

多項ロジットモデルの適合度検定のためのマルコフ基底の計算

数理情報学専攻 48176218 濱田 将樹

指導教員 清 智也 准教授

1 はじめに

多項ロジットモデル [6] は、離散選択実験において用いられる最も代表的なモデルの 1 つである。離散選択実験とは、複数の選択肢の中からどれを選ぶかという人々の選択行動を分析するための実験である [7]。実験で得られる観測データは必ずしも多項ロジットモデルに適合するとは限らないため、モデルの適合度検定を行うことは重要である。適合度検定の p 値の計算手法の 1 つにマルコフ基底を用いた MCMC があり、サンプルサイズが大きくなり、かつ観測値に偏りがある場合には有効な手法である [3, 4]。

本研究では、選択肢がただ 1 つの属性をもつ場合には多項ロジットモデルのマルコフ基底が陽に求められることを示した。さらに、適合度検定の p 値を求めるためにマルコフ基底を列挙せずに MCMC を行うアルゴリズムを提案し、既存の計算手法との比較を行った。

2 多項ロジットモデル

本研究で扱う多項ロジットモデルを記述する。チョイスセットとは実験において回答者に提示する選択肢の集合であり、これが S 個あるとする。各チョイスセット $s = 1, \dots, S$ は J 個の選択肢 $j = 1, \dots, J$ を含み、回答者は選択肢 $j = 1, \dots, J$ の中から最も好ましいものを選択する。チョイスセット s の中の選択肢 j が選ばれる確率 p_{js} は、選択肢がもつ属性を並べたベクトル $\mathbf{x}_{js}$ と各属性が選択確率に与える寄与を並べたパラメータ $\boldsymbol{\beta}$ を用いて

$$p_{js} = \frac{\exp(\mathbf{x}'_{js}\boldsymbol{\beta})}{\sum_{k=1}^J \exp(\mathbf{x}'_{ks}\boldsymbol{\beta})}$$

とモデル化される。実験によって得られる度数データ $\mathbf{y} = (y_{11}, \dots, y_{js}, \dots, y_{JS})'$ は、 y_{js} がチョイスセット s の中の選択肢 j を選んだ人数を表す。

3 マルコフ基底

モデルの適合度検定の p 値の計算のために必要となるマルコフ基底の定義をする。モデルの十分統計量 $\mathbf{t}$ が、度数 $\mathbf{y} \in \mathbb{N}^n$ と行列 A を用いて $\mathbf{t} = A\mathbf{y}$ と表されるとする。

定義 3.1. $\mathcal{B}$ を $\ker A \cap \mathbb{Z}^n$ の部分集合とする。 $\mathcal{B}$ がマルコフ基底であるとは、任意の十分統計量 $\mathbf{t}$ と任意の $\mathbf{y}, \tilde{\mathbf{y}} \in \mathbb{N}^n$ s.t. $\mathbf{t} = A\mathbf{y} = A\tilde{\mathbf{y}}$ に対して、ある $N \in \mathbb{N}$, $\mathbf{z}_j \in \mathcal{B}$, $\varepsilon_j \in \{-1, 1\}$, $j = 1, \dots, N$ が存在して、次の 2 つの条件

$$\begin{aligned} \tilde{\mathbf{y}} &= \mathbf{y} + \sum_{j=1}^N \varepsilon_j \mathbf{z}_j, \\ \mathbf{y} + \sum_{j=1}^n \varepsilon_j \mathbf{z}_j &\geq \mathbf{0}, \quad n = 1, \dots, N \end{aligned}$$

を満たすことである。

マルコフ基底 $\mathcal{B}$ が求められれば、連鎖の各状態 $\mathbf{y}$ において $\mathbf{z} \in \mathcal{B}$ と $\varepsilon \in \{-1, 1\}$ をランダムに選び、 $\mathbf{y} + \varepsilon \mathbf{z} \geq \mathbf{0}$ ならば $\mathbf{y} + \varepsilon \mathbf{z}$ に移動し、そうでなければ $\mathbf{y}$ にとどまるというルールでマルコフ連鎖を構成することにより、適合度検定の p 値を計算することができる。

4 本研究の成果

4.1 属性数 1 の多項ロジットモデルのマルコフ基底

選択肢がただ 1 つの属性をもつときには、離散選択実験によって得られる度数データを構造的ゼロのある 2 元分割表と見なすことができる。分割表の行はただ 1 つの属性の水準を表し、列はチョイスセットを表す。 ℓ 行 s 列のセル (ℓ, s) の値は、チョイスセット s の中に水準 ℓ の選択肢が含まれればその選択肢を選んだ人数であり、そうでなければ構造的ゼロである。

本研究ではこのように作った 2 元分割表のセル確率が多項ロジットモデルに従うことと、2 元分割表の準独立モデルに従うことが等価であることを示した。2 元分割表の準独立モデルのマルコフ基底は [1] によって求められているため、この結果を用いることで属性数 1 の多項ロジットモデルのマルコフ基底を記述することができる。

定義 4.1 ([1]). $1 \leq \ell_1, \dots, \ell_r \leq L$, $1 \leq s_1, \dots, s_r \leq S$ に対して、次数 r のループとは $L \times S$ 整数行列 $M_r(\ell_1, \dots, \ell_r; s_1, \dots, s_r)$ であり、その要素は

$$\begin{aligned} m_{\ell_1 s_1} &= m_{\ell_2 s_2} = \dots = m_{\ell_{r-1} s_{r-1}} = m_{\ell_r s_r} = 1, \\ m_{\ell_1 s_2} &= m_{\ell_2 s_3} = \dots = m_{\ell_{r-1} s_r} = m_{\ell_r s_1} = -1, \end{aligned}$$

その他の要素はすべて 0 である。ただし全ての $n \neq n'$

に対して $\ell_n \neq \ell_{n'}$ かつ $s_n \neq s_{n'}$ である。

定義 4.2 ([1]). ループ $M_r(\ell_1, \dots, \ell_r; s_1, \dots, s_r)$ が自由度 1 であるとは, $L \times S$ 分割表のセルの集合 $\{(\ell, s) \mid \ell \in \{\ell_1, \dots, \ell_r\}, s \in \{s_1, \dots, s_r\}\}$ が次数 $2, \dots, r-1$ のいかなるループのサポートも含まないことである。

定理 4.3. 属性数 1 の多項ロジットモデルの水準数を L , チョイスセットの個数を S とする. これに対して上記のようにして作った $L \times S$ 分割表を考える. 属性数 1 の多項ロジットモデルの一意的極小マルコフ基底はこの 2 元分割表における次数 $2, \dots, \min\{L, S\}$ の自由度 1 の全てのループからなる集合である。

4.2 属性数 1 の多項ロジットモデルの適合度検定に対する提案手法

本研究では属性数 1 の多項ロジットモデルに対して, マルコフ基底を列挙することなく MCMC を構成することで適合度検定の p 値を推定する手法を提案した. 提案手法の概要は, ループ $M_r(i_1, \dots, i_r; s_1, \dots, s_r)$ をランダムに生成することで MCMC を構成するというものである. この手法はマルコフ基底を列挙する必要がないので, マルコフ基底の要素数が膨大になる場合でもメモリ不足にならない。

5 数値実験

以下の 6 つの手法を用いて属性数 1 の多項ロジットモデルの p 値を推定し, その結果を比較する。

1. 提案手法による MCMC(proposed)
2. マルコフ基底を用いた MCMC(makrov)
3. 格子基底を用いた MCMC(lattice)
4. Sequential Importance Sampling を用いたモンテカルロ法 (SIS)
5. カイ二乗検定統計量の漸近分布 (chisq)
6. 尤度比検定統計量の漸近分布 (LR)

このうち, 3. の lattice, 4. の SIS はそれぞれ [5], [2] により提案されたアルゴリズムである. 5. の chisq, 6. の LR はそれぞれカイ二乗検定統計量, 尤度比検定統計量の漸近分布に基づく p 値の計算である. 実験の設定は, 選択肢の個数 $J = 3$, チョイスセットの個数 $S = 7$, 水準数 $L = 5$ とし, チョイスセットは

$$\{1, 2, 3\}, \{1, 2, 5\}, \{1, 3, 4\}, \{1, 3, 5\}, \\ \{2, 3, 4\}, \{2, 4, 5\}, \{3, 4, 5\}$$

の 7 つである. 回答者の人数 (各チョイスセットの観測度数の和) を $N = 3$ とし, 得られた観測度数は

$$\mathbf{y} = (0, 2, 1, 1, 1, 1, 0, 2, 1, 1, 0, 2, 0, 3, 0, 0, 1, 2, 0, 2, 1)'$$

とする ($J \times S = 21$ 次元ベクトル)。

図 1 に, 6 つの手法による p 値の推定値と正確な p 値との絶対誤差を示す. 1.~4. のモンテカルロ法は, 1 回のモンテカルロ法につき 10 秒間の計算を行い, それを 100 回行った下での平均と標準偏差を示した. 図より,

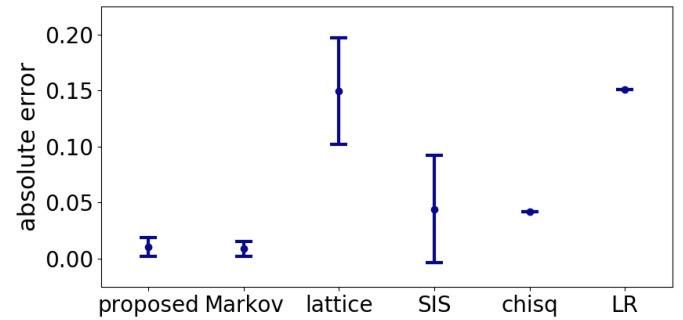


図 1. p 値の推定値と正確な p 値の絶対誤差. モンテカルロ法については 10 秒間の計算を 100 回行い, その平均と標準偏差をプロット。

proposed と markov が最も正確な p 値に近い値を推定できている。

proposed と markov は推定精度の観点からはほぼ同じ性能であるが, proposed はマルコフ基底を列挙する必要がないので, markov に比べ空間計算量を削減することができる点である。

参考文献

- [1] S. Aoki and A. Takemura. Markov chain Monte Carlo exact tests for incomplete two-way contingency tables. *Journal of Statistical Computation and Simulation*, Vol. 75, pp. 787–812, 2005.
- [2] Y. Chen, P. Diaconis, S. P. Holmes, and J. S. Liu. Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, Vol. 100, pp. 109–120, 2005.
- [3] JST CREST 日比チーム. グレブナー道場. 共立出版, 2011.
- [4] P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, Vol. 26, pp. 363–397, 1998.
- [5] H. Hara, S. Aoki, and A. Takemura. Running markov chain without markov basis. In *Harmony of Gröbner Bases and the Modern Industrial Society*, pp. 45–62. 2012.
- [6] D. McFadden. Conditional logit analysis of qualitative choice behavior, In *Frontiers in Econometrics*, pp. 105–142. 1974.
- [7] K. E. Train. Discrete Choice Methods with Simulation. *Cambridge University Press*, 2003.