潜在混合密度回帰モデルによるグループデータ解析

菅澤翔之助　リスク解析戦略研究センター　特任研究員

導入

- データ解析では, 何らかの要因によってグループ分けされたデータ(グループデータ)を扱う場面が頻繁にあり, そのような場合にはグループ間の差異を考慮した解析方法を用いることが望まれる.

  グループデータではグループ内のデータ数は少ないがグループ数は比較的大きいことが一般的である.

- このようなデータに対して頻繁に用いられるモデルとして一般化線形混合モデル(GLMM)がある.

  GLMMは基本的に条件付き平均をモデル化する手法なので, 各グループの他の特徴量(quantile, expectileなど)についても興味がある場合にはあまり有効な手法であるとは言い難い.

- 各グループの密度関数を直接モデル化することができれば, あらゆる特徴量に対して柔軟な推定ができると期待できる.

  各グループ毎に対して何らかのモデルを当てはめるのは, サンプル数が小さいため不安定になる.

目的

グループデータに対して各グループの(条件付き)密度関数を安定的にかつ柔軟に推定するモデルを提案し, その有効性を検証する.

潜在混合密度回帰モデル

アイデア
全グループ共通の潜在的な条件付き密度関数$h_1, \dots, h_G$があり, 各グループの密度関数はこれらの潜在密度関数の有限混合で表現される.

設定
m個のグループにおいてそれぞれ$y_{ij}, \boldsymbol{x}_{ij}, j = 1, \dots, n_i, i = 1, \dots, m$のデータが得られている. このとき, 各グループの条件付き密度関数$f_i(y|\boldsymbol{x})$を推定したい.

モデル
以下のような密度モデルを提案する.

$$f_i(y|\boldsymbol{\pi}_i, \boldsymbol{x}, \boldsymbol{\phi}) = \sum_{g=1}^{G} \pi_{ig} h_g(y|\boldsymbol{x}, \boldsymbol{\phi}_g), \qquad \boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{iG})^t \sim \mathrm{Dir}(\boldsymbol{\alpha}). \tag{1}$$

$\mathrm{Dir}(\boldsymbol{\alpha})$: パラメータ$\boldsymbol{\alpha}$のディリクレ分布
$h_g(\cdot|\cdot, \boldsymbol{\phi}_g)$: p次元未知パラメータ$\boldsymbol{\phi}_g$をもつ潜在密度関数

- 未知パラメータは$\boldsymbol{\phi} = (\boldsymbol{\phi}_1^t, \dots, \boldsymbol{\phi}_G^t)^t, \boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_G)$の$G(p+1)$個.
- グループ毎の差異を潜在密度関数の混合割合$\boldsymbol{\pi}_i$で表現している. (random effectのようなものと解釈できる.)

  G個の潜在密度関数をグループで共通に保持することで安定的な推定が可能.

具体例
1. y_{ij}が連続値: $h_g(y|\boldsymbol{x}, \boldsymbol{\phi}_g)$を平均$\boldsymbol{x}^t\boldsymbol{\beta}_g$, 分散$\sigma_g^2$の正規分布にとる.
2. y_{ij}が離散値: $h_g(y|\boldsymbol{x}, \boldsymbol{\phi}_g)$を平均$\exp(\boldsymbol{x}^t\boldsymbol{\beta}_g)$のポアソン分布にとる.

MCEMアルゴリズムによるパラメータ推定

ラベル変数z_{ij}を導入して, モデル(1)を以下のように階層表現する.

$$f_i(y_{ij}|\boldsymbol{x}_{ij}, (z_{ij} = g)) = h_g(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\phi}_g), \quad P(z_{ij} = g|\boldsymbol{\pi}_i) = \pi_{ig}, \qquad \boldsymbol{\pi}_i \sim \mathrm{Dir}(\boldsymbol{\alpha})$$

この表現における完全尤度関数は

$$L^c(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\pi}) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \prod_{g=1}^{G} \left\{\pi_{ig} h_g(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\phi}_g)\right\}^{w_{ijg}} \prod_{i=1}^{m} p(\pi_i|\boldsymbol{\alpha}),$$

ただし$\boldsymbol{\theta} = (\boldsymbol{\phi}^t, \boldsymbol{\alpha}^t)^t$および$w_{ijg} = I(z_{ij} = g)$である. このモデルにおいてパラメータ推定を行うためのMonte Calro EM (MCEM)アルゴリズムは次のようになる.

1. 初期値$\boldsymbol{\theta}^{(0)}$を設定し, $t = 0$とする.
2. 以下の完全条件付き分布を用いてGibbs samplingによりw_{ijg}, $\boldsymbol{\pi}_i$を生成し, 期待値$E^*[w_{ijg}], E^*[\log \pi_{ig}]$を計算.

$$\Pr(w_{ijg} = 1) = \frac{\pi_{ig} h_g(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\phi}_g^{(t)})}{\sum_{l=1}^{G} \pi_{il} h_l(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\phi}_l^{(t)})}, \quad j = 1, \dots, n_i, \quad g = 1, \dots, G$$

$$\boldsymbol{\pi}_i \sim \mathrm{Dir}(\boldsymbol{\alpha}_i^{\mathrm{pos},(t)}), \quad \alpha_{ig}^{\mathrm{pos},(t)} = \alpha_g^{(t)} + \sum_{j=1}^{n_i} I(w_{ijg} = 1), \qquad g = 1, \dots, G$$

3. 前ステップで計算した$E^*[w_{ijg}], E^*[\log \pi_{ig}]$を利用して, 以下の最大化問題を解き, $\boldsymbol{\theta}^{(t+1)} = (\widehat{\boldsymbol{\phi}}^t, \widehat{\boldsymbol{\alpha}}^t)^t$と更新する.

$$\widehat{\boldsymbol{\phi}}_g = \operatorname*{argmax}_{\boldsymbol{\phi}_g} \sum_{i=1}^{m} \sum_{j=1}^{n_i} E^*[w_{ijg}] \log h_g(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\phi}_g), \qquad g = 1, \dots, G$$

$$\widehat{\boldsymbol{\alpha}} = \operatorname*{argmax}_{\boldsymbol{\alpha}} \left\{ m \log \Gamma\Big(\sum_{g=1}^{G} \alpha_g\Big) - m \sum_{g=1}^{G} \log \Gamma(\alpha_g) - \sum_{g=1}^{G} (\alpha_g - 1) \sum_{i=1}^{m} E^*[\log \pi_{ig}] \right\}.$$

4. アルゴリズムが収束していたら$\boldsymbol{\theta}^{(t+1)}$を推定値として返す. 収束していない場合はステップ2へ戻る.

潜在密度関数の個数Gの選択
応用上Gは何らかの方法で選択する必要があるが, ここではAICによる選択方法を用いる.

$$\mathrm{AIC} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} -2 \log\{f_i^m(y_{ij}|\boldsymbol{x}_{ij}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\alpha}})\} + 2(pG + G),$$

ここでf_i^mは周辺尤度である.

グループ固有の特徴量の推定

グループ固有の特徴量

$$\mu_i = \int H(t) f_i(t|\boldsymbol{x}, \boldsymbol{\phi}) dt = \sum_{g=1}^{G} \pi_{ig} \int H(t) h_g(t|\boldsymbol{x}, \boldsymbol{\phi}_g) dt$$

の推定は以下の形で行うことができる.

$$\widehat{\mu}_i = \sum_{g=1}^{G} \widehat{E[\pi_{ig}|\boldsymbol{y}_i]} \int H(t) h_g(t|\boldsymbol{x}, \widehat{\boldsymbol{\phi}}_g) dt.$$

ここで$\widehat{E[\pi_{ig}|\boldsymbol{y}_i]}$はMCEMアルゴリズムのステップ2のアウトプットから推定できる.

数値実験

データ生成過程

- $m = 50$, $n_i = 30$, $x_{ij} \sim N(0, 1)$
- $G_i \in \{1, 2, 3\}$をそれぞれ$1/3$の確率で選択し, それに応じて以下からy_{ij}を生成.

$$\begin{aligned}
&f_i(y|x) = \phi(y; 1 + x, 1), \quad (G_i = 1),\\
&f_i(y|x) = 0.5\phi(y; -1 - x, 0.5) + 0.5\phi(y; x, 1), \quad (G_i = 2)\\
&f_i(y|x) = 0.2\phi(y; -1 + 1.5x, 0.6) + 0.3\phi(y; 0.5x, 1.2) + 0.5\phi(y; 2, 0.5), \quad (G_i = 3).
\end{aligned}$$

得られたデータセットに対して提案手法(M1), 各グループごとに混合分布を当てはめるモデル(M2), グループ構造を無視して全体に混合分布を当てはめるモデル(M3)を適用.

各手法で推定された密度関数と真の密度関数とのmean integrated squared error (MISE)と, $25, 50, 75\%$分位点のmean squared error (MSE)を50回のくり返しから計算. (Figure 1.)

中央値(50%分位点)以外では提案手法が良い推定値を与えていることがわかる.

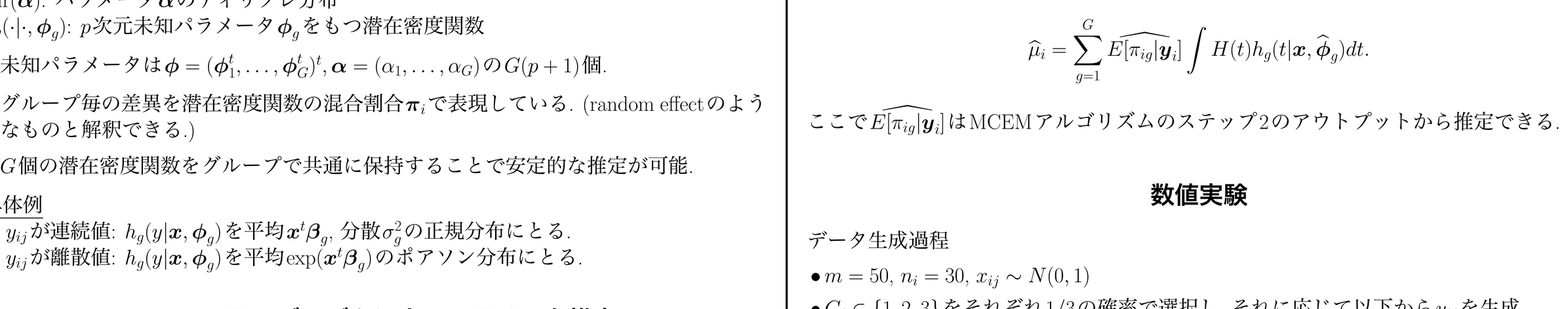

Figure 1: (A): M1に対するM2(赤)およびM3(青)のMISE比の対数値. (B)-(D): M1に対するM2(赤)およびM3(青)の分位点$(25, 50, 75\%)$のMSE比の対数値.