

関数型説明変数を伴う混合効果モデルにおける平滑化パラメータの選択

大阪府立大学 大学院工学研究科 道家悠太
大阪府立大学 大学院工学研究科 林 利治

1 はじめに

本研究の目的は、線形混合効果モデル (LME) を一般化した関数型線形混合効果モデル (FLMM) のパラメータ推定を行う際に、データに応じた適応的な平滑化パラメータの選択を可能にすることである。

2 関数型線形混合効果モデル (FLMM)

線形混合効果モデル (LME) を

$$Y_i = (\alpha + \gamma_i) + (\beta + b_i)X_i + \varepsilon_i \quad (i = 1, 2, \dots, n)$$

と定める (Wu and Zhang [4], 船渡川, 船渡川 [5])。 α, β は固定効果、また γ_i, b_i は変量効果である。

LME の説明変数が共変量 $t \in [0, T]$ の関数なら

$$Y_i = (\alpha + \gamma_i) + \int_0^T (\beta(t) + b_i(t))X_i(t) dt + \varepsilon_i$$

とモデル化できる。これは関数型線形混合効果モデル (functional linear mixed effect model, FLMM) と呼ばれる。推定対象は $\alpha, \gamma_i, \beta(t), b_i(t)$ や ε_i の分散である。

3 FLMM のパラメータ推定

FLMM のパラメータ推定では REML-based EM を用いる (Liu et al. [2])。

FLMM では、関数型の勾配 $\beta(t), b_i(t)$ を推定する際、データへの過度な適合を避けるため罰則を課す。罰則の程度は平滑化パラメータ (λ_β, λ_b) により定められ、その選択は、勾配関数やその他のモデルパラメータの推定精度に大きな影響を与える。

4 平滑化パラメータの選択方法

本研究では、平滑化パラメータを選択する 2 種類の方法を比較する。一方は BIC 基準を用いて (λ_β, λ_b) を同時に選択する方法である。この選択方法は著者が本研究にて推奨している手法である。もう一方は GCV と REML を用いて個別に選択する方法である (Liu et al. [2])。それぞれを選択方法 (A), (B) と呼ぶことにする。

LME の形式で書けるモデルに対しては、最適な平滑化パラメータを選択する方法として、(B) に相当する GCV と REML がよく用いられてきた (Reiss and Odgen [3], Aydin and Memmedli [1], Liu et al. [2])。しかし FLMM において、それらの規準では平滑化パラメータをうまく選択できない。そこで、FLMM のモデルパラメータの推定結果を基に、平滑化パラメータの選択方法 (B) よりも (A) の方が推奨されることを例示する。

5 平滑化パラメータの選択方法の比較実験

5.1 シミュレーションの設定

シミュレーションで用いた FLMM は以下である (Liu et al. [2] 参照)。

$$Y_{ij} = (\alpha + \gamma_i) + \int_0^T (\beta(t) + b_i(t))X_{ij}(t) dt + \varepsilon_{ij}$$

ただし $j = 1, 2, \dots, m_i$ は第 i 観測群における観測の繰り返しを表す。 $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ とした (σ_ε^2 は未知)。

返しを表す。 $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ とした (σ_ε^2 は未知)。

5.2 選択された平滑化パラメータ

選択方法 (A), (B) を比較するために、200 回の実験中に方法 (A), (B) で選択された平滑化パラメータ λ_β の値の分布を以下の表 1 にまとめた。

表 1 $\log(\lambda_\beta)$ の値

	-16	-14	-12	-10	-8	-6	-4	-2	0
(A)	2	1		8	149	38	2		
(B)	41	2	3	16	48				90

表 1 において注目すべき箇所は、各選択方法が選択した割合が高かったパラメータ値 $\log(\lambda_\beta) = -8, 0$ に対応する 2 列である。 $\lambda_\beta = \exp(-8)$, $\exp(0)$ が選ばれたときに推定された勾配関数 $\hat{\beta}(t)$ (図 1, 2) を見ると、 $\lambda_\beta = \exp(-8)$ を選択すれば適切な平滑化が行えるが、 $\lambda_\beta = \exp(0)$ を選択すれば過度な平滑化を行ってしまうことが視覚的にわかる。

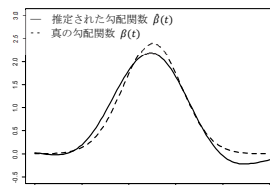


図 1 $\hat{\beta}(t)$ with $\lambda_\beta = \exp(-8)$

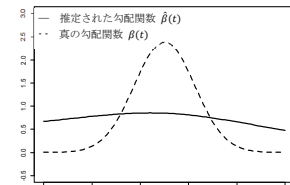


図 2 $\hat{\beta}(t)$ with $\lambda_\beta = \exp(0)$

5.3 勾配関数の推定精度の比較

選択された平滑化パラメータの割合の違いは、 $\beta(t)$ の推定誤差 (RMISE) にも表れている。過度な平滑化を行う値を選択する割合が高い選択方法 (B) では、RMISE は 0.5090 であるのに対して、適切な平滑化が行える値を選択する割合が高い選択方法 (A) では、RMISE が 0.2467 であり、約 50% 推定精度が改善した。

6. 結果のまとめ

結果として、BIC 基準を用いて同時に (λ_β, λ_b) を選択する方が、GCV や REML を用いて個別に選択するよりも、より適切な平滑化を行えることを例示できた (表 1, 図 1, 2 参照)。また、勾配関数の推定精度が上昇したことも示せた。

発表では、 λ_b の選択や関数型説明変数 $X_{ij}(t)$ を離散的な観測から近似する方法についても述べる。

参考文献

- [1] Aydin, D. and Memmedli, M. (2011). *Optimization* **61**, 459–476.
- [2] Liu, B., Wang, L. and Cao, J. (2017). *Computat. Statistics and Data Analysis* **106**, 153–164.
- [3] Reiss, P.T. and Ogden, R. T. (2008). *Journal of the Royal Statistical Society* **71**, 505–523.
- [4] Wu, H. and Zhang, J.T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis*, Wiley.
- [5] 船渡川 伊久子, 船渡川 隆 (2015). 統計解析スタンダード 経時データ解析 朝倉書店.