正則化法による順序ロジットモデルにおける隣接クラスの統合

電気通信大学 大学院情報理工学研究科　永沼 瑞穂
電気通信大学 大学院情報理工学研究科　川野 秀一

1. はじめに

順序関係を持つカテゴリカルデータは，企業の格付けや，アンケートの選好度など広く様々な分野で現れる．このカテゴリカルデータに対しては，クラス数が多くなるについて不要なクラスが生じるといった問題が度々発生する．これらはパラメータ推定の不安定性や，推定結果の解釈時の困難さにつながることが知られている．

本報告では，これらの問題点を克服するために，不要なクラスを統合する方法論を順序ロジットモデル (Agresti, 2010) の枠組みの下で提案する．具体的には，順序ロジットモデルの中でも隣接カテゴリロジットモデルに着目し，スパース正則化に基づき不要なクラスを隣接クラス間に統合する．提案手法の有効性を，モンテカルロ・シミュレーションおよび実データへの適用を通して検証する．

2. 隣接カテゴリロジットモデルとその推定法

順序付きカテゴリカル変数 $G \in \{1, \ldots, J\}$ と p 次元説明変数 $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ からなる，n 組のデータ $\{(g_i, \boldsymbol{x}_i); i = 1, \ldots, n\}$ が得られたとする．ここで，説明変数に関するデータは標準化されているとする．つまり，$\sum_{i=1}^n x_{ih} = 0$，$\sum_{i=1}^n x_{ih}^2 = n$ $(h = 1, \ldots, p)$ が成り立つ．また，データ $\boldsymbol{x}_i$ が得られたとき j 番目のクラスに属する事後確率を $\Pr(G = j|\boldsymbol{x}_i) = \pi_j(\boldsymbol{x}_i)$ $(j = 1, \ldots, J)$ とする．このとき，隣接カテゴリロジットモデルは以下で与えられる．

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})} = \boldsymbol{\beta}_j^\top \boldsymbol{x} + \alpha, \qquad j = 1, \ldots, J-1.$$

ここで，α は切片，$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1}$ は回帰係数ベクトルである．

隣接した各クラスに対して，それらのクラスに属する事後確率が，説明変数 $\boldsymbol{x}$ に依存しない場合にそれらのクラス間を統合することを考える．いまの場合，第 j 番目の回帰係数ベクトルが $\boldsymbol{\beta}_j = \boldsymbol{0}$ となるとき，j 番目のクラスに属する事後確率と $(j+1)$ 番目のクラスに属する事後確率が，説明変数 $\boldsymbol{x}$ に依存しないことがわかる．したがって，隣接している不要なクラスを統合するために，次の最大化問題を考える．

$$\max_{\alpha, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1}} \left\{ \ell(\alpha, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1}) - \lambda \sum_{j=1}^{J-1} \|\boldsymbol{\beta}_j\|_2 \right\}.$$

ここで，$\ell(\alpha, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1})$ は隣接カテゴリロジットモデルに対する対数尤度関数であり，λ は正の値をとる正則化パラメータである．第 2 項目は，隣接したクラスを統合するためのグループ正則化項 (Yuan and Lin, 2006) である．推定値は交互方向乗数法 (Boyd et al., 2011) により構成されるアルゴリズムから求める．その更新式およびモンテカルロ・シミュレーション，実データ解析の結果は当日紹介する．

Agresti, A. (2010). *Analysis of Ordinal Categorical Data (Second Edition)*. Wiley.
Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, **3**, 1–122
Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, **68**, 49–67.