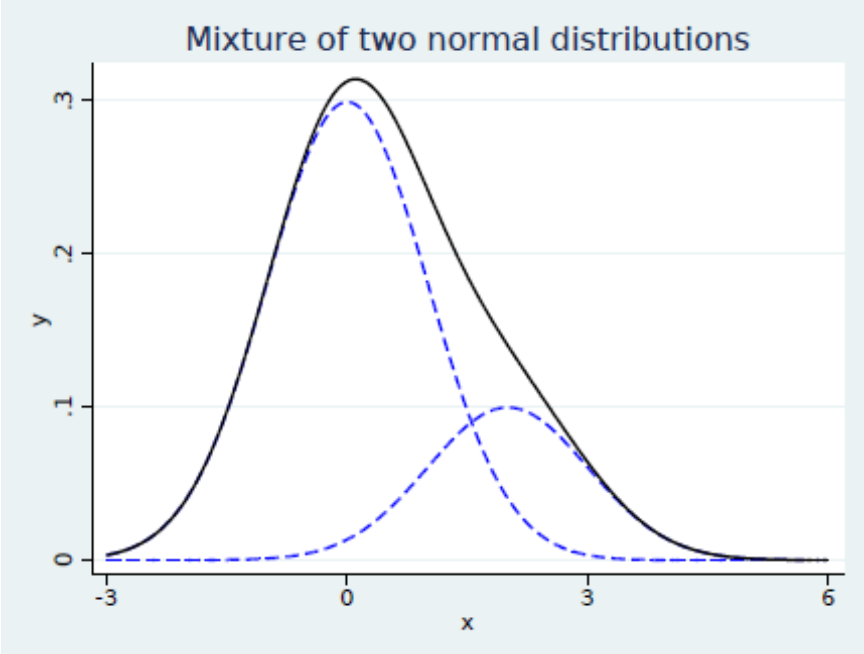
mwp-293

fmm intro - 有限混合モデルの機能概要　【 評価版 】

本 whitepaper では有限混合モデル (FMM: finite mixture models) の機能概要と基本的な用例について紹介します。

1. はじめに

有限混合モデルにおいて中心となる考え方は、観測データがいくつかの観測不能な部分集団 (unobserved subpopulations) からもたらされているとするものです。わかりやすい図にして示すと次のようになります。この図において実線は集団全体として観測された分布を、破線はその基盤をなす 2 つの部分集団の分布密度（ただし観測不能）を表しています。



観測された分布は正規分布に近いもののように見えますが、負のデータよりも正のデータの方が多めであることから、若干非対称な形状となっています。この非対称性は分布が 2 つの正規分布を混合させたものであることに由来します。すなわち右側の分布によって全体としての分布が右方向に歪められているわけです。FMM を使用すると基盤にある 2 つの分布の平均値と分散、及びそれらの分布が全体に占める割合を推定することができます。

より一般的に言うなら、FMM を使うことによって任意個数の部分集団を含む混合状態をモデル化することができるようになります。その場合、部分集団の分布は正規分布でなくても構いません。FMM は線形回帰モデルと一般化線形回帰モデル — 2 値/順序/多項/カウント型応答モデルを含む — の混合を許容すると共に、部分集団に固有の効果をもたらす共変量の存在も許容します。結果としてそれぞれの部分集団に関する推論が可能になると共に、個々の観測データがどの部分集団に属するかを推論することも可能になります。

FMM は高度な柔軟性を有するが故に、観測データの分類やクラスタリングに伴う調整、観測不能な分散不均一性のモデル化といった目的のために、種々の分野で幅広く利用されています。また、等しい分散を持った正規分布を混合させることによって任意の連続分布を近似することができるわけですが、これによって FMM は多峰性の (multimodal) データや歪みを持った (skewed) データ、あるいは非対称性を持ったデータをモデル化する際の有用なツールという位置付けを確保しています。具体的な適用事例については次のような文献を参照ください。

- Jorgensen (2004) - インタネットトラフィックのクラスタリング
- Deb and Trivedi (1997) - 医療サービスの需要
- Schlattmann, Dietz, and Böhning (1996) - 疾病のリスク
- Wedel and DeSarbo (1993) - 消費者リスク
- Jones et al. (2013) - カウント型アウトカム
- McLachlan and Peel (2000), Frühwirth-Schnatter (2006) - 有限混合モデル解説

統計学の視点から言うなら FMM は潜在クラス分析 (LCA: latent class analysis) と関連があると言えます。双方とも顕在（観測）変数からの情報に基づきクラスを特定しようとするものだからです。FMM の場合、単一の従属変数に対する回帰モデル中のパラメータがクラスごとに異なることを許容するのに対し、伝統的な LCA においては複数の従属変数に対して切片項のみのモデルをフィットさせるといった違いがあります。FMM は潜在変数 (latent variable) がカテゴリカルなものであることを前提としたときの構造方程式モデル (SEM: structural equation modeling) であるともいえます。詳細については [SEM] **intro 1** (*mwp-209*), [SEM] **intro 2**, [SEM] **gsem**, Skrondal and Rabe-Hesketh (2004) を参照ください。一方、潜在変数が連続型で顕在変数 (manifest variables) が離散型の場合には項目応答理論 (IRT: item response theory) モデルを用いることができます（[IRT] **irt** (*mwp-238*) 参照)。また潜在変数と顕在変数の双方が連続型である場合には SEM を用いることになります。

マニュアル中では次のような用語が使用されます。

- “クラス”、“グループ”、“タイプ”、“成分” - 観測されない部分集団の意
- “クラス確率”、“成分確率” - 混合中のある成分に帰属する確率

なお、クラス確率については“混合重み”、“混合比率”という用語を使用している文献もあります。

2. 有限混合モデル

FMM は複数の確率密度関数を結合した確率的モデルです。FMM の場合、観測される応答 $\boldsymbol{y}$ は g 種類の異なるクラス $f_1, f_2, \ldots, f_g$ からそれぞれ $\pi_1, \pi_2, \ldots, \pi_g$ の比率でもたらされるものと仮定されます。最も単純な形式を想定するなら g 個の成分からなる混合モデルの密度関数は

$$f(\boldsymbol{y}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y} \,|\, \boldsymbol{x}'\boldsymbol{\beta}_i)$$

のように書けます。ただし π_i は i 番目のクラスに対する確率 $(0 \leq \pi_i \leq 1, \sum \pi_i = 1)$ を、$f_i(\cdot)$ は i 番目のクラスのモデルにおける応答変数の条件付き確率密度関数を意味します。

一方、`fmm` は潜在クラスに対する確率を多項ロジスティック分布によってモデル化します。すなわち i 番目の潜在クラスに対する確率は

$$\pi_i = \frac{\exp(\gamma_i)}{\sum_{j=1}^{g} \exp(\gamma_j)}$$

のように表現されます。ただし γ_i は i 番目の潜在クラスに対する線形予測値を意味します。デフォルトの場合、1 番目の潜在クラスはベースレベルとみなされるため、$\gamma_1 = 0, \exp(\gamma_1) = 1$ という扱いになります。

尤度はそれぞれの潜在クラスからの条件付き尤度に確率による重み付けをしたものの和という形で計算されます。詳細については `[FMM]` **fmm** のセクション“*Methods and formulas*”を参照ください。

3. FMM の用例 - 正規分布の混合

評価版では割愛しています。

4. LCA フレームワークとの関連

評価版では割愛しています。

■