응용통계연구 (2011)
24(5), 833–845
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.5.833

코스피 지수 자료의 베이지안 극단값 분석

윤석훈[1]

[1]수원대학교 통계정보학과

(2011년 9월 접수, 2011년 10월 채택)

요 약

본 논문에서는 1998.01.03부터 2011.08.31까지 수집된 코스피 지수 자료로부터 계산된 일별 로그수익률과 일별 로그손실률에 대한 극단값 통계분석을 수행하였다. 사용된 극단값 통계분석 모형은 포아송-GPD 모형이고 모수의 추정과 극단분위수의 추정은 최대가능도 방법을 적용하였다. 본 논문에서는 또한 포아송-GPD 모형에 추가적으로 모수의 무정보사전분포를 가정한 베이지안 방법을 고려하였다. 여기서는 마르코프 연쇄 몬테칼로 방법을 적용하여 모수와 극단분위수를 추정하였다. 분석 결과 최대가능도 방법과 베이지안 방법에서 모두, 로그수익률 분포의 오른쪽 꼬리는 정규분포보다 짧은 반면, 로그손실률 분포의 오른쪽 꼬리는 정규분포보다 두텁다는 결론이 얻어졌다. 극단값 분석에서 베이지안 방법을 사용할 때의 장점은 정칙조건이 만족되지 않는 경우에도 최대가능도추정량의 전통적 점근성질을 걱정할 필요가 없고 예측의 경우에는 모수의 불확실성과 미래 관측치의 불확실성이 모두 반영되는 효과가 있다는 것이다.

주요용어: 코스피 지수, 극단값 이론, 포아송-GPD 모형, 베이지안 방법, 무정보사전분포, 마르코프 연쇄 몬테칼로 방법.

1. 서론

코스피 지수(KOSPI; Korea Composite Stock Price Index)는 원래 종합주가지수라고 불리우다가 2005년 11월 1일부터 현재의 이름으로 바뀌어 사용되고 있다. 코스피 지수는 증권시장에 상장된 전체 상장기업의 주식 가격 변동을 기준시점과 비교시점을 비교하여 작성한 지표로서, 산출방법은 1980년 1월 4일을 기준시점으로 하여 이 날의 지수를 100으로 정하고 개별 종목의 주가에 상장주식수를 가중한 비교시점의 시가총액을 기준시점의 시가총액에 대비하여 산출한다.

우리나라의 경우 주식시장의 과민 반응에 따른 일시적인 가격 급등락을 방지하고 효율적인 가격 발견을 촉진한다는 정책 목표 아래 개별 주식 거래에 대해 가격제한폭제도를 채택하고 있는데, 이 제도의 변천 과정을 살펴보면 5단계 정액제(1977.02∼1986.12), 17단계 정액제(1986.12∼1995.04), 6% 정율제(1995.04∼1996.11), 8% 정율제(1996.11∼1998.03), 12% 정율제(1998.03∼1998.12), 15% 정율제(1998.12∼현재)와 같다. 즉, 현재 우리나라 주식시장에서 거래되는 개별 주식은 전일 종가 대비 상승 및 하락률이 각각 15%를 초과할 수 없도록 개인투자자 보호를 위해 제한시켜놓은 것인데, 현재 개별 주식에 대해 가격제한폭제도를 명시적으로 채택하고 있는 나라는 한국, 대만, 일본, 중국, 태국 등 주로 아시아 지역의 국가들이다. 이 외에도 안정적인 주식시장을 운용하기 위한 보완장치로서 서킷브레이커(circuit breakers: 코스피 지수의 폭락이 전일 대비 10% 이상인 상태가 1분 이상 지속되면 모든 주식거래를 30분간 중단시키는 제도로서 1998.12부터 도입)와 사이드카(sidecar: 주가지수 선물가격이 전

[1](445-743) 경기도 화성시 봉담읍 와우리 산2-2, 수원대학교 통계정보학과, 부교수. E-mail: syun@suwon.ac.kr

일 종가 대비 5% 이상(코스닥은 6% 이상) 상승 또는 하락하는 상태가 1분간 지속될 때 프로그램 매매 호가의 효력을 5분간 정지시키는 제도로서 선물시장이 급변할 경우 현물시장에 대한 영향을 최소화하기 위하여 주가지수 선물시장을 개설하면서부터 도입) 등의 제도가 추가적으로 더 있다.

본 논문에서는 개별 주식 거래에 대해 12% 이상의 정율제 가격제한폭제도가 채택된 1998년 이후의 코스피 지수 자료만을 대상으로 하여, 코스피 지수에 투자하였을 경우 발생 가능한 일별 수익률과 일별 손실률에 대한 극단값 통계분석을 수행하였다. 여기서, 주된 관심 사항은 수익률 분포와 손실률 분포에서 극단분위수(extreme quantile)의 추정 문제이다. 사용된 극단값 통계분석 모형은 포아송-GPD 모형이고 모수의 추정이나 극단분위수의 추정은 최대가능도(maximum likelihood) 방법을 이용했는데, 일별 수익률 자료의 경우 상위 5.23%가, 그리고 일별 손실률 자료의 경우에는 상위 4.21%가 극단값 분석에 실제로 사용되었다. 또한, 수익률 분포 또는 손실률 분포의 오른쪽 꼬리가 매우 짧은 경우에는 포아송-GPD 모형에서 모수의 최대가능도추정량들이 전통적 점근성질(asymptotic property)을 갖고 있지 못하기 때문에, 이의 대안으로 본 논문에서는 베이지안 방법을 고려하였다.

2. 극단값에 대한 포아송-GPD 모형과 베이지안 방법

시계열 자료의 여러 극단값 통계 분석 방법 중에서 가장 역사가 오래된 방법은 연간 최대값 방법(annual maximum method)이다 (Gumbel, 1958). 이 방법에서는 시계열 자료의 연간 최대값들의 분포 함수 모형으로 다음의 일반화극단값분포함수(generalized extreme value distribution function)를 가정하는데 이 분포를 간단히 기호 GEV(μ, σ, ξ), $\mu \in \mathbb{R}$, $\sigma > 0$, $\xi \in \mathbb{R}$로 나타내기로 한다 (von Mises, 1936).

$$G(x; \mu, \sigma, \xi) := \exp\left[-\left\{ 1 + \frac{\xi(x-\mu)}{\sigma} \right\}_+^{-\frac{1}{\xi}} \right], \quad x \in \mathbb{R}. \tag{2.1}$$

여기서, $z_+ := \max\{z, 0\}$이고, $\xi = 0$인 경우는 항상 $\xi \to 0$일 때의 극한으로 해석한다. 즉, $G(x; \mu, \sigma, 0) = \exp\{-e^{-(x-\mu)/\sigma}\}$. 또한, μ, σ, ξ는 각각 위치모수, 척도모수, 형상모수를 나타내는데, 특별히 ξ는 주어진 시계열 자료의 주변분포의 오른쪽 꼬리의 형태를 결정짓는 중요 모수로서 극단값지수(extreme value index) 또는 꼬리지수(tail index)라고 불리우기도 한다.

연간 최대값 방법은 GEV(μ, σ, ξ) 분포가 대체로 적절히 정규화된 표본 최대값(sample maximum)의 극한분포 (Fisher와 Tippett, 1928; Gnedenko, 1943)라는 장점에도 불구하고 시계열 자료 중 연간 최대값들만을 사용하기 때문에 아주 오랜 기간 동안 관찰된 자료가 아니라면 추정치의 오차가 매우 커질 수 있다는 단점을 갖는다. 따라서, 제한된 기간 동안 관찰된 시계열 자료의 극단값 분석은 시계열 자료의 주변분포의 오른쪽 꼬리에 대한 정보를 가지고 있는 추가적인 자료, 즉 충분히 큰 하나의 분계점(threshold) u를 선택하고 이를 넘어서는 모든 초과값(exceedance)들을 사용하는 것이 보다 효율적인 방법이 될 수 있는데, 이러한 방법을 통상 분계점 방법(threshold method)이라고 부른다 (Davison과 Smith, 1990). 좀 더 구체적으로, 분계점 방법에서는 분계점 u를 넘어서는 초과값들의 초과여분(excess) 분포 함수 모형으로 다음의 일반화파레토분포함수(generalized Pareto distribution function)를 가정하는데 (Pickands, 1975) 이 분포를 간단히 기호 GPD(ϕ, ξ), $\phi > 0$, $\xi \in \mathbb{R}$로 나타내기로 한다.

$$H(y; \phi, \xi) := 1 - \left(1 + \frac{\xi y}{\phi} \right)_+^{-\frac{1}{\xi}}, \quad y > 0. \tag{2.2}$$

여기서, ϕ와 ξ는 각각 척도모수와 형상모수를 나타내는데, 특별히 ξ는 일반화극단값분포함수 (2.1)에서의 형상모수와 일치한다. 분계점 방법에서의 분계점 u를 선택하는 하나의 통상적인 방법은 시계열 자료

의 일별 관측값들의 MRL 그림(mean residual life plot)을 이용하는 것인데, 이에 따르면 적절한 분계점으로, 모든 $t > u$에 대하여 t와 t를 넘어서는 초과값들의 초과여분 산술평균값의 산점도가 근사적으로 직선을 따르는 최소의 u 값으로 선택하는 것이다 (윤석훈, 2010).

일반적으로 분계점 방법에서의 일반화파레토분포함수 모형 (2.2)와 연간 최대값 방법에서의 일반화극단값분포함수 모형 (2.1) 사이의 모수들의 관계는 형상모수 ξ가 동일하다는 것 외에는 특별히 다른 관계는 없다. 따라서, 상대적으로 오차가 작을 것으로 예상되는, 분계점 방법으로 추정한 추정치를 이용하여 연간 최대값 방법에서의 모수를 모두 추정하는 것은 불가능하고, 식 (2.1)의 일반화극단값분포함수를 이용하여 t년(t는 대체로 큰 값) 재발수준(t-year return level), 즉 t년에 한 번 정도 넘어서는 높은 수준을 추정하는 것도 역시 불가능하게 된다. 이를 해결할 수 있는 하나의 일반적인 방법으로 분계점 u를 넘어서는 초과값들에 대하여 2차원 포아송초과점과정(Poisson exceedance point process) 모형을 적용할 수 있는데 (Smith, 1989; 윤석훈, 2009), 여기서는 보다 간편한 형태인 포아송-GPD 모형을 사용하기로 한다 (윤석훈, 2010).

포아송-GPD 모형에서는 분계점 방법에서와 같이 분계점 u를 넘어서는 초과값들의 초과여분 분포 함수 모형으로 GPD(ϕ, ξ) 분포를 가정하고 동시에 u를 넘어서는 연간 초과값들의 개수가 평균이 λ인($\lambda > 0$) 포아송분포를 따른다고 가정한다. 이 경우

$$\mu = u + \frac{\phi(\lambda^{\xi} - 1)}{\xi}, \quad \sigma = \phi\lambda^{\xi} \tag{2.3}$$

로 정의하면, 연간 최대값의 분포 함수의 꼬리가 식 (2.1)의 GEV(μ, σ, ξ) 분포의 꼬리와 일치하는 것을 보일 수 있다 (윤석훈, 2010). 따라서, 포아송-GPD 모형을 사용하면 연간 최대값 방법에서와 달리 보다 많은 자료를 사용하게 되어 추정치의 오차가 작아지게 되고 또한 식 (2.3)의 관계를 이용하여 연간 최대값들의 분포 함수 모형인 식 (2.1)의 모수 추정이 가능해진다. 이렇게 얻어진 추정치들을 사용하면 일반화극단값분포함수를 이용한 t년 재발수준의 추정도 역시 가능하게 된다.

이야기를 구체적으로 전개하기 위하여, 이제 서로 *i.i.d.*(independent and identically distributed)인 일별 관측값들이 m년 동안 관찰되었다고 하고 충분히 큰 하나의 분계점 u가 적절히 선택되었다고 하자. 이 경우, 포아송-GPD 모형에서는 이들 중 u를 넘어서는 초과값들로 $z_1, \ldots, z_n$이 관찰되었다고 할 때 초과값의 전체 개수 n은 평균이 $m\lambda$인 포아송분포를 따르고 초과값들의 초과여분 $z_1 - u, \ldots, z_n - u$는 GPD(ϕ, ξ) 분포를 따른다고 가정하는 것이므로 이들 n개의 초과값들에 기초한 가능도함수(likelihood function) $L(\lambda, \phi, \xi)$는 다음과 같이 주어진다.

$$L(\lambda, \phi, \xi) = e^{-m\lambda} \frac{(m\lambda)^n}{n!} \prod_{i=1}^{n} \left[\frac{1}{\phi} \left\{ 1 + \frac{\xi(z_i - u)}{\phi} \right\}_{+}^{-\frac{1}{\xi} - 1} \right].$$

따라서, 이를 이용하면 모수 λ, ϕ, ξ의 최대가능도추정치(MLE; maximum likelihood estimate)의 계산이 가능하고, 이로부터 식 (2.3)을 이용하여 역시 μ, σ의 최대가능도추정치의 계산이 가능해지며, 또한 이렇게 얻어진 추정치를 다음 식에 대입하면 t년 재발수준 q_t의 최대가능도추정치를 얻을 수 있다.

$$\begin{aligned} q_t &= G^{-1}\left(1 - \frac{1}{t}; \mu, \sigma, \xi\right) \\ &= \mu + \frac{\sigma\left[\{-\log(1 - 1/t)\}^{-\xi} - 1\right]}{\xi}. \end{aligned} \tag{2.4}$$

식 (2.1)의 GEV(μ, σ, ξ) 분포와 식 (2.2)의 GPD(ϕ, ξ) 분포는 분포의 끝점(end-points)이 모수 값에 의존되기 때문에 모수의 최대가능도추정량들이 일반적으로 정칙조건(regularity condition) 하에서 성립

하는 일치성(consistency), 점근효율성(asymptotic efficiency), 점근정규성(asymptotic normality)의 전통적 점근성질을 가지고 있다고 단언할 수 없다. 실제로, Smith (1985)는 $\xi > -0.5$인 경우 최대가능도추정량들은 전통적 점근성질을 가지고 있으나, $-1 < \xi < -0.5$인 경우에는 그렇지 않다는 것과 $\xi < -1$인 경우에는 최대가능도추정량 자체가 존재할 수 없다는 사실을 보였다. 따라서, 이와 비슷한 상황 하에서는 최대가능도추정량의 대안을 생각해야 되는데 베이지안 방법이 그 하나가 될 수 있을 것이다 (Coles와 Powell, 1996).

이제, 위에 소개한 포아송-GPD 모형에 베이지안 방법을 적용해보기로 한다. 베이지안 방법에서는 모수를 확률변수로 가정하고 이에 대한 분포로 우선 사전분포(prior distribution)를 적절히 선택하여 사용한다. 이때 선택된 사전분포는 모수 값에 대한 통계분석자의 주관적인 믿음을 모형화시킨 것으로서 대체적으로 경험적 입증이 곤란한 주관적 확률분포를 의미한다. 베이지안 방법에서는 모수에 대한 이러한 사전 정보와 관찰된 자료의 통계모형(여기서는, 포아송-GPD 모형)으로부터의 정보를 결합하여 베이지안 규칙에 따라 모수의 사후분포(posterior distribution)를 규명하고, 이 사후분포를 이용하여 모수에 대한 추론을 수행한다. 따라서, 베이지안 추론은 관찰된 자료의 양이 적으면 모수의 사전분포에 많이 좌우되고, 반대로 자료의 양이 많으면 자료의 통계모형에 많이 좌우될 것이라는 것을 쉽게 예상해 볼 수 있다. 모수의 사전분포 선택은 모수에 대한 사전 정보의 유무와 사전 정보의 질에 따라 달라질 수 있는데, 여기서는 모수에 대한 구체적인 사전 정보가 전혀 없는 것으로 가정하여 무정보사전분포(noninformative prior distribution)를 사용하기로 한다. 즉, λ, ϕ, ξ가 서로 독립이고 이들의 사전밀도함수(prior density function) $\pi(\lambda, \phi, \xi)$가 다음의 무정보사전분포를 만족한다고 가정한다.

$$\pi(\lambda, \phi, \xi) \propto \frac{1}{\lambda} \times \frac{1}{\phi}, \quad \lambda > 0,\ \phi > 0,\ \xi \in \mathbb{R}.$$

이 경우, (λ, ϕ, ξ)의 사후밀도함수(posterior density function) $\pi(\lambda, \phi, \xi|\text{data})$는 다음과 같이 주어진다.

$$\begin{aligned}\pi(\lambda, \phi, \xi|\text{data}) &\propto \pi(\lambda, \phi, \xi) L(\lambda, \phi, \xi) \\ &\propto \frac{1}{\lambda\phi} L(\lambda, \phi, \xi).\end{aligned}$$

여기서 다시, 모든 모수의 값이 실수값이 되도록

$$\theta_1 = \log \lambda, \qquad \theta_2 = \log \phi, \qquad \theta_3 = \xi$$

로 모수변환을 하면 $\theta = (\theta_1, \theta_2, \theta_3)$의 사후밀도함수 $\pi(\theta|\text{data})$는 다음과 같이 주어진다.

$$\pi(\theta|\text{data}) \propto L\left(e^{\theta_1}, e^{\theta_2}, \theta_3\right), \quad \theta_i \in \mathbb{R},\ i = 1, 2, 3. \tag{2.5}$$

여기서, 사후밀도함수 $\pi(\theta|\text{data})$가 적절한 확률밀도함수가 되기 위한 정규화 상수(normalizing constant)는 일반적으로 정확한 계산이 불가능하여 알 수 없으므로 모수의 사후평균(posterior mean)과 같은 베이지안 추론이 정확히 이루어질 수는 없다. 그러나, 최근 널리 사용되고 있는 마르코프 연쇄 몬테칼로(MCMC; Markov chain Monte Carlo) 방법을 사용하면, 여기서와 같이 정규화 상수를 모를 경우에도 사후분포로부터의 시뮬레이션이 가능해지는데, 이와 같은 방법으로 얻어진 시뮬레이션 값들을 이용하면 사후평균의 추정치를 쉽게 계산할 수 있다. 사후밀도함수 $\pi(\theta|\text{data})$로부터 시뮬레이션 값 $\theta^1, \theta^2, \ldots$들을 얻는 일반적인 MCMC 방법은, 임의의 초기값 θ^0와 하나의 마르코프 연쇄를 정의하면서 쉽게 시뮬레이션 할 수 있는 전이밀도함수(transition density function) $p(\theta^*|\theta^{i-1})$가 주어져 있다고 할 때, $i = 1, 2, \ldots$에 대하여 $p(\theta^*|\theta^{i-1})$로부터 우선 후보 θ^*를 하나 생성한 다음,

$$\alpha\big(\theta^{i-1}, \theta^*\big) := \min\left\{\frac{\pi(\theta^*|\text{data}) p(\theta^{i-1}|\theta^*)}{\pi(\theta^{i-1}|\text{data}) p(\theta^*|\theta^{i-1})}, 1\right\}$$

로 놓고 θ^i를

$$\theta^i = \begin{cases} \theta^*, & \text{with probability } \alpha\left(\theta^{i-1}, \theta^*\right), \\ \theta^{i-1}, & \text{with probability } 1-\alpha\left(\theta^{i-1}, \theta^*\right) \end{cases}$$

의 방법으로(즉, $\alpha(\theta^{i-1}, \theta^*)$는 후보 θ^*를 θ^i의 생성 값으로 채택할 채택 확률을 나타냄) 생성해내는 것이다. 이 경우, 생성된 $\theta^1, \theta^2, \ldots$ 값들은 사용된 전이밀도함수 $p(\cdot\,|\,\cdot)$에 대한 일정한 정칙조건 하에서 정상분포(stationary distribution) $\pi(\theta|\text{data})$를 갖는 하나의 마르코프 연쇄로부터의 관찰치로 간주될 수 있는데, 따라서 초기값 θ^0의 영향력이 없어지는 충분히 큰 값 k가 정해지면 $\theta^{k+1}, \theta^{k+2}, \ldots$ 값들을 이용하여 사후분포에 대한 근사 추론이 가능하게 된다.

본 논문에서는 간단히 후보 θ^*의 생성 전이밀도함수로 $p(\theta^*|\theta^{i-1}) = f(\theta^* - \theta^{i-1})$을 사용하기로 한다. 여기서, f는 평균이 $\mathbf{0}$인 하나의 3변량정규분포의 확률밀도함수를 나타낸다. 이는 곧

$$\theta^* = \theta^{i-1} + \tilde{\theta}, \quad \tilde{\theta} \sim f \tag{2.6}$$

로 표현 가능하므로 이와 같은 종류의 알고리즘을 간단히 확률보행(random walk) Metropolis 알고리즘이라고 부른다 (Albert, 2009; Robert와 Casella, 2004).

극단값 자료 분석을 하는 여러 중요한 이유 중의 하나인 특정 극단 수준에 도달할 미래의 사건 발생 확률을 추정하는 일은 베이지안 방법에서 매우 자연스럽다. 예를 들어, M이 미래의 연간 최대값을 나타내는 확률변수라고 하자. 이제, 과거 자료 data가 주어졌을 때, 관심 모수 θ에 대한 모든 정보는 사후밀도함수 $\pi(\theta|\text{data})$가 가지고 있으므로 M의 사후예측분포함수(posterior predictive distribution function)는

$$P\{M \le x|\text{data}\} = \int_{\mathbb{R}^3} P\{M \le x|\theta\}\pi(\theta|\text{data})\, d\theta \tag{2.7}$$

와 같이 주어진다. 여기서, 우리가 사용하고 있는 사후밀도함수는 식 (2.5)로 주어진 것이고, 따라서 식 (2.7)의 우변은 정확한 계산이 일반적으로 불가능하다. 그러나, 위에 설명한 MCMC 방법을 사용하여, 사후분포로부터 예를 들어 모수의 시뮬레이션 값 $\theta^1, \theta^2, \ldots, \theta^K$를 얻었다고 하면, 식 (2.7)의 우변은 $(1/K)\sum_{i=1}^{K} P\{M \le x|\theta^i\}$로 근사시킬 수 있으므로, 미래 t년 재발수준 q_t는

$$1 - \frac{1}{t} = P\{M \le q_t|\text{data}\} \approx \frac{1}{K}\sum_{i=1}^{K} P\left\{M \le q_t|\theta^i\right\}$$

를 만족하게 된다. 여기서, $P\{M \le q_t|\theta^i\}$의 값은 연간 최대값 분포 함수 모형으로 사용하고 있는 식 (2.1)의 일반화극단값분포함수로 대치하면 결국 q_t의 수치적 계산이 가능해진다. 이렇게 계산된 재발수준 q_t의 값은 모수의 사후분포가 가지고 있는 불확실성과 미래 관측치의 변동성에 의한 불확실성이 모두 반영된 값이므로, 사후분포만으로 추정한 값보다 예측값으로 더욱 적절하게 사용될 수 있다.

3. 코스피 지수 자료의 극단값 분석

본 논문에서 다룬 코스피 지수 자료는 웹 사이트 야후(http://finance.yahoo.com)에서 다운로드한 것으로서 1998년 1월 3일부터 2011년 8월 31일까지 총 13년 8개월 동안의 일별 코스피 종가 자료이다. 이 기간 동안 결측치를 제외한 총 관측치의 개수는 3,421개인데 그림 3.1은 이를 보여준다. 여기서, 우리의 주된 관심사는 코스피 지수 자체보다도 코스피 지수에 투자하였을 경우 얻게 되는(혹은 잃게 되는) 일

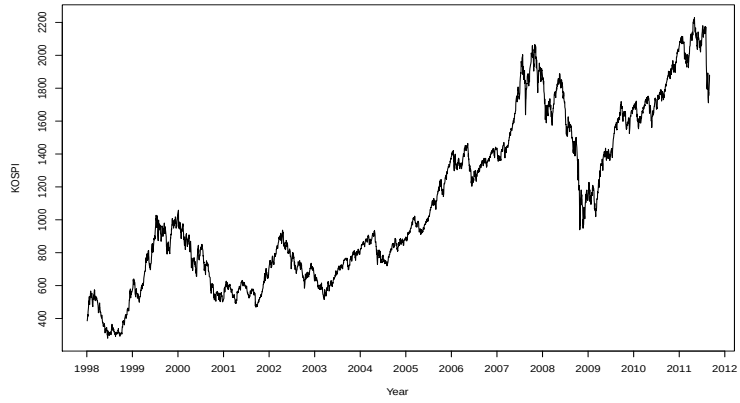

그림 3.1. 일별 코스피 지수

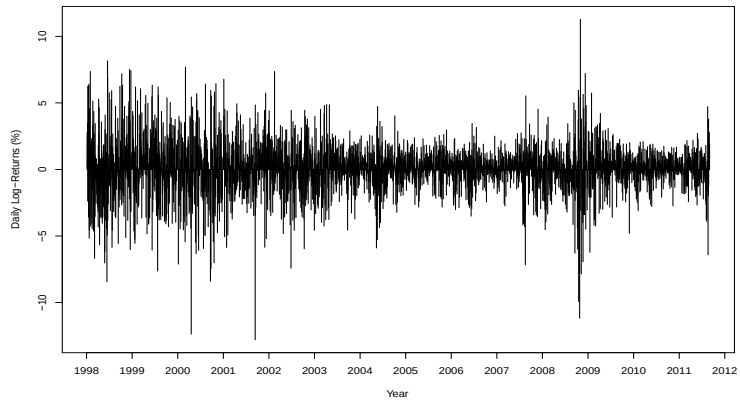

그림 3.2. 코스피 지수의 일별 로그수익률(%)

별 로그수익률(log-returns)과 일별 로그손실률(negative log-returns)이다. 코스피 지수의 t일 로그수익률(단위: %)과 로그손실률(단위: %)은 각각

$$x_t = \pm 100 \times \log\left(\frac{t\text{일 코스피 지수}}{(t-1)\text{일 코스피 지수}}\right)$$

와 같이 계산될 수 있는데, 그림 3.2는 코스피 지수의 일별 로그수익률을 보여준다. 그림 3.1과 달리 그림 3.2에서는 자료 간 상관성이 없어 보이는데, 여기서는 $\{x_t\}$가 *i.i.d.* 변수들의 관측값이라고 가정한다.

우선, 2장의 포아송-GPD 모형으로 일별 로그수익률과 일별 로그손실률의 극단값 통계분석을 하기 위해서는 분계점 방법에서 일반적으로 널리 사용되는 MRL 그림을 이용하여 적절한 분계점 u를 각각 찾아야 하는데, 그림 3.3은 로그수익률의 MRL 그림과 로그손실률의 MRL 그림을 보여준다. 로그수익률의 MRL 그림에서 $t = 7.4$를 넘어서는 로그수익률 5개를 제외하면 t가 $(3, 7.4)$에서 움직일 때 MRL 그림이 (약한) 음의 기울기를 갖는 직선 형태이므로 $u = 3$을 분계점으로 선택하고, 로그손실률의 MRL 그림에서는 $t = 8$을 넘어서는 로그손실률 6개를 제외하면 t가 $(3.5, 8)$에서 움직일 때 MRL 그림이 (약한) 양의 기울기를 갖는 직선 형태이므로 $u = 3.5$를 분계점으로 선택한다. 이 경우 분계점의 초과율은 로그

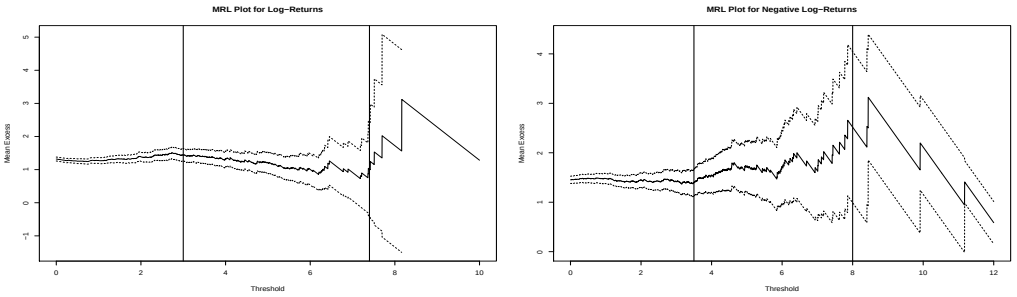

그림 3.3. 코스피 지수의 로그수익률 MRL 그림과 로그손실률 MRL 그림

표 3.1. 로그수익률 자료와 로그손실률 자료에서 분계점 u를 넘어서는 초과값들에 대한 기초 통계량(n: 표본크기, Q_1: 1사분위수, Q_2: 2사분위수, Q_3: 3사분위수, $\bar{x}$: 평균)

자료	u	n	최소값	Q_1	Q_2	$\bar{x}$	Q_3	최대값
로그수익률	3	179	3.01	3.45	4.10	4.45	5.06	11.28
로그손실률	3.5	144	3.50	3.81	4.31	4.90	5.46	12.80

표 3.2. 로그수익률 자료와 로그손실률 자료에서 분계점 u를 넘어서는 초과값들에 대한 포아송-GPD 모형 적합 결과(모수 추정치는 MLE이고 괄호 안은 추정치의 표준오차를 나타냄)

자료	u	λ	ϕ	ξ	μ	σ
로그수익률	3	13.10(0.98)	1.61(0.15)	−0.11(0.06)	6.59(0.25)	1.21(0.13)
로그손실률	3.5	10.54(0.88)	1.15(0.16)	0.18(0.11)	6.88(0.39)	1.77(0.34)

수익률과 로그손실률의 경우 각각 5.23%와 4.21%이다. 이는 곧, 로그수익률과 로그손실률의 극단값 통계분석을 위해 일별 자료 중 각각 상위 5.23%와 4.21%만의 자료를 이용하겠다는 의미이다.

표 3.1은 일별 로그수익률 자료와 일별 로그손실률 자료에서 각각의 분계점 u를 넘어서는 초과값들에 대한 기초 통계량을 보여준다. 표에서의 최대값 11.28%와 12.80%는 각각 2008.10.30과 2001.09.12에 발생한 것이다.

표 3.2는 일별 로그수익률 자료와 일별 로그손실률 자료에서 각각의 분계점 u를 넘어서는 초과값들에 대한 포아송-GPD 모형의 최대가능도추정법에 의한 적합 결과를 보여 준다. 표에 의하면, 분계점 u를 넘어서는 초과값들의 개수는 로그수익률의 경우 연평균 13.10개 정도이고 로그손실률의 경우에는 연평균 10.54개 정도임을 알 수 있다. 또한, 꼬리지수 ξ의 값은 로그수익률의 경우 작은 값이지만 음수의 값으로 추정되고 있는데 반하여, 로그손실률의 경우에는 역시 작은 값이지만 양수의 값으로 추정되고 있음을 알 수 있다. 이는 다시 말하여, 정규분포와 비교할 때 로그수익률의 분포는 오른쪽 꼬리가 비교적 짧은 반면, 로그손실률의 분포는 반대로 오른쪽 꼬리가 다소 두터운 편에 속한다는 것을 의미한다. 즉, 이와 같은 현상은, 주식 투자로 수익을 볼 때는 일별 수익률이 비교적 미미하지만 손실을 볼 때는 일별 손실률이 생각 외로 매우 커질 수도 있음을 뜻하는 것으로 해석할 수 있을 것이다. 표 3.2에는 또한 식 (2.3)을 이용하여 계산된 GEV(μ, σ, ξ) 분포의 위치모수 μ와 척도모수 σ의 최대가능도추정치가 포함되어 있다.

표 3.3은 일별 로그수익률 자료와 일별 로그손실률 자료에서 각각의 분계점 u를 넘어서는 초과값들에 대한 포아송-GPD 모형에서의 10년, 20년 재발수준에 대한 최대가능도추정법에 의한 추정 결과를 보여준다. 표에는 또한 프로파일 로그 가능도에 기초하여 계산된 95% 신뢰구간이 포함되어 있다. 여기서, 로그손실률의 경우 q_{10}의 신뢰구간 상한과 q_{20}의 신뢰구간 상한은 모두 15를 초과한 값으로 나타나 있

표 **3.3.** 로그수익률 자료와 로그손실률 자료에서 분계점 u를 넘어서는 초과값들에 대한 포아송-GPD 모형에서의 t년 재발수준 q_t의 추정 결과(SE: 표준오차, CI: 프로파일 로그 가능도에 기초하여 계산된 신뢰구간)

자료	u	q_{10}		q_{20}	
		MLE(SE)	95% CI	MLE(SE)	95% CI
로그수익률	3	9.00(0.53)	(8.22, 10.63)	9.65(0.66)	(8.74, 11.81)
로그손실률	3.5	11.81(1.80)	(9.50, 18.19)	13.86(2.72)	(10.55, 24.31)

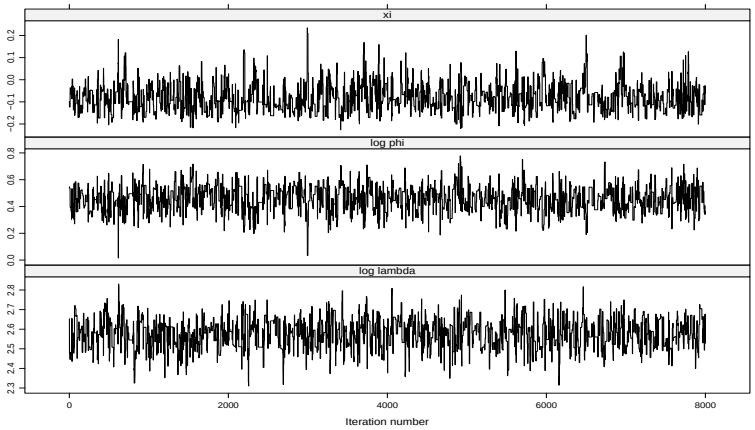

그림 3.4. 일별 로그수익률 자료에서 MCMC 방법으로 사후분포를 시뮬레이션한 결과

는데 앞으로도 계속 현재처럼 개별 주식 거래에 대해 15% 정율제의 가격제한폭제도를 유지한다면 이 두 신뢰구간의 상한은 15로 조정하는 것이 마땅하다.

이제 2장의 마지막 부분에 소개되어 있는 베이지안 방법, 즉 포아송-GPD 모형에 모수의 무정보사전분포를 가정하여 식 (2.5)로 유도된 변환 모수 $\theta = (\theta_1 = \log\lambda, \theta_2 = \log\phi, \theta_3 = \xi)$에 대한 사후밀도함수 $\pi(\theta|\text{data})$를 이용하여 모수를 추정하는 방법을 생각하기로 한다. 여기서 사용할, 모수의 사후분포에 대한 요약 방법은 식 (2.6)으로 주어지는 확률보행 Metropolis 알고리즘을 이용한 MCMC 방법이다. 이를 위한 모수의 초기값 θ^0로는 표 3.2의 MLE를 사용하도록 한다. 즉, 일별 로그수익률 자료의 경우 초기값으로 $\theta^0 = (\log 13.10, \log 1.61, -0.11)$를 선택한다. 또한, 적절한 정칙조건 하에서 모수의 사후분포는 평균이 사후최빈값(posterior mode)이고 분산/공분산 행렬이 사후최빈값에서 계산된 Fisher의 관측정보행렬(observed information matrix)의 역행렬인 다변량정규분포로 근사시킬 수 있는데, 이를 감안하여 여기서는 식 (2.6)에 사용된 f로 $f \sim N_3(\mathbf{0}, 4V)$를 선택하기로 한다. 여기서, $N_3(\cdot,\cdot)$는 3변량 정규분포를 의미하고, V는 사후최빈값에서 계산된 Fisher의 관측정보행렬의 역행렬을 나타낸다.

이와 같은 MCMC 방법을 사용하여 식 (2.5)로 주어진 θ의 사후밀도함수로부터 10,000개의 θ값을 시뮬레이션하였는데, 추적 결과 후보 θ^*의 채택률은 전체적으로 대략 일별 로그수익률 자료와 일별 로그손실률 자료의 경우 각각 18%와 19%인 것으로 나타났다. 이 중에서 초기값 θ^0의 영향력을 배재하기 위하여 처음 2,000개를 제외시키고 나머지 8,000개의 값들을 그려보았는데, 그림 3.4는 일별 로그수익률 자료에 대한 결과이다. 그림에서 보듯 시뮬레이션 값들은 모수의 사후분포 영역을 고루 터치하고 있는 것으로 보이고, 또한 특정한 패턴없이 대체적으로 확률잡음(random noise)처럼 느껴진다. 그림 3.5는 동일한 시뮬레이션 값들의 자기상관 그림을 보여주는데, 시차 1의 자기상관은 높으나 시차가 커지면서 자기상관 값이 비교적 빠르게 소멸되고 있음을 알 수 있다.

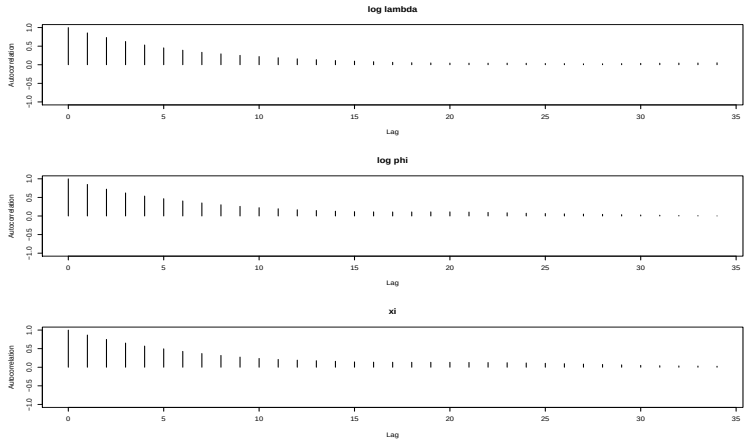

그림 3.5. 일별 로그수익률 자료에서 MCMC 방법으로 사후분포를 시뮬레이션한 값들의 자기상관 그림

표 3.4. 로그수익률 자료와 로그손실률 자료에서 MCMC 방법으로 사후분포를 시뮬레이션한 모수 값들의 요약 통계(Mean: 표본평균, SD: 표본표준편차, Naive SE: 독립표본 가정하에서 표본평균의 표준오차, Batch SE: Batch로 구분하여 추정한 표본평균의 표준오차, CI: 신용구간)

자료	모수	Mean	SD	Naive SE	Batch SE	95% CI
로그수익률	$\log\lambda$	2.58	0.07	0.0008	0.0029	(2.43, 2.72)
	$\log\phi$	0.46	0.09	0.0011	0.0036	(0.27, 0.64)
	ξ	−0.08	0.06	0.0007	0.0026	(−0.18, 0.06)
	λ	13.21	0.98	0.0109	0.0385	(11.36, 15.13)
	ϕ	1.59	0.15	0.0017	0.0057	(1.31, 1.90)
	μ	6.68	0.27	0.0030	0.0110	(6.21, 7.24)
	σ	1.29	0.16	0.0018	0.0068	(1.07, 1.71)
	q_{10}	9.35	0.70	0.0078	0.0301	(8.38, 11.22)
	q_{20}	10.12	0.92	0.0103	0.0394	(8.89, 12.64)
로그손실률	$\log\lambda$	2.35	0.08	0.0009	0.0032	(2.19, 2.52)
	$\log\phi$	0.11	0.14	0.0016	0.0049	(−0.17, 0.38)
	ξ	0.22	0.12	0.0013	0.0043	(0.01, 0.48)
	λ	10.55	0.88	0.0099	0.0336	(8.91, 12.39)
	ϕ	1.13	0.16	0.0018	0.0055	(0.84, 1.47)
	μ	6.99	0.43	0.0048	0.0154	(6.29, 7.96)
	σ	1.92	0.43	0.0048	0.0154	(1.34, 2.98)
	q_{10}	12.85	2.63	0.0294	0.0962	(9.70, 19.53)
	q_{20}	15.62	4.31	0.0482	0.1586	(10.80, 26.77)

표 3.4는 일별 로그수익률 자료와 일별 로그손실률 자료에서 MCMC 방법으로 사후분포를 시뮬레이션한 최종 8,000개 θ값들의 요약 통계를 보여주는데, 그림 3.4와 그림 3.5에서 나타난 MCMC 수행 결과 값들의 행태를 잘 설명해주고 있다. 예를 들어, 로그수익률 자료에서 모수 ξ의 사후평균 추정치는 −0.08이다. 이 시뮬레이션한 표본이 독립표본이라고 가정하면 이 추정치의 표준오차는 0.0007이 된다. 하지만 그림 3.5의 자기상관을 고려하여, 전체 8,000개의 시뮬레이션 값들을 크기가 50인 총 160개의 batch로 나누고 batch 별로 각각 표본평균을 구한 후 이들을 이용하여 사후평균 추정치의 표준오차를 추정하면 0.0026이 되는데 이 값이 보다 정밀한 값이라고 할 수 있을 것이다. 표에서 95% CI는 상하

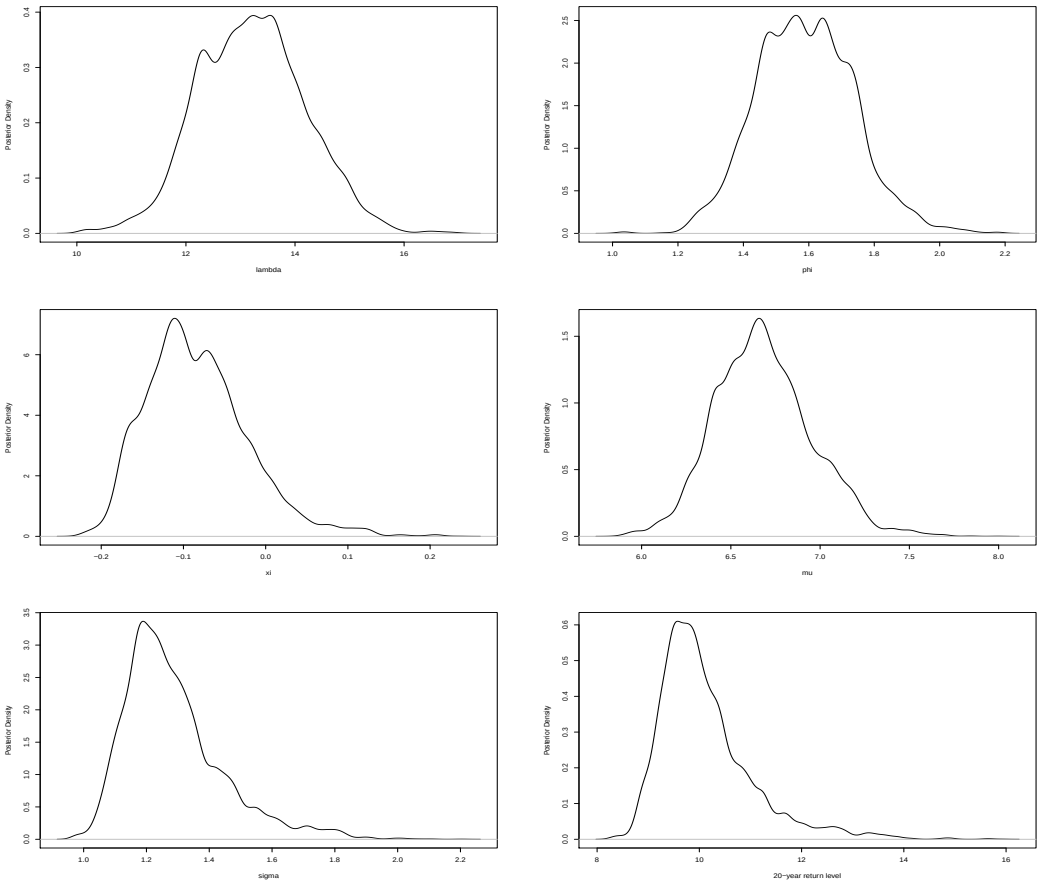

그림 3.6. 일별 로그수익률 자료에서 모수 λ, ϕ, ξ, μ, σ와 20년 재발수준 q_{20}의 추정 사후밀도함수

2.5% 표본분위수로 추정한 모수의 95% 신용구간(credible interval)을 나타낸다. 표에는 로그 변환 전으로 환원한 모수와 식 (2.3)으로 계산되는 GEV 분포 모수, 그리고 식 (2.4)로 계산되는 10년, 20년 재발수준에 대해 각각 θ에 대한 위 8,000개 시뮬레이션 값들을 적절히 변환하여 계산한 요약 통계 결과가 역시 포함되어 있다.

그림 3.6은 표 3.4의 일별 로그수익률 자료에 대한 모수 추정에 사용된 각 모수별 8,000개의 MCMC 시뮬레이션 값들을 이용하여 추정한, 모수 λ, ϕ, ξ, μ, σ와 20년 재발수준 q_{20}의 사후밀도함수를 보여준다.

그림 3.7은 일별 로그수익률 자료에서 재발수준 q_t를 재발기간(return period) t(단위: 년)의 함수로 추정한 곡선을 보여주는데(그림에서 수평축은 로그 눈금 사용), 실선은 t년 재발수준 q_t의 사후중앙값(posterior median), 긴 쇄선(long dashed line)은 사후평균, 그리고 위, 아래 두개의 쇄선은 95% 신용구간을 각각 나타낸다. 그림에 따르면, 재발수준의 사후평균이 사후중앙값보다 약간 높게 나타나고 있음을 알 수 있다.

그림 3.8은 일별 로그수익률 자료에서 2장의 마지막 부분에 소개되어 있는 베이지안 예측 방법으로 추정한 t년 재발수준 q_t의 예측 곡선을 보여준다(그림에서 수평축은 로그 눈금 사용). 여기서, 실선은 q_t의 사후예측, 긴 쇄선은 사후평균, 그리고 쇄선은 MLE를 각각 나타낸다. 그림에 따르면, 재발수준

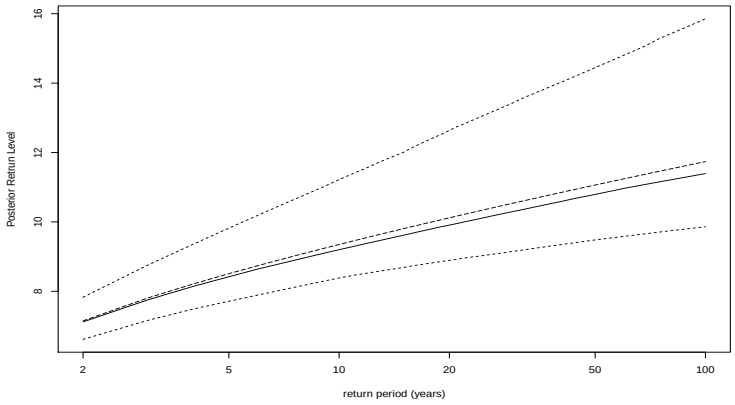

그림 3.7. 일별 로그수익률 자료에서 t년 재발수준 q_t의 추정 곡선(실선: 사후중앙값, 긴 쇄선: 사후평균, 쇄선: 95% 신용구간)

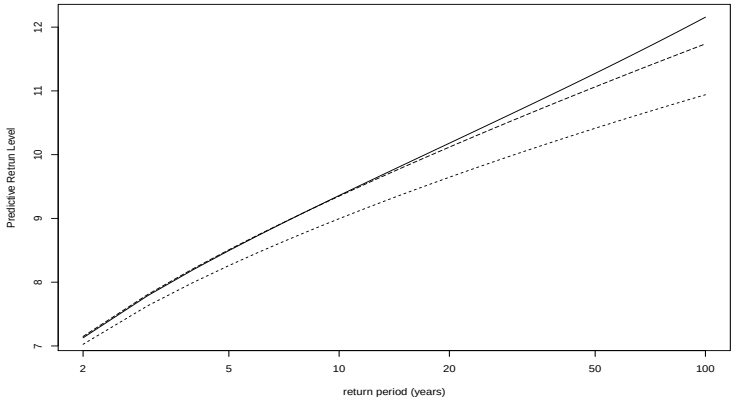

그림 3.8. 일별 로그수익률 자료에서 t년 재발수준 q_t의 예측 곡선(실선: 사후예측, 긴 쇄선: 사후평균, 쇄선: MLE)

의 MLE값이 가장 낮게 나타나고 있고, 반면에 사후분포가 가지고 있는 불확실성과 미래 관측치의 불확실성이 모두 반영된 사후예측값이 대체적으로 가장 높게 나타나고 있음을 알 수 있다.

4. 결론

본 논문에서는 코스피 지수 자료에서 일별 로그수익률과 일별 로그손실률의 극단값에 대한 통계분석을 수행하였다. 사용된 극단값 통계분석 모형은 포아송-GPD 모형으로서, 일별 로그수익률의 경우 꼬리지수 ξ의 MLE가 -0.11의 음수로 나타났고, 반면에 일별 로그손실률의 경우에는 ξ의 MLE가 0.18의 양수로 나타났다. 이는 로그수익률 분포의 오른쪽 꼬리는 정규분포보다 비교적 짧은 반면, 로그손실률 분포의 오른쪽 꼬리는 정규분포보다 다소 두텁다고 하는 상반된 결과가 얻어졌음을 의미한다. 오늘날 많은 금융자료에서 분포의 꼬리 부분이 정규분포보다 두텁다고 하는 현상이 자주 나타나고 있는 것과는 대조적으로 여기서 분석한 코스피 지수 자료의 로그수익률 분포는 오히려 정규분포보다 짧다는 특이한 결과가 얻어진 것이다. 이와 같은 특이한 결과는 현재 개별 주식 거래에 대해 가격제한폭을 15% 정율제로 운용하고 있는 것과도 무관하지 않을 것으로 판단된다. 또한 이와 같은 현상은, 일별 손실률은 때로 매

우 커질 수 있으나 일별 수익률은 대체적으로 그리 크게 발생되는 경우가 거의 드물다는 것을 의미한다.

본 논문에서는 또한 포아송-GPD 극단값 통계 모형에 모수의 무정보사전분포를 가정한 베이지안 방법을 고려하였다. 이 경우, 사후평균으로 추정한 꼬리지수 ξ의 추정치로 일별 로그수익률에서는 -0.08이, 그리고 일별 로그손실률에서는 0.22가 각각 얻어졌는데, 이는 ξ의 MLE와 비교했을 때 엇비슷한 값이므로 결국 상호 일관성있는 결과가 얻어진 것이라고 할 수 있다. 물론 모수의 사전분포를 달리 택하면 사후평균값이 영향을 받게 될 것이다. 그러나, 정칙조건이 만족되지 않는 경우에도 베이지안 방법에서는 최대가능도 방법에서와 달리 추정치의 전통적 점근성질을 걱정할 필요가 없고 예측의 경우에도 모수의 불확실성과 미래 관측치의 불확실성이 모두 자연스럽게 반영되는 장점이 존재한다.

참고문헌

윤석훈 (2009). 원/달러 환율 투자 손실률에 대한 극단분위수 추정, <한국통계학회논문집>, **16**, 803–812.

윤석훈 (2010). 국제현물원유가의 일일 상승 및 하락률의 극단값 분석, <응용통계연구>, **23**, 835–844.

Albert, J. (2009). *Bayesian Computation with R*, 2nd ed., Springer, New York.

Coles, S. G. and Powell, E. A. (1996). Bayesian methods in extreme value modelling: A review and new developments, *International Statistical Review*, **64**, 119–136.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion), *Journal of the Royal Statistical Society, Series B*, **52**, 393–442.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Proceedings of the Cambridge Philosophical Society*, **24**, 180–190.

Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire, *Annals of Mathematics*, **44**, 423–453.

Gumbel, E. J. (1958). *Statistics of Extremes*, Columbia University Press, New York.

Pickands, J. (1975). Statistical inference using extreme order statistics, *Annals of Statistics*, **3**, 119–131.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*, Springer, New York.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases, *Biometrika*, **72**, 67–90.

Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone (with discussion), *Statistical Science*, **4**, 367–393.

von Mises, R. (1936). La distribution de la plus grande de n valeurs. Reprinted in *Selected Papers II*, American Mathematical Society, Providence, R.I. (1954), 271–294.

A Bayesian Extreme Value Analysis of KOSPI Data

Seokhoon Yun[1]

[1]Department of Applied Statistics, University of Suwon

(Received September 2011; accepted October 2011)

Abstract

This paper conducts a statistical analysis of extreme values for both daily log-returns and daily negative log-returns, which are computed using a collection of KOSPI data from January 3, 1998 to August 31, 2011. The Poisson-GPD model is used as a statistical analysis model for extreme values and the maximum likelihood method is applied for the estimation of parameters and extreme quantiles. To the Poisson-GPD model is also added the Bayesian method that assumes the usual noninformative prior distribution for the parameters, where the Markov chain Monte Carlo method is applied for the estimation of parameters and extreme quantiles. According to this analysis, both the maximum likelihood method and the Bayesian method form the same conclusion that the distribution of the log-returns has a shorter right tail than the normal distribution, but that the distribution of the negative log-returns has a heavier right tail than the normal distribution. An advantage of using the Bayesian method in extreme value analysis is that there is nothing to worry about the classical asymptotic properties of the maximum likelihood estimators even when the regularity conditions are not satisfied, and that in prediction it is effective to reflect the uncertainties from both the parameters and a future observation.

Keywords: KOSPI, extreme value theory, Poisson-GPD model, Bayesian method, noninformative prior distribution, Markov chain Monte Carlo method.

[1]Associate Professor, Department of Applied Statistics, University of Suwon, Suwon 445-743, Korea.
E-mail: syun@suwon.ac.kr