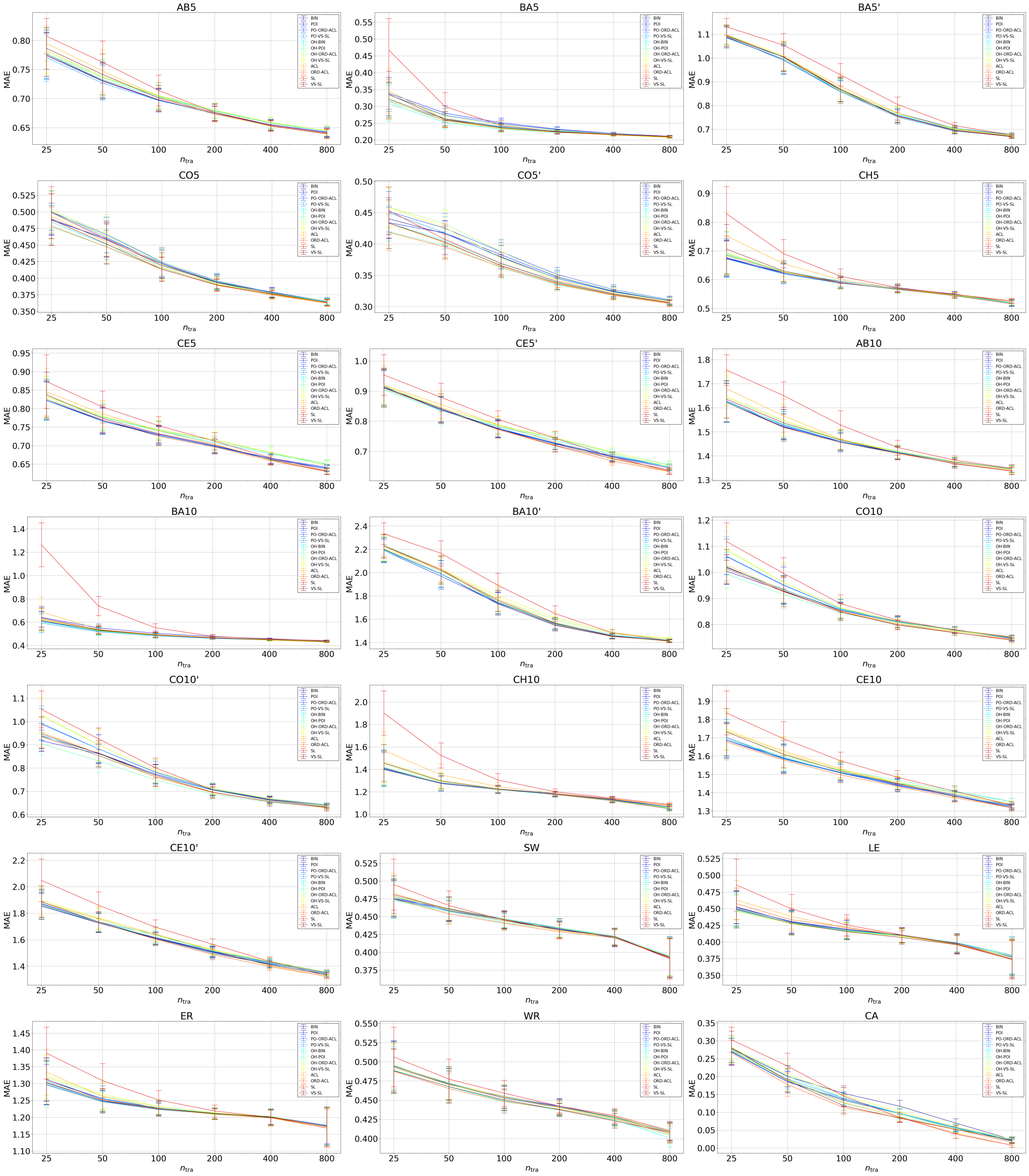


Figure 9: It shows errorbar-plots of mean and STD of the test MAEs of all methods for each dataset.

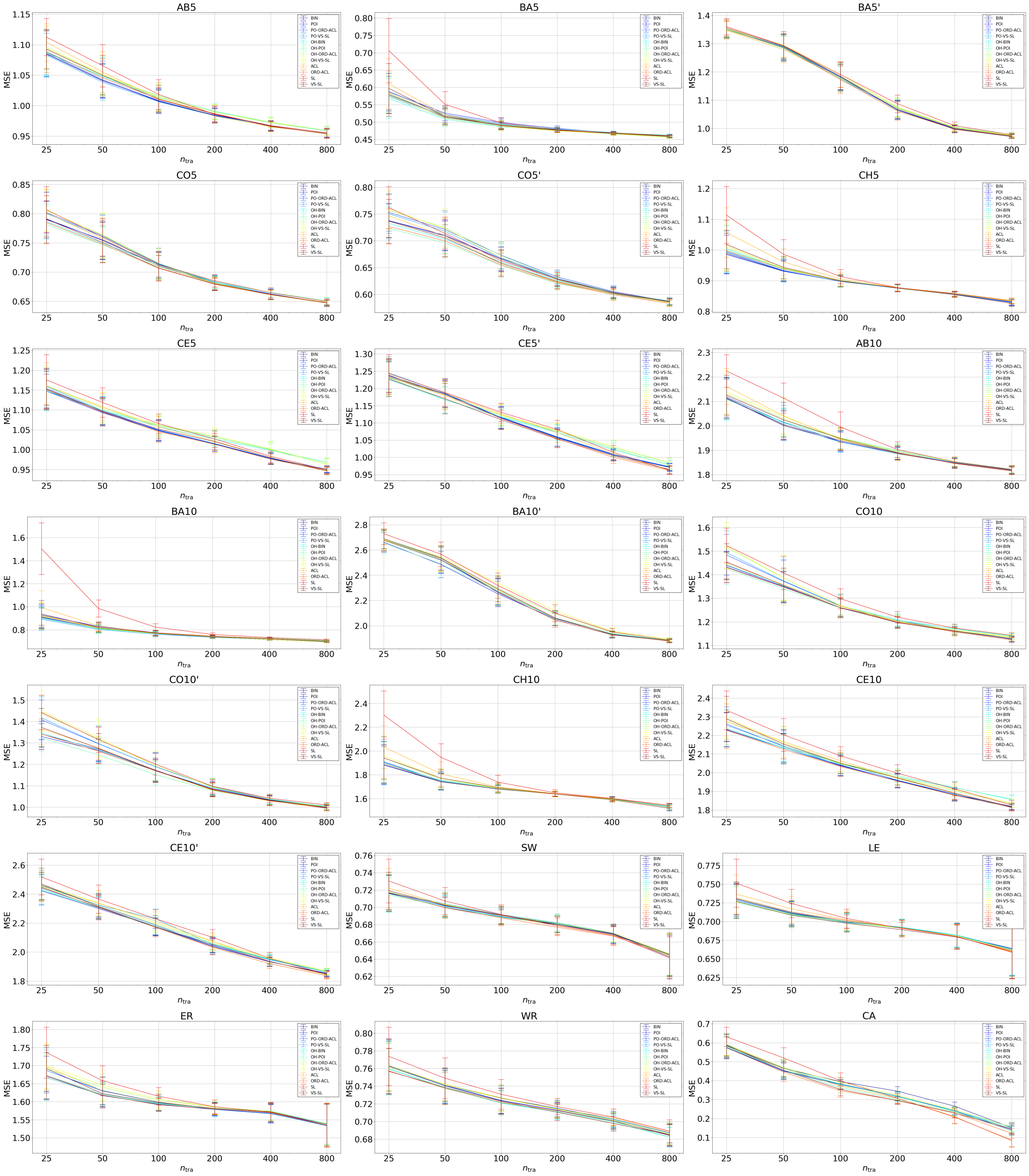

AB5, BA5, BA5', CO5, CO5', CH5, CE5, CE5', AB10, BA10, BA10', CO10, CO10', CH10, CE10, CE10', SW, LE, ER, WR, CA

Figure 10: It shows errorbar-plots of mean and STD of the test MSEs of all methods for each dataset.