ベイズセミパラメトリックモデルを用いた因果の方向性の検討

宮崎　慧 [1]　星野　崇宏 [2]　繁桝　算男 [1]
[1] 東京大学大学院　総合文化研究科　[2] 名古屋大学大学院　経済学研究科

1　はじめに

近年パラメータに任意の形状の分布を仮定した上でベイズ MCMC 推定を可能にするディリクレ過程事前分布 (Ferguson, 1973) が注目されている．また一般に誤差変数の正規性の仮定を除くと，パスの方向性が反対のモデルが同値モデルではなくなり，モデル比較が可能になる．そこで本研究では，誤差項が非正規という仮定のもとで原因変数と結果変数を入れ替えたモデルをディリクレ過程を用いて表現し，シミュレーション研究により，周辺尤度を用いたパスの方向性に関するモデル選択の性能を評価した．

2　モデル

データは同一であるが，説明変数と従属変数を入れ替えた 2 つのシンプルな単回帰モデル

$$y_i = \beta x_i + e_i \quad , \quad x_i = \beta' y_i + e_i' \tag{1}$$

を考える．誤差変数は非正規であることを仮定する．

ディリクレ過程について，Ishwaran & Zarepour (2000) は L 次元の有限ディリクレ過程事前分布を用いて，確率変数 $\boldsymbol{Y}$ の分布を

$$\boldsymbol{Y} \ \sim \ \sum_{l=1}^{L} p_l f(\cdot|\theta_l) \tag{2}$$

とし，L が大のときには十分ディリクレ過程事前分布を近似することを示した．ここで，$p_l = \prod_{k=1}^{l-1}(1-V_k)V_l$，$V_1, V_2, \cdots,$ は独立な $Be(1,\alpha)$ 確率変数である．また Ishwaran & James (2001) はこの有限ディリクレ過程事前分布を用いたモデルでのパラメータの事後分布を求めるためのアルゴリズムとして Blocked Gibbs Sampler を提唱した．

事前分布について，本研究では誤差項に有限ディリクレ過程事前分布を設定する．

$$\boldsymbol{e} \quad \sim \quad DP_L(\gamma, N(\boldsymbol{\mu}, \boldsymbol{\sigma})) \tag{3}$$

$$\beta \quad \sim \quad N(\beta_0, \sigma_\beta^2) \tag{4}$$

DP_L は L 次元の有限ディリクレ過程を表す．各ハイパーパラメータは固定．γ は他のコンポーネントへの遷移のし易さを表すパラメータである．

3　アルゴリズムおよびモデル選択

本研究では Ishwaran & James (2001) により提案された Blocked Gibbs Sampler アルゴリズムを用いてパラメータ推定を行った．以下，$\cdots$ は他のパラメータの意味で用いる．

$$p(\beta|\cdots) \ \propto \ \prod_{i}^{N} p(y_i|k_i = l, \beta, \mu_l, \sigma_l) \times p(\beta) \tag{5}$$

$$p(\mu_l|\cdots) \ \propto \ \prod_{i}^{N} p(y_i|k_i = l, \mu_l, \sigma_l^2, \beta) p(\mu_l) \tag{6}$$

$$p(\sigma_l|\cdots) \ \propto \ \prod_{i}^{N} p(y_i|k_i = l, \mu_l, \sigma_l^2, \beta) p(\sigma_l^2) \tag{7}$$

各サンプルが所属するコンポーネント $\boldsymbol{k}$ の発生

$$p(k_i|\cdots) \ \sim \ \sum_{l=1}^{L} \pi_{li}\delta_l(\cdot) \quad , \quad \pi_{li} = \frac{p_l \sigma_l^{-1} \exp\left[-\frac{1}{2\sigma_l^2}(e_i - \mu_l)^2\right]}{\sum_l^L p_l \sigma_l^{-1} \exp\left[-\frac{1}{2\sigma_l^2}(e_i - \mu_l)^2\right]} \tag{8}$$

$\delta_l(\cdot)$ は l 上でのみ 1 の値を取る測度である．
サンプルの各コンポーネントへの所属確率 $\boldsymbol{p}$ の発生：$\boldsymbol{p}$ の完全条件付分布は以下にある一般ディリクレ分布から発生する．

$$p_l = \prod_{m=1}^{l-1}(1-V_m)V_l \quad , \quad V_l \sim \mathrm{Beta}(a_l + M_l, b_l + \sum_{m=l+1}^{L} M_m) \tag{9}$$

M_l は l 番目のコンポーネントに所属するサンプル数である．
Chib (1995) によりギブスサンプラーで得られた標本を用いた周辺尤度計算方法が提案されている．またディリクレ過程混合モデリングを用いた際の周辺尤度計算法が Basu & Chib (2003) により提案されている．本研究では誤差変数にディリクレ過程を仮定しているため，Basu & Chib の方法を用いて周辺尤度の計算を行った．

4 シミュレーション

まず $\boldsymbol{x}$ を説明変数，$\boldsymbol{y}$ を従属変数とし，データを発生した．誤差項はコンポーネント数 2 の正規混合分布から発生した．真値は $\boldsymbol{\mu} = (-2.0, 2.0)^t$, $\boldsymbol{\sigma} = (0.5, 2.0)^t$ である．ハイパーパラメータは $\beta_0 = 0, \sigma_\beta^2 = 1000, u_{0,l} = 0, V_{0,l} = 100, f_{0,l} = 2, G_{0,l} = 1.0$ に設定し，サンプル数は 200 とした．回帰係数の真値は 2.0 とした．さらに $\boldsymbol{x}$ について正規なデータの場合と非正規なデータの場合に分け，それぞれに対し説明変数と従属変数を入れ替えて解析を行った．$\boldsymbol{x}$ が正規データのときは標準正規分布から，非正規データにするときは標準正規分布から発生した変数 $\boldsymbol{v} = (v_1, \cdots, v_N)^t$ について以下のような変換を施した．

$$x_i = \begin{cases} \log(v_i) & \text{if } (v_i > 0) \\ -\log(|v_i|) & \text{if } (v_i < 0) \end{cases} \tag{10}$$

サンプリング回数を 5000 回とし，最初の 2000 回を Burn-in とした．50 セットデータ発生し，解析したのち，周辺尤度を計算した．結果は表 1 に掲載した．特に説明変数を 3 乗した非線形回帰モデルから発生したデータに対し，パスの方向性についてモデル選択を行うと，正しいパスを選択する回数は 2 割に満たないことが分かる．そもそもパスの方向性の決定と因果の方向性の決定は本質的に別次元の問題であるにも関わらず，誤差変数が非正規であることを利用してパスの方向性から因果関係を断定しようとすると，大きな誤謬を招く可能性があることを本結果は示唆している．

表 1: 周辺尤度によるモデル選択の結果（括弧内は選択された回数を表す）

データ発生のモデル	$y_i = \beta x_i + e_i$（$\boldsymbol{x}$ は正規）	$y_i = \beta x_i + e_i$（$\boldsymbol{x}$ は非正規）	$y_i = \beta x_i^3 + e_i$
モデルのペア 1	$y_i = \beta_{111} x_i + e_i$ (8) $x_i = \beta_{112} y_i + e_i'$ (42)	$y_i = \beta_{121} x_i + e_i$ (16) $x_i = \beta_{122} y_i + e_i'$ (34)	$y_i = \beta_{131} x_i + e_i$ (8) $x_i = \beta_{132} y_i + e_i'$ (42)
モデルのペア 2	$y_i = \beta_{211} x_i^3 + e_i$ (0) $x_i = \beta_{212} y_i^{1/3} + e_i'$ (50)	$y_i = \beta_{221} x_i^3 + e_i$ (0) $x_i = \beta_{222} y_i^{1/3} + e_i'$ (50)	$y_i = \beta_{231} x_i^3 + e_i$ (0) $x_i = \beta_{232} y_i^{1/3} + e_i'$ (50)

参考文献

Basu, S. & Chib, S. (2003). Marginal likelihood and Bayes factors for Dirichlet Process mixture models. *Journal of the American Statistical Association*, **98**, 224-235.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the Americal Statistical Association*, **90**, 1313-1321.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209-230.

Ishwaran, H. & James, L.F. (2001). Gibbs sampling methods for Stick-Breaking priors. *Journal of the American Statistical Association*, **96**, 161-173.

Ishwaran, H. & Zarepour, M. (2000). Markov Chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, **87**, 371-390.