Limited Dependent Variables を含む連立方程式モデル系のベイズ推定*1

Bayesian Estimation of Simultaneous Equations System with Limited Dependent Variables*1

岡田直也*2・菊池輝*3・北村隆一*4

*By Naoya Okada*2・Akira Kikuchi*3・and Ryuichi Kitamura*4*

1.はじめに

通勤時の経路選択や出発時刻の選択など繰り返しの交通行動の分析が必要となる場面は多い．繰り返し行動を分析する際，多く用いられるのが連立方程式モデルである．しかしながら連立方程式モデルを用いて未知パラメータを推定する際には，問題点が存在する．まず扱う時点数が増加するにつれ，推定の際に複雑な積分計算が必要となり計算自体が困難となる．また従属変数間に相互依存関係が存在する場合にも，推定は困難となる．その様な問題点を克服した推定手法としてベイズ推定が挙げられる．ベイズ推定手法では，MCMC 法を用いることで複雑な積分計算を不要としており，近年様々な分野で用いられている．

しかしベイズ推定手法にも，問題点は存在する．観測データにおいて，その実現値が観測不可能な場合にはその推定の実行自体が不可能なことである．そのため選択効用の様に，従属変数の値が正負しか観測されていないといった場合等には，ベイズ推定を行うことが出来ない．

本研究では，上記の問題点に対して，既存のベイズ推定手法に加えて潜在変数をシミュレートする過程を設けた手法を提案する．

また simulated data を用いて推定を行うことによって，本研究で提案する推定手法についてその妥当性の検証を行う．

2.推定手法

以下に本研究で提案する推定アルゴリズムを示す．$\mathbf{y}_i$ を内生変数ベクトル，$\mathbf{X}_i$ を外生変数行列，$\mathbf{B}$ と $\boldsymbol{\gamma}$ をそれぞれ係数行列，係数ベクトル，$\boldsymbol{\varepsilon}_i$ を誤差項ベクトルとすると，同時方程式モデルは一般に，以下の式で表される．

$$\mathbf{y}_i = \mathbf{B}\mathbf{y}_i + \mathbf{X}_i\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\varepsilon}_i \sim \text{i.i.d.}\, N(0,\boldsymbol{\Sigma}) \quad (1)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \text{L} & \sigma_{1g} \\ \text{M} & \text{O} & \text{M} \\ \sigma_{g1} & \text{L} & \sigma_{gg} \end{pmatrix}$$

また，g は方程式の数，k は説明変数の数，n をケースの数とし，i は1～n の値をとるものとする．

本研究で提案する推定手法を示すに当たり，まず既存の推定手法について触れる．以降では簡単のため $\mathbf{B}=\mathbf{0}$ とし，右辺に内生変数を含まないSUR（Seemingly Unrelated Regressions）モデルであるとする．SURモデルは，FIMLを用いた場合，同時方程式モデルの推定と一致するという特徴がある．このとき，SURモデルの尤度関数は次式で表される．

$$f(\mathbf{y}\,|\,\boldsymbol{\gamma},\boldsymbol{\Sigma}) \propto \prod_{i=1}^{n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\gamma})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\gamma})\right\} \quad (3)$$

ここで未知パラメータ $(\boldsymbol{\gamma},\boldsymbol{\Sigma})$ の事前分布を，

$$\boldsymbol{\gamma} \sim N(\mathbf{c}_0,\mathbf{C}_0),\ \boldsymbol{\Sigma} \sim IW(n_0,\mathbf{S}_0) \quad (4)$$

とする．ここに，$\mathbf{c}_0$ は $gk\times 1$ のベクトル，$\mathbf{C}_0$ は $gk\times gk$ の行列，IW は $(n_0,\mathbf{S}_0)$ をパラメータとする逆ウィッシャート分布を表し，$\mathbf{S}_0$ は $g\times g$ の行列である．逆ウィッシャート分布の密度関数は，

$$h(\boldsymbol{\Sigma}\,|\,n_0,\mathbf{S}_0) \propto |\boldsymbol{\Sigma}|^{-\frac{n_0+g+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{S}_0^{-1}\boldsymbol{\Sigma}^{-1}\right)\right\} \quad (5)$$

で与えられる．このとき，事後分布は，以下のように導かれる．

---

*1 キーワード：ベイズ推定, simulated data

*2学生員, 京都大学大学院工学研究科都市社会工学専攻

*3 正員, 工博, 京都大学大学院工学研究科都市社会工学専攻

*4正員, Ph.D., 京都大学大学院工学研究科都市社会工学専攻

（京都市西京区京都大学桂4-C1-2,

TEL 075-383-3240, FAX 075-383-3236）

$$\boldsymbol{\gamma} \mid \boldsymbol{\Sigma}, \mathbf{y} \sim N(\mathbf{c}_1, \mathbf{C}_1), \ \boldsymbol{\Sigma} \mid \boldsymbol{\gamma}, \mathbf{y} \sim IW(n_1, \mathbf{S}_1) \quad (6)$$

$$\mathbf{c}_1 = \mathbf{C}_1\left(\mathbf{C}_0^{-1}\mathbf{c}_0 + \sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{y}_i\right)$$

$$\mathbf{C}_1^{-1} = \mathbf{C}_0^{-1} + \sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i$$

$$n_1 = n_0 + n$$

$$\mathbf{S}_1^{-1} = \mathbf{S}_0^{-1} + \sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\gamma})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\gamma})'$$

以上のようにパラメータの事後分布を得ることができた．この事後分布から確率標本をサンプリングする，つまり乱数を発生させることができれば，得られた標本の収束値を未知パラメータの推定値とみなせる．事後分布から確率標本を得る有効な方法として，マルコフ連鎖モンテカルロ法（MCMC法）があるが，本研究ではMCMC法の1つであるギブス・サンプリングを用いることとする．以下に式（6）で与えられた事後分布に即して，ギブス・サンプリングを行う手順を示す．

1) 初期値を以下のように選定し，$t = 1$とする．

$$(\boldsymbol{\gamma}^{(0)}, \boldsymbol{\Sigma}^{(0)}) = (\gamma_{11}^{(0)}, \gamma_{12}^{(0)}, \mathrm{L}, \gamma_{gk}^{(0)}; \sigma_{11}^{(0)}, \sigma_{12}^{(0)}, \mathrm{L}, \sigma_{gg}^{(0)})$$

2) $\boldsymbol{\Sigma}^{(t-1)}$に基づいて，$(\mathbf{c}_1^{(t)}, \mathbf{C}_1^{(t)})$を次のように更新し，得られる多変量正規分布から$\boldsymbol{\gamma}^{(t)}$をサンプリングする．

$$\mathbf{c}_1^{(t)} = \mathbf{C}_1^{(t)}\left(\mathbf{C}_0^{-1}\mathbf{c}_0 + \sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{(t-1)^{-1}}\mathbf{y}_i\right) \quad (7)$$

$$\mathbf{C}_1^{(t)^{-1}} = \mathbf{C}_0^{-1} + \sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Sigma}^{(t-1)^{-1}}\mathbf{X}_i \quad (8)$$

3) 得られた$\boldsymbol{\gamma}^{(t)}$を用いて，逆ウィッシャート分布のパラメータを次のように更新する．

$$\mathbf{S}_1^{(t)^{-1}} = \mathbf{S}_0^{-1} + \sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\gamma}^{(t)})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\gamma}^{(t)})' \quad (9)$$

このとき$\boldsymbol{\Sigma}^{(t)}$の密度関数は

$$h(\boldsymbol{\Sigma}^{(t)} \mid n_1, \mathbf{S}_1^{(t)}) \propto \mid \boldsymbol{\Sigma} \mid^{-\frac{n_1+g+1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\mathbf{S}_1^{(t)^{-1}}\boldsymbol{\Sigma}^{(t)^{-1}}\right)\right\} \quad (10)$$

と与えられる．$\boldsymbol{\Sigma}^{(t)}$のサンプリングにあたっては，逆ウィッシャート分布ではなく，密度関数

$$h(\mathbf{Z}^{(t)} \mid n_1, \mathbf{S}_1^{(t)}) \propto \mid \boldsymbol{\Sigma} \mid^{\frac{n_1-g-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\mathbf{S}_1^{(t)}\mathbf{Z}^{(t)}\right)\right\} \quad (11)$$

を持つウィッシャート分布からのサンプリングを行う[4]．ここに，$\mathbf{Z}^{(t)} = \boldsymbol{\Sigma}^{(t)^{-1}}$である．

4) $\boldsymbol{\gamma}^{(t)}$，$\Sigma^{(t)}$の値を記録し，$t = t + 1$とする．

2)から4)の手順を繰り返し，初期値の影響が十分に緩和された後，4)で記録した$(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\Sigma}^{(t)})$の平均値をもって，未知パラメータの推定値とする．以上が既存の推定手法となっている．

本研究で提案する推定手法は，従属変数$\mathbf{y}_i$の実現値が得られていない場合を対象としており， 2)3)の過程で事後分布のパラメータ$\mathbf{c}_1$，$\mathbf{S}_1$を求めることが出来ない．そこで$\mathbf{y}_i$をsimulateするため以下の過程を上述の手順1)に加える．ここで観測データから$\mathbf{y}_i$に関する制約条件が得られているとする．

1-a)得られた$\boldsymbol{\Sigma}^{(t-1)}$を用いて，誤差項$\boldsymbol{\varepsilon}_i$を$N(0, \boldsymbol{\Sigma}^{(t-1)})$からサンプリングする．

1-b) 得られた$\boldsymbol{\varepsilon}_i$，$\boldsymbol{\gamma}^{(t-1)}$に基づいて，$\mathbf{y}_i$の実現値$\mathbf{y}^*{}_i$を求める．

$$\mathbf{y}^*{}_i = \mathbf{X}_i\boldsymbol{\gamma}^{(t-1)} + \boldsymbol{\varepsilon}_i$$

1-c) $\mathbf{y}^*{}_i$が制約条件を満たしているかどうかを確認する．満たしていない場合1-a)，1-b)を繰り返す．

以上の過程によって，$\mathbf{y}_i$の実現値を得た後，上述の手順を踏むことによって，$\mathbf{y}_i$の実現値が得られていない場合にも，ベイズ推定を行うことが可能となる．なお今回は上記の様に，制約条件を満たすという点にのみ着目して実現値の simulate を行った．しかしより適切な誤差項$\boldsymbol{\varepsilon}_i$のサンプリング方法については，Geweke [3] により研究がなされているのでそちらを参照されたい．

3.検証項目

(1)対象モデル

本研究では，bivariate binary probit model，tobit model を対象とした検証を行う．対象となるモデルを以下に示す．

$$\begin{cases} y_{1i}^{*} = \gamma_{11} x_{11i} + \gamma_{12} x_{12i} + \varepsilon_{1i} \\ Z_{1i} = \begin{cases} 1, & \text{if } y_{1i}^{*} > 0 \\ 0, & \text{if } y_{1i}^{*} \le 0 \end{cases} \\ y_{2i}^{*} = \theta Z_{1i} + \gamma_{21} x_{21i} + \gamma_{22} x_{22i} + \varepsilon_{2i} \\ Z_{2i} = \begin{cases} 1 & \text{if } y_{2i}^{*} > 0 \\ 0 & \text{if } y_{2i}^{*} \le 0 \end{cases} \end{cases}$$

$$\boldsymbol{\varepsilon}_i \sim \text{i.i.d. } N(0, \boldsymbol{\Sigma})$$

$$\begin{cases} y_{1i} = \gamma_{11} x_{11i} + \gamma_{12} x_{12i} + \varepsilon_{1i} \\ y_{2i}^{*} = \theta y_{1i} + \gamma_{21} x_{21i} + \gamma_{22} x_{22i} + \varepsilon_{2i} \\ Z_i = \begin{cases} y_{2i}^{*} & \text{if } y_{2i}^{*} > 0 \\ 0 & \text{if } y_{2i}^{*} \le 0 \end{cases} \end{cases}$$

$$\boldsymbol{\varepsilon}_i \sim \text{i.i.d. } N(0, \boldsymbol{\Sigma})$$

(2)simulated data 作成方法

本研究では，simulated data を用いて推定手法の妥当性を検証する．つまり未知パラメータに対して真値を設定し，各データセットを作成する．そして作成したデータセットを用い推定を行い，結果を真値と比較することで推定手法の妥当性の検証を行った．以下に設定した真値および simulated data の作成方法を示す．

a) 設定した真値

$$(\gamma_{11}, \gamma_{12}, \theta, \gamma_{21}, \gamma_{22}) = (1.0, 1.0, 0.5, 1.0, 1.0)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{pmatrix}$$

b) simulated data 作成方法

以下の過程を繰り返し行うことで推定を行うためのデータセットを作成した．

1) 誤差項$\boldsymbol{\varepsilon}_i$を$N(0, \boldsymbol{\Sigma})$からサンプリングする．
2) $\mathbf{X}_i$を$N(0,1)$からサンプリングする．
3) 得られた$\boldsymbol{\varepsilon}_i$，$\boldsymbol{\gamma}$に基づいて，$\mathbf{y}_i$の値を求める．

(3)検証項目

提案した推定手法の妥当性を検証するに当たり，以下の３項目に着目し，検証を行った．

a) 推定値と真値の比較

推定結果とデータ作成の際に設定した真値の比較を行う．

b) 推定結果に対するサンプル数の影響

上述の通りギブス・サンプリングでは，事後分布の推定を行うため，事後分布から繰り返しサンプリングを行う．その際得られたサンプルの初期のものは，初期値の影響が残っているとして棄却し，残りのサンプルの平均値を求め推定結果とする．そこで棄却・採用したサンプル数の推定結果に対する影響を調べるため複数の棄却・採用サンプル数について推定を行い，推定結果を比較する．

c) サンプルの収束度合い

ギブス・サンプリングでは，事後分布から繰り返しサンプリングを行い，その収束値を推定値として用いる．そのため得られたサンプルが収束しているかどうか検定を行うことが必要となる．そこで得られたサンプルの最初 10%と最後 50%の平均値に差がないことを帰無仮説とした仮説検定を行う．その際に用いられる P 値を geweke_P 値と呼ぶ．

4.検証結果

(1) bivariate binary probit model

bivariate binary probit model における推定結果を表 1 に示す．

表―1 probit model 推定結果

(採用数，棄却数)		γ 11	γ 12	θ	γ 21	γ 22
(10000，1000)	推定値	0.95	1.04	0.52	1.01	0.95
	geweke	0.507	0.494	0.501	0.511	0.496
(10000，10000)	推定値	0.94	1.03	0.53	1.01	0.95
	geweke	0.531	0.541	0.5	0.477	0.493
(50000，1000)	推定値	0.94	1.03	0.53	1.01	0.95
	geweke	0.524	0.51	0.493	0.525	0.527
(50000，10000)	推定値	0.95	1.04	0.53	1.01	0.95
	geweke	0.482	0.475	0.508	0.511	0.515
(100000，1000)	推定値	0.95	1.04	0.53	1.01	0.95
	geweke	0.512	0.491	0.502	0.465	0.524
(100000，10000)	推定値	0.94	1.04	0.53	1.01	0.95
	geweke	0.501	0.571	0.438	0.534	0.513

(採用数，棄却数)		σ11	σ12	σ21	σ22
(10000，1000)	推定値	1.04	0.8	0.8	1.01
	geweke	0.492	0.493	0.493	0.525
(10000，10000)	推定値	1.03	0.8	0.8	1.01
	geweke	0.493	0.463	0.463	0.505
(50000，1000)	推定値	1.03	0.8	0.8	1.01
	geweke	0.498	0.545	0.545	0.509
(50000，10000)	推定値	1.03	0.8	0.8	1.01
	geweke	0.498	0.519	0.519	0.505
(100000，1000)	推定値	1.03	0.8	0.8	1.01
	geweke	0.492	0.522	0.522	0.525
(100000，10000)	推定値	1.03	0.8	0.8	1.01
	geweke	0.493	0.623	0.623	0.505

表よりどの場合においても真値に近い値が推定出来ており，サンプル数による推定結果の大きな違いも見られない．また geweke__P も有意水準を満たしておらず，得られたサンプルは収束しているといえる．

(2) tobit model

tobit model における推定結果を表 2 に示す．

表より，ほとんどの未知パラメータにおいて真値に近い値が推定出来ているといえる．しかしながら共分散 σ_{12} σ_{21} においては真値と推定値に違いがみられる．またサンプル数による推定結果の違いも見られず，geweke__P 値から得られたサンプルも収束しているといえる．

表―2　tobit model 推定結果

(採用数，棄却数)		γ11	γ12	θ	γ21	γ22
(10000，1000)	推定値	1.02	1.09	0.51	0.99	0.98
	geweke	0.497	0.496	0.486	0.493	0.499
(10000，10000)	推定値	1.02	1.09	0.51	0.99	0.98
	geweke	0.506	0.491	0.495	0.5	0.512
(50000，1000)	推定値	1.02	1.09	0.51	0.99	0.98
	geweke	0.492	0.493	0.483	0.489	0.486
(50000，10000)	推定値	1.02	1.09	0.51	0.99	0.98
	geweke	0.502	0.506	0.505	0.499	0.493
(100000，1000)	推定値	1.02	1.09	0.51	0.99	0.98
	geweke	0.494	0.496	0.486	0.493	0.499
(100000，10000)	推定値	1.02	1.09	0.51	0.99	0.98
	geweke	0.502	0.491	0.495	0.5	0.512

(採用数，棄却数)		σ11	σ12	σ21	σ22
(10000，1000)	推定値	1.04	0.56	0.56	1.01
	geweke		0.503	0.503	
(10000，10000)	推定値	1.03	0.56	0.56	1.01
	geweke		0.51	0.51	
(50000，1000)	推定値	1.03	0.56	0.56	1.01
	geweke		0.507	0.507	
(50000，10000)	推定値	1.03	0.56	0.56	1.01
	geweke		0.503	0.503	
(100000，1000)	推定値	1.03	0.56	0.56	1.01
	geweke		0.503	0.503	
(100000，10000)	推定値	1.03	0.56	0.56	1.01
	geweke		0.51	0.51	

5. おわりに

既存のベイズ推定の過程に実現値を simulate する過程を設けることで，実現値の得られていないデータからも推定を行えるよう改善した．また bivariate binary probit model と tobit model を対象とし，simulated data を用い推定手法の妥当性の検証を行った．結果として bivariate binary probit model を対象とした場合には，係数パラメータ，誤差項の共分散行列共に真値に近いパラメータを推定出来ることが確認出来た．また tobit model を対象とした推定については，係数パラメータの推定については真値に近い値が推定出来ていた．しかしながら誤差項の共分散行列については，共分散の推定値と真値が大きくずれる場合があることも確認出来た．その原因については今後考察が必要となる．

参考文献

1)和合肇：ベイズ計量経済分析，東洋経済，2005

2)伊庭幸人，種村正美，大森裕浩，和合肇
佐藤整尚，高橋明彦：計算統計II，マルコフ連鎖モンテカルロ法とその周辺，岩波書店，2005

3) Geweke, J. (1991) Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints and the evaluation of constraint probabilities. Presented at Computing Science and Statistics: the Twenty-third Symposium of the Interface, Seattle, April 22-24.