

Bayesian Inference for Autoregressive Models with Skewed Exponential Power Errors

Hyunnam Ryu^a · Dal Ho Kim^{a,1}

^aDepartment of Statistics, Kyungpook National University

(Received September 1, 2014; Revised October 23, 2014; Accepted October 27, 2014)

Abstract

An autoregressive model with normal errors is a natural model that attempts to fit time series data. More flexible models that include normal distribution as a special case are necessary because they can cover normality to non-normality models. The skewed exponential power distribution is a possible candidate for autoregressive models errors that may have tails lighter(platykurtic) or heavier(leptokurtic) than normal and skewness; in addition, the use of skewed exponential power distribution can reduce the influence of outliers and consequently increases the robustness of the analysis. We use SIR algorithm and grid method for an efficient Bayesian estimation.

Keywords: Autoregressive model, Bayesian p -value, skewed exponential power distribution, Gibbs sampler, robust.

1. 서론

실제 자료를 다룰 때 정규성 가정에 위배되는 경우가 종종 발생하며, 시계열 자료의 경험적 연구에서 두꺼운 꼬리 분포를 가지는 경우가 흔히 있다. 최근 통신 및 경제 네트워크의 발달로 방대한 양의 시계열 자료의 획득이 가능해 졌고, 이러한 실제 시계열 자료들은 두꺼운 꼬리를 가지는 경우가 많다. 이처럼 정규성 가정에 위배되는 경우 정규성 가정을 완화하기 위한 방법으로 두꺼운 꼬리를 수용하는 유연한 모형을 고려할 수 있다. 특별한 경우로 정규분포를 포함하는 유연한 모형은 정규성을 가지는 모형 뿐만 아니라 정규성에 벗어나는 모형까지 포괄할 수 있기 때문에 이를 도입하는 것이 현실적으로 매우 유용하다. 지수멱 분포(exponential power distribution; EPD)는 Box와 Tiao (1992)에 의하여 제안되었으며, 정규분포보다 얇거나 두꺼운 꼬리를 가지는 자기회귀모형의 오차를 위한 유연한 분포로 고려할 수 있다. 실제로 지수멱 분포는 이상치의 영향을 줄일 수 있으며 좀 더 로버스트한 추론을 가능하게 한다. 지수멱 분포의 밀도함수는 식 (1.1)과 같이 나타낼 수 있다.

$$f(x|\mu, \sigma, p) = \frac{1}{2(2p)^{\frac{1}{2p}-1}\Gamma\left(\frac{1}{2p}\right)\sigma} \exp\left(-\frac{1}{2p}\left|\frac{x-\mu}{\sigma}\right|^{2p}\right), \quad (1.1)$$

¹Corresponding author: Department of Statistics, Kyungpook National University, 80 Daehakro, Bukgu, Daegu, 702-701, Korea. E-mail : dalkim@knu.ac.kr

여기서 $\mu \in \mathbf{R}$, $\sigma > 0$, $p > 0$ 이며, 이를 $X \sim \text{EP}(\mu, \sigma, p)$ 로 나타내자. 지수멱 분포는 3개의 모수를 가지는 데, μ 는 위치모수, σ 는 척도모수, 그리고 p 는 첨도를 결정하는 형상모수이다. 만약 모수 p 가 고정되어 있다면, 분포는 위치-척도 분포족에 속하게 된다. 따라서 모수 p 는 분포 꼬리의 두께와 관련이 있다. 특히, $0 < p < 1$ 이면 정규분포 보다 두꺼운 꼬리를 가지는 분포가 되며, $p > 1$ 이면 정규분포 보다 얇은 꼬리를 가지는 분포가 된다. 또한 지수멱 분포의 특별한 경우로서 $p = 0.5$ 일 때 라플라스 분포, $p = 1$ 일 때 정규분포를 가지며, $p \rightarrow \infty$ 일 때 균등분포가 된다.

이와 관련하여 최근 로버스트 관점에서 첨도 뿐만 아니라 왜도까지 고려한 모형 및 이를 사용한 연구가 활발하다. 이 중 대칭성을 가지는 확률밀도함수에 대해서 왜도를 반영하는 모형을 고려할 수 있다. 만약 $f(\cdot)$ 이 0에 대하여 대칭인 밀도함수 이고, $G(\cdot)$ 가 $G'(\cdot)$ 이 대칭인 밀도함수를 만족하는 연속인 분포 함수라면, 식 (1.2)와 같은 형태의 왜도를 반영하는 분포를 제안 할 수 있다.

$$2G(\lambda y)f(y), \quad -\infty < y < \infty, \lambda \in \mathbf{R}. \quad (1.2)$$

지수멱 분포(EPD) 또한 0에 대하여 대칭이므로, 식 (1.2)에서 $f(\cdot)$ 형태로 지수멱 분포를 고려하고 $G(\cdot)$ 의 형태로 식 (1.3)을 반영 할 수 있다.

$$G(y) = \Phi \left\{ \text{sgn}(y) \frac{|y|^\psi}{\sqrt{\psi}} \right\}. \quad (1.3)$$

이러한 형태의 분포를 비대칭 지수멱 분포(skewed exponential power distribution; SEPD)라 하며, 두꺼운 꼬리와 왜도를 동시에 반영 할 수 있다. 비대칭 지수멱 분포의 확률밀도함수는 식 (1.4)과 같다 (DiCiccio와 Monti, 2004; Azzalini, 1985,1986 참조).

$$\begin{aligned} f_{\text{SEP}}(y; \mu, \sigma, \lambda, p) &= 2\Phi \left\{ \text{sgn}(\lambda \frac{y-\mu}{\sigma}) \frac{|\lambda (\frac{y-\mu}{\sigma})|^p}{\sqrt{p}} \right\} f_{\text{EP}}(y; \mu, \sigma, p) \\ &= \frac{1}{(2p)^{\frac{1}{2p}-1} \Gamma(\frac{1}{2p}) \sigma} \Phi \left\{ \text{sgn} \left(\lambda \frac{y-\mu}{\sigma} \right) \frac{|\lambda (\frac{y-\mu}{\sigma})|^p}{\sqrt{p}} \right\} \exp \left(-\frac{1}{2p} \left| \frac{y-\mu}{\sigma} \right|^{2p} \right), \end{aligned} \quad (1.4)$$

여기서 $\mu \in \mathbf{R}$, $\sigma > 0$, $\lambda \in \mathbf{R}$, $p > 0$ 이며, 이를 $X \sim \text{SEP}(\mu, \sigma, \lambda, p)$ 로 나타내자. 비대칭 지수멱 분포는 4개의 모수를 가지는 데, μ 는 위치모수, σ 는 척도모수, 그리고 λ 는 왜도, p 는 첨도를 결정하는 형상모수이다. 특히, $\lambda = 0$ 이면 지수멱 분포를 가지며 $p = 1$ 이면 비대칭 정규분포 그리고 $\lambda = 0, p = 1$ 인 경우 정규분포가 된다.

Figure 1.1에서는 $p = 1.3$ 일 때 $\lambda = 0, 0.5, 1, 1.5$, $\lambda = 1.3$ 일 때 $p = 1, 0.8, 1.5, 2$ 를 가지는 비대칭 지수멱 분포의 밀도함수를 나타낸 것이다. 이 때 λ 가 양수 일때만을 고려하였지만, 음수의 경우에는 0에 대해서 대칭인 형태로 고려할 수 있다.

이 논문에서는 정규성을 위반하는 경우에 대해서 정규분포의 대안으로 두꺼운 꼬리와 왜도를 반영하는 왜곡된 지수멱 분포를 고려하였으며, 이를 베이지안 관점에서 추론하였다.

2. 자기회귀모형을 위한 베이지안 추론

2.1. 비대칭 지수멱 오차를 가지는 자기회귀모형

t 시점의 관측치를 y_t 라 할 때, 1차 자기회귀모형(AR(1))은 식 (2.1)과 같다.

$$y_t = \alpha + \rho y_{t-1} + e_t, \quad t = 1, \dots, n, \quad (2.1)$$

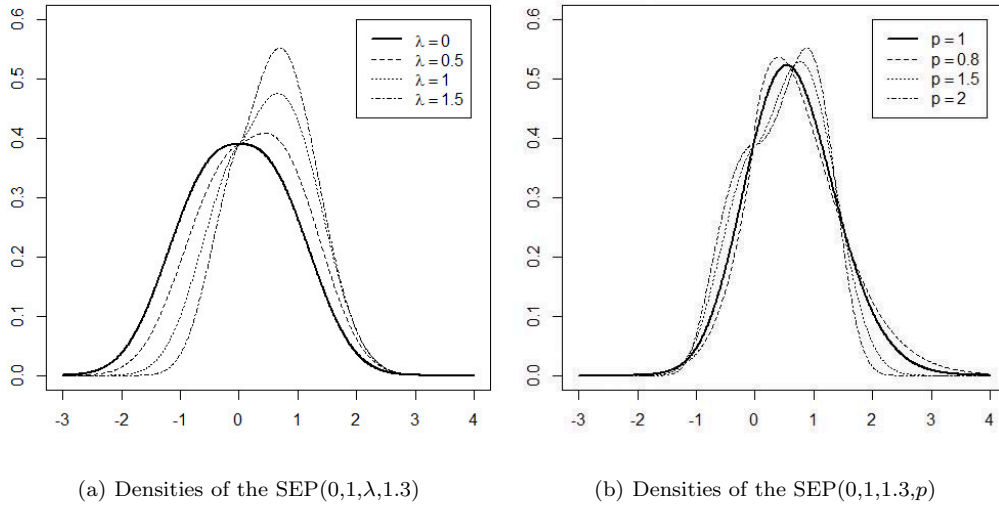


Figure 1.1. Various cases of the SEP

여기서 α 는 절편, ρ 는 자기회귀 계수이며, $e_t \stackrel{iid}{\sim} \text{SEP}(0, \sigma, \lambda, p)$ 이다. 그러면 y_t 에 대한 분포는 식 (2.2)과 같다.

$$y_t | \alpha, \rho, \sigma, \lambda, p \sim \text{SEP}(\alpha + \rho y_{t-1}, \sigma, \lambda, p), \quad t = 1, \dots, n. \tag{2.2}$$

$\mathbf{y} = (y_1, \dots, y_n)'$ 는 n 개의 관측치이며, 이에 대한 우도함수는 식 (2.3)과 같다.

$$L(\alpha, \rho, \sigma, \lambda, p | \mathbf{y}) \propto \frac{1}{(2p)^{n(\frac{1}{2p}-1)} (\Gamma(\frac{1}{2p}))^n \sigma^n} \prod_{t=1}^n \Phi \left\{ \text{sign} \left(\lambda \left(\frac{y_t - \alpha - \rho y_{t-1}}{\sigma} \right) \right) \left| \lambda \left(\frac{y_t - \alpha - \rho y_{t-1}}{\sigma} \right) \right|^p \right\} \times \exp \left\{ - \sum_{t=1}^n \frac{1}{2p} \left| \frac{y_t - \alpha - \rho y_{t-1}}{\sigma} \right|^{2p} \right\}. \tag{2.3}$$

효율적인 계산을 위하여, 감마함수에 대한 근사를 고려하면 식 (2.4)와 같다.

$$\Gamma(z) \approx (2\pi)^{\frac{1}{2}} z^{z-\frac{1}{2}} \exp(-z) \exp(\phi(z)), \quad 0 < \phi(z) < \frac{K}{z}, \quad K : \text{상수}. \tag{2.4}$$

베이저안 추론을 위해 주어진 모수에 고려되는 사전분포는 식 (2.5)과 같다.

$$\begin{aligned} \alpha &\sim U(-10^3, 10^3), \\ \rho &\sim U(-1, 1), \\ \pi(\sigma, \lambda, p) &\propto \frac{1}{(1 + \sigma^2)^2} \frac{1}{(1 + \lambda^2)} \frac{1}{(1 + p)^2}. \end{aligned} \tag{2.5}$$

이 때, α 에 대한 사전분포는 무정보적 사전분포이며, 1차 자기회귀계수의 정상성을 고려하여 ρ 의 사전 분포는 -1 과 1 의 범위를 가지는 균등분포를 고려하였다. 또한 사전분포 $\pi(\sigma, \lambda, p)$ 는 적절성을 가진다.

이에 따른 $\alpha, \rho, \sigma, \lambda, p$ 의 결합 사후분포 $\pi(\alpha, \rho, \sigma, \lambda, p|\mathbf{y})$ 는 식 (2.6)과 같다.

$$\begin{aligned} \pi(\alpha, \rho, \sigma, \lambda|\mathbf{y}) &\propto (\sigma^2)^{-\frac{n}{2}} (1 + \sigma^2)^{-2} (1 + \lambda^2)^{-1} (1 + p)^{-2} (2p)^{\frac{n}{2}} \\ &\times \exp \left\{ -\frac{1}{2p} \sum_{t=1}^n \left(\left| \frac{y_t - \alpha - \rho y_{t-1}}{\sigma} \right|^{2p} - 1 \right) \right\} \\ &\times \prod_{t=1}^n \Phi \left\{ \text{sign} \left(\lambda \left(\frac{y_t - \alpha - \rho y_{t-1}}{\sigma} \right) \right) \frac{\left| \lambda \left(\frac{y_t - \alpha - \rho y_{t-1}}{\sigma} \right) \right|^p}{\sqrt{p}} \right\}. \end{aligned} \quad (2.6)$$

주어진 모형에서 λ, p 에 대한 조건부 사후분포로부터 난수를 얻기가 쉽지 않으므로, 비대칭 지수떡 오차를 가지는 자기회귀모형에 대한 효율적인 베이지안 추론을 하기 위하여 SIR(Sample Importance Resampling) 알고리즘과 격자망(grid)방법을 고려하도록 하자.

2.2. 베이지안 추론을 위한 계산

비대칭 지수떡 오차를 가지는 자기회귀모형에 대한 베이지안 추론을 하기 위하여, SIR 알고리즘을 고려하도록 하자. SIR 알고리즘은 목표분포와 비슷한 형태를 가지는 분포로부터 난수를 생성하여, 목표분포로부터의 가중치를 두고 난수를 재표집하는 방법이다 (Rubin, 1987 참고). 비대칭 지수떡 오차를 가지는 자기회귀모형에 대한 결합 사후 분포 식 (2.6)로부터 λ 와 p 가 주어졌을 때, α, ρ, σ 에 대한 조건부사후분포 식 (2.7)을 얻을 수 있다.

$$\pi(\alpha, \rho, \sigma|\lambda, p, \mathbf{y}) = \int_0^\infty \int_{-\infty}^\infty \pi(\alpha, \rho, \sigma, \lambda, p|\mathbf{y}) d\lambda dp. \quad (2.7)$$

만약 $\lambda = 0, p = 1$ 를 고려한다면, 주어진 자기회귀모형에 대해서 분포 식 (2.8)을 가진다.

$$y_t|\alpha, \rho, \sigma^2 \sim N(\alpha + \rho y_{t-1}, \sigma^2), \quad (2.8)$$

$$\pi(\alpha, \rho, \sigma^2) \propto \frac{1}{\sigma^2}. \quad (2.9)$$

그러므로, $\lambda = 0, p = 1$ 일 때, α, ρ, σ 에 대한 난수를 생성하기 위하여 사전분포로 식 (2.9)를 가진다면 계산이 편리하다. α, ρ, σ 에 대한 사후분포를 $\pi^*(\alpha, \rho, \sigma^2|\mathbf{y})$ 라 한다면, $\pi^*(\alpha, \rho, \sigma^2|\mathbf{y})$ 로부터 M^* 개의 난수를 생성한다. 이 때, M^* 는 목표분포로부터 생성하고자 하는 난수의 크기보다 훨씬 크다고 하자. 그러면 $\pi^*(\alpha, \rho, \sigma^2|\mathbf{y})$ 으로부터의 α, ρ, σ 의 조건부사후분포는 닫힌 형태이며 알려진 분포를 따르므로 깃스 표집기(Gibbs sampler)를 통해 난수를 생성할 수 있다. 생성된 난수에서 목표분포로부터의 난수를 재표집하기 위하여 고려되는 가중치는 식 (2.10)으로 나타낼 수 있다.

$$w_h \propto \frac{\int_0^1 \int_{-1}^1 \pi(\alpha^{(h)}, \rho^{(h)}, \sigma^{(h)}, \phi_1, \phi_2|\mathbf{y}) d\phi_1 d\phi_2}{\pi^*(\alpha^{(h)}, \rho^{(h)}, \sigma^{(h)}|\mathbf{y})}, \quad h = 1, \dots, M^*, \quad (2.10)$$

여기서 λ 와 p 는 $\lambda = \phi_1/\sqrt{1-\phi_1^2}, p = \sqrt{1/\phi_2} - 1$ 로 재모수화 하였으며, ϕ_1 와 ϕ_2 의 범위는 $-1 < \phi_1 < 1$ 와 $0 < \phi_2 < 1$ 을 만족한다. 재표집은 기껏해야 10%정도의 비율로 비복원 추출을 하도록한다. 예를 들면, 목표분포로부터 1000개의 난수를 생성하고자 한다면 $M^* = 10,000$ 이상의 난수를 생성해야한다. 또한, SIR 알고리즘을 적용하기 위해 고려해야할 중요한 성질은 유계성(boundedness)를 가져야 한다는 것인데, 만약 유계성을 가진다면 $w_h \approx 1/M^*$ 을 만족하며 지배적인 가중치가 없을 것이다.

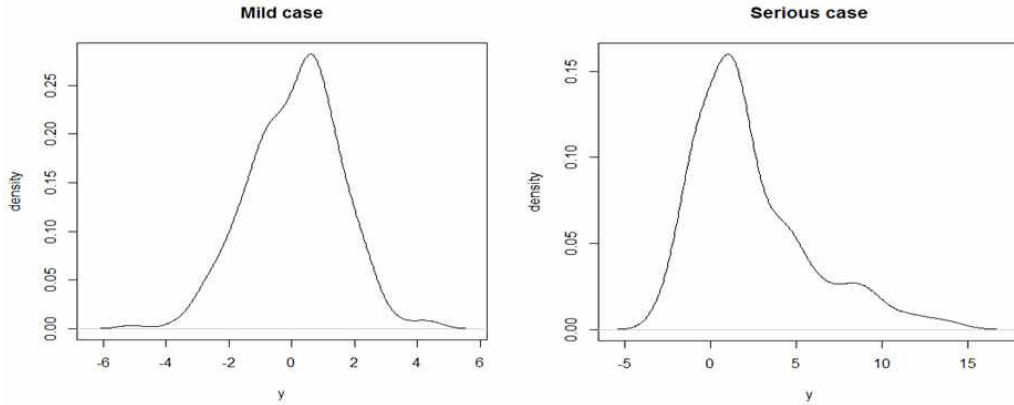


Figure 3.1. Density of the simulated data

SIR 알고리즘을 통해서 α, ρ, σ 에 대한 난수를 생성하였다면, 식 (2.11)을 통해 ϕ_1 와 ϕ_2 에 대한 난수를 생성할 수 있다.

$$\pi(\phi_1, \phi_2 | \alpha, \rho, \sigma, \mathbf{y}) = \pi(\phi_1 | \phi_2, \alpha, \rho, \sigma, \mathbf{y}) \pi(\phi_2 | \alpha, \rho, \sigma, \mathbf{y}). \tag{2.11}$$

이는 $\pi(\phi_2 | \alpha, \rho, \sigma, \mathbf{y})$ 로 부터 ϕ_2 를 생성하고, 이로부터 $\pi(\phi_1 | \phi_2, \alpha, \rho, \sigma, \mathbf{y})$ 에서 ϕ_1 를 생성하는 것으로 구체화 할 수 있다. 먼저, $\pi(\phi_2 | \alpha, \rho, \sigma, \mathbf{y})$ 로 부터 ϕ_2 ($0 < \phi_2 < 1$)을 생성하기 위하여 다음 절차를 따르는 격자망 방법을 고려하도록 한다.

1. 0에서 1사이 구간을 100개의 구간으로 분할하고, 각 구간의 경계점을 $I_k, k = 1, \dots, 101$ 라 하자.
2. 각 구간의 중간값을 계산한다($M_k, k = 1, \dots, 100$).
3. 각 중간값(M_k)에 대한 조건부분포 함수값을 구한다($a_k = \pi(M_k | \alpha, \rho, \sigma, \mathbf{y}), k = 1, \dots, 100$).
4. $A = \sum_{k=1}^{100} a_k$ 라고 할 때, $b_k = a_k/A$ 를 계산한다.
5. $U(0, 1)$ 로 부터 난수 u_1 을 생성한다.
6. 1.에서 구한 구간 중에서 $b_k \leq u_1 < b_{k+1}$ 을 만족하는 k 번째 구간을 선택한다.
7. $U(I_k, I_{k+1})$ 로 부터 u_2 를 생성하여 ϕ_2 라 한다.

비슷한 방법으로 $\pi(\phi_1 | \phi_2, \alpha, \rho, \sigma, \mathbf{y})$ 로부터 ϕ_1 를 생성할 수 있다. 단, ϕ_1 는 -1에서 1사이 구간을 적용한다.

3. 수치적 연구

3.1. 오염된 오차를 가지는 시계열 자료

아래와 같은 1차 자기회귀모형을 고려해 보자.

$$y_t = \alpha + \rho y_{t-1} + e_t, \tag{3.1}$$

Table 3.1. Bayesian estimates and HPD credible intervals under AR(1): $\delta = 1$

parameter	SEP				Normal			
	Mean	SD	95% HPD		Mean	SD	95% HPD	
ρ	0.6982	0.0704	0.5553	0.8398	0.6914	0.0732	0.5476	0.8349
σ^2	1.1914	0.1755	0.8485	1.5411	1.0621	0.1545	0.7756	1.3719
λ	0.0522	0.1680	-0.2857	0.3683				
p	1.2001	0.1199	0.9746	1.4445				

Table 3.2. Bayesian estimates and HPD credible intervals under AR(1): $\delta = 8$

parameter	SEP				Normal			
	Mean	SD	95% HPD		Mean	SD	95% HPD	
ρ	0.7265	0.0536	0.6174	0.8264	0.7951	0.0614	0.6754	0.9189
σ^2	6.0853	0.8804	4.5207	7.9123	6.9334	1.0006	5.0955	8.9435
λ	0.3035	0.1910	-0.0267	0.7088				
p	0.8319	0.0618	0.7133	0.9518				

여기서 e_t 는 독립이면서 동일한 분포를 따르며, 각 e_t 는 오염된 정규분포 $0.9 * N(0, 1) + 0.1 * N(\delta, 1)$ 를 따른다고 하자. e_t 의 분포를 근사하기 위하여 지수적 분포를 사용하였을 때 ρ 의 추정치의 성질을 알아보기 위해서 두 가지 경우의 모의실험을 고려해보도록 하자($\delta = 1$ 일 때의 약한 오염인 경우와 $\delta = 8$ 일 때의 심각한 오염인 경우). 모의실험은 $n = 100$ 이고, $\alpha = 0$, $\rho = 0.7$ 일 때를 고려하였다 (Figure 3.1 참조).

여기서 SIR 알고리즘을 위한 $\pi^*(\alpha, \rho, \sigma^2 | \mathbf{y})$ 로 부터의 난수에 대해서, 각 2개의 체인에서 1000번의 변인과정을 거친 5000개의 난수를 생성하여 10,000개($M^* = 10,000$)의 후보난수를 고려한다. 이 중 가중치를 반영하여 10%에 해당하는 1,000개의 난수를 재표집하도록 한다. 여기서 비대칭 지수적 오차를 가지는 경우와 정규 오차를 가지는 모형을 가정하였을 때, 서로 다른 오차를 가지는 AR(1) 모형의 베이지안 추정값을 비교하고자 한다.

이 모의실험에서 우리의 관심은 오차에 어떤 분포를 적합시켰을 때, 실제 모수 ρ 와 가깝게 추정되는지에 있다. 두 경우에 대해 관심모수 ρ 의 사후평균과 사후표준편차, 95% HPD 구간은 Table 3.1-Table 3.2와 같다. 약한 오염 상태에서는 비대칭 지수적 오차를 가질 경우에 ρ 의 추정치가 0.6982, 정규 오차를 가질 경우의 추정치가 0.6914로 두 경우 모두 실제 모수 값인 0.7과 가장 가깝게 나타났으며, 95% HPD 구간을 확인해 보았을 때, 두 경우 모두 실제모수 0.7을 포함하고 있다. 심각한 오염 상태(즉, 약한 오염 상태보다 두꺼운 꼬리의 경우)에서는 비대칭 지수적 오차를 가질 경우에 ρ 의 추정치가 0.7265으로 정규 오차를 가지는 경우와 비교하여 실제 모수 값인 0.7과 더 가깝게 나타났다. 약한 오염상태의 경우 비대칭 지수적 오차를 가지는 경우와 정규 오차의 경우가 큰 차이가 없는 것으로 나타났으며, 추정된 λ 와 p 또한 각각 0.0522와 1.2001로 정규분포의 경우와 거의 비슷함을 알 수 있다. 반면에 심각한 오염 상태에서는 추정된 λ 와 p 가 각각 0.3035, 0.8319로 λ 는 0보다 크며, p 는 1보다 작게 나타남을 알 수 있다.

3.2. 이상치를 가지는 시계열 자료

1720년 부터 1970년 까지의 태양 흑점의 연평균 갯수를 고려하도록 한다. 자료는 <http://www.ngdc.noaa.gov/stp/SOLAR/solar.html>에서 이용가능하다. 태양 흑점 갯수 자료에 대하여 보통 2차 자기회귀모형(AR(2))을 고려하며 이는 식 (3.2)로 나타낼 수 있다.

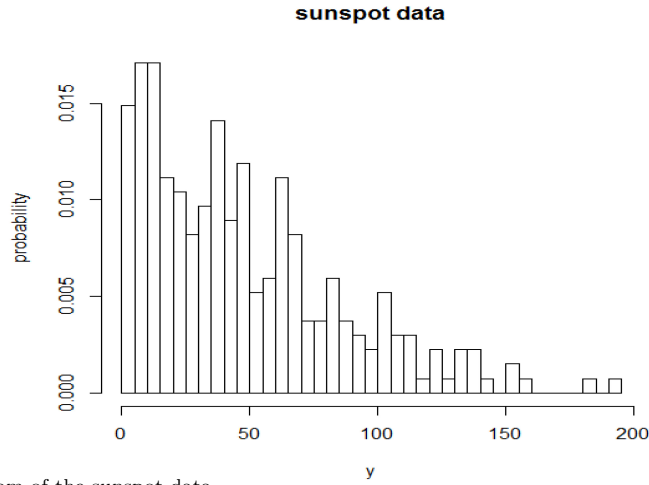


Figure 3.2. Histogram of the sunspot data

Table 3.3. Bayesian estimates and HPD credible intervals under AR(2)

parameter	SEP				Normal			
	Mean	SD	95% HPD		Mean	SD	95% HPD	
α	15.5914	2.1652	11.4056	19.7712	15.3798	1.7202	12.0133	18.8005
ρ_1	1.3888	0.0422	1.3060	1.4761	1.3878	0.0447	1.3007	1.4776
ρ_2	-0.7014	0.0433	-0.7874	-0.6134	-0.6983	0.0453	-0.7885	-0.6100
σ^2	271.3599	31.6944	210.4090	332.3888	262.8848	23.8604	218.2910	310.1391
λ	0.0241	0.2108	-0.3968	0.4539				
p	1.0377	0.1143	0.8160	1.2647				

$$y_t = \alpha + \rho_1 y_{t-1} + \rho_2 y_{t-2} + e_t. \tag{3.2}$$

적합된 모형에 대한 모수의 최대우도추정량은 $\hat{\rho}_1 = 1.3836$, $\hat{\rho}_2 = -0.6963$, $\hat{\sigma}^2 = 259$ 이다. Figure 3.2는 몇개의 이상점을 가지는 태양 흑점 자료의 히스토그램이다.

2절에서 언급한 AR(1)에 대한 추정 문제를 AR(2)로 확장 시키기 위해 정상성을 고려하여 AR(1) 모형에서의 ρ 에 대한 사전분포 $U(-1, 1)$ 대신에 ρ_1 과 ρ_2 에 대한 사전분포로 각각 $U(-2, 2)$ 와 $U(-1, 1)$ 을 고려하도록 한다. AR(2) 모형에 대한 추론도 3.1절과 마찬가지로 여기서 SIR 알고리즘을 위한 $\pi^*(\alpha, \rho_1, \rho_2, \sigma^2 | \mathbf{y})$ 로 부터의 난수에 대해서, 각 2개의 체인에서 1000번의 번인과정을 거친 5000개의 난수를 생성하여 10,000개($M^* = 10,000$)의 후보난수를 고려한다. 이 중 가중치를 반영하여 10%에 해당하는 1,000개의 난수를 재표집하도록 한다. 여기서 비대칭 지수떡 오차를 가지는 경우와 정규 오차를 가지는 모형을 가정하였을 때, 서로 다른 오차를 가지는 AR(2) 모형의 베이시안 추정값을 비교하고자 한다. 이 자료에서 우리의 관심은 어떤 분포로부터의 결과가 더 잘 적합하는지와 예측이 충분히 효율적인지에 있다. 관심 모수 ρ_1 과 ρ_2 의 사후 평균과 사후표준편차, 95% HPD 구간은 Table 3.3와 같다.

우리가 고려한 여러가지 모형을 비교하기 위하여, 베이시안 사후예측 유의확률(Bayesian posterior predictive p -value; PPP)을 고려하도록 하자. 자료의 사후예측 반복을 이용한 베이시안 형태는 자료와 미지의 모수에 의존하는 검정통계량의 사용을 가능하게 하며, 통계량과 모수 값의 불일치 측도의 사용이

Table 3.4. Posterior predictive p -values for possible models

Model	Posterior predictive p -value
SEP	0.5310
Normal	0.5630

가능하다. 이 베이지안 사후 예측 유의확률은 프리컨티스트 관점의 유의확률의 사후평균으로 이해할 수 있다 (Meng, 1994). $\theta_{i,t} = \alpha_i + \rho_1 y_{t-1} + \rho_2 y_{t-2}$ 과 σ_i 는 각 i 번째 반복에서 얻어진다고 하자. 그리고 \mathbf{y}_{obs} 는 관측치의 벡터형태, \mathbf{y}_{new} 는 사후 예측값의 벡터형태라고 하자. 그러면, $d(\mathbf{y}, \boldsymbol{\theta}_i)$ 는 i 번째 불일치 측도라 할 수 있고, $d(\mathbf{y}, \boldsymbol{\theta}_i)$ 는 식 (3.3)과 같다.

$$d(\mathbf{y}, \boldsymbol{\theta}_i) = \sum_{t=1}^n \frac{(y_t - \theta_{i,t})^2}{\sigma_i^2}, \quad i = 1, \dots, \# \text{ of replication.} \quad (3.3)$$

그러므로, 자료에 대한 불일치측도는 $d(\mathbf{y}_{\text{obs}}, \boldsymbol{\theta})$ 와 $d(\mathbf{y}_{\text{new}}, \boldsymbol{\theta})$ 이고, $d(\mathbf{y}_{\text{obs}}, \boldsymbol{\theta})$ 보다 $d(\mathbf{y}_{\text{new}}, \boldsymbol{\theta})$ 가 클 사후 확률을 PPP라고 한다. 만약 제안된 모형이 잘 적합된다면, 계산된 PPP는 0.5에 가깝게 나타날 것이다. 이 자료의 적합결과에 대한 PPP는 Table 3.4과 같다. 그 결과, 몇 개의 이상점을 포함하는 태양 흑점 자료는 비대칭 지수떡 분포가 가장 잘 적합하는 것으로 나타났다.

4. 결론

본 논문에서는 오차에 보다 유연한 형태의 분포를 설정한 자기회귀모형에서의 베이지안 추론을 고려하였다. 비대칭 지수떡 분포족은 정규분포, 지수떡분포, 비대칭 정규분포를 포함하며, 정규분포보다 두껍거나 얇은 꼬리를 가지는 경우, 정규분포보다 비뚤어진 경우를 포괄한다. 자기회귀모형에서의 오차항에 이러한 비대칭 지수떡 분포를 이용함으로써 관측된 자료가 두꺼운 꼬리를 가지거나 비뚤어진 형태의 분포 특성을 가진다면 모수 λ 와 p 에 의하여 다른 모형보다 왜도와 첨도 특성을 더 잘 반영할 수 있으며, 이로 인해 정규성 가정을 완화 할 수 있음을 알 수 있다. 또한 제안된 분포에 대하여 SIR 알고리즘과 격자망 방법을 사용함으로써 효율적인 베이지안 추론을 가능하게 하였다.

References

- Azzalini, A. (1985). A Class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, **12**, 171–178.
- Azzalini, A. (1986). A further results on a class of distributions which includes the normal ones, *Statistica*, **46**, 199–208.
- Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*, New York : Wiley.
- DiCiccio, T. J. and Monti, A. C. (2004). Inferential aspects of the skew exponential power distribution, *Journal of the American Statistical Association*, **99**, 439–450.
- Meng, X. L. (1994). Posterior predictive p -values, *The Annals of Statistics*, **22**, 1142–1160.
- Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A non-iterative sampling/importance resampling alternative to the data augmentation algorithm for creating a Few imputations when fractions of missing information are modest: The SIR algorithm, *Journal of the American Statistical Association*, **82**, 543–546.

비대칭 지수멩 오차를 가지는 자기회귀모형에서의 베이지안 추론

류현남^a · 김달호^{a,1}

^a경북대학교 통계학과

(2014년 9월 1일 접수, 2014년 10월 23일 수정, 2014년 10월 27일 채택)

요약

시계열 자료를 위한 가장 기본적인 모형인 자기회귀모형을 고려한다. 흔히 시계열 자료에서 정규성 가정이 위배되는 경우가 발생하며, 정규성 가정을 완화하기 위한 방법으로 두꺼운 꼬리를 가지는 분포 또는 비대칭 분포를 고려할 수 있다. 비대칭 지수멩 분포의 사용은 비뚤림이 있는 두꺼운 꼬리를 가지는 자기회귀모형의 이상치의 영향을 줄이고 로버스트한 추론을 할 수 있도록 한다. 본 논문에서는 자기회귀모형에 대한 오차항에 정규분포 보다 첨도와 왜도에 유연함을 가지는 분포를 고려함으로써 정규성 가정을 완화하여 추론하고자 하였다. 정규분포의 대안으로 비대칭 지수멩 분포를 고려하였으며 정규분포의 결과와 비교 하여 비대칭 지수멩 분포의 로버스트함을 보였다. 또한 주어진 분포에 대한 효율적인 베이지안 추론을 하기 위하여 SIR 알고리즘과 격자망 방법을 고려하였다.

주요용어: 자기회귀모형, 베이지안 p -값, 비대칭 지수멩 분포, 김스 표본자, 로버스트.

¹교신저자: (702-701) 대구광역시 북구 대학로 80, 경북대학교 통계학과. E-mail : dalkim@knu.ac.kr