A-014

サイズの異なる時系列カテゴリカルデータにおける組み合わせ最適問題の解法の比較

武石 悠希 †　山崎 綾一郎 †　髙井 健人 ‡　山岸 祐己 †§¶　周 律旋 ¶
君島 真沙実 ¶　高林 貴仁 ¶
† 静岡理工科大学 ‡ 株式会社ジーニー § 浜松医科大学 ¶ 株式会社良品計画

1 はじめに

EC サイトでは，消費者の評価が時間と共に動的に変わるため，その変動を正確かつ迅速に捉える手法の構築が重要である．本研究では，評価傾向の変化をモデル化する手法としてレジームスイッチングモデルに注目し，異なるサンプルサイズを持つジャンル別レビューデータに適用した．特に，分割数の推定において，様々な大きさのデータに対し適切な分割数を設定することを目的に，近似解と厳密解とで精度や実行速度に差が存在するか，また候補を複数持つことで極端な解を排除できるかということを，近似解と厳密解における複数の情報量基準を用いた分割数とその結果の違いについて比較・検討を行った．

2 分析手法

複数の状態に変化しうる時系列データを $\mathcal{D} = \{(s_1, t_1), \cdots, (s_N, t_N)\}$ とする．ここで，s_n と t_n は J カテゴリの状態と n 番目の観測時間をそれぞれ表す．観測数を $|D| = N$ とすると，$t_1 \leq \cdots \leq t_n \leq \cdots \leq t_N$ となる．タイムステップを n とし，タイムステップの集合を $\mathcal{N} = \{1, 2, \ldots, N\}$ と定義する．また，k 番目のレジームの開始時刻を $T_k \in \mathcal{N}, \mathcal{T}_K = \{T_0, \ldots, T_k, \ldots, T_{K+1}\}$ をスイッチングタイムステップ集合とし，便宜上 $T_0 = 1$ および $T_{K+1} = N + 1$ とする．すなわち，$T_1, \ldots, T_K$ は推定される個々のスイッチングタイムステップであり，$T_k < T_{k+1}$ を満たすものとする．そして，$\mathcal{N}_k$ を k 番目のレジーム内のタイムステップ集合とし，各 $k \in \{0, \ldots, K\}$ に対して $\mathcal{N}_k = \{n \in \mathcal{N} \mid T_k \leq n < T_{k+1}\}$ と定義する．このとき，全体のタイムステップ集合は $\mathcal{N} = \mathcal{N}_0 \cup \cdots \cup \mathcal{N}_K$ である．

いま，各レジームの状態分布が J カテゴリの多項分布に従うと仮定する．$\mathbf{p}_k$ を k 番目のレジームにおける多項分布の確率ベクトルとし，確率ベクトルの集合を $\mathcal{P}_K = \{\mathbf{p}_0, \ldots, \mathbf{p}_K\}$ とすると，$\mathcal{T}_K$ が与えられたときの対数尤度関数は以下のように定義できる．

$$\mathcal{L}(\mathcal{D}; \mathcal{P}_K, \mathcal{T}_K) = \sum_{k=0}^{K} \sum_{n \in \mathcal{N}_k} \sum_{j=1}^{J} s_{n,j} \log p_{k,j}, \tag{1}$$

ここで，$s_{n,j}$ は $s_n \in \{1, \ldots, J\}$ を

$$s_{n,j} = \begin{cases} 1 & \text{if } s_j = j, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

のように変換したダミー変数である．各レジーム $k = 0, \ldots, K$ と各状態 $j = 1, \ldots, J$ に対する式 (1) の最尤推定量は $\hat{p}_{k,j} = \sum_{n \in \mathcal{N}_k} s_{n,j} / |\mathcal{N}_k|$ のように与えられる．これらの推定量を式 (1) に代入すると，以下の式が導ける．

$$L(D; \hat{\mathcal{P}}_K, \mathcal{T}_K) = \sum_{k=0}^{K} \sum_{n \in \mathcal{N}_k} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{k,j}. \tag{3}$$

したがって，スイッチングタイムステップの検出問題は，式 (3) を最大化する $\mathcal{T}_K$ の探索問題に帰着できる．しかし，式 (3) だけでは $\mathcal{T}_K$ の導入によって，どれだけ尤度が改善されたかという直接的な評価をすることができないため，尤度比最大化問題として目的関数を構築し直す．もし，レジームスイッチングのような変化が存在しない，すなわち $\mathcal{T}_0 = \emptyset$ と仮定すると，式 (3) は次のように表される．

$$L(\mathcal{D}; \hat{\mathcal{P}}_0, \mathcal{T}_0) = \sum_{n \in \mathcal{N}} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{0,j}, \tag{4}$$

ここで，$\hat{p}_{0,j} = \sum_{n \in \mathcal{N}} s_{n,j} / N$ である．よって，K 個のスイッチングを持つ場合と，スイッチングを持たない場合の対数尤度比は以下のように与えられる．

$$LR(\mathcal{T}_K) = L(D; \hat{\mathcal{P}}_K, \mathcal{T}_K) - L(D; \hat{\mathcal{P}}_0, \mathcal{T}_0). \tag{5}$$

最終的に，この問題は上記の $LR(\mathcal{T}_K)$ を最大化する $\mathcal{T}_K$ の探索問題に帰着できるが，式 (5) を網羅的に解くと計算量は $O(N^K)$ となる．貪欲法と局所改善法を利用して高速に近似解を求める解法も存在するが [2]，ここでは動的計画法を用いて効率的に厳密解を求めることを考える．

閉区間 $[1, n]$ で k 個のスイッチングを持つ場合の最大対数尤度を $V(n, k)$ とすると，次の漸化式が成り立つ．

$$V(n,k) = \max_{1 \leq m < n} \{V(m, k-1) + \sum_{j=1}^{J} S(m+1, n, j) \log \frac{S(m+1, n, j)}{n - m + 1}\}. \tag{6}$$

Regime Segmentation of Categorical Series Data Using Dynamic Programming
†Yuki TAKEISHI †Ryoichiro YAMAZAKI ‡Kento TAKAI †§¶Yuki YAMAGISHI ¶Lyuxuan ZHOU ¶Masami KIMIZIMA ¶Takahito TAKABAYASHI
†Shizuoka Institute of Science and Technology
‡Geniee, Inc.
§Hamamatsu University School of Medicine
¶Ryohin Keikaku Co., Ltd.

ここで，$S(m,n,j) = \sum_{i=m}^{n} s_{i,j}$ である．この漸化式を再帰的に解き，$V(N,K)$ を求めることで，時系列データ全体に対しての最大対数尤度を得ることができる．また，$V(n,k)$ を求める過程で，各ステップにおいて最適な分割点を記録することで最適なスイッチングタイムステップ集合 $\mathcal{T}_k$ を得ることができる（ただし，$k \in \{1,\ldots,K\}$）．

このアルゴリズムの実装上の時間計算量は $O(N^2(K+J))$ で，空間計算量は $O(NK)$ であり，愚直な解法よりも効率的に解くことが可能である．

3 評価実験とまとめ

情報量基準の AIC, BIC, MDL, さらにエルボー法の自動化（L method [1]）の 4 つの基準と，2 つの解法（近似解 (*approx.*) と厳密解 (*exact*)）で比較を行う．なお，最大分割数 K は今回 15 に設定し，実験では無印良品のネットストアにおけるレビューデータセット $N = 390,993$（レビューを 100 以上有するジャンル 105 種）を用いた．2 つの解法における各基準が選択した最適な分割数 $\hat{k}$ の分布を図 1, 2 に，サンプルサイズ N ごとの $\hat{k}$ の移動平均を図 3 に，解法ごとの計算時間を図 4 にそれぞれ示す．

L method によるエルボー法は，情報量基準とは異なる $\hat{k}$ の推移を示し，サンプルサイズが小さい時に検出力を持つことから，情報量基準の他に解の候補を得るうえで新たな基準として有用である．また，計算時間については，サンプルサイズが小さい時には動的計画法による厳密解の方が高速であり，精度の観点からも使用に適している．しかし，サンプルサイズが増加するにつれて，貪欲法がベースの近似解と速度が逆転し，その後は指数関数的に差が広がるため，ある程度の大きさを持つサンプルに対しては，近似解を使用することが実務上では有用であると言える．

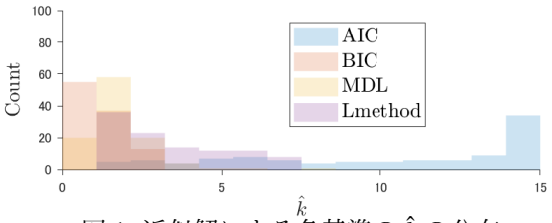

図 1: 近似解による各基準の $\hat{k}$ の分布

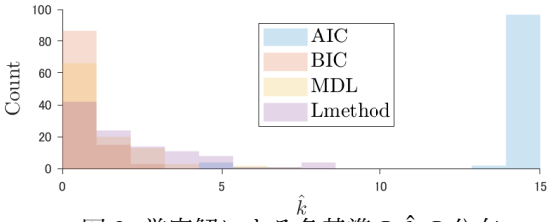

図 2: 厳密解による各基準の $\hat{k}$ の分布

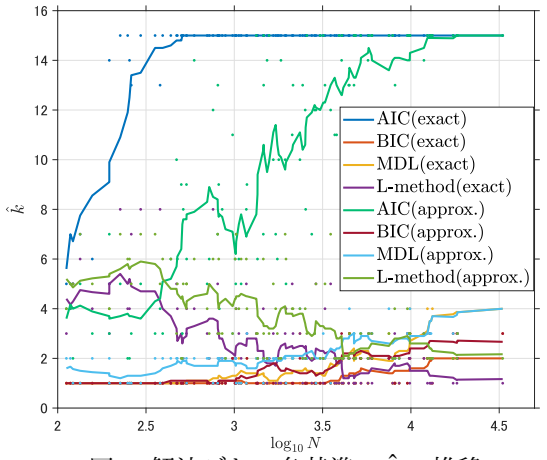

図 3: 解法ごとの各基準の $\hat{k}$ の推移

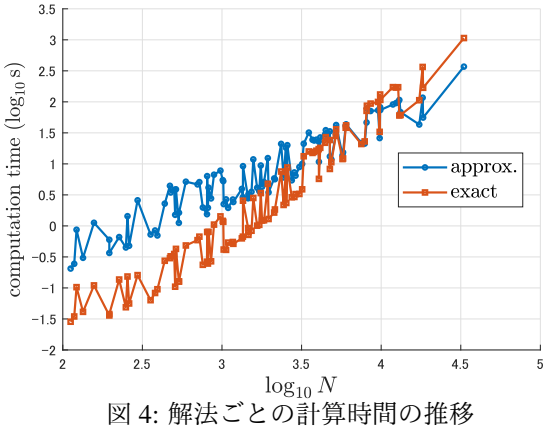

図 4: 解法ごとの計算時間の推移

参考文献

[1] Stan Salvador and Philip Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *16th IEEE International Conference on Tools with Artificial Intelligence*, 12 2004.

[2] Yuki Yamagishi and Kazumi Saito. Visualizing switching regimes based on multinomial distribution in buzz marketing sites. In *Foundations of Intelligent Systems - 23rd International Symposium, ISMIS 2017*, volume 10352 of *Lecture Notes in Computer Science*, pages 385–395. Springer, 2017.