# Structured Mixture of Continuation-ratio Logits Models for Ordinal Regression

Jizhou Kang

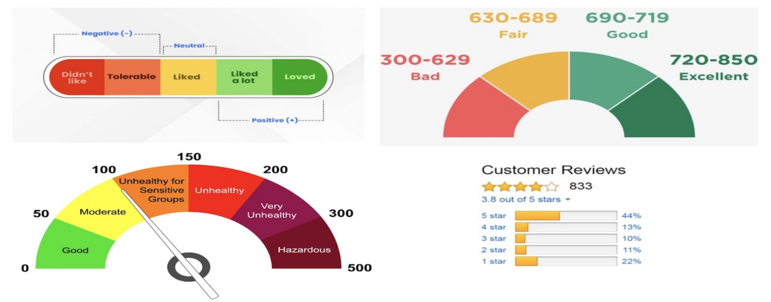
Aug 9th, 2022

Department of Statistics, UC Santa Cruz

Joint work with Prof. Athanasios Kottas

## Background

- Ordinal data are widely encountered in many fields, such as econometrics, business, biomedical and social sciences.




- Ordinal responses together with features (covariates) form up an ordinal regression problem.

# Motivation

- In ordinal regression problems, flexible inference methods need to capture the covariate-response relationship, as well as incorporate the ordinal discrete nature of the responses;

- The covariate-response relationship is depicted by the probability response curves;

# Motivation

- In ordinal regression problems, flexible inference methods need to capture the covariate-response relationship, as well as incorporate the ordinal discrete nature of the responses;
- The covariate-response relationship is depicted by the probability response curves;
- Challenges:
  - Non-standard response distributions;
  - Non-standard regression relationships between the ordinal response categories and covariates;
  - In terms of computation, the proposed method should have tractable inference algorithm.

## Literature review

- Discretize latent continuous variable: probit regression;
- Seeking for more flexibility, semiparametric and nonparametric models have been developed;
- To inference the probability, one need to do integral with respect to the regions defined by the cut-off points;

# Literature review

- Discretize latent continuous variable: probit regression;
- Seeking for more flexibility, semiparametric and nonparametric models have been developed;
- To inference the probability, one need to do integral with respect to the regions defined by the cut-off points;
- Model discrete distribution directly: logits regression family;
- Members including adjacent-categories logits, cumulative logits, and continuation-ratio logits;
- Modeling probability parameters directly, which makes inference easier.

# Summary of existing literatures

| | Discretize latent continuous variable | Model discrete distribution directly |
|---|---|---|
| **Parametric model** | Probit regression (Albert & Chib, 1993) | Logits regression family<br>**Continuation-ratio logits models** (Tutz, 1991) |
| **Semiparametric model** | Relaxing normality assumption (Newton et al., 1996), linearity assumption (Mukhopadhyay & Gelfand, 1997), or both (Chib & Greenberg, 2010) | Replacing systematic component with Gaussian process (Linderman et al., 2015); Adding random effects term with DP prior to systematic component (Tang and Duan, 2012) |
| **Nonparametric model** | **Bayesian density estimation** for the joint distribution of covariates and responses, for categorical variables (Shahbaba & Neal, 2009; Dunson & Bhattacharya, 2010), and ordinal variables (DeYoreo & Kottas, 2018) | Common-atoms DDP model for specific types of problems (Fronczyk and Kottas, 2014) |

# Objectives

- Flexible ordinal regression made easy;

  1. Flexible: allow general forms for ordinal response distribution, ordinal regression relationship;
  2. Flexible: clear prior specification strategy, efficient posterior inference algorithm, easy to work with expressions for quantities of interest;

# Objectives

- Flexible ordinal regression made easy;
  1. Flexible: allow general forms for ordinal response distribution, ordinal regression relationship;
  2. Flexible: clear prior specification strategy, efficient posterior inference algorithm, easy to work with expressions for quantities of interest;
- Study the trade-off between model flexibility and implementation difficulty, and seek a balance between them;

# Objectives

- Flexible ordinal regression made easy;
  1. Flexible: allow general forms for ordinal response distribution, ordinal regression relationship;
  2. Flexible: clear prior specification strategy, efficient posterior inference algorithm, easy to work with expressions for quantities of interest;
- Study the trade-off between model flexibility and implementation difficulty, and seek a balance between them;
- A unified modeling framework that is applicable for different kinds of ordinal regression problems.

# Proposed Methodology

# Extending parametric model to nonparametric model

- The multinomial distribution has the continuation-ratio logits representation:

$$\mathbf{Y} \sim Mult(\mathbf{Y}|m, \pi_1, \cdots, \pi_C) \Longleftrightarrow$$
$$\mathbf{Y} \sim Bin(Y_1|m_1, \varphi(\theta_1)) \cdots Bin(Y_{C-1}|m_{C-1}, \varphi(\theta_{C-1})) = K(\mathbf{Y}|\mathbf{m}, \boldsymbol{\theta})$$

  - $m_j = m$ if $j = 1$ and $m_j = m - \sum_{k=1}^{j-1} y_k$ for $j = 2, \cdots, C-1$;
  - $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{C-1})$ and $\theta_j = \mathbf{x}^T \boldsymbol{\beta}_j$;
  - Denote the expit function as $\varphi(x) = \frac{\exp(x)}{1+\exp(x)}$ ;

# Extending parametric model to nonparametric model

- The multinomial distribution has the continuation-ratio logits representation:

$$\mathbf{Y} \sim Mult(\mathbf{Y}|m, \pi_1, \cdots, \pi_C) \Longleftrightarrow$$
$$\mathbf{Y} \sim Bin(Y_1|m_1, \varphi(\theta_1)) \cdots Bin(Y_{C-1}|m_{C-1}, \varphi(\theta_{C-1})) = K(\mathbf{Y}|\mathbf{m}, \boldsymbol{\theta})$$

  - $m_j = m$ if $j = 1$ and $m_j = m - \sum_{k=1}^{j-1} y_k$ for $j = 2, \cdots, C-1$;
  - $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{C-1})$ and $\theta_j = \mathbf{x}^T \boldsymbol{\beta}_j$;
  - Denote the expit function as $\varphi(x) = \frac{\exp(x)}{1+\exp(x)}$ ;

- Generalize the model via Bayesian nonparametric mixture modeling. Placing a covariate-dependent nonparametric prior:

$$\mathbf{Y}|G_{\mathbf{x}} \sim \int K(\mathbf{Y}|\mathbf{m}, \boldsymbol{\theta}) dG_{\mathbf{x}}(\boldsymbol{\theta}) = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) K(\mathbf{Y}|\mathbf{m}, \boldsymbol{\theta}_\ell(\mathbf{x}))$$

# Model structure illustration

- The continuation-ratio logits structure offers a sequential mechanism to allocate the ordinal response $Y$.

Generate $\mathcal{H}_1 \sim Bern(\Delta_1), \Delta_1 = \varphi(\mathbf{x}^T \boldsymbol{\beta}_1), \boldsymbol{\beta}_1 \sim N(\boldsymbol{\mu}_1, \Sigma_1)$

$\mathcal{H}_1 = 1$ $\quad$ $\mathcal{H}_1 = 0$

$Y = 1$ $\quad$ Generate $\mathcal{H}_2 \sim Bern(\Delta_2), \Delta_2 = \varphi(\mathbf{x}^T \boldsymbol{\beta}_2), \boldsymbol{\beta}_2 \sim N(\boldsymbol{\mu}_2, \Sigma_2)$

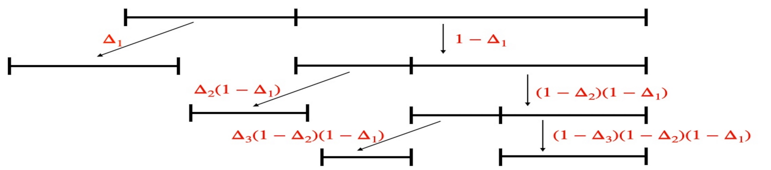
$\mathcal{H}_2 = 1$ $\quad$ $\mathcal{H}_2 = 0$

$Y = 2$ $\quad$ Generate $\mathcal{H}_3 \sim Bern(\Delta_3), \Delta_3 = \varphi(\mathbf{x}^T \boldsymbol{\beta}_3), \boldsymbol{\beta}_3 \sim N(\boldsymbol{\mu}_3, \Sigma_3)$

$\mathcal{H}_3 = 1$ $\quad$ $\mathcal{H}_3 = 0$

$Y = 3$ $\quad$ $\cdots$

- The stick-breaking weights are also determined by it.



## General modeling framework

- The general logit stick-breaking process (LSBP) model for ordinal regression:

$$\mathbf{Y}|G_{\mathbf{x}} \sim \int K(\mathbf{Y}|\mathbf{m}, \theta) dG_{\mathbf{x}}(\theta), \quad G_{\mathbf{x}} = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \delta_{\boldsymbol{\theta}_\ell(\mathbf{x})}$$

## General modeling framework

- The general logit stick-breaking process (LSBP) model for ordinal regression:

$$\mathbf{Y}|G_{\mathbf{x}} \sim \int K(\mathbf{Y}|\mathbf{m}, \boldsymbol{\theta})dG_{\mathbf{x}}(\boldsymbol{\theta}), \quad G_{\mathbf{x}} = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x})\delta_{\boldsymbol{\theta}_\ell(\mathbf{x})}$$

  - The weights $\omega_\ell(\mathbf{x})$ are generated by LSBP: $\omega_1(\mathbf{x}) = \varphi(\mathbf{x}^T\boldsymbol{\gamma}_1)$ and $\omega_\ell(\mathbf{x}) = \varphi(\mathbf{x}^T\boldsymbol{\gamma}_\ell)\prod_{h=1}^{\ell-1}(1-\varphi(\mathbf{x}^T\boldsymbol{\gamma}_h))$, $\ell = 2, 3, \cdots$;

# General modeling framework

- The general logit stick-breaking process (LSBP) model for ordinal regression:

$$\mathbf{Y}|G_{\mathbf{x}} \sim \int K(\mathbf{Y}|\mathbf{m}, \boldsymbol{\theta})dG_{\mathbf{x}}(\boldsymbol{\theta}), \quad G_{\mathbf{x}} = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x})\delta_{\boldsymbol{\theta}_\ell(\mathbf{x})}$$

  - The weights $\omega_\ell(\mathbf{x})$ are generated by LSBP: $\omega_1(\mathbf{x}) = \varphi(\mathbf{x}^T\boldsymbol{\gamma}_1)$ and $\omega_\ell(\mathbf{x}) = \varphi(\mathbf{x}^T\boldsymbol{\gamma}_\ell)\prod_{h=1}^{\ell-1}(1 - \varphi(\mathbf{x}^T\boldsymbol{\gamma}_h))$, $\ell = 2, 3, \cdots$;
  - The atoms, $\boldsymbol{\theta}_\ell(\mathbf{x}) = \{\theta_{j\ell}(\mathbf{x}) : j = 1, \cdots, C-1\}$, have linear regression structure, $\theta_{j\ell}(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}_{j\ell} \overset{ind.}{\sim} N(\mathbf{x}^T\boldsymbol{\mu}_j, \mathbf{x}^T\Sigma_j\mathbf{x})$, and to be independent across $\ell$;

## General modeling framework

- The general logit stick-breaking process (LSBP) model for ordinal regression:

$$\mathbf{Y}|G_{\mathbf{x}} \sim \int K(\mathbf{Y}|\mathbf{m}, \boldsymbol{\theta}) dG_{\mathbf{x}}(\boldsymbol{\theta}), \quad G_{\mathbf{x}} = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \delta_{\boldsymbol{\theta}_\ell(\mathbf{x})}$$

  - The weights $\omega_\ell(\mathbf{x})$ are generated by LSBP: $\omega_1(\mathbf{x}) = \varphi(\mathbf{x}^T \boldsymbol{\gamma}_1)$ and $\omega_\ell(\mathbf{x}) = \varphi(\mathbf{x}^T \boldsymbol{\gamma}_\ell) \prod_{h=1}^{\ell-1}(1 - \varphi(\mathbf{x}^T \boldsymbol{\gamma}_h))$, $\ell = 2, 3, \cdots$;
  - The atoms, $\boldsymbol{\theta}_\ell(\mathbf{x}) = \{\theta_{j\ell}(\mathbf{x}) : j = 1, \cdots, C-1\}$, have linear regression structure, $\theta_{j\ell}(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}_{j\ell} \overset{ind.}{\sim} N(\mathbf{x}^T \boldsymbol{\mu}_j, \mathbf{x}^T \Sigma_j \mathbf{x})$, and to be independent across $\ell$;
  - Hyperprior: $\boldsymbol{\gamma}_\ell \overset{i.i.d.}{\sim} N(\boldsymbol{\gamma}_0, \Gamma_0)$ and $\boldsymbol{\mu}_j|\Sigma_j \overset{ind.}{\sim} N(\boldsymbol{\mu}_j|\boldsymbol{\mu}_{0j}, \Sigma_j/\kappa_{0j})$, $\Sigma_j \overset{ind.}{\sim} IW(\Sigma_j|\nu_{0j}, \Lambda_{0j}^{-1})$.

# Simplified models

- **Common-weights**: defining the weights through the stick-breaking process that defines DP, we obtain the linear DDP model;
  - $G_{\mathbf{x}} = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\boldsymbol{\theta}_\ell(\mathbf{x})}$;
  - $\eta_\ell \overset{i.i.d.}{\sim} Beta(1, \alpha)$, $\omega_1 = \eta_1$ and $\omega_\ell = \eta_\ell \prod_{h=1}^{\ell-1}(1 - \eta_\ell)$, for $\ell = 2, 3, \cdots$;
  - $\boldsymbol{\theta}_\ell(\mathbf{x})$ are defined as in the general model;

# Simplified models

- **Common-weights**: defining the weights through the stick-breaking process that defines DP, we obtain the linear DDP model;
  - $G_{\mathbf{x}} = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\boldsymbol{\theta}_\ell(\mathbf{x})}$;
  - $\eta_\ell \overset{i.i.d.}{\sim} Beta(1, \alpha)$, $\omega_1 = \eta_1$ and $\omega_\ell = \eta_\ell \prod_{h=1}^{\ell-1}(1 - \eta_\ell)$, for $\ell = 2, 3, \cdots$;
  - $\boldsymbol{\theta}_\ell(\mathbf{x})$ are defined as in the general model;

- **Common-atoms**: we formulate the common-atoms LSBP model;
  - $G_{\mathbf{x}} = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \delta_{\boldsymbol{\theta}_\ell}$;
  - $\theta_{j\ell} | \mu_j, \sigma_j^2 \overset{ind.}{\sim} N(\mu_j, \sigma_j^2)$;
  - $\omega_\ell(\mathbf{x})$ are determined as in the general model.

# Flexible ordinal regression relationships

## Model properties

The general LSBP model allows flexible estimate of the probability response curves.

1. Marginal regression relationships:

$$Pr(\mathbf{Y} = j | G_{\mathbf{x}}) = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \left\{ \varphi(\theta_{j\ell}(\mathbf{x})) \prod_{k=1}^{j-1} [1 - \varphi(\theta_{k\ell}(\mathbf{x}))] \right\}$$

2. Conditional regression relationships:

$$Pr(\mathbf{Y} = j | \mathbf{Y} \geq j, G_{\mathbf{x}}) = \sum_{\ell=1}^{\infty} w_{j\ell}(\mathbf{x}) \left\{ \varphi(\theta_{j\ell}(\mathbf{x})) \right\}$$

where $w_{j\ell}(\mathbf{x}) = \frac{\omega_\ell(\mathbf{x}) \prod_{k=1}^{j-1} [1 - \varphi(\theta_{k\ell}(\mathbf{x}))]}{\sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{x}) \prod_{k=1}^{j-1} [1 - \varphi(\theta_{k\ell}(\mathbf{x}))]}$.

- Both have a weighted sum form with locally adjustable weights.

# Model implementation

- **Prior specification:**
  - Conjugate hyperprior, $\boldsymbol{\gamma}_\ell \overset{i.i.d.}{\sim} N(\boldsymbol{\gamma}_0, \Gamma_0)$, $\boldsymbol{\mu}_j|\Sigma_j \overset{ind.}{\sim} N(\boldsymbol{\mu}_{0j}, \Sigma_j/\kappa_{0j})$, $\Sigma_j \overset{ind.}{\sim} IW(\nu_{0j}, \Lambda_{0j}^{-1})$.
  - "Baseline choice" of hyperparameter: $\boldsymbol{\mu}_{0j} = \gamma_0 = \mathbf{0}_p$, $\Sigma_j = \Gamma_0 = 10^2\mathbf{I}_p$, and $\kappa_{0j} = \nu_{0j} = p + 2$;
  - Under the baseline prior, $E(Pr(\mathbf{Y} = j|G_{\mathbf{x}})) \equiv 2^{-j}$, $j = 1, \cdots, C-1$, and $E(Pr(\mathbf{Y} = C|G_{\mathbf{x}})) \equiv 2^{-(C-1)}$.

# Model implementation

- **Prior specification:**
  - Conjugate hyperprior, $\gamma_\ell \overset{i.i.d.}{\sim} N(\gamma_0, \Gamma_0)$, $\mu_j | \Sigma_j \overset{ind.}{\sim} N(\mu_{0j}, \Sigma_j / \kappa_{0j})$, $\Sigma_j \overset{ind.}{\sim} IW(\nu_{0j}, \Lambda_{0j}^{-1})$.
  - "Baseline choice" of hyperparameter: $\mu_{0j} = \gamma_0 = \mathbf{0}_p$, $\Sigma_j = \Gamma_0 = 10^2 \mathbf{I}_p$, and $\kappa_{0j} = \nu_{0j} = p + 2$;
  - Under the baseline prior, $E(Pr(\mathbf{Y} = j | G_\mathbf{x})) \equiv 2^{-j}$, $j = 1, \cdots, C-1$, and $E(Pr(\mathbf{Y} = C | G_\mathbf{x})) \equiv 2^{-(C-1)}$.

- **Posterior inference:**
  - Blocked Gibbs sampler; Truncating $G_\mathbf{x}$ at a large enough level $L$ and introducing latent configuration variable $\mathcal{L}_i$ for $i = 1, \cdots, n$;
  - Pólya-Gamma augmentation; Introduce two groups of Pólya-Gamma latent variables for the weights and atoms;
  - Same structure for the weights and atoms; Same sampling strategy;
  - All model parameters can be sampled via Gibbs sampling.

# Data Examples

# Synthetic data examples

- In both experiments, $n$ pairs of ordinal response and covariate $(\mathbf{Y}_i, x_i)$ are generated, where $x_i \overset{i.i.d.}{\sim} Unif(x_i | -10, 10)$ such that with the intercept, the covariate vector is $\mathbf{x}_i = (1, x_i)^T$;

- **First experiment**: We generate $n = 100$ responses by first sampling a latent continuous variable $\tilde{y}_i$ from normal distribution, then discretizing $\tilde{y}_i$ with cut-off points to get the ordinal response $\mathbf{Y}_i$;

- **Second experiment**:
  - We generate data from $\mathbf{Y} \sim \sum_{k=1}^{3} \omega_k(\mathbf{x}) K(\mathbf{Y}|\mathbf{m}, \boldsymbol{\theta}_k(\mathbf{x}))$;
  - The true probability response curves have nonstandard shape.
  - Perform the experiment with 200 simulated data and with 800 simulated data.

# First experiment result

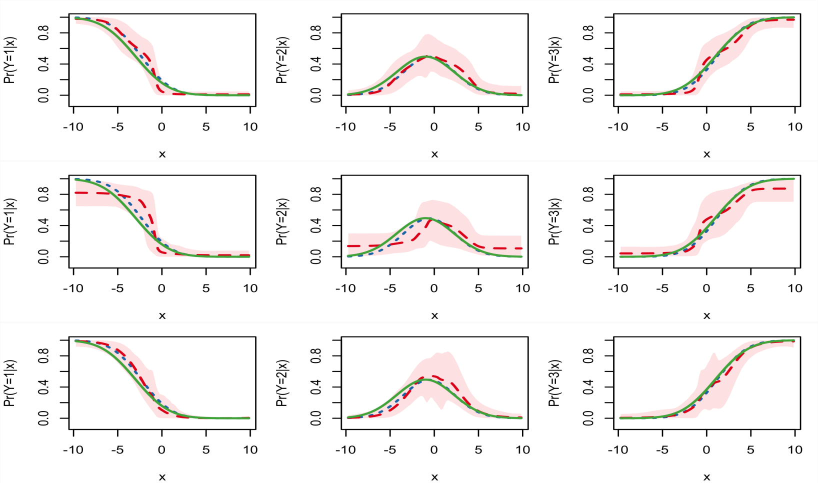



**Figure 1:** Green solid lines: true response probabilities. Red dashed lines and shadow area: nonparametric model; Blue dotted lines: probit regression model.

# Second experiment result

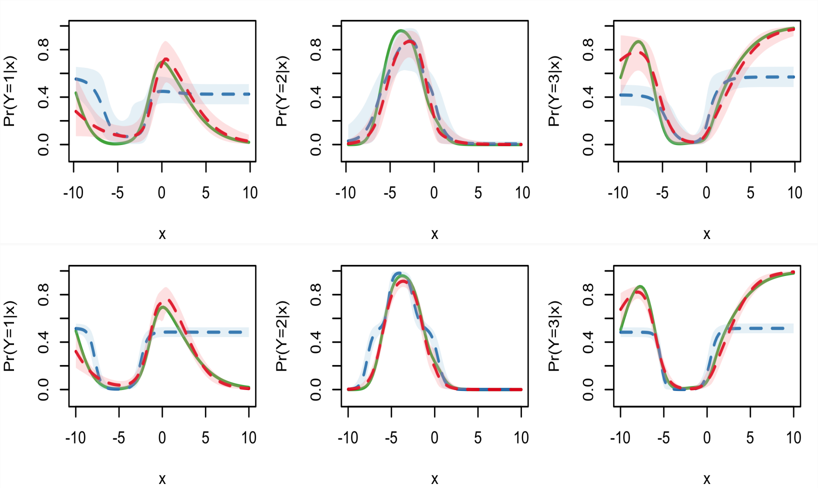



**Figure 2:** Green solid lines: the true response probabilities; Red line and shadow area: the general LSBP model; Blue line and shadow area: the linear DDP model.

# How it works?

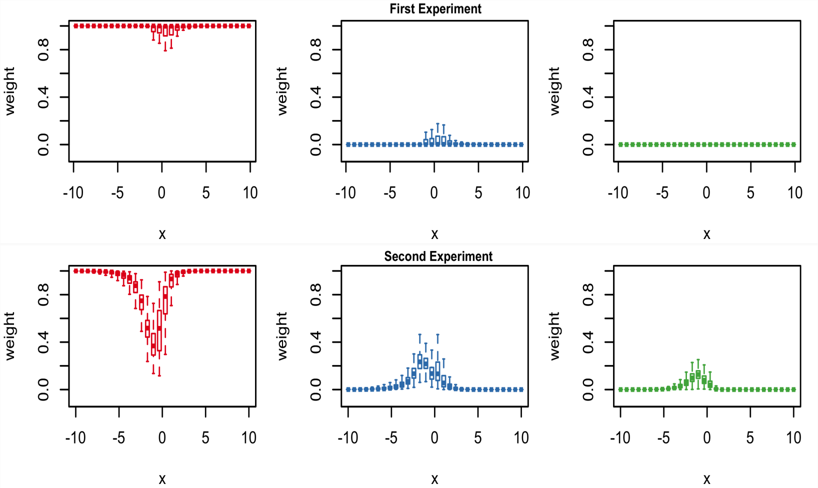



**Figure 3:** Box plots for posterior samples of the three largest mixing weights (order decreases from left to the right) under the general model.

# Real Data Application (Credit Ratings of U.S. Companies)

- Standard and Poor's (S&P) credit ratings for 921 U.S. firms;
- For each firm, a credit rating on a seven-point ordinal scale is available, along with five characteristics;
- Combined the first two categories and the last two categories to produce an ordinal response with five levels;
- The covariates are: (1) book leverage $X_1$, (2) earnings before interest and taxes divided by total assets $X_2$, (3) standardized log-sales $X_3$, (4) retained earnings divided by total assets $X_4$, (5) working capital divided by total assets $X_5$;
- Quantities of interest: the first and second order marginal probability curves $Pr(\mathbf{Y} = j | G_{\mathbf{x}}; \mathbf{x}_{\mathbf{s}})$ for $j = 1, \cdots, 5$ and $\mathbf{s} \in \{1, \cdots, 5\}$.

# First order marginal probability curves

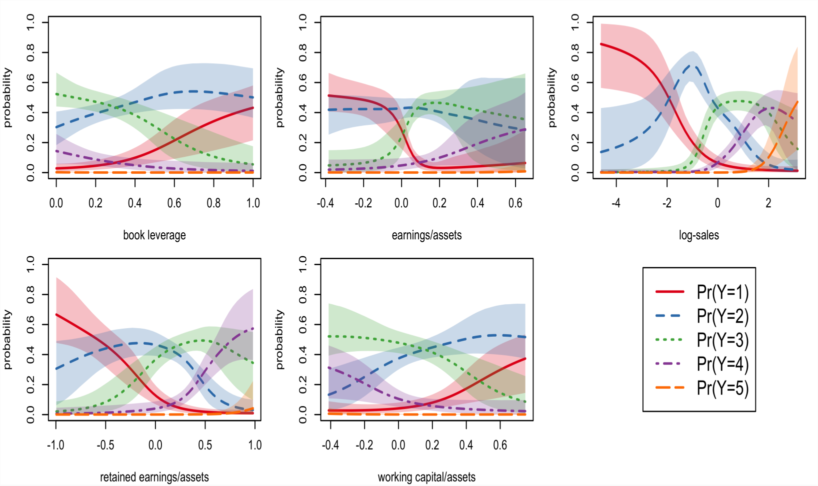



**Figure 4:** First order marginal probability curves. All five ordinal response curves are displayed in a single panel for each covariate.

# Second order marginal probability surfaces

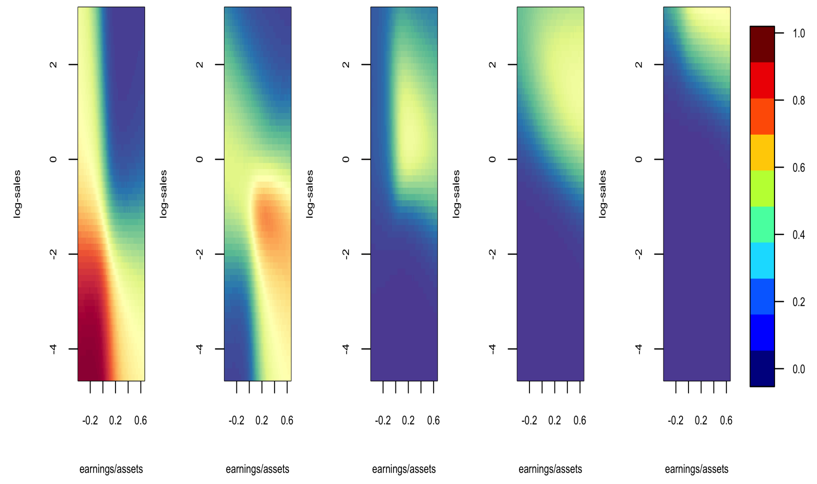



**Figure 5:** Second order marginal probability curves. The corresponding credit rating decreases from left to right.

# Developmental toxicology data (DYME)

- Data structure for Segment II designs (exposure after implantation):
  - Data at dose levels, $x_i$, $i = 1, \cdots, N$, including a control group (dose$= 0$);
  - $n_i$ pregnant laboratory animals (dams) at dose level $x_i$;
  - For the $j$-th dam at dose $x_i$:
    - $m_{ij}$: number of implants;
    - $R_{ij}$: number of resorptions and prenatal deaths ($R_{ij} \leq m_{ij}$);
    - $y_{ij}$: number of live pups with a malformation.

- Focus on the dose-response curves on the clustered categorical endpoints, embryolethality $R_{ij}$, fetal malformation for live pups $y_{ij}$, and combined negative outcomes among implants $R_{ij} + y_{ij}$.

# Estimation of dose-response curves

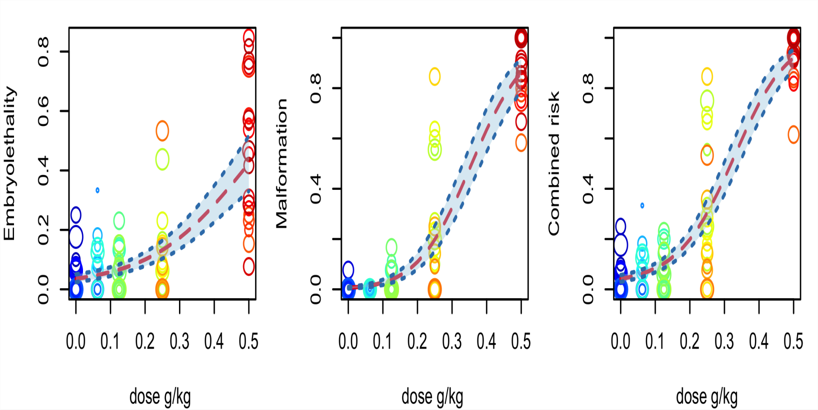



**Figure 6:** DYME data. The coordinates of the circle are the dose level and the proportion of the specific endpoint: non-viable fetuses among implants (left panel); malformations among live pups (middle panel); combined negative outcomes among implants (right panel).

# Concluding Remarks

# Summary and Discussion

- We propose a unified toolbox for ordinal regression by directly modeling the discrete response distribution. The virtues of the proposed models rely on the following key ingredients:
  1. Continuation-ratio logits representation;
  2. Pólya-Gamma data augmentation technique;
  3. Logit stick-breaking process prior.
- The principal advantages of the proposed models include their modularity and extensibility:
  1. Modularity: the models can be readily embedded in a more complex framework;
  2. Extensibility: there are numerous possible extensions of the proposed models. For example, replacing the binomial distribution in the mixing kernel with logistic-normal-binomial distributions to allow more overdispersion.

# MANY THANKS!

## I am happy to answer any questions.

**Jizhou Kang and Athanasios Kottas (2022), "Structured mixture of continuation-ratio logits models for ordinal regression", submitted, arXiv: 1234.56789.**