한국통계학회 논문집 제1권1호
1994년 12월 pp. 94-101

공변량을 갖는 패널자기회귀 과정에 대한 베이즈추정

신 민 웅[1], 신 기 일[1]

요약

본 논문은 패널(panel) 자기회귀 모형에서 자기회귀 계수의 추정을 베이지안 방법으로 접근하였는데, 이 때 특별히 Gibbs Sampling 방법을 이용하여 사후분포를 계산하였다. 또한 모의 실험을 통하여 자기회귀계수를 Gibbs Sampling 방법으로 추정한 베이지안 추정치가 non-Bayesian 방법으로 구한 추정치보다 더 우월함을 보였다.

1. 서론

베이지안 추론을 하는데 필요한 주변사후밀도함수(marginal posterior density function)를 계산하는데 어려움은 실제 자료에 베이지안기법을 응용하는데 장애가 되어 왔다. 이러한 어려움을 극복하는데 Gibbs Sampler 기법이 최근에 널리 쓰여지고 있다. 즉, Gibbs Sampler 기법을 사용하므로서, 우리는 어려운 계산을 피할수 있다. Gibbs Sampler 방법은 Geman 과 Geman(1984)의 논문으로 출발한이래, 최근에 Gelfand, Hill, Racin-Poon과 Smith(1990)가 정규분포에서 베이지안 사후분포를 계산하는데 Gibbs Sampler을 사용하였다. Zeger와 Karim(1991)은 랜덤효과를 갖는 일반선형 모형에서 베이지안 사후분포를 계산하는데 Gibbs Sampling을 사용하였다. Albert와 Chib(1993)은 관찰할수 없는 잠재변수(latent variable)를 이용하여 이진반응(binary response) 자료의 베이지안 분석에 Gibbs Sampler를 사용하였다. Park(1994)는 중도절단자료에서의 역추정문제에 Gibbs Sampling 방법을 사용하였다. Gibbs Sampling은 Markov Chain의 성질을 기본으로 하여 밀도함수를 계산하지 않고 주변분포로 부터 간접적으로 확률변수를 생성하는 기법이다. Gibbs Sampler에 대해서는 Casella와 George(1992)의 논문에 상세히 설명되어 있다.

Gibbs Sampler 같은 기법을 사용하므로서 어려운 계산을 피할 수 있으므로 실제적인 문제에 많이 응용될수 있다. Gibbs Sampler는 대부분 베이지안 모형에서 응용된다. 시계열 분석에서는 McCulloch 와 Tsay(1993)가 Gibbs Sampler를 사용하여서 평균과 분산에 대한 베이지안 추론을 하였는데, 특기 할 점은 시계열에 Probit모형을 첨가하여 주어진 시점에서 Shift가 일어나는 확률이 설명변수와 관련되도록 하였다. 또한 McCulloch와 Tsay(1994)는 이상점과 결측치가 있는 자기회귀모형에서, level-shift 모형을 다루는데 Gibbs Sampler가 유용히 사용됨을 보였다. 그들은 Gibbs Sampler는 전통적인 우도방법보다 우월한 점이 많음과 전통적인 우도방법은 많은 양의 계산이 필요한데 Gibbs Sampler는 그러한 계산이 용이하게 실행됨을 보여 주었다.

우리는 공변량(covariate)을 갖는 패널 자기모형에서 랜덤 자기회귀 모계수를 추정할 때, Gibbs Sampler 방법을 통해 사후분포를 용이하게 계산하였다.

제 2절에서는 공변량을 갖는 패널자기회귀 모형에 대하여 언급하고, 제3절에서는 자기회귀계수

1)(449-791) 경기도 용인군 모현면 왕산리 한국외국어대학교 통계학과

의 사후분포를 계산하기 위한 Gibbs Sampling 방법과 이에 관련한 이론을 다루었다. 제 4절에서는 모의 실험으로 자기회귀계수를 Gibbs Sampling방법으로 추정하였다.

2. 패널 자기회귀모형

대부분의 시계열분석에서는 하나의 긴 시계열이 분석의 대상이 된다. 그러나 현실적으로 짧은 기간 동안 많은 개체를 조사하게 되는 경우가 많다. 예컨데 많은 환자들을 며칠간 계속하여 혈압을 측정하거나, 여러 지역에서 여러 달 동안 경제지표를 조사하였을 경우이다. 즉 우리는 짧은 시계열 자료들의 패널을 갖고 있을 때가 많다. 그러면 각각의 시계열 자료로부터의 정보를 통합(pool)하여, 개개의 모수들을 추정하는데 효율을 높일 수가 있다.

서로 독립인 AR(1)모형

$$Y_{it} - \phi_i Y_{it-1} = a_{it}\ , \quad i = 1, \cdots, n, \quad t = 1, \cdots, T \tag{2.1}$$

를 생각하자. 단, Y_{it}는 i번째 시계열의 t번째 관찰치이고, a_{it}는 $N(0,\sigma^2)$으로 부터의 *i.i.d.*인 오차, 그리고 T는 i번째 시계열로 부터의 관찰치의 수이다, 우리는 σ^2을 기지라고 가정한다. 우리는 기지의 공변량 X_i ($p \times 1$ 벡터)가 ϕ_i와 선형으로 관계되어 있다고 가정한다.

즉, 랜덤모수 ϕ_i는 공변량 X_i와 다음과 같은 관계가 있다고 가정한다.

$$E[\phi_i \mid X_i, \beta] = X_i{}^T\beta, \qquad i = 1, \cdots, n \tag{2.2}$$

단, β는 i번째 자기회귀모수로 ϕ_i에 대한 공변량 X_i에 대응되는 초모수(hyperparameter)들의 벡터이다. 초모수 β는 모든 각각의 시계열에 대해 공통인 모수이다. 즉, ϕ_i가 다음과 같은 절단(truncated)정규분포를 한다고 가정한다.

$$\phi_i \mid X_i, \beta \sim N(X_i{}^T\beta, \tau^2), \quad i = 1, \cdots, n, \text{ (독립)} \tag{2.3}$$
$$\text{-1의 왼쪽과 1의 오른쪽으로 절단}$$

여기서 τ^2은 기지라고 가정한다. 베이지안 접근에서는 모수 ϕ와 β가 확률변수이다. $f(\phi,\beta)$는 ϕ와 β에 대한 결합사전 분포를 나타낸다. 우리는 사후분포 $f(\phi,\beta \mid Y)$를 유도하려고 한다. 여기서, $Y = (Y_1, \cdots, Y_n)^T$ 이고, $Y_i = (Y_{i1}, \cdots, Y_{iT})^T$이다. 그러나 $f(\phi,\beta \mid Y)$의 계산은 매우 복잡하므로 다음절에서 Gibbs Sampler 기법을 이용한다.

3. Gibbs Sampler와 조건부 분포

Gibbs Sampler는 원하는 사후분포를 추정하는 Monte Carlo 방법이다. 이 방법을 간략히 소개

한다. 세 변수 U, V 그리고 W를 생각하자. 조건부 분포를 $[U \mid V, W]$, $[V \mid U, W]$ 그리고 $[W \mid U, V]$로 나타내고, 결합분포는 $[U, V, W]$로 나타내자. Gibbs Sampler는 다음과 같이 $[U, V, W]$로 부터 랜덤변량(variate)을 생성시키는 방법이다. $[U \mid V^{(0)}, W^{(0)}]$로 부터 $U^{(1)}$을 뽑는데, 여기서 $U^{(0)}, V^{(0)}, W^{(0)}$는 주어진 임의의 초기치이다. 다음은 $[V \mid U^{(1)}, W^{(0)}]$로 부터 $V^{(1)}$를 뽑고, 계속해서 $[W \mid U^{(1)}, V^{(1)}]$로 부터 $W^{(1)}$을 뽑는다. 이와같이 계속 반복을 하여 B번 반복을 하였다면, 우리는 $[U^{(B)}, V^{(B)}, W^{(B)}]$를 얻는다. Geman과 Geman(1984)은 적당한 조건 아래서 결합분포 $[U^{(B)}, V^{(B)}, W^{(B)}]$가 $B \to \infty$일 때에 $[U, V, W]$로 수렴함을 보였다.

우리가 생각하는 공변량을 갖는 패널자기회귀 모형에서는 결합분포 $[\phi, \beta \mid Y]$와 그 주변분포 $[\phi \mid Y]$, $[\beta \mid Y]$를 구하고자 한다. 여기서 $\phi = (\phi_1, \cdots, \phi_n)$이다. 이 결합분포는 조건부 분포 $[\phi \mid \beta^{(k)}, Y]$와 $[\beta \mid \phi^{(k)}, Y]$로 부터 표본추출하므로서 얻을 수 있다. 이 접근의 우아함은 이 각각의 조건부 분포로 부터 상대적으로 쉽게 결합분포를 구할 수 있다는데 있다.

3.1 $[\beta \mid \phi^{(k)}, Y]$

Gibbs Sampling 알고리즘을 써서 β의 주변 사후분포를 계산하는데는 ϕ에 대한 조건부 β의 사후분포만 필요하다. ϕ가 주어졌을 때 β의 조건부 사후밀도는 다음의 정규선형모델 $\phi = X\beta + \varepsilon$으로부터의 사후밀도이다. 단, $X^T = [X_1^T, \cdots, X_n^T]$ 이고 ε은 $N(0, I\tau^2)$인 정규분포를 한다. 여기서 I는 단위행렬이다.

우리는 $[\beta \mid \phi^{(k)}, Y]$로 부터 $\beta^{(k+1)}$을 생성하고자 한다. β의 사전분포가 확산분포(diffuse)일 때, 표준 선형 이론에 따르면

$$\beta \mid \phi^{(k)}, Y \sim N(\hat{\beta}_\phi, (X^T X)^{-1}\tau^2) \tag{3.1}$$

이다. 단, $\hat{\beta}_\phi = (X^T X)^{-1} X^T \phi^{(k)}$이다.

3.2 $[\phi \mid \beta^{(k)}, Y]$

이제 다음 단계로 $[\phi \mid \beta^{(k)}, Y]$로 부터 $\phi^{(k+1)}$을 생성하려고 한다. $\beta^{(k)}$가 주어졌을 때 ϕ_i의 조건부 사후 분포는 다음과 같다.

$$\phi_i \mid \beta^{(k)}, Y \sim N(\delta_i, \sigma^2\tau^2/a_i) \quad i = 1, \cdots, n \quad \text{(독립)} \tag{3.2}$$
$$\text{-1의 왼쪽과 1의 오른쪽으로 절단}$$

단, $\delta_i = [\sigma^2 (X_i{}^T \beta^{(k)}) + \tau^2 \sum_{t=1}^{T} Y_{it} Y_{it-1}]/a_i$ 그리고 $a_i = \sigma^2 + \tau^2 \sum_{t=1}^{T} Y^2{}_{it-1}$ 이다.

(3.2) 식은 Kim 과 Basawa(1992) 가 구한 자기회귀계수 ϕ_i 와 관찰치 Y_i 의 결합밀도함수로 부

터 유도 될 수 있으며 자기회귀계수 ϕ_i 와 관찰치 Y_i 의 결합밀도함수는 다음과 같다.

$$f(Y_i,\phi_i)=f(Y_i|\phi_i)f(\phi_i)$$

$$=(2\pi\sigma^2)^{-\frac{T}{2}}(2\pi\tau^2)^{-\frac{1}{2}}\exp\{-\frac{a_i}{2\sigma^2\tau^2}[\phi_i-\frac{\tau^2\sum_{i=1}^{T}Y_{it}Y_{it-1}+\sigma^2(X_i^T\beta)}{a_i}]^2\}$$

$$\exp\{\frac{-\sum_{i=1}^{T}(Y_{it}-(X_i^T\beta)Y_{it-1})^2}{2a_i}\}\exp\{-\frac{\tau^2}{2\sigma^2 a_i}[\sum_{i=1}^{T}Y_{it}^2-(\sum_{i=1}^{T}Y_{it}Y_{it-1})^2]\} \qquad (3.3)$$

우리는 많은 실제문제에서 흔히 β의 사전분포를 확산분포(diffuse)로 간주한다. β의 초기값이 이미 주어졌다고 가정할 때 (3.1)과 (3.2)의 분포로 부터 Gibbs알고리즘의 한 사이클(Cycle)로 ϕ 와 β를 생성한다. β의 초기값, $\beta^{(0)}$는 최대우도 추정값이나 최소자승추정치를 쓸수 있다.

4. 모의 실험

공변량을 갖는 패널 자기회귀 모형에 대한 Gibbs Sampler 는 SAS로 실행하였다. 모의실험에서 우리는 다음과 같은 공변량을 갖는 자기회귀 모형을 생각하였다.

$$Y_{it}=\phi_i Y_{it-1}+\varepsilon_{it}, \qquad i=1,...25, \quad t=1,...30.$$

그리고

$$E[\phi_i|X_i,\ \beta]=X_i^T\beta, \quad i=1,...,25.$$

여기서 공변량에 대한 회귀계수 β_0, β_1 이 위의 식을 만족하도록 β_0, β_1, 그리고 ϕ_i 를 정하였다. 정해진 공변량 회귀계수는 $\beta_0=-1.2, \beta_1=0.05$ 이고 기지인 공변량 X_i 와 ϕ_i, $i=1,...,25$ 는 <표4.1>과 같다.

모의 실험을 간편하게 하기위하여 일반성을 잃지않고 $\sigma^2=1$ 그리고 $\tau^2=.02$ 라고 가정하였다.

각각의 자료 집합에서 200 개의 표본을 추출하고 이중 앞의 100 개를 버리고 뒤부분에서 얻어진100 개의 표본을 사용하였다. 위에서 정의된 시계열이 정상성 조건을 만족해야 하기때문에 $|\phi_i|>1$, $|X_i^T\beta|>1$ 인 경우에 있어서는 Gibbs sampler 에서 제외시켰다. 여기서 우리는 ϕ_i 를 두 가지 방법으로 추정하였다. 그 첫번째는 Gibbs Sampler 에서 얻어진 ϕ_i 의 표본으로부터 ϕ_i 를 추정하는 것이다.

<표4.1> 기지인 공변량 X_i 와 회귀계수 ϕ_i

X_i		ϕ_i
X_{0i},	X_{1i}	
1	10.0	-.70
1	10.6	-.67
1	11.2	-.64
1	11.8	-.61
1	13.5	-.525
1	13.8	-.51
1	15.7	-.415
1	16.2	-.39
1	16.7	-.365
1	17.9	-.305
1	21.5	-.125
1	22.4	-.08
1	23.7	-.015
1	24.8	.04
1	25.5	.075
1	25.7	.085
1	25.8	.09
1	26.3	.115
1	26.4	.12
1	27.4	.17
1	27.6	.18
1	29.9	.295
1	31.5	.375
1	36.4	.62
1	36.7	.635

두번째는 식 (3.2) 에서 정의된 δ_i 가 ϕ_i 의 사후분포에서의 기대치 이므로 δ_i 를 이용하여 ϕ_i 를 추정하는 것이다. 표 4.2, 와 표 4.3 은 모의실험 결과이다.

얻어진 100 개의 표본에서 ϕ_i, $i=1,...25$, β_0, β_1 그리고 (3.2) 에서 정의 된 δ_i, $i=1,...,25$ 의 평균과 표본표준편차를 구하였다. 10 개의 자료집합이 생성되었고 각 자료집합에서 구한 평균과 표준편차를 다시 평균하였다. 표 4.2 에서 θ 는 ϕ_i 의 참값을 $\bar{\theta}$ 는 10개의 자료집합에서 얻어진 ϕ_i 의 전체 평균을 $s_{\bar{\theta}}$ 는 ϕ_i의 표준편차의 평균 그리고 표 4.3 에서 $\bar{\theta}$ 는 10 개의 자료 집합에서 얻어진 δ_i 들의 전체평균 그리고 $S_{\bar{\theta}}$ 는 δ_i 의 표준편차의 평균을 의미한다. 또한 *OLS* 는 조건부 최소자승 추정량($Y_{i0}=0$, $i=1,...25$, 이라 가정한 상태에서의 *OLS*) 그리고 $R\text{-}Bias = \left| \frac{\theta-\bar{\theta}}{\theta} \right|$ 는 상대편의를 나타낸다.

<표4.2> Gibbs Sampler 에 의한 모의 실험(ϕ_i 사용)

θ	*OLS*	$\bar{\theta}$	$s_{\bar{\theta}}$	*R-bias*	비 고
-.700	-.716	-.707	.105	.008	*
-.670	-.681	-.676	.119	.008	*
-.640	-.647	-.645	.102	.008	*
-.610	-.613	-.610	.108	.000	*
-.525	-.519	-.515	.123	.020	*
-.510	-.503	-.505	.136	.011	*
-.415	-.402	-.401	.133	.034	*
-.390	-.376	-.382	.140	.022	*
-.365	-.340	-.349	.121	.043	*
-.305	-.286	-.290	.137	.048	*
-.125	-.089	-.104	.142	.165	
-.080	-.038	-.05	.132	.364	
-.015	.038	.021	.141	2.390	
.040	.103	.078	.139	.943	
.075	.146	.121	.140	.609	
.085	.158	.132	.137	.555	
.090	.164	.135	.137	.497	
.115	.194	.166	.143	.444	
.120	.200	.169	.143	.407	
.170	.262	.224	.142	.316	
.180	.274	.236	.142	.310	
.295	.414	.371	.132	.258	
.375	.507	.461	.128	.229	
.620	.751	.720	.105	.161	
.635	.763	.730	.103	.149	
-1.2	-1.278	-1.204	.244	.003	
.05	.0559	.0508	.00787	.016	

여기서 *는 *R-bias*가 .05 이내인 경우를 표시함.

모의실험 결과 Gibbs Sampler 는 일반적으로 *OLS*보다 더 좋은 추정값을 주는 것으로 나타났다. 물론 시계열 자료분석에서 조건부 *OLS*가 가장 좋은 추정량은 아니다. 그러나 많이 사용되고 있는 추정량 중의 하나이다. 이는 공변량이 갖고 있는 정보를 *OLS*는 이용하지 않는 반면 Gibbs Sampler는 이용하기 때문에 당연한 결과라고 보아야 할것이다. 또한 표 4.2 와 표 4.3 을 비교해 볼때 δ_i 의 표준 편차가 일반적으로 작게 나오기 때문에 직접 ϕ_i 를 사용하는 것보다 δ_i 를 사용하는 것이 더 좋을 것이다. 이번 모의 실험의 결과에서 Gibbs Sampler에서 사용한 패널의 크기와 초기치와는 밀접한 관계가 있음이 나타났다. 즉 *OLS* 가 "-" 부호에서는 매우 좋게 나온 반면 "+" 부호에서는 나쁘게 나왔다. 이때 δ_i 의 식을 고쳐보면

<표4.3> Gibbs Sampler 에 의한 모의 실험(δ_i 사용)

θ	*OLS*	$\bar{\theta}$	$s_{\bar{\theta}}$	*R-bias*	비 고
-.700	-.716	-.709	.067	.013	*
-.670	-.681	-.675	.071	.008	*
-.640	-.647	-.642	.074	.003	*
-.610	-.613	-.609	.077	.001	*
-.525	-.519	-.518	.083	.012	*
-.510	-.503	-.503	.084	.015	*
-.415	-.402	-.404	.088	.027	*
-.390	-.376	-.378	.089	.031	*
-.365	-.340	-.352	.089	.036	*
-.305	-.286	-.290	.091	.050	*
-.125	-.089	-.100	.094	.201	
-.080	-.038	-.051	.095	.357	
-.015	.038	.019	.096	2.296	
.040	.103	.080	.096	1.004	
.075	.146	.119	.096	.589	
.085	.158	.130	.096	.534	
.090	.164	.136	.096	.511	
.115	.194	.164	.096	.427	
.120	.200	.170	.096	.414	
.170	.262	.226	.096	.331	
.180	.274	.238	.096	.320	
.295	.414	.369	.094	.250	
.375	.507	.460	.091	.226	
.620	.751	.718	.072	.159	
.635	.763	.732	.070	.153	
-1.2	-1.278	-1.204	.244	.003	
.05	.0559	.0508	.00787	.016	

여기서 *는 *R-bias*가 .05 이내인 경우를 표시함.

$\delta_i^{(k)} = \frac{\sigma^2}{a_i} X^T\beta^{(k-1)} + (1 - \frac{\sigma^2}{a_i})OLS$이 된다. 즉, δ_i 는 *OLS* 와 $X^T\beta^{(k)}$,($(k-1)$번째 Gibbs Sampler에서 나온값)의 weighted sum 으로 나타난다. 따라서 Gibbs sampler 에 의해 추정된 값도 이와 같은 추세를 보인다. 물론 패널의 크기가 충분히 크고 사이클의 반복횟수를 늘린다면 이 문제는 해결되리라 본다. 또한 $|\phi_i| < 0.1$ 인 상황에서는 상대적으로 큰 상대편의를 나타내고있다. 따라서 이러한 경우에 있어서는 사용에 주의를 요한다.

5. 토의

우리는 공변량을 갖는 패널 자기회귀 모형의 랜덤 모수를 추정하기 위한 Gibbs Sampler를 설명하였다. 그리고 모의실험을 한 결과 non-Bayesian방법으로 구한 추정치보다 더 우월한 베이지안 추정치를 얻어 현실적으로 의미있는 수치를 얻었다.

우리가 앞으로 더 연구하고 개선해야 할 점은 다음과 같다. 우리는 모의실험을 간략히 하기 위해 자기회귀 모형에서 오차의 분산을 기지인 상수로 놓고, 또 랜덤 자기회귀 분포의 분산을 기지인 상수로 놓았다. 그러나 실제문제에 있어서 이러한 분산들이 미지인 경우는 이들 분산을 랜덤 모수로 놓고 Gibbs Sampler방법을 써야 한다. 그리고 정도(precision)가 더 높은 추정치를 얻기 위해서는 자료의 집합을 더 많이 늘이고 사이클의 반복 횟수도 더 늘여야 한다. 그리고 적절한 실제 자료를 찾아서 Gibbs Sampler를 응용해 보는 것이다.

참고문헌

[1] Albert J., and Chib, S, (1993). Bayesian Method for Binary and Polychotomous Response Data, *Journal of the American Statistical Association,* Vol. 88, 669-679.

[2] Casella, G., and George, E. (1992). Explaining the Gibbs Sampler, *American Statistician,* Vol. 46, 167-174.

[3] Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M. (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling, *Journal of the American Statistical Association,* Vol. 85, 972-985.

[4] Geman, S., and Geman, D. (1984). Stochastic Relaxation ; Gibbs Distributions and the Bayesian Restoration of Images, *IEEE Translations on Pattern Analysis and Machine Intelligence,* Vol. 6, 721-741.

[5] Li, W.K. and Hui, Y.V. (1983). Estimation of random coefficient autoregreesive process: An Empirical Bayes approach, *Journal of Time Series Analysis,* Vol. 4, 89-94.

[6] Mcculloch, R.E., and Tsay, R.S. (1994). Bayesian Analysis of Autoregreesive Time Series via the Gibbs Sampler, *Journal of Time Series Analysis,* Vol. 15, No2, 235-250.

[7] Mcculloch, R.E., and Tsay, R.S. (1993). Bayesian Inference and Prediction for Mean and Variance Shifts in Autoregreesive Time Series, *Journal of the American Statistical Association,* Vol. 88, 968-978.

[8] Park, N.H., Lee, S., Lee N,Y., Park Y. and Lee, S.H. (1994). On the Calibration Problem with Censored Data, *The Korean Journal of Applied Statistics,* Vol. 7, 1-17.

[9] Kim, Y.W. and Basawa, I.V. (1992). Empirical Bayes Estimation for First-Order Autoregressive Process, *Australian Journal of Statistics,* Vol. 4, No2, 89-94.

[10] Zeger, S.L. and Karim, M.R. (1991). Generalized Linear Models with Random Effects; A Gibbs Sampling Approach, *Journal of the American Statistical Association,* Vol. 86, 79-86.