clustered data に対するグループ化有限混合モデル

菅澤翔之助（東京大学空間情報科学研究センター)

1　はじめに

clustered data とは、何らかの属性（地域、個人、種、学校など）ごとに情報が得られているデータを指し、様々な科学分野で登場する。このようなデータに対する統計解析の目標としては、cluster ごとの異質性を考慮しながら、興味のある応答変数を何らかの説明変数で説明することである。本研究では、その目標を実現するための、柔軟で比較的解釈が容易な有限混合モデルを提案する。

2　グループ化有限混合モデル

$i = 1, \ldots, m$ を cluster index とし、$j = 1, \ldots, n_i$ を within-cluster の index とする。このとき、データとして y_{ij}（応答変数）および x_{ij}（共変量）が得られているとする。目的は cluster 毎に異なる条件付き密度（確率）関数 $f_i(y|x)$ を推定することである。そのために、Sugasawa et al. (2018) では、以下のような latent mixture model を提案した。

$$f_i(y|x) = \sum_{k=1}^{L} \pi_{ik} h_k(y|x), \quad i = 1, \ldots, m.$$

ここで $\pi_i = (\pi_{i1}, \ldots, \pi_{iL})$ は cluster-dependent な混合割合であり、$h_1, \ldots, h_L$ が全ての cluster に共通する latent regression model である。上記のモデルでは、混合比によって cluster 間の潜在的な構造の異質性を表現しており、latent model を cluster 間で共通に取ることで、within-cluster のサンプル数 n_i が小さくても、$f_i(y|x)$ を安定的に推定することを可能にしている。Sugasawa et al. (2018) では π_i が共通な Dirichlet 分布に従う構造を仮定し、EM アルゴリズムによる推定方法を開発した。しかし、この方法は π_i にパラメトリックな構造を仮定していることや、結果の解釈性の問題などが欠点として挙げられる。

本研究 (Sugasawa, 2018) では、π_i にパラメトリックな構造を仮定する代わりに、以下のように、クラスターが有限個のグループに分割できる構造を想定する。

$$f_i(y|x) = \sum_{k=1}^{L} \pi_{g_i k} h_k(y|x), \quad i = 1, \ldots, m.$$

ここで、$g_i \in \{1, \ldots, G\}$ はグルーピングを表すパラメータである。このモデルは EM アルゴリズムによって容易に推定することが可能である。当日は、推定方法や、漸近的な性質および実データ解析の結果について報告する。

参考文献

- Sugasawa, S., Kobayashi, G. and Kawakubo, Y. (2018). Latent mixture modeling for clustered data. *Statistics and Computing*, to appear.
- Sugasawa, S. (2018). Grouped heterogeneous mixture modeling for clustered data. arXiv.