

スライドのダウンロードはこちらから



<https://okuno.net/slides>

An interpretable neural network-based
non-proportional odds model for ordinal regression.

Accepted to JCGS

(<https://doi.org/10.1080/10618600.2024.2321208>)


Akifumi Okuno^{1,2} and Kazuharu Harada^{3,1}

¹ISM ²RIKEN AIP ³Tokyo Medical University

Abstract

ニューラルネットを利用して、線形の係数に変化する非比例オッズモデルを定義した。さらに予測モデル全体が単調性を満たすための十分条件を示した。和文での解説：<https://doi.org/10.51094/jxiv.549>

自己紹介

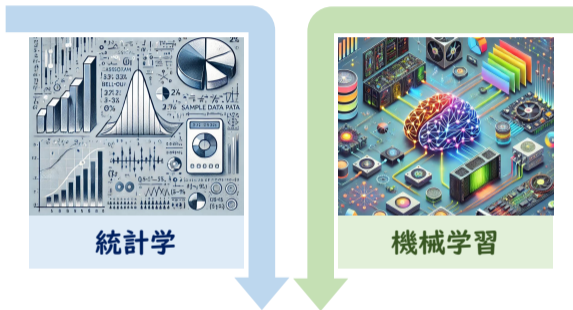
 オクノ アキフミ
奥野彰文 博士 (情報学, 京都大学)

<https://okuno.net>



- ▶ 専門：数理統計，統計的機械学習 (特に手法開発とその理論解析).
- ▶ 最近の興味：科学応用を含めた異分野連携研究.
- ▶ 学生時代の指導教員：下平英寿教授 (<https://stat.sys.i.kyoto-u.ac.jp/>)

- ▶ 基本的に何でもやりますが、統計学 × 機械学習がメインテーマです。



小サンプルでもきちんと動く & 機械学習の便利ツールに相乗りができる。
統計学と機械学習手法研究の良い所を組み合わせたいと思っています。

Problem: CCP estimation

Purpose of this study

is to model the *conditional cumulative probability (CCP)*:

$$\mathbb{P}(H \leq u \mid X = \mathbf{x})$$

(or equivalently, $\mathbb{P}(H > u \mid X = \mathbf{x})$).

- ▶ Example (real-estate):

$$\mathbb{P}(\text{House-price} \leq 100,000[\text{dollars}] \mid \text{House-age} = 10[\text{years}]) = 0.7$$

- ▶ **Monotonicity** is required:

$$\mathbb{P}(\text{H.-price} \leq 10,000 \mid \text{H.-age} = 10) \leq \mathbb{P}(\text{H.-price} \leq 20,000 \mid \text{H.-age} = 10)$$

- ▶ We want to interpret the interaction between X and H .

Outline

1 Background

- Ordinal Discrete Response
- Proportional Odds Model (POM)
- Non-Proportional Odds Model (NPOM)
- Difficulties in NPOM

2 Proposal

- Neural Network-based NPOM (N³POM)
- Monotonicity of N³POM
- Monotonicity Preserving Stochastic (MPS) Algorithm

3 Numerical Experiments

- Synthetic Dataset Experiments
- Real-world Dataset Experiments

4 Conclusion

Background

Ordinal Discrete Response

$G \in \{1, 2, \dots, J\}$: ordinal discrete response associated with the covariate $X \in \mathbb{R}^d$.

Dataset	Response G	Covariate X
car	Rating $\{1, 2, \dots, 5\}$	maintenance, #doors, #persons, ...
heart-disease	Diagnosis $\{0, 1, \dots, 4\}$	Age, Chol, Ca, Thal, ...
wine-quality	Quality $\{1, 2, \dots, 9\}$	acidity, sugar, density, pH, alcohol, ...
shogi, chess, ...	Rank $\{1, 2, \dots, 9\}$?????

- ▶ Typically, $J \leq 10$ is considered (not so much!).

Conventional: *Proportional Odds Model (POM)*

$$\mathbb{P}_{\text{POM}}(G \leq j \mid X = \mathbf{x}) = \sigma(\alpha_j + \langle \boldsymbol{\beta}, \mathbf{x} \rangle), \quad \theta = (\{\alpha_j\}_j, \boldsymbol{\beta}).$$

$$\ell_{\text{POM}}(\theta) = \sum_{i=1}^n \log \underbrace{\{\mathbb{P}_{\text{POM}}(G \leq g_i \mid X = \mathbf{x}_i) - \mathbb{P}_{\text{POM}}(G \leq g_i - 1 \mid X = \mathbf{x}_i)\}}_{=\mathbb{P}_{\text{POM}}(G=g_i|X=\mathbf{x}_i)}$$

👍 Monotone! (as long as $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_J$).

- ▶ $\alpha_{j+1} = \alpha_j + \exp(\gamma_j)$,
- ▶ $\alpha_{j+1} = \alpha_j + \gamma_j^2$,
- ▶ $\alpha_{j+1} = \alpha_j + |\gamma_j|, \dots$ can be implemented easily.

👎 $\boldsymbol{\beta}$ is constant for $j = 1, 2, \dots, J$.

Is β really constant?

Impact of each covariate may be different depending on the degree of response:

- ▶ In restaurant-ratings,
 - ▶ good hygiene (cleanliness) seems more important for expensive restaurants,
 - ▶ good wine-selection seems more important for expensive restaurants, ...
- ▶ In house-price evaluation,
 - ▶ distance to convenience store seems less important for expensive houses, etc...

Coefficients β **seems not constant** for some practical cases.
(see Long and Freese (2006); Williams (2016))

Conventional: *Non-Proportional Odds Model (NPOM)*

$$\mathbb{P}_{\text{NPOM}}(G \leq j \mid X = \mathbf{x}) = \sigma(\alpha_j + \langle \boldsymbol{\beta}_j, \mathbf{x} \rangle), \quad \theta = (\{\alpha_j\}_j, \{\boldsymbol{\beta}_j\}_j).$$

👍 Coefficients vary across the threshold $j = 1, 2, \dots, J$.

👎 Three difficulties:

(D-1) monotonicity is not guaranteed (i.e., $\mathbb{P}_{\text{NPOM}}(G \leq j \mid \mathbf{x}) \not\leq \mathbb{P}_{\text{NPOM}}(G \leq j + 1 \mid \mathbf{x})$),

(D-2) adjacent proximity is not guaranteed (i.e., $\boldsymbol{\beta}_j \not\approx \boldsymbol{\beta}_{j+1}$),

(D-3) cannot consider continuous response.

Several approaches are proposed to address each difficulty.

Solution to (D-1) monotonicity

NPOM is expected to be monotone, if NPOM gets closer to POM (as POM is monotone).

- ▶ Define $\boldsymbol{\beta}_j = \boldsymbol{\beta} + \boldsymbol{\delta}_j$ and penalize $\lambda \|\boldsymbol{\delta}_j\|$. ($\lambda \rightarrow \infty$ indicates $\boldsymbol{\beta}_j \rightarrow \boldsymbol{\beta}$ (POM)).

(See, e.g., Wurm et al. (2021), Lu et al. (2022), Tutz and Berger (2022))

- 🗨 Biased if the coefficient is λ is large.
- 🗨 Not monotone if the coefficient is λ is small.
- 🗨 No solution to choose λ .

Solution to (D-2) adjacent proximity

Fused penalty.

- ▶ $\sum_{j=1}^{J-1} \lambda_j \|\boldsymbol{\beta}_{j+1} - \boldsymbol{\beta}_j\|_2^2$, (see, e.g., Tutz and Gertheiss (2016), Ugba et al. (2021)).

- 🗨 Cannot be simply incorporated into the monotone models.

Solution to (D-3) continuous response

Continuous extension of NPOM.

Satoh et al. (2016) proposed a continuous model:

$$\mathbb{P}_{\text{Satoh}}(H \leq u \mid X = \mathbf{x}) = \sigma(a_{\text{Satoh}}(u) + \langle \mathbf{b}_{\text{Satoh}}(u), \mathbf{x} \rangle)$$

with polynomial functions $a_{\text{Satoh}}(u)$, $\mathbf{b}_{\text{Satoh}}(u)$.

👎 Functions a_{Satoh} , $\mathbf{b}_{\text{Satoh}}$ are restrictive.

👎 Monotonicity is not guaranteed.

Baumann et al. (2021) and Kook et al. (2022b) consider a model

$$\mathbb{P}_{\text{DCTM}}(H \leq u \mid X = \mathbf{x}) = \sigma(\langle \mathbf{b}_{\text{DCTM}}(u), \mathbf{c}_{\text{DCTM}}(\mathbf{x}) \rangle + e_{\text{DCTM}}(\mathbf{x})),$$

with the *Bernstein basis* expansion (Farouki, 2012) for $\mathbf{b}_{\text{DCTM}}(u)$ and estimate $\mathbf{c}_{\text{DCTM}}(\mathbf{x})$ by *deep NN*.

- 👍 Monotone if $\mathbf{c}_{\text{DCTM}}(\mathbf{x})$ is suitably estimated.
- 👎 Not compatible with our setting, as we fix $\mathbf{c}_{\text{DCTM}}(\mathbf{x}) = \mathbf{x}$ for interpretability.
 - ▶ If $\mathbf{c}_{\text{DCTM}}(\mathbf{x}) = \mathbf{x}$, the DCTM model cannot be monotone for both \mathbf{x} , $-\mathbf{x}$ simultaneously.

We appreciate Dr. Kook for reference information and kind comments!

Proposal

To address these difficulties (D-1)–(D-3) simultaneously, we propose:

 *Neural Network-based NPOM (N³POM)*

$$\mathbb{P}_{\text{N}^3\text{POM}}(H \leq u \mid X = \mathbf{x}) = \sigma(a(u) + \underbrace{\langle \mathbf{b}(u), \mathbf{x} \rangle}_{\text{NN}}), \quad (u \in [1, J]).$$

- ▶ $a(u)$: piece-wise linear increasing function.
- ▶ $\mathbf{b}(u)$: neural network (one-hidden-layer perceptron).

 Functions a , \mathbf{b} are expressive enough!

 Interpretable! (by considering $\mathbf{b}(u)$)

 Monotone? ( yes, we provide a sufficient condition!)

Details

- ▶ Piece-wise linear increasing function (parameterized by $\phi, \{\varphi_r\}$):

$$a(u) = \phi + \sum_{r=2}^R |\varphi_r| \left[\frac{u - j_{r-1}}{j_r - j_{r-1}} \right], \text{ where } \llbracket z \rrbracket = \begin{cases} 0 & (z < 0) \\ z & (z \in [0, 1]) \\ 1 & (z > 1) \end{cases}.$$

- ▶ Neural network $\mathbf{b}(u) = (b_1(u), b_2(u), \dots, b_J(u))$ (parameterized by $\{w_{k,\ell}^{(2)}, w_{k,\ell}^{(1)}, v_k^{(2)}, v_{k,\ell}^{(1)}\}$):

$$b_k(u) = \sum_{\ell=1}^L w_{k,\ell}^{(2)} \sigma(w_{k,\ell}^{(1)} u + v_{k,\ell}^{(1)}) + v_k^{(2)}.$$

Ordinal Continuous Response

$H \in [1, J]$: ordinal continuous response associated with the covariate $X \in \mathbb{R}^d$.

Dataset	Response H	Covariate X
fuel efficiency	miles per gallon	horsepower, weight, acceleration, ...
housing price	dollars per unit area	houseage, distance to station, ...
cement property	compressive strength	cement, water, age, ...
survival analysis	lifespan	?????
Tabe-log (食べログ)	score [1, 5]	?????

- ▶ Also, discrete response with large J can be regarded as (pseudo-)continuous.

Monotonicity

Proposition 1:

$\mathbb{P}_{\text{N}^3\text{POM}}(H \leq u \mid X = \mathbf{x})$ is monotone for any fixed $\mathbf{x} \in \mathbb{R}^d \Rightarrow \mathbf{b}(u)$ is constant.

This prop. clarifies that **non-proportional models cannot be monotone for entire \mathbb{R}^d** .

We instead consider the monotonicity over a restricted region

$$\mathcal{X}_2(\eta) := \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 \leq \eta\} \quad (\subset \mathbb{R}^d).$$

- ▶ For instance, house-price is expected to be no more than $\$10^{10}$.
- ▶ Typically, $\eta := \max_j \|\mathbf{x}_j\|_2$.

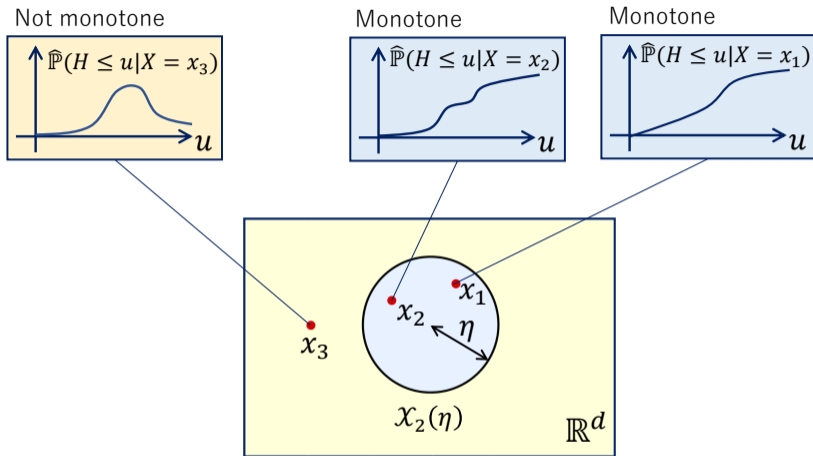


Figure: Our strategy is to guarantee the monotonicity for any fixed $\mathbf{x} \in \mathcal{X}_2(\eta)$.

With $\rho_\infty^{[1]} := \sup_{z \in \mathbb{R}} |\rho'(z)|$, consider an inequality

$$(\star) \quad \text{minimum slope of } a(u) \geq \eta \cdot \rho_\infty^{[1]} \cdot \underbrace{\sqrt{\sum_{k=1}^d \left\{ \sum_{\ell=1}^L |w_{k,\ell}^{(2)} w_{k,\ell}^{(1)}| \right\}^2}}_{\text{neural network weights}}.$$

Proposition 2:

If (\star) is satisfied, $\mathbb{P}_{\text{N}^3\text{POM}}(H \leq u \mid X = \mathbf{x})$ is **monotone** for any fixed $\mathbf{x} \in \mathcal{X}_2(\eta)$.

- ▶ There exists a **trade-off** between η (area of $\mathcal{X}_2(\eta)$) and the fluctuation of the NN \mathbf{b} . (see next page!)

Interpretation of the condition (★)

$$(\star) \quad \text{minimum slope of } a(u) \geq \eta \cdot \rho_{\infty}^{[1]} \cdot \underbrace{\sqrt{\sum_{k=1}^d \left\{ \sum_{\ell=1}^L |w_{k,\ell}^{(2)} w_{k,\ell}^{(1)}| \right\}^2}}_{\text{neural network weights}}.$$

- ▶ Small η (small covariate region) \Rightarrow NN weights $w_{k,\ell}^{(1)}, w_{k,\ell}^{(2)}$ can be large.
- ▶ Large η (large covariate region) \Rightarrow Small NN weights $w_{k,\ell}^{(1)}, w_{k,\ell}^{(2)}$
- ▶ $\eta \rightarrow \infty$ ($\mathcal{X}(\infty) = \mathbb{R}^d$) \Rightarrow Constant NN (i.e., $w_{k,\ell}^{(1)} w_{k,\ell}^{(2)} \rightarrow 0$) compatible with Proposition 1.

(Conventional) Minibatch Gradient Ascent Algorithm

- ▶ Log-likelihood to be maximized:

$$\ell(\theta) := \frac{1}{n} \sum_{i=1}^n \left\{ \log \sigma^{[1]}(a(h_i) + \langle \mathbf{b}(h_i), \mathbf{x}_i \rangle) + \log(a^{[1]}(h_i) + \langle \mathbf{b}^{[1]}(h_i), \mathbf{x}_i \rangle) \right\},$$

- ▶ Initialization: $\theta^{(0)}$.
- ▶ For each iteration $t = 1, 2, \dots$,
 - ▶ Uniformly resample minibatch $I^{(t)}$ of size N from $\{1, 2, \dots, n\}$.
 - ▶ Minibatch Gradient Ascent: $\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \nabla \ell^{(t)}(\theta^{(t)})$, with

$$\ell^{(t)}(\theta) = \frac{1}{M} \sum_{i \in I^{(t)}} \left\{ \log \sigma^{[1]}(a(h_i) + \langle \mathbf{b}(h_i), \mathbf{x}_i \rangle) + \log(a^{[1]}(h_i) + \langle \mathbf{b}^{[1]}(h_i), \mathbf{x}_i \rangle) \right\}$$

(satisfying $\mathbb{E}_{I^{(t)}}[\ell^{(t)}(\theta)] = \ell(\theta)$).

- ▶ However, monotonicity is not guaranteed.

Monotonicity Preserving Stochastic (MPS) Algorithm

Repeat the following steps until convergence:

- (1) Compute a single step of minibatch gradient ascent algorithm.
- (2) With the coefficient

$$c := \min \left\{ 1, \frac{\text{minimum slope of } a(u)}{\eta \cdot \rho_{\infty}^{[1]} \cdot \sqrt{\sum_{k=1}^d \left\{ \sum_{\ell=1}^L |w_{k,\ell}^{(2)} w_{k,\ell}^{(1)}| \right\}^2}} \right\},$$

replace $w_{k,\ell}^{(2)}, w_{k,\ell}^{(1)}, v_{k,\ell}^{(1)}$ by $\sqrt{c}w_{k,\ell}^{(2)}, \sqrt{c}w_{k,\ell}^{(1)}, \sqrt{c}v_{k,\ell}^{(1)}$, to satisfy (★).

 Monotonicity is guaranteed!

Summary: Training N³POM

Input: observations $\{(h_i, \mathbf{x}_i)\}_{i=1}^n$ and optimization parameters.

- (1) Round $[h_i] \in \{1, 2, \dots, J\}$.
- (2) Apply existing NPOM¹ to $([h_i], \mathbf{x}_i)$ and obtain discrete coefficients $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_J$.
- (3) (**NN Initialization**) Distillate the coefficients so that $\mathbf{b}(j) \approx \boldsymbol{\beta}_j$ ($j = 1, 2, \dots, J$).
- (4) (**NN Training**) compute MPS algorithm.

Output: $\overset{\color{red}\curvearrowright}{a}(u), \overset{\color{red}\curvearrowright}{\mathbf{b}}(u)$.

¹*serp* package, that penalizes $\|\boldsymbol{\beta}_{j+1} - \boldsymbol{\beta}_j\|_2^2$ (Ugba et al., 2021).

NN Initialization (1/2)

► How to distill the NPOM coefficients $\hat{\beta}_{jk}$ for initialization?

we employ an NN using sigmoid activation function $\rho(z) = 1/(1 + \exp(-z))$ and $L > d$; with a sufficiently large constant T (e.g., $T = 10$, satisfying $\rho(-T) \approx 0, \rho(+T) \approx 1$), we define

$$v_k^{(2)} = \frac{1}{J} \sum_{j=1}^J \hat{\beta}_{jk},$$

$$v_{k,\ell}^{(1)} = \begin{cases} -T\ell & (\ell \in \{1, 2, \dots, J\}) \\ 0 & (\text{Otherwise}) \end{cases},$$

$$w_{k,\ell}^{(1)} = \begin{cases} T & (\ell \in \{1, 2, \dots, J\}) \\ 0 & (\text{Otherwise}) \end{cases},$$

$$w_{k,\ell}^{(2)} = \begin{cases} \frac{\hat{\beta}_{\ell k} - v_k^{(2)} - \sum_{\ell B=1}^{\ell-1} w_{k,\ell B}^{(2)}}{\rho(0)} & (\ell \in \{1, 2, \dots, J\}), \quad \forall k \in \{1, 2, \dots, d\}. \\ 0 & (\text{Otherwise}) \end{cases}$$

NN Initialization (2/2)

$$\begin{aligned} b_k(j) &= \sum_{\ell=1}^L w_{k,\ell}^{(2)} \rho \left(w_{k,\ell}^{(1)} j + v_{k,\ell}^{(1)} \right) + v_k^{(2)} \\ &= \sum_{\ell=1}^J w_{k,\ell}^{(2)} \rho(T(j-\ell)) + v_k^{(2)} \\ &\approx \sum_{\ell=1}^J w_{k,\ell}^{(2)} \{ \rho(-\infty) \mathbb{1}(j < \ell) + \rho(0) \mathbb{1}(\ell = j) + \rho(\infty) \mathbb{1}(\ell < j) \} + v_k^{(2)} \\ &= w_{k,j}^{(2)} \rho(0) + \sum_{\ell=1}^{j-1} w_{k,\ell}^{(2)} + v_k^{(2)} \\ &= \frac{\overset{\text{red}}{\beta}_{jk} - v_k^{(2)} - \sum_{\ell=1}^{j-1} w_{k,\ell}^{(2)}}{\rho(0)} \rho(0) + \sum_{\ell=1}^{j-1} w_{k,\ell}^{(2)} + v_k^{(2)} = \overset{\text{red}}{\beta}_{jk}. \end{aligned}$$

Adaptation to the discrete responses

- ▶ Discrete responses $\{g_i\}$ are not sufficient to fully train the continuous model $f_u(\mathbf{x})$.
- ▶ With $e_1, e_2, \dots, e_n \sim U[-0.5, 0.5]$, we define a random perturbation operator:

$$\mathfrak{C}(g_i) := \arg \min_{j \in [1, J]} |(g_i + e_i) - j|, \quad (i \in \{1, 2, \dots, n\}).$$

- ▶ Although heuristic, this perturbation operator is explainable in the context of OLS regression:
the estimated regression function in OLS converges to the conditional expectation $f_*(X) = \mathbb{E}[G | X] = \mathbb{E}[\mathfrak{C}(G) | X]$; it does not cause any bias.
- ▶ Numerically better for NN training.

Weak derivatives (1/4)

For $r^\dagger \in \{2, 3, \dots, R\}$, $k^\dagger \in [d]$, $\ell^\dagger \in [L]$ and $u \in \mathcal{U}$, we have

$$\frac{\partial}{\partial \phi} f_u(\mathbf{x}_i) = 1,$$

$$\frac{\partial}{\partial \varphi_{r^\dagger}} f_u(\mathbf{x}_i) = \text{sign}(\varphi_{r^\dagger}) \left\lfloor \frac{u - j_{r^\dagger-1}}{j_{r^\dagger} - j_{r^\dagger-1}} \right\rfloor,$$

$$\frac{\partial}{\partial v_{k^\dagger, \ell^\dagger}^{(1)}} f_u(\mathbf{x}_i) = x_{ik^\dagger} \frac{\partial}{\partial v_{k^\dagger, \ell^\dagger}^{(1)}} b_{k^\dagger}(u) = x_{ik^\dagger} w_{k^\dagger, \ell^\dagger}^{(2)} \rho^{[1]}(w_{k^\dagger, \ell^\dagger}^{(1)} u + v_{k^\dagger, \ell^\dagger}^{(1)}),$$

$$\frac{\partial}{\partial v_{k^\dagger}^{(2)}} f_u(\mathbf{x}_i) = x_{ik^\dagger} \frac{\partial}{\partial v_{k^\dagger, \ell^\dagger}^{(2)}} b_{k^\dagger}(u) = x_{ik^\dagger},$$

$$\frac{\partial}{\partial w_{k^\dagger, \ell^\dagger}^{(1)}} f_u(\mathbf{x}_i) = x_{ik^\dagger} \frac{\partial}{\partial w_{k^\dagger, \ell^\dagger}^{(1)}} b_{k^\dagger}(u) = x_{ik^\dagger} w_{k^\dagger, \ell^\dagger}^{(2)} u \rho^{[1]}(w_{k^\dagger, \ell^\dagger}^{(1)} u + v_{k^\dagger, \ell^\dagger}^{(1)}),$$

$$\frac{\partial}{\partial w_{k^\dagger, \ell^\dagger}^{(2)}} f_u(\mathbf{x}_i) = x_{ik^\dagger} \frac{\partial}{\partial w_{k^\dagger, \ell^\dagger}^{(2)}} b_{k^\dagger}(u) = x_{ik^\dagger} \rho(w_{k^\dagger, \ell^\dagger}^{(1)} u + v_{k^\dagger, \ell^\dagger}^{(1)}).$$

Weak derivatives (2/4)

$$\frac{\partial}{\partial \phi} \ell_{\zeta}(\theta) = \sum_{i=1}^n \zeta_i \frac{\sigma^{[2]}(f_{h_i}(\mathbf{x}_i))}{\sigma^{[1]}(f_{h_i}(\mathbf{x}_i))},$$

$$\begin{aligned} \frac{\partial}{\partial \varphi_{r^{\dagger}}} \ell_{\zeta}(\theta) &= \text{sign}(\varphi_{r^{\dagger}}) \sum_{i=1}^n \zeta_i \frac{\sigma^{[2]}(f_{h_i}(\mathbf{x}_i))}{\sigma^{[1]}(f_{h_i}(\mathbf{x}_i))} \left\lfloor \frac{h_i - j_{r^{\dagger}-1}}{j_{r^{\dagger}} - j_{r^{\dagger}-1}} \right\rfloor \\ &\quad + \frac{\text{sign}(\varphi_{r^{\dagger}})}{j_{r^{\dagger}} - j_{r^{\dagger}-1}} \sum_{i=1}^n \zeta_i \mathbb{1}(h_i \in \mathcal{U}_{r^{\dagger}-1}) \frac{1}{f_{h_i}^{[1]}(\mathbf{x}_i)}, \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial v_{k^{\dagger}, \ell^{\dagger}}^{(1)}} \ell_{\zeta}(\theta) &= \sum_{i=1}^n \zeta_i \frac{\sigma^{[2]}(f_{h_i}(\mathbf{x}_i))}{\sigma^{[1]}(f_{h_i}(\mathbf{x}_i))} x_{ik^{\dagger}} w_{k^{\dagger}, \ell^{\dagger}}^{(2)} \rho^{[1]}(w_{k^{\dagger}, \ell^{\dagger}}^{(1)} h_i + v_{k^{\dagger}, \ell^{\dagger}}^{(1)}) \\ &\quad + \sum_{i=1}^n \zeta_i \frac{1}{f_{h_i}^{[1]}(\mathbf{x}_i)} x_{ik^{\dagger}} w_{k^{\dagger}, \ell^{\dagger}}^{(2)} w_{k^{\dagger}, \ell^{\dagger}}^{(1)} \rho^{[2]}(w_{k^{\dagger}, \ell^{\dagger}}^{(1)} h_i + v_{k^{\dagger}, \ell^{\dagger}}^{(1)}), \end{aligned}$$

Weak derivatives (3/4)

$$\frac{\partial}{\partial v_{k^\dagger}^{(2)}} \ell_\zeta(\boldsymbol{\theta}) = \sum_{i=1}^n \zeta_i \frac{\sigma^{[2]}(f_{h_i}(\mathbf{x}_i))}{\sigma^{[1]}(f_{h_i}(\mathbf{x}_i))} x_{ik^\dagger},$$

$$\begin{aligned} \frac{\partial}{\partial w_{k^\dagger, \ell^\dagger}^{(1)}} \ell_\zeta(\boldsymbol{\theta}) = & \sum_{i=1}^n \zeta_i \frac{\sigma^{[2]}(f_{h_i}(\mathbf{x}_i))}{\sigma^{[1]}(f_{h_i}(\mathbf{x}_i))} x_{ik^\dagger} w_{k^\dagger, \ell^\dagger}^{(2)} h_i \rho^{[1]}(w_{k^\dagger, \ell^\dagger}^{(1)} h_i + v_{k^\dagger, \ell^\dagger}^{(1)}) \\ & + \sum_{i=1}^n \zeta_i \frac{1}{f_{h_i}^{[1]}(\mathbf{x}_i)} \left\{ x_{ik^\dagger} w_{k^\dagger, \ell^\dagger}^{(2)} w_{k^\dagger, \ell^\dagger}^{(1)} h_i \rho^{[2]}(w_{k^\dagger, \ell^\dagger}^{(1)} h_i + v_{k^\dagger, \ell^\dagger}^{(1)}) \right. \\ & \left. + x_{ik^\dagger} w_{k^\dagger, \ell^\dagger}^{(2)} \rho^{[1]}(w_{k^\dagger, \ell^\dagger}^{(1)} h_i + v_{k^\dagger, \ell^\dagger}^{(1)}) \right\}, \end{aligned}$$

Weak derivatives (4/4)

$$\begin{aligned} \frac{\partial}{\partial w_{k^\dagger, \ell^\dagger}^{(2)}} \ell_\zeta(\boldsymbol{\theta}) &= \sum_{i=1}^n \zeta_i \frac{\sigma^{[2]}(f_{h_i}(\mathbf{x}_i))}{\sigma^{[1]}(f_{h_i}(\mathbf{x}_i))} x_{ik^\dagger} \rho(w_{k^\dagger, \ell^\dagger}^{(1)} h_i + v_{k^\dagger, \ell^\dagger}^{(1)}) \\ &\quad + \sum_{i=1}^n \zeta_i \frac{1}{f_{h_i}^{[1]}(\mathbf{x}_i)} x_{ik^\dagger} w_{k^\dagger, \ell^\dagger}^{(1)} \rho^{[1]}(w_{k^\dagger, \ell^\dagger}^{(1)} h_i + v_{k^\dagger, \ell^\dagger}^{(1)}). \end{aligned}$$

- ▶ We implemented these weak derivatives manually:
<https://github.com/oknakfm/N3POM/blob/main/scripts/functions.R>
- ▶ Modern technology: “automatic differentiation”

Numerical Experiments

Synthetic Dataset Experiments

- ▶ $a_*(u) = 2u - 9$, $\mathbf{b}_*(u) = (-1 + m_1 u^2, 1 + m_2 u^2)$,
- ▶ $\mathbf{x}_i = (r_i \cos \theta_i, r_i \sin \theta_i)$, $r_i \sim U[0, 1]$, $\theta_i \sim U[0, 2\pi)$, $h_i \sim \sigma(a_*(u) + \langle \mathbf{b}_*(u), \mathbf{x}_i \rangle)$.
- ▶ Number of hidden units: $L = 50$.

Model	Optimizer	Response	$(m_1, m_2) = (0.05, -0.05)$	
			MSE(b_1)	MSE(b_2)
N ³ POM	MPS	h_i	0.066 (0.158)	0.120 (0.152)
N ³ POM	MPS	$\mathfrak{e}([h_i])$	0.116 (0.101)	0.162 (0.092)
N ³ POM	MPS	$[h_i]$	0.516 (0.061)	0.531 (0.034)
POM	polr	$[h_i]$	0.516 (0.024)	0.514 (0.022)
NPOM	(ridge) oNet	$[h_i]$	0.230 (0.041)	0.234 (0.080)
NPOM	(elastic) oNet	$[h_i]$	0.220 (0.184)	0.292 (0.197)
NPOM	(lasso) oNet	$[h_i]$	0.203 (0.197)	0.273 (0.205)
NPOM	serp	$[h_i]$	0.174 (0.075)	0.195 (0.086)

- ▶ N³POM is the best even for other settings $(m_1, m_2) = (0.05, 0.05), \dots$

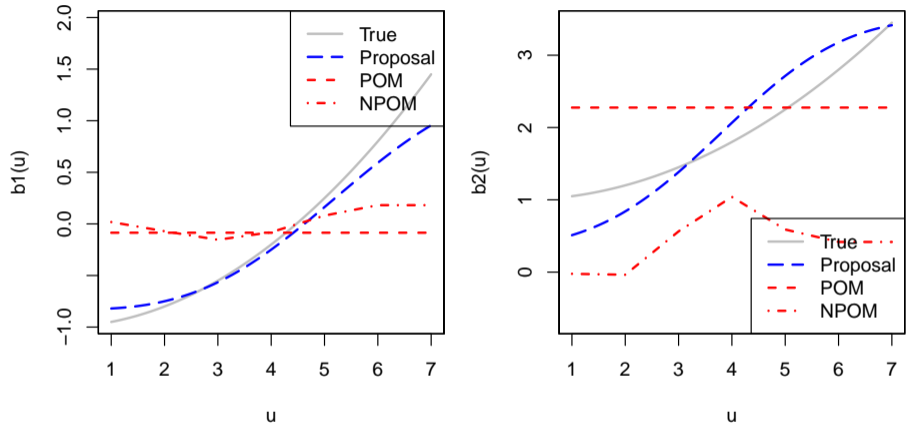


Figure: Illustration of POM, NPOM, and N³POM (Left: $m_1 = 0.5$, Right: $m_2 = 0.5$).

Real-world Dataset Experiments

(Preprocessing²)

- ▶ Covariates $\{\mathbf{x}_i\}_{i=1}^n$ are normalized (centered and scaled).
- ▶ Responses $\{h_i\}_{i=1}^n$ are Affine-transformed to take value in $[1, 10]$ ($J = 10$).

(NN training)

- ▶ We repeat 5000 iterations for MPS algorithm.
- ▶ With different random seeds, we compute 10 trials of MPS algorithm (and plot the results of the 10 trials).

Interpretation: we monitor $\mathbf{s}(u) = -\mathbf{b}(u)$:

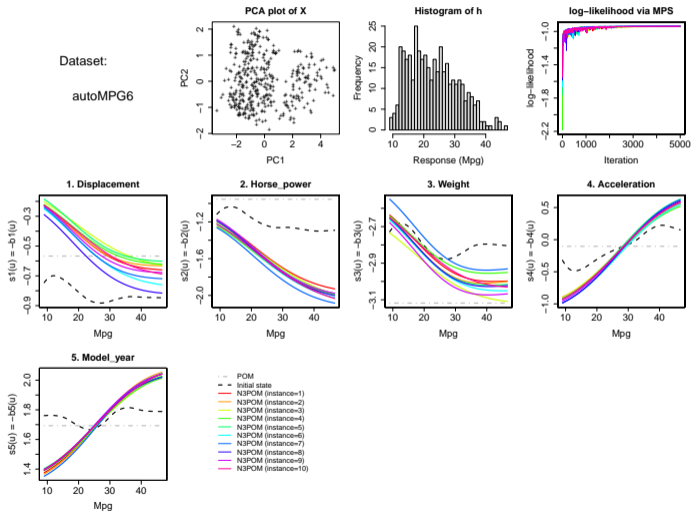
$$\text{logit}(\mathbb{P}_{\text{N}^3\text{POM}}(H > u \mid X = \mathbf{x})) = r(u) + \langle \mathbf{s}(u), \mathbf{x} \rangle,$$

with $r(u) = -a(u)$, $\mathbf{s}(u) = -\mathbf{b}(u)$.

²Datasets are collected from UCI Machine Learning repository (Dua and Graff, 2017)

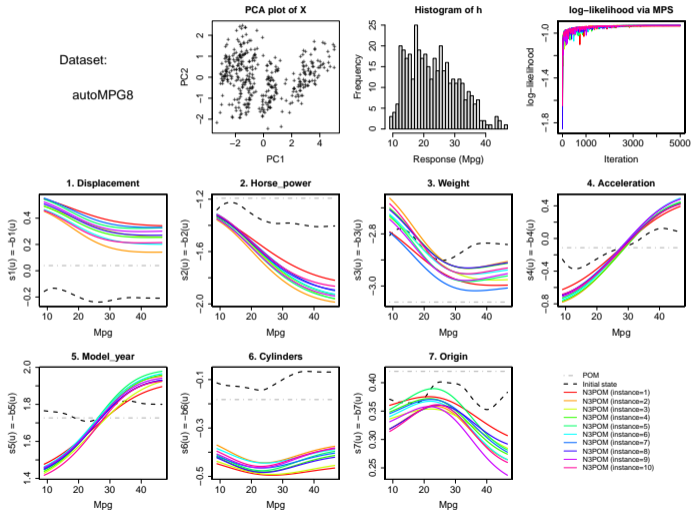
autoMPG6 ($n = 392, d = 5$)

- ▶ Response = fuel efficiency (miles per gallon)



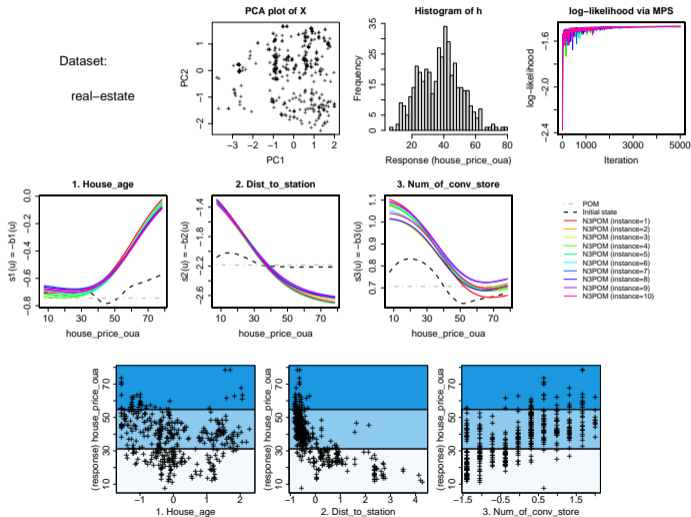
autoMPG8 ($n = 392, d = 7$)

- ▶ Response = fuel efficiency (miles per gallon)



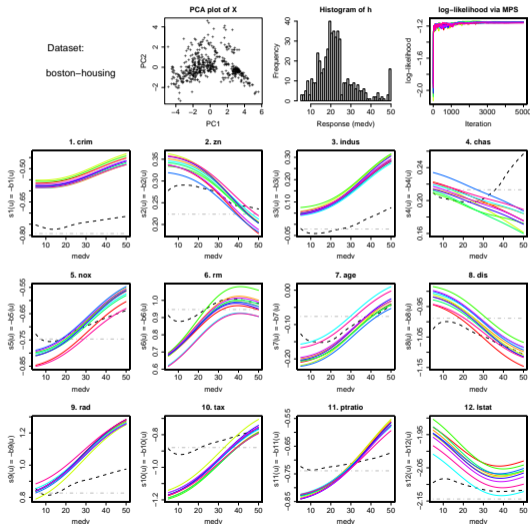
real-estate ($n = 413, d = 3$)

- ▶ Response = house price of unit area



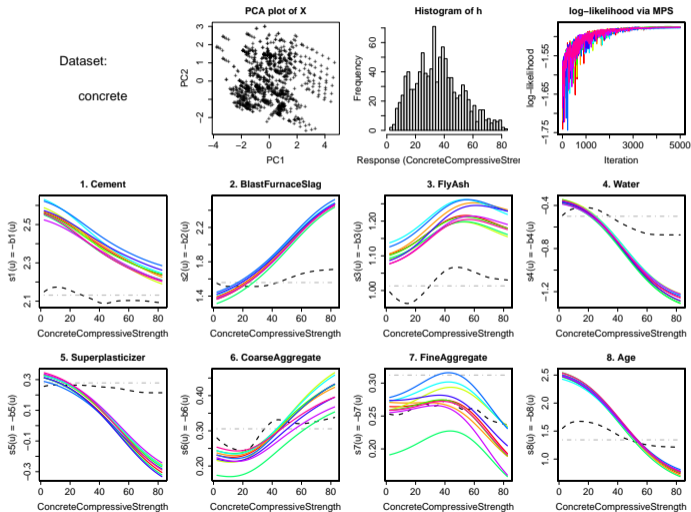
boston-housing ($n = 506, d = 12$)

- ▶ Response = median value of houseprice



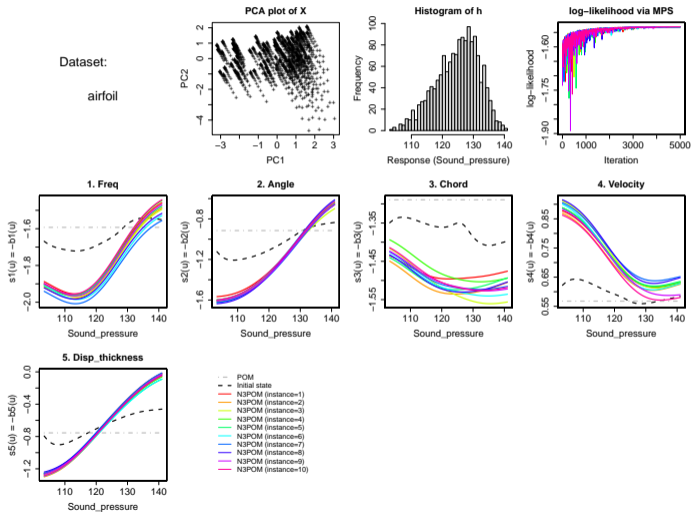
concrete ($n = 1030, d = 8$)

► Response = compressive strength



airfoil ($n = 1503, d = 5$)

► Response = sound pressure



Conclusion

Conclusion

- (1) We proposed N³POM and MPS algorithm with theoretical guarantee.
- (2) N³POM demonstrates better performance than conventional NPOM.
- (3) N³POM is applied to several real-world datasets.

▶ **Paper:** <https://doi.org/10.1080/10618600.2024.2321208>

▶ **Contact Info:** A. Okuno (okuno@ism.ac.jp)



We appreciate Kohei Hattori, Keisuke Yano, Shuichi Kawano, Kei Hirose, and Lucas Kook for helpful discussions.

宣伝

計算技術による学際的統計解析ワークショップ（まだ参加登録可能です）

- ▶ 統計学 + 数値解析 + 計算代数 + . . .
- ▶ 日時：2025年2月17日・18日
- ▶ 場所：統計数理研究所
- ▶ 講演者（敬称略）：寺田吉壺（大阪大学），相島健助（法政大学），柳下翔太郎（統計数理研究所），廣瀬慧（九州大学），深作亮也（九州大学），今倉暁（筑波大学），佐藤峻（東京大学），加葉田雄太郎（長崎大学），松田孟留（東京大学）。
- ▶ <https://okuno.net/events/ISACT2025>



References I

- Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. (2021). Neural additive models: Interpretable machine learning with neural nets. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.
- Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. John Wiley & Sons.
- Agresti, A. and Tarantola, C. (2018). Simple ways to interpret effects in modeling ordinal categorical data. *Statistica Neerlandica*, 72(3):210–223.
- Archer, K. J., Hou, J., Zhou, Q., Ferber, K., Layne, J. G., and Gentry, A. E. (2014). ordinalgmifs: An R package for ordinal regression in high-dimensional data settings. *Cancer Informatics*, 13:CIN.S20806.
- Baumann, P. F. M., Hothorn, T., and Rügamer, D. (2021). Deep conditional transformation models. In Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J., and Lozano, J. A., editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 3–18. Springer International Publishing.

References II

- Bennett, S. (1983). Log-logistic regression models for survival data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 32(2):165–171.
- Cardoso, J. S. and da Costa, J. F. P. (2007). Learning to classify ordinal data: The data replication method. *J. Mach. Learn. Res.*, 8:1393–1429.
- Chu, W. and Ghahramani, Z. (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6(35):1019–1041.
- Cockerham, W. C. (2016). *International Encyclopedia of Public Health*. Academic Press.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314.
- Dua, D. and Graff, C. (2017). UCI machine learning repository.
- Farouki, R. T. (2012). The bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design*, 29(6):379–419.
- Garcia, T. P., Marder, K., and Wang, Y. (2019). Time-varying proportional odds model for mega-analysis of clustered event times. *Biostatistics*, 20(1):129–146.

References III

- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.
<http://www.deeplearningbook.org>.
- Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102.
- Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796.
- Herzog, L., Kook, L., Gotschi, A., Petermann, K., Hansel, M., Hamann, J., Durr, O., Wegener, S., and Sick, B. (2022). Deep transformation models for functional outcome prediction after acute ischemic stroke. *Biometrical Journal*.
- Howell, D. C. (2010). *Statistical methods for psychology (7th. ed.)*. Thomson Wadsworth, Belmont, CA.
- Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Inc., 2nd edition.

References IV

- Kook, L., Baumann, P. F., Dürr, O., Sick, B., and Rügamer, D. (2022a). Estimating conditional distributions with neural networks using r package deeptrafo. *arXiv preprint arXiv:2211.13665*.
- Kook, L., Herzog, L., Hothorn, T., Durr, O., and Sick, B. (2022b). Deep and interpretable regression models for ordinal outcomes. *Pattern Recognition*, 122:108263.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22:140.
- Liu, Q., Shepherd, B. E., Li, C., and Harrell, Jr, F. E. (2017). Modeling continuous response variables using ordinal regression. *Statistics in Medicine*, 36(27):4316–4335.
- Long, J. S. and Freese, J. (2006). *Regression models for categorical dependent variables using Stata*, volume 7. Stata press.
- Lu, F., Ferraro, F., and Raff, E. (2022). Continuously generalized ordinal regression for linear and deep models. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 28–36. SIAM.

References V

- McCullagh, P. (1980). Regression models for ordinal data. *Journal of Royal Statistical Society. Series B (Methodological)*, 42(2):109–127.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall CRC, London.
- Peterson, B. and Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. *Journal of Royal Statistical Society. Series C (Applied Statistics)*, 39(2):205–217.
- Pöbnecker, W. and Tutz, G. (2016). A general framework for the selection of effect type in ordinal regression. Technical report, Ludwig-Maximilians-Universität München.
- Satoh, K., Tonda, T., and Izumi, S. (2016). Logistic regression model for survival time analysis using time-varying coefficients. *American Journal of Mathematical and Management Sciences*, 35(4):353–360.

References VI

- Sick, B., Hathorn, T., and Durr, O. (2021). Deep transformation models: Tackling complex regression problems with neural network based transformation models. In *Proceedings of the 25th International Conference on Pattern Recognition*, pages 2476–2481. IEEE Computer Society.
- Thas, O., Neve, J. D., Clement, L., and Ottoy, J.-P. (2012). Probabilistic index models. *Journal of Royal Statistical Society. Series B (Methodological)*, 74(4):623–671.
- Tutz, G. and Berger, M. (2022). Sparser ordinal regression models based on parametric and additive location-shift approaches. *International Statistical Review*, 90(2):306–327.
- Tutz, G. and Gertheiss, J. (2016). Regularized regression for categorical data. *Statistical Modelling*, 16(3):161–200.
- Ugba, E. R., Mörlein, D., and Gertheiss, J. (2021). Smoothing in ordinal regression: An application to sensory data. *Stats*, 4(3):616–633.

References VII

- Vargas, V. M., Gutiérrez, P. A., and Hervás, C. (2019). Deep ordinal classification based on the proportional odds model. In Ferrández Vicente, J. M., Álvarez-Sánchez, J. R., de la Paz López, F., Toledo Moreo, J., and Adeli, H., editors, *From Bioinspired Systems and Biomedical Applications to Machine Learning*, pages 441–451. Springer International Publishing.
- Vargas, V. M., Gutierrez, P. A., and Hervas-Martinez, C. (2020). Cumulative link models for deep ordinal classification. *Neurocomputing*, 401:48–58.
- Williams, R. (2006). Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *Stata Journal*, 6:58–82.
- Williams, R. (2016). Understanding and interpreting generalized ordered logit models. *The Journal of Mathematical Sociology*, 40(1):7–20.
- Wurm, M. J., Rathouz, P. J., and Hanlon, B. M. (2021). Regularized ordinal regression and the ordinalNet R package. *Journal of Statistical Software*, 99(6).
- Zahid, F. M. and Tutz, G. (2013). Proportional odds models with high-dimensional data structure. *Int. Stat. Rev.*, 81(3):388–406.

Appendix

Non-linear extension of NPOM (Kook et al., 2022b)

Ordinal regression is generalized to non-linear models, leveraging

- ▶ Gaussian process (Chu and Ghahramani, 2005),
- ▶ neural networks (Cardoso and da Costa, 2007; Vargas et al., 2019, 2020).

Complex intercept (CI) model in Kook et al. (2022b) considers

$$f_j^{(\text{CI})}(\mathbf{x}) = \alpha_0(\mathbf{x}) + \sum_{c=1}^j \exp(\gamma_c(\mathbf{x}))$$

with non-linear functions $\{\gamma_c\}_{c=1}^J$. This model can be regarded as a response-dependent prediction model (i.e., a non-linear extension of NPOM). While this CI model is guaranteed to be monotone for increasing $j = 1, 2, \dots, J$ and they have high expressive power, the model cannot be extended to continuous response (as we cannot estimate uncountable many functions $\{\gamma_c\}_{c \in [1, J]}$), and it is hard to interpret the relationship between the covariates and the response, through the estimated non-linear functions γ_c .

Marginal effect (Agresti and Tarantola, 2018)

We may employ a marginal effect (Agresti and Tarantola, 2018):

$$\frac{\partial}{\partial \mathbf{x}} \mathbb{P}_{N^3\text{POM}}(H \leq u \mid X = \mathbf{x})$$

when considering the influence to the CCP directly. However, the marginal effect differs depending on the covariate \mathbf{x} , and it tends to (excessively) shrink the influence to 0 in the tail region ($u \approx 0, u \approx J$) as $\sigma^{[1]}(-\infty) = \sigma^{[1]}(+\infty) = 0$; unlike the simple coefficients $\mathbf{s}(u)$, marginal effect cannot capture the tendency whether the influence of the covariate increases or decreases (as u increases), because of the shrinking behavior in the tail region. In the case of real-estate dataset above, the interesting coefficient $s_k(u)$ of house-age that shrinks to 0 as $u \nearrow J$, cannot be detected when using the marginal effect, as almost all marginal effects shrink to 0 as $u \nearrow J$ (regardless of how the coefficient is important in the tail region). As a future work, it would be worthwhile to consider a more interpretable score for evaluating the influence of the covariates in the context of ordinal regression.