順序のあるカテゴリカルデータに対する ノンパラメトリックベイズモデリング

アステラス製薬　上野 真依
大阪大・基礎工　狩野 裕

1. 概要

順序のあるカテゴリカルデータとは，例えば“ 商品の満足度 ”項目を“ 非常に満足 ”,“ 満足 ”,“ どちらでもない ”,“ 不満 ”,“ 非常に不満 ”などのカテゴリーに分けて被験者に選択させることで得られるデータを言い，各カテゴリーに含まれる被験者数をクロス表にまとめることで分割表が得られる．順序のあるカテゴリカルデータは本来連続量（e.g., 心理学的連続体）であって測定機構が十分でないためカテゴリーとして観測されると考えることがある．このとき，データの背後に隠れている連続量を潜在変数とし，その潜在変数が正規分布に従うと仮定して解析する方法がよく用いられる．しかしデータが角に集中した分割表や中心がスパースな分割表に対してこの仮定は適切ではない．より柔軟にモデリングするために，Kottas et al.(2005) は正規分布の混合で潜在変数をモデリングする方法を提案した．彼らは混合の構成要素数を指定する必要のないディリクレ過程混合を用いてモデリングを行った．しかしこのモデルはデータの次元が増えるとパラメータ数が膨大になってしまう．そこで本発表では相関パラメータをすべて 0 という制約をおいた正規分布の混合を用いるモデルを提案する．このような制約をおくことで分散共分散行列のパラメータ数を節約することが可能となり，その結果高次元データの解析に強くなる．また提案モデルにおいてもディリクレ過程混合を用いることで構成要素数を定める必要がない．本発表では，人口データによる数値比較を通して，Kottas et al.(2005) のモデルと提案モデルの振る舞いを比較し，高次元における提案モデルの有用性を議論する．

2. 提案モデル

順序のあるカテゴリカル変数 $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})'(i = 1, \ldots, n)$ に対して，以下を満たす潜在変数 $\boldsymbol{Z}_i$ が，共分散がすべて 0 の p 次元正規分布の混合分布に従うと仮定する．

$$y_{ij} = l \ \text{ if } \ \gamma_{j,l-1} < Z_{ij} \leq \gamma_{j,l} \ (j = 1, \ldots, p, \ l = 1, \ldots, d_j).$$

ここで $-\infty = \gamma_{j,0} < \gamma_{j,1} < \cdots < \gamma_{j,d_j-1} < \gamma_{j,d_j} = \infty \ (j = 1, \ldots, p)$ は cutoff と呼ばれ，提案モデルにおいて任意に固定することができる．この混合分布をディリクレ過程混合を用いて以下のようにモデリングを行う．

$$\begin{aligned}
&\boldsymbol{Z}_i \overset{i.i.d.}{\sim} f, \ f(\cdot|G) = \int p_{N_p}(\cdot|\boldsymbol{m}, \boldsymbol{S}) dG(\boldsymbol{m}, \boldsymbol{S}), \\
&G|M, \boldsymbol{\lambda}, \boldsymbol{\sigma}^2, \boldsymbol{\beta} \sim DP(M, G_0), \\
&G_0(\boldsymbol{m}, \boldsymbol{S}) = \prod_{j=1}^{p} N(m_j|\lambda_j, \sigma_j^2) IG(\tau_j^2|\alpha, \beta_j),
\end{aligned}$$

ただし $p_{N_p}(\cdot|\boldsymbol{m}, \boldsymbol{S})$ は平均 $\boldsymbol{m}$，分散共分散行列 $\boldsymbol{S}$ をもつ p 次元正規分布の密度関数，$N(\cdot|\lambda_j, \sigma_j^2)$ は平均 λ_j，分散 σ_j をもつ 1 変量正規分布，$IG(\cdot|\alpha, \beta_j)$ はパラメータ α, β_j をもつ逆ガンマ分布を表す．ハイパーパラメータ $M, \lambda_j, \sigma_j^2, \beta_j (j = 1, \ldots, p)$ に対してハイパープライヤーをそれぞれ $M \sim Ga(a_0, b_0)$, $\lambda_j \sim N(q, Q^2)$, $\sigma_j^2 \sim \mathrm{IG}(b, B)$, $\beta_j \sim Ga(c, d)$ とおく．

参考文献

Kottas, A., Müller, P., and Quintana, F. (2005), Nonparametric Bayesian Modeling for Multivariate Ordinal Data, *Journal of Computational and Graphical Statistics*, 14:3, 610-625.
上野真依 (2015). 順序のあるカテゴリカルデータに対するノンパラメトリックベイズモデリング．大阪大学修士論文．