論文解説：順序回帰における柔軟性とドメイン制約のトレードオフ

奥野彰文 [*1,2], 原田和治 [3]

[1] 統計数理研究所 統計思考院　[2] 理化学研究所 AIP センター
[3] 東京医科大学 医療データサイエンス分野

要旨

本稿はニューラルネットを利用した順序回帰に関する我々の英文プレプリント: Okuno and Harada (2023) の解説です．解説の平易さを優先するため，より厳密な記述については原著論文をご参照ください．

キーワード: 非比例順序回帰，ニューラルネット，解釈性，単調性，生存時間解析

1　研究背景

1.1　順序回帰の目的

1 以上 J 以下の値を取る応答変数 $H \in [1, J]$ と，対応する共変量 $X \in \mathbb{R}^d$ のペアがいくつか観測できているとしましょう．例えばレストランの評価などを想像してください．$J = 5$ とすると，応答変数の観測値 $h_i = 5$ はあるレストラン i が最高の評価であることを表し，$h_i = 1$ は最低の評価であることを表します．典型的なレストランには $h_i = 2.8$ などがつけられるでしょう．$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id}) \in \mathbb{R}^d$ はそのレストランの共変量で，例えば x_{i1} が駅からの距離，x_{i2} が従業員数を表します．

本稿で扱う順序回帰とは，共変量 X が条件付けられているときの応答変数 H の条件付き累積確率分布関数

$$\mathbb{P}(H \leq u \mid X = \boldsymbol{x}) \tag{1}$$

を推定する問題，およびその推定を介して共変量 X と応答変数 H の間の関係を見つける問題を指します．先ほどの例では，あるレストランの共変量から評価がどのくらいかを推定し，その過程で例えば駅からの距離がレストランの評価に良い影響あるいは悪い影響を与えているのか，などが知りたいわけです．

1.2　生存時間解析との関わり

順序回帰ととても似た問題に生存時間解析がありますが，順序回帰では (主に) 推定の対象とする関数が異なります．生存時間解析では主にハザード関数

$$\lambda(u \mid X = x) = -\frac{\mathrm{d}}{\mathrm{d}u} \log\{1 - \mathbb{P}(H \leq u \mid X = \boldsymbol{x})\}$$

を推定していて，ここで (比例) ハザード関数を $\lambda_{\mathrm{Cox}}(u \mid X = \boldsymbol{x}) = \lambda_0(u) \exp(\langle \boldsymbol{\gamma}, \boldsymbol{x} \rangle)$ とするのが Cox モデルです．順序回帰で推定した累積分布関数 (1) を微分することでハザード関数が導出できますが，一般にパラメトリックモデルの積分は難しいので，逆にハザード関数から累積分布関数を導出するのは困難です．したがって順序回帰はユーザが扱いやすい関数（累積分布関数) を推定していると考えることもできますが，累積分布関数には単調性の制約があり，整合的なパラメトリックモデルの設定が重要となります．

1.3　比例/非比例オッズモデル

順序回帰に話を戻しましょう．1.1 節での設定を考えると，順序回帰のモデリングでは以下の 2 つの要件：

1. (解釈性): X と H の関係が解釈できる
2. (単調性): (1) の推定量は u について単調増加する

を満たすことが求められます．上記の要件を考えながら，本節では既存の順序回帰モデルについて紹介します．

応答変数 H は本来連続の値をとる場合でも，離散値に丸めたものを想定することで解析が容易になります．そこで，ここではまず，$H \in [1, J]$ ではなく離散の応答 $G \in \{1, 2, \ldots, J\}$ を考えます．このとき，ロジット関数 $\mathrm{logit}(z) = \log z - \log(1 - z)$ と内積 $\langle \cdot, \cdot \rangle$, および離散応答の閾値 $j \in \{1, 2, \ldots, J\}$ を用いて

$$\mathrm{logit}(\mathbb{P}_{\mathrm{POM}}(G \leq j \mid X = \boldsymbol{x})) = \alpha_j + \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle$$

で定義されるのが比例オッズモデル (proportional odds

* 責任著者，okuno@ism.ac.jp

model; POM) であり，$\theta = (\{\alpha_j\}, \boldsymbol{\beta})$ が推定するパラメータです．POM は x に関して線形なので，各共変量に対応する係数 $\boldsymbol{\beta} \in \mathbb{R}^d$ の値を見ることで，どの共変量が応答 G の値に強い影響を与えているのかがわかり，解釈性を持ちます．また $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_J$ という制約をかけると，(任意に x を固定したとき)j に関しての単調性を満たします[a]．

一方で，POM の係数 $\boldsymbol{\beta}$ は応答 j に依らず常に定数です．先ほどのレストランの例では，例えば評価の高いレストランと評価の低いレストランでどちらも駅からの距離の影響力が同じであるということを仮定していることになり，現実的にはやや不自然です[b]．このような問題に対処するために，

$$\mathrm{logit}(\mathbb{P}_{\mathrm{NPOM}}(G \leq j \mid X = \boldsymbol{x})) = \alpha_j + \langle \boldsymbol{\beta}_j, \boldsymbol{x} \rangle$$

で定義されるのが非比例オッズモデル (Non-POM; NPOM) です．NPOM は解釈性に優れ，また POM より柔軟ですが，係数 $\boldsymbol{\beta}_j$ が応答 j に依存して変わるので，α_j が単調に増加してもモデル全体では単調性を持たないことがあります．また $\boldsymbol{\beta}_j \in \mathbb{R}^d$ を各 j ごとに独立に推定するので，共変量の影響力を表す係数 $\boldsymbol{\beta}_j$ が j について連続に変化しないことも起こりえます．このような問題に対処するために，例えば $\boldsymbol{\beta}_j = \boldsymbol{\beta} + \boldsymbol{\delta}_j$ とし，単調性や連続性に反しない程度にハイパーパラメータを変更しながら $\boldsymbol{\delta}_j$ を小さく推定するなど，ヒューリスティックな対処法が研究されてきました．

2　本研究の貢献

本研究では，ヒューリスティックを排除して NPOM の単調性と解釈性をどう保証すればよいか考えました．

これまでの研究は主に離散の応答 G を考えていましたが，我々の研究ではまず，もともとの連続な応答変数 $H \in [1, J]$ を扱い，NPOM を連続に拡張した Neural Network-based NPOM (N^{3}POM):

$$\mathrm{logit}(\mathbb{P}_{\mathrm{N^3POM}}(H \leq u \mid X = \boldsymbol{x})) = a(u) + \langle \boldsymbol{b}(u), \boldsymbol{x} \rangle$$

を提案しています．$u \in [1, J]$ は連続応答の閾値，$a : [1, J] \to \mathbb{R}$ は連続で区分線形な単調増加関数，$\boldsymbol{b} : [1, J] \to \mathbb{R}^d$ は連続なニューラルネットです．

N^{3}POM の単調性

単調性に関連して，我々はまず次の定理を示しました．

定理 1. N^{3}POM が任意の $\boldsymbol{x} \in \mathbb{R}^d$ について (u に関して) 単調増加するのは，$\boldsymbol{b}(u)$ が定数関数の場合に限る．

上記の定理 1 は，$\boldsymbol{x}$ がユークリッド空間全域 $\mathbb{R}^d$ を動くとき，単調性を保証できるような線形のオッズモデルは POM に限ることを示しています．著者の知る限りこの事実を明示的に記した既存研究はありませんが，同様の定理が NPOM にも成り立ちます．つまり応答変数に応じて係数 $\boldsymbol{b}(u)$(または $\boldsymbol{\beta}_j$) を変化させることは非常に難しいことがここで明確に示され，非比例モデルを研究するためにはやや悲観的な結果が得られた一方で，例えばレストランから駅までの距離が (多くの場合は)10km を超えることがないように，実際の共変量 $\boldsymbol{x}$ が値をとるのは有界な集合上であるとも考えられます．そこで，共変量 $\boldsymbol{x}$ が値をとりうる集合を，適当な $\eta > 0$ について

$$\mathcal{X}_2(\eta) := \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_2 \leq \eta\} \tag{2}$$

に限定し，ある不等式

$$(a(u) \text{ の傾きの最小値}) \geq \eta \cdot (\boldsymbol{b}(u) \text{ の複雑度}) \tag{3}$$

を考える[c]と，以下の定理が示せました．

定理 2. 不等式 (3) を満たすとする．任意の $\boldsymbol{x} \in \mathcal{X}_2(\eta)$ を固定したとき，N^{3}POM は u について単調増加する．

つまり，係数関数 $\boldsymbol{b}(u)$ の変動が切片関数 $a(u)$ の変動に比べて小さければ，切片関数 $a(u)$ の単調性により，N^{3}POM の単調性が担保できることが定理 2 により示されました．定理 2 で満たすべき不等式 (3) を考えると，N^{3}POM が単調であるためには，モデルの柔軟性 (つまり $\boldsymbol{b}(u)$ の複雑度) とドメイン制約 (つまり η の大きさ) にトレードオフがあることもわかります．定理 2 は非比例モデルを正当化するために喜ばしい結果でもあり，不等式 (3) を満たすようニューラルネットの重みを反復修正することで，単調性を担保しながら N^{3}POM を学習する Monotinicity-Preserving Stochastic (MPS) アルゴリズムを提案しています．

N^{3}POM の解釈性

N^{3}POM はその定義から，係数関数 $\boldsymbol{b}(u) \in \mathbb{R}^d$ の値により各応答の閾値 u における共変量の影響力を読み取ることができて，解釈性も担保されていることがわかりま

[a] 実装では，適当なパラメータ $\{\gamma_j\}$ を使って $\alpha_{j+1} = \alpha_j + |\gamma_j|$ などとします．

[b] $\boldsymbol{\beta}_j$ が応答 j によらず一定であることを平行性仮定と呼び，平行性仮定が成り立つかの検定などが 2000 年以前に盛んに研究されました．語句の一貫性のため，本原稿ではこの性質を平行性ではなく比例性と呼びます．

[c] $\boldsymbol{b}(u)$ の複雑度の定義など，具体的な不等式についてはOkuno and Harada (2023) 式 (9) を参照してください．

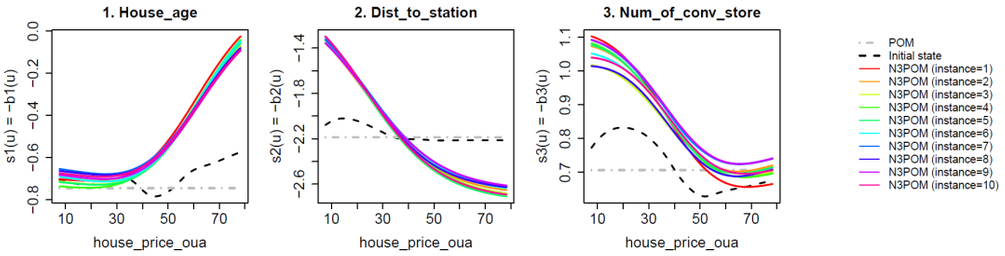

図 1: 乱数を 10 回変更して推定した N^3POM の係数関数 $\hat{\boldsymbol{s}}(u) = -\hat{\boldsymbol{b}}(u)$.

す．解釈のため，$H \leq u$ の確率ではなく $H > u$ の確率を考えると，$c(u) = -a(u), \boldsymbol{s}(u) = -\boldsymbol{b}(u)$ を用いて

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

となり，共変量 $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ の各要素 x_j に対応する係数 $s_j(u) = -b_j(u)$ が大きいとき，応答変数 h を大きくする影響力が強いことがわかります．

数値実験

UCI Machine Learning Repository[d]から取得した実データを用いて，実際に N^3POM を学習してみましょう．一例として，ここでは住宅価格をまとめた real-estate データセットの結果を示します[e]．

real-estate データセットには $n = 413$ 物件の情報が保存されており，各物件 i の，応答変数 h_i が単位面積当たりの住宅価格 (house price of unit area) を，共変量 x_{i1} が築年数, x_{i2} が駅からの距離, x_{i3} が周辺にあるコンビニの数を表しています．このデータを標準化[f]した後，提案した MPS アルゴリズムにより中間素子数 $L = 50$ で中間層 1 層のパーセプトロン $\boldsymbol{b}(u)$ と区分線形関数 $a(u)$ を学習します．推定された係数関数 $\hat{\boldsymbol{s}}(u) = -\hat{\boldsymbol{b}}(u)$ は図 1を見てください．

まず “1. House_age” は築年数に関する係数関数を表しています．応答変数つまり住宅価格の閾値 u が小さいとき係数関数は -0.7 あたりの値をとるので，安価な住宅では築年数が住宅価格にネガティブな影響を与えることが推察できます．一方で，u が大きいとき係数関数がほぼ 0 になるので，高価な住宅に関しては，築年数は住宅価格にあまり影響を与えなさそうです．実際のデータでもこの傾向を読み取ることができて，価格帯が高いところでは古いが高価な住宅がいくつも存在し，築年数と住宅価格の相関が小さくなっていました．

次に，“2. Dist_to_station” は駅からの距離に関する係数関数を表しています．全体として負の値をとるので，駅からの距離が遠いほど住宅価格にネガティブな影響を与えることが推察できます．安価な住宅より高価な住宅であるほど，住宅価格へのネガティブな影響は大きいようです．データセットを調べてみると，高価な住宅がほぼすべて駅近くにあり，駅の距離と住宅価格は反比例しているので，この観察も正しそうです．

最後に，“3. Num_of_conv_store” はコンビニの数に関する係数関数を表しています．全体として住宅価格にポジティブな影響を与えていますが，安価な住宅であるほどコンビニの数が重視されていて，高価な住宅であるほど影響力が下がることが読み取れます．

3 今後の展望

本稿では N^3POM を提案し，単調性に関する理論および最適化のアルゴリズムを提案しました．一方で，本研究にも様々な課題が残されています．

まず 1 つ目の課題として，本研究で推定された係数関数 $\hat{\boldsymbol{b}}(u)$ の信頼度を評価できていません．順序回帰は生存時間解析などとも密接にかかわっているので，医療データなどへの応用が連想されますが，信頼度を評価できないことが大きなボトルネックとなっています．

2 つ目の課題に高次元データへの適応があります．N^3POM では共変量 $\boldsymbol{x}$ が集合 (2) に属していることを仮定しますが，η は $\|\boldsymbol{x}\|_2$ の上限なので，$\boldsymbol{x}$ の次元が上がれば η は増大します．このとき不等式 (3) を無理に満たそうと思えば，ニューラルネットによる係数関数 $\boldsymbol{b}(u)$ の複雑度を小さくせねばならず，$\boldsymbol{b}(u)$ は定数関数に近付きます．このように，現状の不等式では共変量の次元が大きくなると N^3POM を POM に近付けないといけないので，より緩和された不等式の導出が望まれます．

参考文献

Okuno, A. and Harada, K. (2023). An interpretable neural network-based non-proportional odds model for ordinal regression. https://arxiv.org/abs/2303.17823.

[d] https://archive.ics.uci.edu/

[e] https://github.com/oknakfm/N3POM にコードを公開しています．より整理されたコードを今後提供予定です．

[f] X の平均と分散が $0, 1$ となるよう調整し，H は $[1, J]$ に値をとるよう調整しています．