碩 士 學 位 論 文

MCEM을 이용한
정규혼합분포 추정과 조건부 VaR

高麗大學校 大學院
應用統計學科
李 昇 燦

2015년 6월 1일

全 秀 榮 敎 授 指 導
碩 士 學 位 論 文

MCEM을 이용한
정규혼합분포 추정과 조건부 VaR

이 論文을 理學 碩士學位 論文으로 提出함.

2015년 6월 1일

高 麗 大 學 校 大 學 院
應 用 統 計 學 科
이 승 찬 (印)

李 昇 燦의 理學 碩士學位論文
審査를 完了함.

2015년 6월 1일

委 員 長 (印)

委 員 洪 勝 萬 (印)

委 員 任 成 秀 (印)

요 약

정규혼합분포는 정규분포의 혼합으로 이루어져있으며 위험관리 분야에서 많이 이용되고 있는 극단값분포에 대한 대안으로 활용할 수 있다.

본 연구는 KOSPI200수익률의 정규혼합분포를 MCEM(Monte Carlo Expectation Maximization) 알고리즘(Mcculloch, 1997)을 이용하여 추정하고 이에 따라 조건부 VaR를 구한다. 본 연구에서는 각 분포의 성분을 결측값으로 보고 결측값을 Metropoils-Hastings 알고리즘(Metropolis et al., 1953; Hastings, 1970)을 통해 생성하며 EM 알고리즘을 통해 각 분포의 모수를 결정 한다.

2, 3, 그리고 4 components 모형에 대한 모의실험에 MCEM 알고리즘을 적용한 결과 각 모형을 잘 찾아주고 있음을 알 수 있었다. 본 연구의 실증 자료 표본은 2008년 외환위기 1년 후인 2009년 10월 15일부터 2014년 10월 14일 동안의 일별 자료이다. MCEM 알고리즘을 통해 분포의 추정을 실시한 결과 대다수의 분포는 2개의 성분을 갖고 있는 정규혼합분포로 나타났다. 실 자료 중 1000개의 표본기간을 1일씩 연속적으로 이동하여 총 238개의 정규혼합분포를 추정하였고 조건부 VaR의 추정을 실시한 결과 1%, 5%, 95%, 그리고 99% 모두에서 유의함을 보여 조건부 VaR를 잘 추정하고 있음을 알 수 있었다. 따라서 본 연구 결과 극단값 분포를 활용하여 구하는 것보다는 데이터가 따르는 정확한 분포를 통하여 위험 측정값을 구하는 것이 다중최빈값을 갖는 경우에서 더 활용성이 높으며 충분히 리스크 측정도구로서 사용될 수 있음을 보였다.

목 차

그 림 목 차

표 목 차

제 1장. 서론

1.1 연구배경 및 목적

1990년대 이후 한국금융시장의 위기는 1997년 외환위기와 2008년의 금융위기라 할 수 있다. 이러한 금융위기를 극복하기 위해서 시장분석을 통한 위험관리가 최우선 사항이라고 할 수 있다. 경제에는 다양한 형태의 위험이 존재하고 있다. 본 연구에서 다루는 위험은 시장위험(market risk)이라 불리는데 가장 기본적이며 가장 광범위한 것이다. 위와 같은 시장위험을 분석하는 것은 위험관리의 그 첫 번째라고 할 수 있다. 그에 따라 본 연구는 우리나라의 대표적인 주가지수인 주가지수 200(KOSPI 200)을 대상으로 시장위험 측정에 관하여 논의 한다.

위험의 측정값으로는 잘 알려진 VaR(Value at Risk)와 ES(Expected Shortfall) 등이 있다. VaR는 '정상적인시장(normal market) 여건 하에서 주어진 신뢰수준(confidence level)으로, 목표기간(target period) 동안에 발생할 수 있는 최대 손실금액(maximum loss)'이라고 정의한다. 예를 들어 내가 어떤 투자자산을 갖고 있는데 보유기간 1주일, 신뢰수준95%의 VaR가 10억원이라 하자. 이는 '나의 투자자산 포지션의 가치에 영향을 미치는 어떤 리스크 요인의 변화로 인해 1주일 동안에 발생할 수 있는 손실이 10억원보다 작다는 사실을 95% 신뢰수준에서 확신할 수 있다'는 의미이다. 다시 말하면 '1주일 동안에 10억원보다 큰 손실이 발생할 확률이 5%'라는 의미이다.

VaR의 정의를 구체적으로 살펴보면, 첫째, 정상적인 시장이다. 이는 누구나가 인정하는 보편적인 시장을 의미한다. 두 번째가 신뢰수준이다. 신뢰수준이란 관찰값의 분포에서 특정 관찰 값이 포함될 확률을 의미한다. 세 번째가 목표기간이다. 이는 해당자산 또는 포트폴리오의 보유기간으로 통상 금융기관은 영업일기준으로 계상한다. 일주일은 영업일기준으로는 5일이며, 1년은 250일이다. 목표기간은 보유포지션의 헷지(hedge)기간이나 계약해지 기간 등 원래포지션의 유동성과 연관하여 설정하는 것이 좋다 (서영수, 2012).

VaR의 추정을 위해서 다양한 분포가 이용되어 왔고, 그 중 가장 간단한 방법은 정규분포를 이용하는 것이지만 주식수익률의 분포는 정규분포를 따르지 않는 다는 사실은 매우 잘 알려져 있다. 주요 근거로는 왜도와 첨도가 정규분포의 경우와 다르기 때문이며, 그에 대한 대안으로 가장 대표적인 것이 분포혼합(mixture of

distribution)과 극단값 분포(extreme value distribution)이다.

본 연구에서는 KOSPI 200수익률의 분포를 구하기 위해서 정규혼합분포(normal mixture)를 이용한다. 정규혼합분포는 분포혼합의 하나로 이중 가장 많이 사용되는 것이다. 정규혼합분포를 따르는 확률변수는 몇 개의 성분으로 분류될 수 있는데 각 성분은 각각의 정규분포를 따른다. 즉, 정규혼합분포란 것은 정규분포의 혼합이다. 게다가 개별 성분이 정규분포를 따른다고 하더라도 정규혼합분포는 일반적으로 정규분포와는 다르다. 그러나 이 성분들의 조합에 대한 분포는 우리가 알 수 있는 것이 아니다. 따라서 이 정규혼합분포를 만들어주기 위해 본 연구에서는 MCEM(Monte Carlo Expectation Maximization) 알고리즘(Mcculloch, 1997)의 방법을 사용한다.

정규혼합분포를 적합 시킬 때는 일반적으로는 EM(Expectation Maximization) 알고리즘(Dempster et al., 1977)을 사용하고 있다. 이 알고리즘의 장점은 기존의 근사접근이나, 수치해석적인 방법에 비해 상대적으로 프로그래밍하기 쉽고, 우도함수 또는 로그우도함수를 최대화시키는 MLE로의 단조수렴 추정치를 생성해 낸다는 것이다. 기존의 연구는 보편적인 EM알고리즘을 이용한 경우가 많았다. 그러나 EM알고리즘은 몇 가지의 한계를 가지고 있으며 이 한계를 극복하기 위한 여러 가지 방법 중에 MCEM의 방법으로 혼합분포를 추정하여 보도록 하였다.

정규혼합분포의 유용성은 위험 관리에서 얼마나 적합한지가 중요한 문제가 된다. 이를 위해서 조건부 VaR(conditional Value at Risk)에 대한 사후검정을 수행한다. 사후검정으로는 본 연구에서는 McNeil and Frey(2000)를 따르기로 한다.

본 논문의 구성으로 2장에서는 자료의 소개와 KOSPI200의 기술통계량을 제시하고 그에 따른 알고리즘의 특징들을 소개한다. 3장에서는 극단값분포와 정규혼합분포에 대하여 알아보고 4장에서는 EM알고리즘과 MCEM의 방식을 소개하고 5장에서는 모의실험, 6장에서는 실증분석의 결과와 사후분석을 살펴본다. 그리고 7장에서는 논문의 결론을 정리한다.

제 2장. 자료의 소개

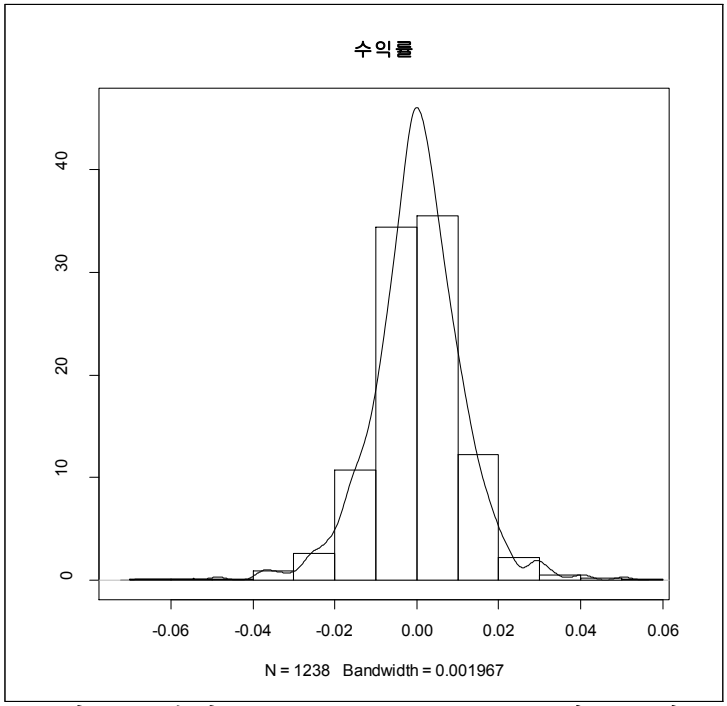
2.1 KOSPI 200

본 연구는 한국거래소(Korea Exchange)의 주가지수 200을 이용한다. 표본은 2009년 10월 15일부터 2014년 10월 14일 동안의 일별 자료이다. 본 자료는 2008년 금융위기 1년 후의 자료부터 현재까지의 자료이다. 또한 VaR 측정 시 고려해야 할 중요 요소인 보유기간은 1일 종가 지수를 기준으로 분석하였다. 사실 보유 기간은 필요에 따라 달라질 수 있는데 하루 단위로 선정하는 이유는 변동성이 관찰되면서 BIS(The Bank for International Settlements)에서 요구하는 것이 2주 VaR임에도 불구하고 대부분의 금융회사들은 내부위험제어의 목적으로 하루 동안 손실을 막을 수 있는 VaR를 적용하기 때문이다. 실제 VaR를 추정할 때 분석의 대상인 주가지수의 일별 수익률은 연속 복리 수익률(continuously compounded returns)로 이는 다음의 식 (2.1)과 같다.

$$r_t = \ln\left(\frac{p_t}{p_{t-1}}\right) \tag{2.1}$$

(r_t : 수익률, p_t : $KOSPI200$의 t시점의 종가지수)

실무적으로 가격보다는 수익률이 보다 더 통계적으로 의미를 갖게 되고, 주어진 가격에 대한 상대적인 변화를 측정하기 위하여 절대 수익률 보다는 상대 수익률과 연속 복리 수익률이 선호된다. 따라서 본 논문에서는 연속 복리 수익률을 사용하여 1일 초과 시에 연속 복리 수익률의 경우 단순히 일일 연속 복리 수익률을 합하면 되기 때문에 편리하게 사용할 수 있다. (이희찬, 2012)



<그림 2.1> 수익률(2009.10.15.~2014.10.14.)자료 그래프

<그림 2.1>은 2009년 10월 15일부터 2014년 10월 14일까지의 KOSPI 200 수익률의 히스토그램이다. <그림 2.1>의 히스토그램을 살펴보면 수익률 분포의 평균이 거의 0에 가깝고, 좌우 비대칭적인 모양을 나타내고 있고, 표시된 곡선은 수익률 자료의 정규분포의 모형을 비교분석하기 위하여 나타낸 그림이다. 위와 같이 수익률의 분포를 단순 일봉(uni-modal)의 정규분포 모형이라고 할 경우 꼬리가 두터운 분포의 형태를 나타내고 있어 극단적인 수익률을 설명하기가 힘들어 진다. 따라서 수익률의 분석에서 적합한 분포를 찾는 일은 대단히 중요한 과제이다.

2.2 기술 통계량

<표 2.2> 실제 자료 기술 통계량

평균	분산	왜도	첨도	Jarque - Bera Test	최대	최소	중간값
9.45E-05	0.000131	-0.35437	6.385599	620.8496 (0.0)	0.050572	-0.06649	0.000206

주) 괄호 안은 p-value

수익률의 기술 통계량을 봤을 때 평균은 0.0000945이며 분산은 0.000131로 나타나고 있으며 특히 중요한 왜도는 좌우 비대칭이 이루어지지 않고 있음을 보여주고 있고, 첨도는 3이상의 초과첨도를 나타내고 있다. 또한 Jarque-Bera Test(Jarque, C. M., Bera, A. K., 1981)를 통하여 수익률의 분포가 정규분포를 따르지 않는다는 정확한 증거로 나타나고 있다. <그림2.1>에서 나타난 두꺼운 꼬리로 인하여 혼합 분포 모형이 필요한 증거가 됨을 보여줄 수 있다.

제 3장. 극단값분포와 정규혼합분포

3.1 극단값분포

수익률이 정규분포를 따른다고 볼 수 없기 때문에 이에 적합한 분포를 찾는 일은 위험관리 분야에서 중요한 과제가 되어 왔다. 여기에는 여러 가지 방법이 도입되었는데 그 중 가장 대표적인 방법은 극단값분포를 이용하는 것이다. 극단값분포는 이미 여러 학문분야에서 이용되었고 최근엔 금융 분야에서도 교과서에 포함될 정도로 보편화 되었다(McNeil, Frey and Embrechts, 2005).

극단값 분포란 극단값(extreme value)의 분포를 말한다. 예를 들어 수익률이 있을 때 이 중 극단적인 값, 즉 최대값(maximum) 또는 최소값(minimum)의 분포를 극단값분포라 한다. 비교하자면 수익률의 '평균'이 하나의 분포를 갖는 확률변수인 것처럼 수익률의 '극단값'도 하나의 분포를 갖는 확률변수이며(물론 평균이 수익률의 함수인 것처럼 극단값도 수익률의 함수이다.), 이를 극단값분포라고 한다. 따라서 극단값분포를 이용하는 것은 수익률분포 자체를 다루는 것이 아니라 수익률의 함수, 즉 극단값의 분포를 다루는 것이다.

극단값분포가 널리 활용되게 된 것은 수익률의 경우 첨도가 중요하기 때문이다. 즉, 극단적인 수익률이 빈번하게 발생하며 이로 인해 수익률 분포의 꼬리가 두툼하기 때문이다. 이를 분석하기에 극단값분포는 매우 유용하지만 분명한 것은 극단값분포가 수익률의 분포전체를 다루지 않는다는 것이다. 따라서 극단값분포를 이용할 경우 왜도를 명시적으로 고려하지 않을 뿐만 아니라 꼬리가 아닌 부분의 분포도 명시적으로 고려하지 않는다. 단지 극단값분포는 꼬리에 대한 분석에 적합한 것이다.

따라서 극단값분포에 대한 대안은 수익률의 분포 자체를 분석하는 것이다. 현재 가장 많이 이용되는 것은 분포혼합이며 이 중 대표적인 것이 정규혼합분포이다. 정규혼합분포는 정규분포의 혼합으로 이루어진 하나의 분포다. 따라서 정규혼합분포를 이용하여 수익률의 분포를 추정하면 이는 그 자체로 수익률의 분포를 다루는 것이다.

3.2 정규혼합분포

혼합분포(mixture distribution)란 여러 분포의 혼합으로 이루어진 분포를 말한다. 따라서 혼합분포를 따르는 확률변수는 몇 개의 성분(component)으로 분류될 수 있고 각 성분은 각자의 분포를 따른다. 정규혼합분포(normal mixture distribution)는 혼합분포의 한 종류일 뿐이며, 각 성분은 정규분포를 따른다. (윤종인, 2011)

$Y=(Y_1, Y_2, \cdots, Y_n)$를 수익률이라고 하고 Y가 K개 정규분포혼합을 따른다고 하자. 자료 Y는 K개의 성분으로 분류할 수 있고 각 성분은 $\theta_1, \theta_2, \ldots, \theta_K$를 모수(parameter)로 갖는 K개 정규분포를 각각 따른다고 가정한다. 여기서 $\theta_i = (\mu_i, \sigma_i^2), i = 1, \cdots, K$를 따른다.

개별 관측치 y_i의 밀도는 다음과 같이 표현할 수 있다.

$$f(y_i|\pi,\theta) = \sum_{k=1}^{K} \pi_k f(y_i|\theta_k), \vec{\pi} = (\pi_1, \cdots, \pi_k), \vec{\theta} = (\theta_1, \cdots, \theta_k) \qquad (3.1)$$

여기에서 π_k는 Y_i가 $N(\mu_k, \sigma_k^2)$를 따를 확률을 의미한다. 따라서 Y의 확률밀도함수는 다음과 같다.

$$f(y|\pi,\theta) = \prod_{i=1}^{n} f(y_i|\pi,\theta) = \prod_{i=1}^{n} [\sum_{k=1}^{K} \pi_k f(y_i|\theta_k)] \qquad (3.2)$$

이때 $i = 1, 2, \cdots, n$에 대해서 확률변수 $Z_i = (Z_{i1}, Z_{i2}, \cdots, Z_{ik})$를 정의한다. Z_i는 K개의 원소로 이루어져 있으며 Z_i의 k번째 원소 Z_{ik}는 y_i가 $N(\mu_k, \sigma_k^2)$를 따를 경우 1의 값을 가지며 그렇지 않을 경우 0의 값을 갖는다. 따라서 Z_i는 다항분포(multinomial distribution)를 따르고 다항분포의 모수를 $\vec{\pi} = (\pi_1, \pi_2, \cdots, \pi_K)$라고 할 때 $Z_i \sim multinomial(\vec{\pi})$로 나타낼 수 있다. 따라서 식(3.2)에서 언급한 π_k는 확률변수 Z_i를 이용하면 $P(Z_{ik}=1 \mid \theta)$ 에 해당되는 개념이다.

Z를 분류(classification)라고 하는데 Y와 Z를 모두 이용하면 식 (3.2)의 함수는 다음과 같은 확률밀도함수로 바뀌게 된다.

$$f(y,Z|\pi,\theta)=\prod_{i=1}^{n}f(y_i|Z_i,\theta)p(Z_i|\theta)=\prod_{i=1}^{n}\prod_{k=1}^{K}[\pi_k f(y_i|\theta_k)]^{z_{ik}} \quad (3.3)$$

여기에서 z_{ik}는 확률변수 Z_{ik}의 실현 값이다.

혼합분포의 추정에서 중요한 특징은 분류 Z를 구하는 절차와 모수를 추정하는 절차가 동시에 필요하다는 데 있다. 물론 이 작업이 그렇게 쉽지는 않다. 왜냐하면 식 (3.3)에서 분류 Z는 관측할 수 있는 변수가 아니기 때문이다. 그렇기 때문에 대부분의 경우 EM 알고리즘을 사용하여 연구를 진행하고 있으나, 본 연구에서는 이 Z의 분포를 가정하지 않고 바로 찾아 분류할 수 있는 MCEM(Monte Carlo Expectation Maximization) 알고리즘을 이용하여 이 문제를 해결하기로 한다.(Mcculloch, 1997)

제 4장. 정규혼합분포 추정을 위한 MCEM 알고리즘

4.1 MCEM 알고리즘

먼저 Dempster et al.(1977)에 의해 제안된 Expectation Maximization (EM) 알고리즘을 알아본다. EM 알고리즘은 다양한 불완전 자료(incomplete data)로부터 최대 우도추정치(MLE)를 반복적인 기법을 통해 구할 수 있는 방법으로 어떤 정보가 알려져 있지 않은 경우 가장 그럴듯하게 모델을 추정할 때 사용하는 효과적인 방식이다. 이 반복적 알고리즘의 가장 큰 장점은 기존의 근사접근이나, 수치해석적인 방법에 비해 상대적으로 프로그래밍 하기가 쉽고, 우도함수 또는 로그우도함수를 최대화시키는 MLE로의 단조수렴 추정치를 생성해 낸다는 것이다. 알고리즘의 각 반복은 Expectation 단계(E-step)와 Maximization 단계 (M-step)로 구성되어 있기 때문에 이것을 EM알고리즘이라고 부르며, 관련이론의 단순성과 일반성을 가지고 있다. 또한 다양한 분야에 대해 적용이 가능하기 때문에 많은 주목을 받아왔다. 특히 MLE가 쉽게 계산되어지는 지수 족에서의 완전자료인 경우, EM 알고리즘의 M-step의 계산은 마찬가지로 쉽게 계산되어진다. 이제부터 E-step과 M-step이 어떻게 유도되어 EM알고리즘을 구성하는지 알아보도록 하겠다.

먼저 EM 알고리즘을 설명하기 위해서 다음을 가정한다. n개의 표본을 고려할 때, n_1는 관측이 되었고, $n_2 = n - n_1$개는 관측이 되지 않았다고 하였을 때, 관측된 표본을 $Y = (Y_1, Y_2, \cdots, Y_{n_1})$, 그렇지 않은 표본을 $Z = (Z_1, Z_2, \cdots, Z_{n_2})$라고 가정한다. 각각의 y_i는 서로 독립인 확률밀도함수 $g(y|\theta), \theta \in \Theta$를 가지고, Y와 Z는 서로 독립을 가정한다. 그리고 관측된 표본과 관측되지 않은 표본들의 결합 확률 밀도함수를 $f(y,z|\theta)$라고 정의한다. 그리고 관측된 표본이 주어진 상황에서 모든 전체표본들$(x = (y,z))$의 조건부 확률밀도함수를 $t(x|y,\theta) = \frac{f(y,z|\theta)}{g(y|\theta)}$ 로 정의할 수 있다.

관측된 표본들의 우도함수는 $L(\theta|y) = g(y|\theta)$와 동일하다. 그리고 모든 데이터의 우도함수는 $L(\theta|x) = f(y,z|\theta)$ 와 같이 쓸 수 있다.

EM 알고리즘은 $L(\theta|x)$를 이용하여 관측된 표본의 우도함수 $L(\theta|y)$를 최대화 하는 것이다. $t(x|y,\theta) = \frac{f(y,z|\theta)}{g(y|\theta)}$를 이용하여 약간의 계산을 하면 최종적으로 다음과 같은 식이 성립한다.

$$\log L(\theta|y) = E_{\theta_0}[\log L(\theta|x)|\theta_0, y] - E_{\theta_0}[\log t(x|y,\theta)|\theta_0, y] \qquad (4.1)$$

식 (4.1)의 오른쪽 첫 번째 식을 $Q(\theta|\theta_0, y) = E_{\theta_0}[\log L(\theta|x)|\theta_0, y]$로 정의하고, 이 Q 함수를 EM 알고리즘의 E-step라 한다. 그리고 $\arg max Q(\theta|\theta_0, y)$를 EM 알고리즘의 M-step이라고 정의한다. 이 때 Q 함수를 최대화시키는 θ_0가 우리가 원하는 해가 된다. $E_{\theta_0}[\log t(x|y,\theta)|\theta_0, y]$는 고려하지 않아도 된다. (McLachlan, G., Pell, D., 2000)

EM 알고리즘을 정규혼합분포 모형에 적용하기 위해서 전체 component가 J인 정규혼합분포 모형에서 $f_j(y_i|\mu_i, \sigma_j)$를 j번째 component인 정규분포라고 한다면, 추정하고자하는 전체모수들은 $\theta = (\pi_1, \cdots, \pi_J, \mu_1, \cdots, \mu_J, \sigma_1, \cdots, \sigma_J)$ 가 된다.

본 연구에 EM을 적용하기 위하여 결측값라는 개념을 생각하고자 한다. 즉, 각각의 관측치 y_i는 J개의 component들 중 하나에 속해 있는데, 그 관측치가 어디에 속해 있는지 모르기 때문에 이를 결측값으로 간주하고, 이때 이 결측값(missing)를 $z = (z_1, \cdots, z_J)$라는 확률변수로 생각할 수 있다.

E-step을 적용하기 위해서 먼저 관측된 표본 y_i가 주어졌을 때, z의 조건부 분포를 고려하면

$$E_{\theta_0}(z_i|\theta', y_i) = P(z_i = j|\theta', y_i) = \frac{\pi_j f_j(y_i|\mu_j, \sigma_j)}{\sum_{s=1}^{J} \pi_s f_s(y_i|\mu_s, \sigma_s)} \qquad (4.2)$$

를 유도할 수 있고, Q 함수는 다음과 같이 정의 할 수 있다.

$Q(\theta|\theta') = E(L(\theta|x)|y, \theta')$가 되고, $t_{ij}(\theta') = P(z_i = j|\theta', y_i)$로 정의 되며

$$Q(\theta|\theta') = \sum_{i=1}^{n}\sum_{s=1}^{J} t_{is}(\theta') \log \pi_s f_s(y_i|\mu_s, \sigma_s)) \qquad (4.3)$$

를 계산할 수 있다. 그리고 이 Q 함수를 최대화(M-step)하는 θ를 $\theta^{(k+1)}$이라하고 이를 반복하여 계산한다. 위의 내용을 바탕으로 $k+1$번째 해는 아래와 같이 나타날 수 있다.

EM 알고리즘 : 정규혼합분포 모형

1. E-step : $i=1,\cdots,n,\ j=1,\cdots,J$ 라고 할 때,

$$t_{ij}(\theta^{(k)})=\frac{\pi_j^{(k)} f_j(y_i|\mu_j^{(k)},\sigma_j^{(k)})}{\sum_{s=1}^{J}\pi_s^{(k)} f_s(y_i|\mu_s^{(k)},\sigma_s^{(k)})} \text{ 를 계산}$$

2. M-step :

$\theta^{(k+1)}=(\pi_1^{(k+1)},\cdots,\pi_J^{(k+1)},\mu_1^{(k+1)},\cdots,\mu_J^{(k+1)},\sigma_1^{(k+1)},\cdots,\sigma_J^{(k+1)})$를 설정하고,

$$\pi_j^{(k+1)}=\frac{1}{n}\sum_{i=1}^{n} t_{ij}(\theta^{(k)}),$$

$$\mu_j^{(k+1)}=\frac{\sum_{i=1}^{n} t_{ij}(\theta^{(k)})y_i}{\sum_{i=1}^{n} t_{ij}(\theta^{(k)})},$$

$$\sigma_j^{2^{(k+1)}}=\frac{\sum_{i=1}^{n} t_{ij}(\theta^{(k)})(y_i-\mu_j^{(k+1)})^2}{\sum_{i=1}^{n} t_{ij}(\theta^{(k)})} \text{를 계산.}$$

수렴할 때까지 E-step과 M-step을 반복하면 $L(\theta|y)$를 국소적으로 최대화 하는 θ를 얻는다.

EM 알고리즘은 E-step의 계산이 비교적 간단하다면, 우도함수가 단조증가 하는 특징을 가지고 있어 안정성을 보장한다. 그렇지만 결측값 값이 조건부로 들어간 우도함수의 기댓값을 구하는 E-step은 고차의 적분을 수반하여 쉽게 계산되지 않는 문제점도 갖고 있으며 기본적으로 결측값 값의 분포를 실제로는 알지 못하는 경우가 다분히 발생하기 때문에 정확한 결론에 도달하지 못하는 문제점을 안고 있다.

다음으로 MCEM 알고리즘을 알아본다. MCEM(Monte Carlo Expectation Maximization) 알고리즘은 EM 알고리즘의 단점인 E-step이 고차의 적분을 수반하게 되면 계산되지 않는 경우가 발생한다는 것을 보완하기 위해 제안된 알고리즘이다(Wei et al., 1990). MCEM 알고리즘은 EM 알고리즘의 E-step의 기댓값 계산에 필요한 적분과정을 몬테카를로(Monte Carlo)방법으로 해결하여 결측 자료를 구

하고 이로부터 모수를 추정하는 방법이다.

몬테카를로 방법이란 통계적 문제를 난수(Random number)를 사용한 무작위적인 표본을 이용하여 해결하는 방법이다. 즉, 변수의 관계가 확실하여 예측치를 정확하게 찾을 수 있는 확정모형과는 달리, 대부분의 모형들은 많은 부분이 결과를 정확하게 예측할 수 없는 확률모형이다. 일반적으로 확정모형에서는 분석적 해를 찾는 것이 가능하다. 그러나 확률모형에서는 분석적인 방법으로 해를 찾는 것이 불가능한 경우가 많다. 이 경우에는 수치적으로 일련의 난수를 반복적으로 발생해서 모의실험을 하면 해를 찾을 수 있는데 이것이 몬테카를로 방법이다. 몬테카를로 방법의 장점 중의 하나는 계산이 다른 수학적 방법에 비해 간단하다는 것을 들 수 있다. (Christian and George, 2004; Charles, 1997)

본 연구를 위해 Mcculloch(1997)가 제안한 방법인 Metropolis-Hastings 알고리즘을(Metropolis et al., 1953; Hastings, 1970) 이용한다. 결측값 값으로 나타나는 $z^* = (z_1, z_2, \cdots, z_{k-1}, z_k^*, z_{k+1}, \cdots, z_q)$ 의 값을 생성해 낸다. 그 첫 단계로 채택 확률(acceptance rate), α값을 구하여 주어야 한다.

$$\alpha = \min\left\{1, \frac{f_{z|y}(z^*|y,\theta,\pi)h_z(z)}{f_{z|y}(z|y,\theta,\pi)h_z(z^*)}\right\} \tag{4.4}$$

여기서 $h_z = f_z$ 라고 할 때

$$\frac{f_{z|y}(z^*|y,\theta,\pi)h_z(z)}{f_{z|y}(z|y,\theta,\pi)h_z(z^*)} = \frac{\prod_{i=1}^{n} f_{y_i|z}(y_i|z^*,\theta,\pi)f_z(z^*)f_z(z)}{\prod_{i=1}^{n} f_{y_i|z}(y_i|z,\theta,\pi)f_z(z)f_z(z^*)}$$

$$= \frac{\prod_{i=1}^{n} f_{y_i|z}(y_i|z^*,\theta,\pi)}{\prod_{i=1}^{n} f_{y_i|z}(y_i|z,\theta,\pi)} \tag{4.5}$$

위 Metropolis 알고리즘을 EM에 적용한 MCEM 알고리즘은 다음과 같다.

MCEM 알고리즘

1. 모수의 초기 값을 생성한다.

2. N개의 $z^{(1)}, z^{(2)}, \cdots, z^{(N)}$ 값을 $f_{z|y}(z|y, \theta^{(m)}, \pi^{(m)})$ 식에서 Metropolis-Hastings 알고리즘을 통하여 생성한다.

3. 생성된 결측값 값을 이용하여 $E[\ln f_{y|z}(y|z, \theta, \pi)|y]$ 의 값이 최대가 되는 모수 $\theta^{(m+1)}, \pi^{(m+1)}$ 값을 선택한다. 즉, $\theta^{(m+1)}, \pi^{(m+1)}$ 은 $\frac{1}{N}\sum_{k=1}^{N} \ln f_{y|z}(y|z^{(k)}, \theta, \pi)$ 의 값을 최대화 시킨다.

4. 만약 성공적으로 모수 값들이 수렴하게 된다면, $\theta^{(m+1)}, \pi^{(m+1)}$은 MLE $\hat{\theta}, \hat{\pi}$으로 수렴하게 된다. 즉, 수렴을 할 때 까지 위의 방식을 계속 시도하게 된다.

결국 처음 결측값 값의 분포를 가정하지 않은 상태에서 오직 난수 생성을 통하여 결측값 값을 생성시켜주는 장점이 있다고 할 수 있다. 즉, EM 알고리즘에서 E-step전에 MCMC기법을 사용하여 결측값 값들을 생성시키는 것이다.

4.2 정규혼합분포추정을 위한 MCEM 알고리즘

본 논문은 정확한 분포추정을 통하여 VaR을 구하여 리스크를 측정하는 것이 목적이다. 그래서 정확한 분포 추정엔 성분의 수가 매우 중요한 포인트가 된다. 따라서 먼저 성분수 s가 주어졌다고 가정하자.

정확한 수익률의 VaR을 구하기 위해서 정확한 분포 추정이 필요하다. 따라서 본 논문에서는 수익률 자료를 분석하기 위해, MCEM 알고리즘을 적용시켜 본다.

1. 결측값 데이터 z는 $z \sim unif(p_i)$ 로 생성한다.

   여기서 MH 알고리즘을 이용하여 π_i를 생성한 뒤 $p_i = \frac{\pi_k f(Y|\theta_k)}{\sum_{l=1}^{k} \pi_l f(Y|\theta_l)}$ 로 각각의 자료의 확률 값을 이용하여 분포를 생성한다.

2. 생성된 z를 이용하여 $E[\ln f_{y|z}(y|z,\theta_s,\pi_s)|y]$의 값이 최대화 되는 모수를 선택한다. 즉, $\frac{1}{N}\sum_{k=1}^{N} \ln f_{y|z}(y|z^{(k)},\theta_s,\pi_s)$ 이 최대화 되는 모수의 값을 선택한다.

   현재 반복m에서 $\hat{\pi}_s^{(m)}, \hat{\mu}_s^{(m)}, \hat{\sigma}^{2(m)}_s$ 의 값이 주어졌다고 하자.
   반복 $m+1$에서는 다음과 같다.
   $l = 1,\cdots,s$ 에 대하여,

$$\hat{\pi}_l^{(m+1)} = \frac{\sum_{i=1}^{n} \hat{p}(z_{is}=1)}{\sum_{i=1}^{n}\sum_{j=1}^{k} \hat{p}(z_{ij}=1)} = \frac{n_s \hat{\pi}_s}{n_1\hat{\pi}_1 + \cdots + n_k\hat{\pi}_k} \tag{4.6}$$

$$\hat{\mu}_l^{(m+1)} = \frac{\sum_{i=1}^{n} y_i^{z_{is}} \hat{p}(z_{is}=1)}{\sum_{i=1}^{n} \hat{p}(z_{is}=1)} = \frac{\hat{\pi}_s \sum_{i=1}^{n} y_i I(z_{is})}{n_s \hat{\pi}_s} \tag{4.7}$$

$$\widehat{\sigma^2}_l^{(m+1)} = \frac{\sum_{i=1}^{n}(y_i - \widehat{\mu_s})^2 \hat{p}(z_{is}=1)}{\sum_{i=1}^{n}\hat{p}(z_{is}=1)} = \frac{\pi_s \sum_{i=1}^{n}(y_i - \widehat{\mu_s})^2 I(z_{is})}{n_s \widehat{\pi_s}} \quad (4.8)$$

3. $MSE(= \sqrt{\sum_{l=1}^{s}(\hat{\mu}_l^{(m+1)} - \hat{\mu}_l^{(m)})^2 + (\widehat{\sigma^2}_l^{(m+1)} - \widehat{\sigma^2}_l^{(m)}) + (\hat{\pi}_l^{(m+1)} - \hat{\pi}_l^{(m)})}$

를 계산하며 수렴할 때 까지 1 과 2를 반복한다. 수렴하는 그 값은 MLE로 받아들인다.

4.3 성분수의 선택과 수렴의 문제

본 논문에서는 아카이케정보기준(Akaike information criterion : AIC ; Akaike , 1974)과 베이지안 정보기준(Bayesian Information Criterion :BIC ; Schwarz, 1978)을 이용하여 성분의 수 K를 검정한다(McLachlan and Pell, 2000, pp.175~220). 특정의 성분의 수에서 얻은 모수의 추정치가 $\widehat{\pi^0}$, $\widehat{\theta^0}$ 일 때 AIC와 BIC는 다음과 같다.

$$\begin{aligned} AIC &= -2l(y,z|\widehat{\pi^0},\widehat{\theta^0})+d \\ BIC &= -2l(y,z|\widehat{\pi^0},\widehat{\theta^0})+d\log n \end{aligned} \tag{4.10}$$

여기에서 n 과 d는 각각 표본의 수와 모수의 수이다.

아카이케정보기준(AIC)은 주어진 복잡한 데이터의 모델의 균형을 이루는 적합도 모형을 선택하는 방식이다. 우도함수는 최우도함수가 될 때 확률분포를 가장 잘 추정한다고 알려져 있으며 확률분포를 표현하는 모수의 개수가 많을수록 과추정하는 경향이 있기 때문에 같은 수준의 우도함수에서는 더 적은 개수의 모수를 가진 분포가 더 작은 AIC를 갖는다. 식 (4.10)에 따라 AIC가 가장 작은 분포가 데이터를 가장 잘 표현한 분포이다.

AIC의 방법은 표본의 크기가 큰 경우에도 과도 추정하는 경향을 가지고 있어서, 이런 단점을 고려하여 Schwarz(1978)는 베이지안 논리를 활용하여 분포형이 지수족이고 독립적으로 동일하게 분포된 관측치에 대한 모델의 차원을 구하는 정보판단기준 BIC를 제시하였다. BIC의 값을 계산하기 위해서는 AIC 방법과 같이 최우도함수가 구해져야 한다. 또한 BIC의 값도 식 (4.10)에 따라서 BIC가 가장 작은 분포가 데이터를 가장 잘 표현한 분포이다.

Akaike는 모의실험을 통하여 BIC는 자기회귀의 차수를 과대추정하지 않는 것으로 주장하였다. 그리고 이 최소 BIC 추정량은 일치성을 가짐이 증명되었다.

제 5장. 모의 실험

5장에서는 MCEM 알고리즘을 이용하여 임의 생성된 자료가 얼마나 잘 분류가 되는지 알아보고자 한다. 모의실험을 위한 통계 패키지로는 R-software 3.1.2버전을 사용하였다.

5.1 자료 생성

본 논문의 모의실험을 위해 아래와 같이 자료를 생성한다.

1. c1~N(-0.5,1), c2~N(0.5,1) 를 따르는 자료 각각 400개, 600개를 임의로 생성한다.

2. c1~N(-2,4^2), c2~N(0,1), c3~N(2,4^2)을 따르는 자료 각각 150개, 700개, 150개를 임의로 생성한다.

3. c1~N(-5,0.5^2), c2~N(-2,0.5^2), c3~N(2,0.5^2), c4~N(5,0.5^2)을 따르는 자료 각각 250개씩 임의로 생성한다.

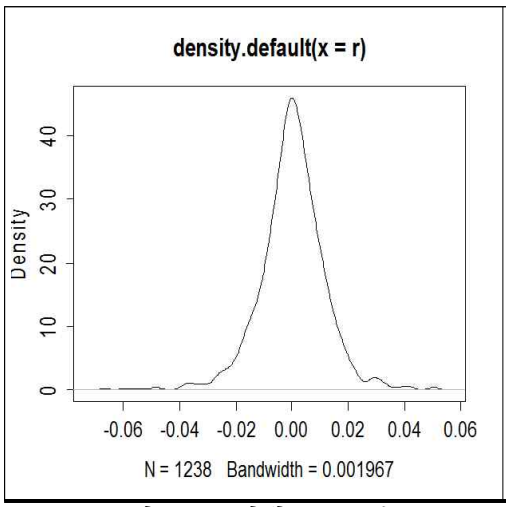
이를 이용하여 MCEM 알고리즘이 성분을 잘 분류하는지를 알아보도록 하였다. 아래 <그림 5.1>은 실제 자료의 분포모형이고, <그림 5.2>, <그림 5.3>, <그림 5.4>는 모의실험 추출결과 분포모형이다.

<표 5.1> 모의실험

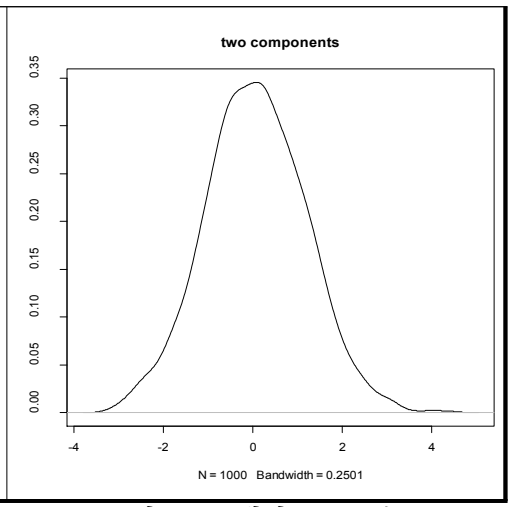
k=2	μ_1	μ_2	σ_1	σ_2	π_1	π_2						
	-0.5	0.5	1	1	0.4	0.6						

k=3	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3	π_1	π_2	π_3			
	-2	0	2	4	1	4	0.15	0.7	0.15			

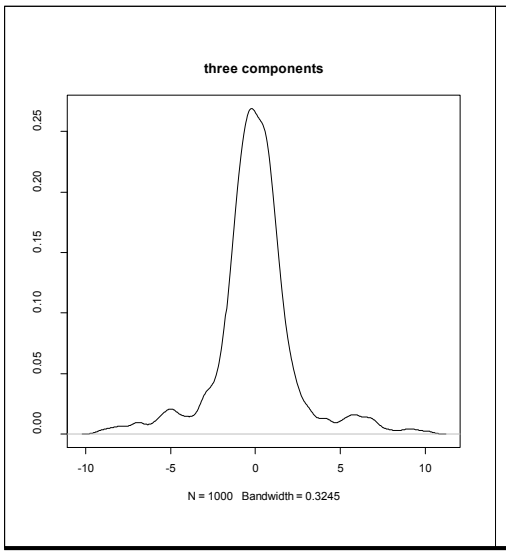
k=4	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4	π_1	π_2	π_3	π_4
	-5	-2	2	5	0.5	0.5	0.5	0.5	0.25	0.25	0.25	0.25



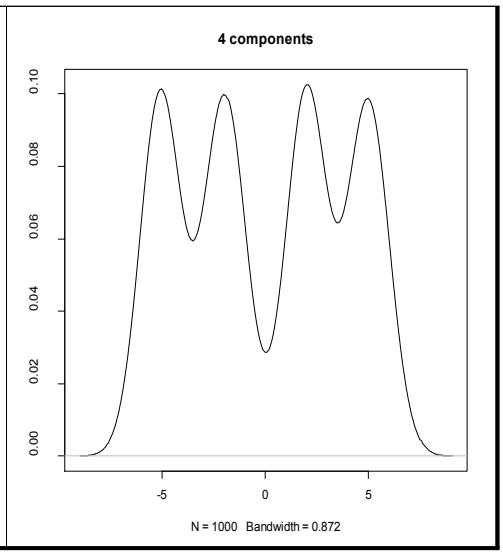
<그림 5.1> 실제 date 분포



<그림 5.2> 생성 data 분포



<그림 5.3> 생성 date 분포



<그림 5.4> 생성 data 분포

최대한 실제와 같은 모형과 비슷한 자료를 통하여 모의실험 결과를 살펴보도록 하자. EM 알고리즘이 보편적으로 사용되는 방법이지만, 본 논문에서는 결측값 데이터, Z의 분포를 가정하지 않는 MCEM 알고리즘을 시행해보았다. 우선적으로 위 알고리즘을 통해 임의로 생성된 모형들의 분포가 정확하게 나누어지는지를 확인해 보도록 한다.

5.2 모의실험 결과

5.2.1 각 생성 data분포의 정규성 검정

<표 5.2> 각 생성 data분포의 정규성 검정

2 components	skewness	kurtosis	Jarque Bera Test
	-0.2132	0.094448386 [excess]	8.018(0.01815)

3 components	skewness	kurtosis	Jarque Bera Test
	0.17625	3.3954826 [excess]	489.2123(2.2e-16)

4 components	skewness	kurtosis	Jarque Bera Test
	-0.2459	-0.37023657 [excess]	15.6617(0.0003973)

주) 괄호 안은 p-value

위와 같은 결과에 의하여 2 components, 3 components, 4 components 모두 왜도와 첨도 그리고 Jarque Bera Test(Jarque, C. M., Bera, A. K., 1981)에 의해서 정규분포를 따른다고 말 할 수가 없다. 따라서 우리는 정규분포 혼합 모형이라는 가정 하에 MCEM 알고리즘을 통하여 결과가 본 연구에서 생성 데이터의 모수를 따르는지 확인해 보도록 하자.

2 components, 3 components, 4 components로 생성된 분포들에 각각 MCEM 알고리즘을 적용한 결과들을 정리한 것이 **<표 5.3>, <표 5.4>, <표 5.5>**이다. 표의 가장 첫 번째 열은 각 분포의 모수의 참값이다. 각각의 분포는 2components, 3components, 4components 용의 프로그램으로 모의실험 한 결과이다.

5.2.2 두 개의 정규분포 혼합 모형

<표 5.3> 2 components

Components	참값	2	3	4
logL		-1531.79	-1535.84	-1538.57
AIC		3073.576	3087.672	3099.144
BIC		3098.115	3126.934	3153.129
μ_1	-0.5	-0.52363	-0.15755	-0.3965
μ_2	0.5	0.475495	0.046308	-0.07247
μ_3			0.51374	0.265238
μ_4				0.607401
σ_1	1	0.981621	1.622285	1.038405
σ_2	1	1.02114	0.968031	1.0448
σ_3			1.631248	1.058491
σ_4				1.152468
π_1	0.4	0.406561	0.083108	0.280784
π_2	0.6	0.593439	0.83116	0.25689
π_3			0.085733	0.238228
π_4				0.224098

두 개의 정규분포가 혼합(2 components)되어 있음을 가정하여 자료를 생성시킨 후 MCEM 알고리즘을 적용해 보았다. 가정된 분포를 얼마만큼이나 MCEM이 잘 예측하는지 살펴보고자 한다. <표 4.2>에 의하면 2개의 분포가 혼합된 모형에서는 역시 2 components를 가정한 알고리즘에서의 AIC, BIC가 3073.576, 3098.151로 제일 작게 나타나고 3 components과 4 components의 모형 모두의 AIC, BIC 보다 각각 10이상의 차이를 보이며 올바르게 성분이 분류가 되었음을 확인할 수 있다. 또한 실제의 θ, π의 값과 추정치 $\hat{\theta}, \hat{\pi}$의 값도 매우 비슷하게 추정되었다.

5.2.3 세 개의 정규분포 혼합 모형

<표 5.4> 3 components

Components	참값	2	3	4
logL		-2337.46	-2117.95	-2315.33
AIC		4684.911	4251.905	4652.655
BIC		4709.45	4291.167	4706.64
μ_1	-2	-1.43756	-1.87862	-1.46688
μ_2	0	1.075649	-0.02336	-0.34103
μ_3	2		1.825071	0.419985
μ_4				1.740581
σ_1	4	2.108523	3.668633	2.328197
σ_2	1	2.103329	1.025493	1.622556
σ_3	4		3.87668	1.693587
σ_4				2.055158
π_1	0.15	0.440048	0.149816	0.309222
π_2	0.7	0.559952	0.703581	0.237553
π_3	0.15		0.146603	0.215457
π_4				0.237768

세 개의 정규분포가 혼합되어 있는 형태로 생성된 데이터에 대하여 MCEM 알고리즘을 사용하여 분석하여 보았다. 5.2.2의 분석과 마찬가지로 세 개 정규분포혼합모형의 데이터는 3 components 모형이 가장 적합한 결론을 보여주고 있다. 이는 역시 AIC, BIC가 4251.905와 4291.167로 제일 작아 다른 모형보다 모형 적합도가 매우 우수하게 나타남을 알 수 있다. 또한 실제의 θ, π의 값과 추정치 $\hat{\theta}, \hat{\pi}$의 값도 매우 비슷하게 추정되어 있다.

5.2.4 네 개의 정규분포 혼합 모형

<표 5.5> 4 components

Components	참값	2	3	4
logL		-2582.2507	-2857.9143	-2524.2436
AIC		5174.5013	5731.8287	5070.4872
BIC		5199.0401	5771.0907	5124.4725
μ_1	-5	-3.5148863	-2.2082324	-5.0703569
μ_2	-2	3.4910862	0.015837735	-1.9570426
μ_3	2		2.166155	1.9896444
μ_4	5			5.0028276
σ_1	0.5	1.6332179	3.8515248	0.47795332
σ_2	0.5	1.5857552	1.8716997	0.50980564
σ_3	0.5		3.8378242	0.4664023
σ_4	0.5			3.0595715
π_1	0.25	0.5	0.35329142	0.25019162
π_2	0.25	0.5	0.29423353	0.24980639
π_3	0.25		0.35247505	0.25085828
π_4	0.25			0.24914371

위 실험은 네 개의 정규분포가 혼합되어 있는 형태로 생성된 데이터에 대하여 MCEM 알고리즘을 사용하여 분석해 보았다. 여기에서는 위의 모의실험과는 다르게 데이터는 4개의 정규분포가 혼합이 되었고 AIC와 BIC를 통하여 보았을 때 2 components 와 4 components의 모형 적합도 결과는 더욱 확실히 구분된 결과가 나타났다. 5070.4872과 5124.4725으로 가장 작고 다른 모형에 비해 모형 적합도가 매우 우수하게 나타남을 보여주고 있다.

5.2.5 모의실험 결과

모의실험의 결과의 자료들에 따라서 AIC, BIC를 살펴본 바와 같이 정규혼합분포에서의 모형 추정결과는 MCEM 알고리즘에 의하여 성분이 매우 잘 분리되며 모수 추정도 잘 된다는 것을 볼 수 있다. 또한 실제 데이터가 갖는 모형의 형태가 정규분포를 따르지 않음은 Jarque Bera Test를 통해서 보였다 (표<2.2>). 따라서 본 모의실험에서 사용된 MCEM 알고리즘을 통하여 실제 데이터를 이용한 분석에서도 믿을 수 있는 결과를 도출 할 수 있을 것이라고 예상이 된다.

제 6장. 실 증 분 석

6.1 KOSPI 200 자료

본 논문의 실증분석에 사용되는 자료는 증권거래소에서 제공하는 자료 중 2009년 10월 15일부터 2014년 10월 14일까지의 KOSPI 200의 자료이다. 그러나 본 연구는 McNeil and Frey(2000)와 비슷하게 1000개의 수익률로 이루어진 표본을 이용하여 각 표본에 대해 정규혼합분포를 추정하였다. 즉 가장 먼저 2009년 10월 15일부터 이후 1000개의 수익률을 이용하여 추정하였고 그 다음 자료는 2009년 10월 16일부터 1000개의 수익률을 이용하여 추정하였다. 연속적으로 표본기간을 이동할 경우 총 238번의 정규혼합분포를 추정하게 된다.

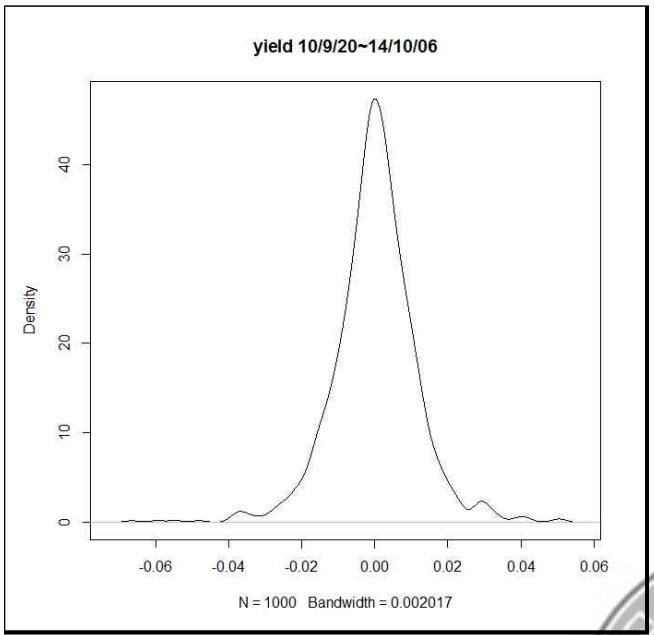
이와 같은 방식을 취한 이유는 뒤에 제시하게 되는 바와 같이 조건부 VaR을 추정하여 이로부터 사후검정을 하려 했기 때문이다. 물론 추정결과가 너무 많기 때문에 이들을 모두 제시할 수는 없을 것이다. 다만 예시를 위하여 2010년 9월 20일부터 2014년 10월 6일까지의 표본기간을 이용한 추정결과를 제시한다.

추정을 위해서 가장 먼저 성분의 수를 정해야 한다. 238개의 자료에 대해서
이 구간 외 자료들은 238개 모두가 2개의 성분이 적합한 것으로 나타났다. 그 중 위의 표본기간의 예로 들어 보았을 때, AIC, BIC는 각각 성분 2개, 3개, 4개의 순서로 (-6071.92, -6065.85, -6000.64), (-6047.38, -6026.59, -5946.65) 로 나타난다. 따라서 이 구간의 성분의 개수는 2라고 정할 수 있다.

아래의 **<표 6.1>**에는 각 성분의 개수에 맞는 MCEM 알고리즘을 적용하여 도출된 값이다.

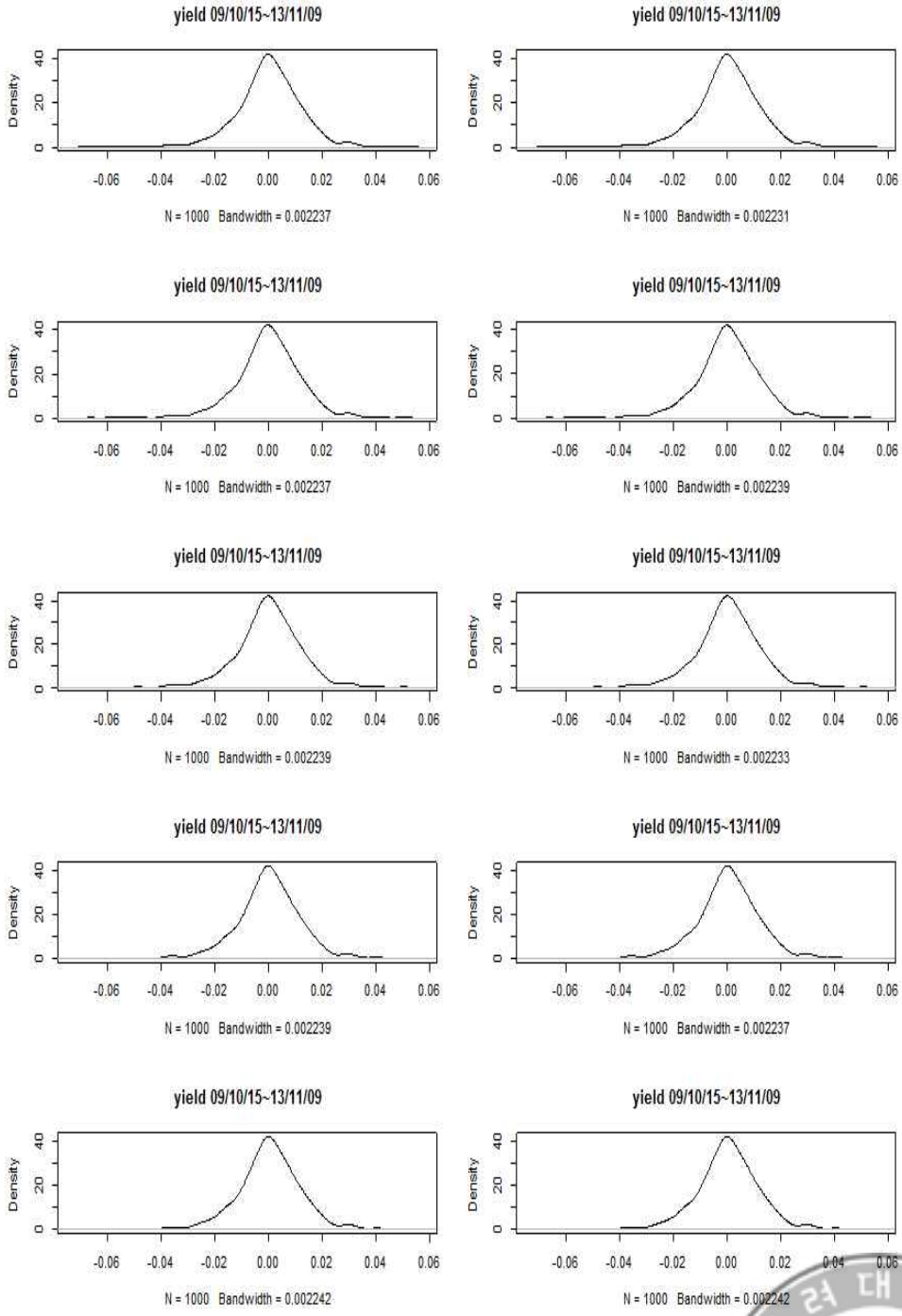
<표 6.1> 2010년 9월 20일부터 2014년 10월 6일까지

Components	2	3	4
logL	3040.961	3040.927	3011.318
AIC	-6071.92	-6065.85	-6000.64
BIC	-6047.38	-6026.59	-5946.65
μ_1	0.000814	-0.00131	0.998574
μ_2	-0.00057	5.27E-05	0.999907
μ_3		0.001374	1.001208
μ_4			1.002477
σ_1	0.011649	0.01171	0.011746
σ_2	0.011461	0.011524	0.01143
σ_3		0.011304	0.011297
σ_4			0.024185
π_1	0.37877	0.103657	0.259158
π_2	0.62123	0.792581	0.424186
π_3		0.103762	0.25787
π_4			

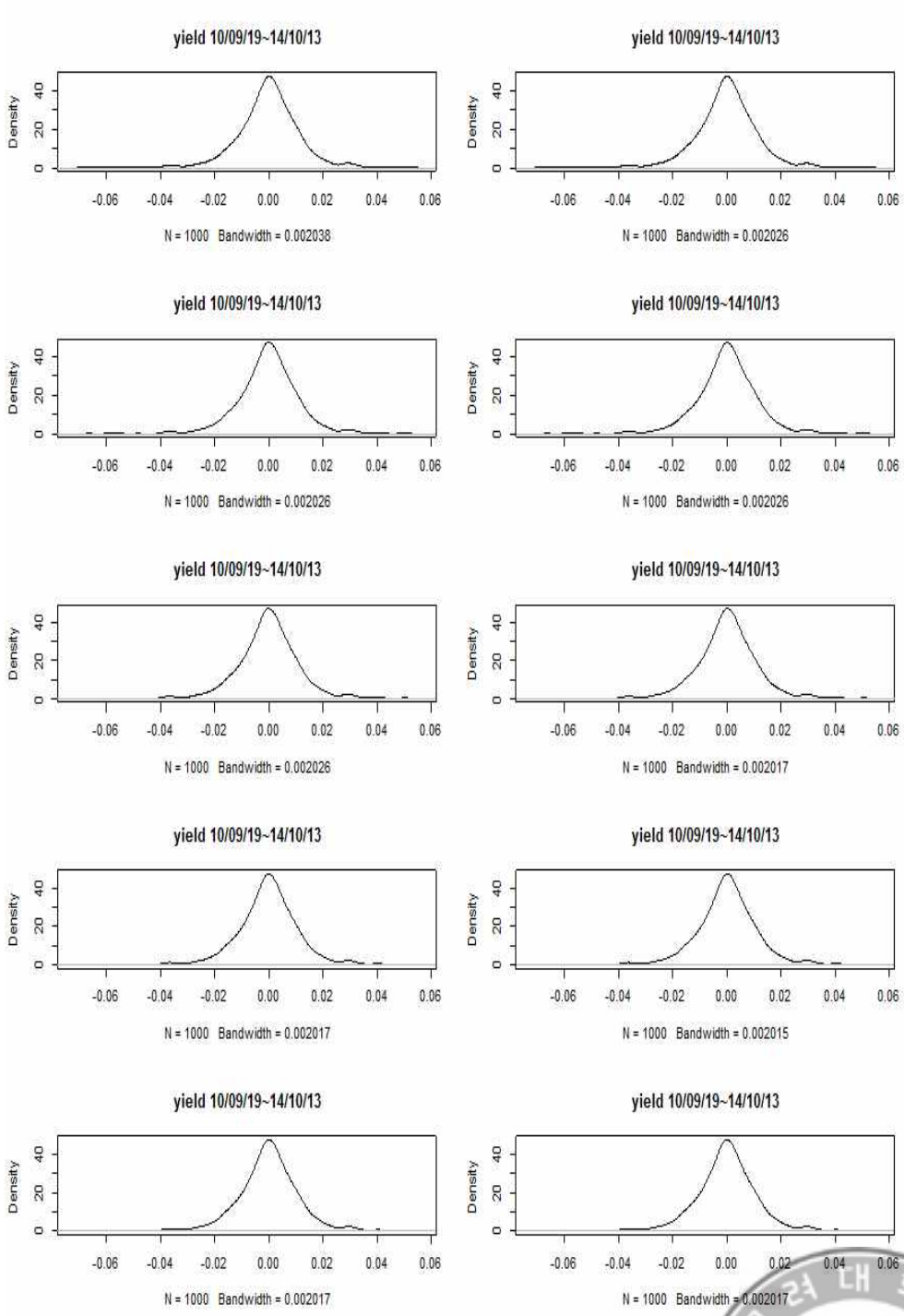


<그림 6.1> 2010년 9월 20일~ 2014년 10월 6일

모수 π_k는 각 성분에 속할 확률이 되고 μ_k는 각 성분분포의 평균이며 σ_k는 각 성분분포의 표준편차가 된다. 이 기간의 분포모형을 2개의 성분을 갖는 정규혼합분포라고 가정할 시에 두터운 꼬리(fat tail)부분을 설명하는 유용한 방법임에 틀림없다.



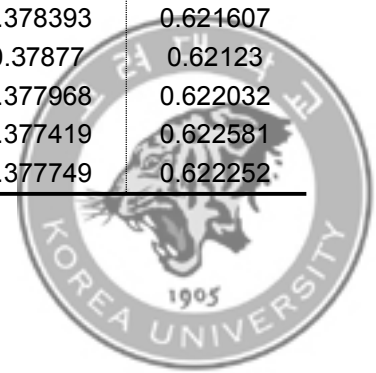
<그림 6.2> 2009년 10월 15일 ~ 2013년 11월 9일



<그림 6.3> 2010년 9월 19일부터 ~ 2014년 10월 13일

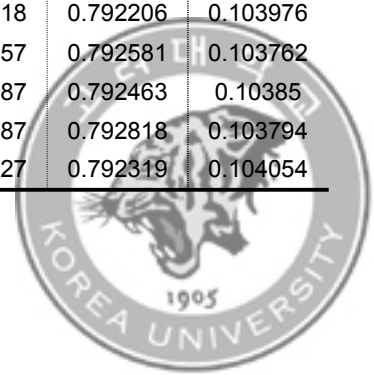
<표 6.2> 2 components MCEM 알고리즘

k=2	AIC	BIC	logL	μ_1	μ_2	σ_1	σ_2	π_1	π_2
1	-5959.08	-5934.55	2984.542	0.000687	-0.00075	0.012435	0.01207	0.377062	0.622938
2	-5959.62	-5935.08	2984.81	0.000744	-0.00081	0.012382	0.012089	0.377174	0.622826
3	-5959.34	-5934.8	2984.67	0.000707	-0.00077	0.012368	0.012104	0.377192	0.622808
4	-5959.39	-5934.85	2984.694	0.00073	-0.00079	0.0124	0.012085	0.377014	0.622986
5	-5959.4	-5934.86	2984.699	0.000687	-0.00077	0.01239	0.01209	0.378006	0.621994
6	-5960.89	-5936.35	2985.444	0.000702	-0.00081	0.01236	0.012092	0.377062	0.622938
7	-5959.59	-5935.05	2984.796	0.00071	-0.00078	0.01242	0.012073	0.377573	0.622427
8	-5960.24	-5935.7	2985.119	0.000724	-0.00078	0.012399	0.012077	0.37684	0.62316
9	-5959.87	-5935.33	2984.933	0.000736	-0.00078	0.012384	0.012087	0.377347	0.622653
10	-5963.87	-5939.33	2986.933	0.000668	-0.00077	0.012344	0.012074	0.37722	0.62278
					⋮				
229	-6072.83	-6048.3	3041.418	0.000727	-0.0006	0.011642	0.01146	0.37715	0.62285
230	-6073.51	-6048.97	3041.754	0.000748	-0.0006	0.011624	0.011464	0.377904	0.622096
231	-6073.48	-6048.95	3041.742	0.000697	-0.00056	0.01165	0.011453	0.377846	0.622154
232	-6071.73	-6047.19	3040.866	0.000745	-0.00056	0.011641	0.011471	0.378633	0.621367
233	-6071.61	-6047.07	3040.806	0.000747	-0.00055	0.011697	0.011444	0.377818	0.622182
234	-6072.05	-6047.52	3041.027	0.000801	-0.00056	0.011669	0.011451	0.378393	0.621607
235	-6071.92	-6047.38	3040.961	0.000814	-0.00057	0.011649	0.011461	0.37877	0.62123
236	-6072.11	-6047.57	3041.056	0.000791	-0.00054	0.011682	0.011444	0.377968	0.622032
237	-6070.84	-6046.3	3040.418	0.000833	-0.00053	0.011649	0.011472	0.377419	0.622581
238	-6070.82	-6046.28	3040.41	0.000803	-0.00051	0.011653	0.011473	0.377749	0.622252



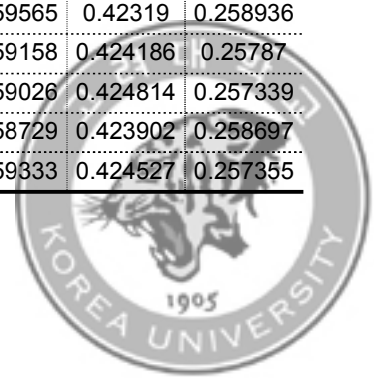
<표 6.3> 3 components MCEM 알고리즘

k=3	AIC	BIC	logL	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3	π_1	π_2	π_3
1	-5953.01	-5913.75	2984.505	-0.00128	0.00022	0.001627	0.012623	0.012176	0.011936	0.103317	0.79322	0.103463
2	-5953.43	-5914.17	2984.715	-0.00133	0.000238	0.001676	0.012497	0.012178	0.011981	0.103751	0.791579	0.104671
3	-5953.22	-5913.96	2984.612	-0.00133	0.00021	0.001755	0.012454	0.012191	0.011928	0.103607	0.792333	0.10406
4	-5952.79	-5913.53	2984.394	-0.00125	0.000204	0.001737	0.012387	0.012196	0.011979	0.103289	0.792273	0.104437
5	-5953.37	-5914.1	2984.683	-0.00132	0.000238	0.001629	0.012504	0.012199	0.01185	0.103559	0.792603	0.103838
6	-5954.66	-5915.4	2985.331	-0.00125	0.000244	0.001703	0.012423	0.012186	0.011927	0.103477	0.792226	0.104297
7	-5953.12	-5913.86	2984.56	-0.00118	0.00022	0.001593	0.012482	0.012191	0.011924	0.103828	0.791733	0.104439
8	-5954.48	-5915.22	2985.239	-0.00136	0.000222	0.001687	0.01262	0.012173	0.011862	0.103024	0.793172	0.103804
9	-5953.51	-5914.24	2984.753	-0.0013	0.00021	0.001697	0.012413	0.012197	0.011907	0.103505	0.792279	0.104216
10	-5957.73	-5918.47	2986.866	-0.00125	0.000235	0.001647	0.012444	0.012159	0.011931	0.103471	0.791798	0.104731
						⋮						
229	-6066.77	-6027.51	3041.384	-0.0012	0.000103	0.001363	0.011702	0.011522	0.011313	0.104084	0.79211	0.103806
230	-6067.78	-6028.52	3041.891	-0.00127	9.1E-05	0.00141	0.011796	0.011494	0.011372	0.103932	0.792188	0.10388
231	-6067.48	-6028.22	3041.74	-0.00128	9.14E-05	0.001394	0.01176	0.011492	0.011432	0.103401	0.793377	0.103222
232	-6065.58	-6026.32	3040.789	-0.00128	6.9E-05	0.001354	0.011672	0.011522	0.011393	0.102922	0.793124	0.103954
233	-6065.3	-6026.04	3040.649	-0.00125	5.88E-05	0.001387	0.011741	0.01152	0.011346	0.103661	0.792525	0.103814
234	-6065.88	-6026.62	3040.942	-0.00128	5.4E-05	0.001324	0.011686	0.011535	0.01124	0.103818	0.792206	0.103976
235	-6065.85	-6026.59	3040.927	-0.00131	5.27E-05	0.001374	0.01171	0.011524	0.011304	0.103657	0.792581	0.103762
236	-6065.86	-6026.6	3040.931	-0.00116	2.95E-05	0.001274	0.011744	0.011521	0.011332	0.103687	0.792463	0.10385
237	-6064.91	-6025.64	3040.453	-0.00132	1.76E-05	0.001316	0.011754	0.011529	0.011315	0.103387	0.792818	0.103794
238	-6064.67	-6025.4	3040.333	-0.00135	2.43E-05	0.001305	0.011675	0.011533	0.011351	0.103627	0.792319	0.104054



<표 6.4> 4 components MCEM 알고리즘

k=4	AIC	BIC	logL	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4	π_1	π_2	π_3
1	-5866.81	-5812.83	2944.407	0.001407	-4.1E-05	-0.00151	-0.00282	0.012443	0.012124	0.011912	0.025475	0.258595	0.42395	0.258553
2	-5876	-5822.01	2948.999	0.0014	-3.3E-05	-0.00153	-0.00296	0.012482	0.012087	0.011912	0.025565	0.257613	0.423814	0.259673
3	-5879.43	-5825.44	2950.713	0.00147	-4.7E-05	-0.00154	-0.00294	0.012485	0.012098	0.01191	0.02543	0.258689	0.423407	0.258527
4	-5872.31	-5818.32	2947.154	0.001502	-7.7E-05	-0.0015	-0.00305	0.012447	0.012123	0.011915	0.025447	0.25788	0.424581	0.258301
5	-5873.96	-5819.97	2947.979	0.001435	-6.8E-05	-0.0015	-0.00295	0.012502	0.01209	0.011913	0.025487	0.258216	0.423238	0.259252
6	-5871.28	-5817.29	2946.64	0.001422	-7.4E-05	-0.00158	-0.00286	0.012481	0.01208	0.011899	0.025401	0.25805	0.424307	0.258357
7	-5874.91	-5820.92	2948.454	0.00144	-6.2E-05	-0.00149	-0.00304	0.012561	0.012063	0.011887	0.025453	0.25884	0.423537	0.258437
8	-5863.14	-5809.15	2942.568	0.001457	-7.1E-05	-0.0015	-0.00288	0.0125	0.012071	0.011955	0.025547	0.258471	0.423319	0.258896
9	-5872.11	-5818.13	2947.056	0.001434	-4.6E-05	-0.00146	-0.00301	0.012514	0.012064	0.011897	0.025556	0.258333	0.422964	0.259581
10	-5873.27	-5819.28	2947.635	0.001425	-6.6E-05	-0.00152	-0.00292	0.012479	0.012072	0.011865	0.0254	0.258246	0.424425	0.25817
							⋮							
229	-6000.61	-5946.63	3011.307	0.001381	2.57E-05	-0.00125	-0.00248	0.011733	0.011429	0.011292	0.02431	0.258301	0.424257	0.258812
230	-6006.92	-5952.94	3014.462	0.001493	3.66E-06	-0.00126	-0.00254	0.0117	0.011406	0.011339	0.024216	0.259834	0.422904	0.258208
231	-6008.14	-5954.16	3015.072	0.001426	4.89E-05	-0.00124	-0.00256	0.011751	0.011394	0.011303	0.024238	0.258168	0.423715	0.259104
232	-5997.72	-5943.73	3009.858	0.001474	5.16E-05	-0.00124	-0.00249	0.011716	0.011412	0.011338	0.024272	0.259483	0.422876	0.258557
233	-5999.54	-5945.55	3010.769	0.001456	7.04E-05	-0.00123	-0.00251	0.011811	0.011411	0.011278	0.024185	0.258102	0.423906	0.259094
234	-6005.14	-5951.16	3013.57	0.001426	9.37E-05	-0.00118	-0.00259	0.011729	0.011416	0.011322	0.024445	0.259565	0.42319	0.258936
235	-6000.64	-5946.65	3011.318	0.001426	9.25E-05	-0.00121	-0.00248	0.011746	0.01143	0.011297	0.024185	0.259158	0.424186	0.25787
236	-6003.87	-5949.88	3012.933	0.001448	0.000121	-0.00123	-0.00247	0.011723	0.011449	0.011299	0.024215	0.259026	0.424814	0.257339
237	-5996.45	-5942.46	3009.223	0.001446	0.000157	-0.00121	-0.00242	0.011711	0.011435	0.011339	0.024366	0.258729	0.423902	0.258697
238	-6010.07	-5956.08	3016.033	0.001446	0.000144	-0.00118	-0.00253	0.011763	0.01142	0.011307	0.024227	0.259333	0.424527	0.257355



6.2 VaR과 사후검정

우선 본 논문에서는 기준값을 넘는다고 간주되는 초과치(exceedances)의 수를 정하는 방법을 사용할 것이다(McNeil and Frey, 2000). 즉 표본이 1000개 일 때 100개를 초과치로 간주하고, 정규혼합분포를 추정했던 것과 같은 방식으로 1000개의 수익률로 이루어진 표본을 이용하여 각 표본에 대해 정규혼합분포를 추정하였다. 238개의 표본기간에 대하여 추정하였고 곧 이는 238개의 조건부 VaR을 얻게 되었다. 이는 McNeil and Frey(2000)에서 사용한 방식을 통해 사후검정을 수행하고 **<표 6.5>**의 결과를 도출해 내었다.

<표 6.5> 조건부 VaR

VaR	1%	5%	95%	99%
Expected	12	62	62	12
mixture normal	9(0.273)	49(0.095)	50(0.118)	10(0.337)

※ 괄호안은 p-value

사후검정방법을 간단히 설명하면 첫째, t기까지의 자료를 이용하여 구한 VaR 추정치와 t+1기의 실제수익률을 비교한다. 그런데 VaR의 유의수준 q하에서 실제수익률이 VaR 추정치보다 더 클 확률은 1-q이다 . 둘째, t+1기의 실제수익률이 VaR추정치보다 클 때 1의 값을 갖고 그렇지 않을 경우 0의 값을 갖는 확률변수 I_t는 1-q를 모수로 갖는 베르누이분포(Bernoulli distribution)를 따르게 된다. 셋째 표본기간에 대하여 $\sum I_t$를 구하면 이 확률변수는 포본수와 1-q를 모수로 갖는 이항분포(binomial distribution)를 따르게 된다.

즉, 가장 먼저 2009년 10월 15일부터 2013년 10월 22일까지의 1000개의 수익률을 이용하여 추정하였고 그 다음 자료는 2009년 10월 16일부터 2013년 10월 23일까지의 1000개의 수익률을 이용하여 추정하였다. 총 자료의 개수는 1239개의 수익률 자료가 있기 때문에 연속적으로 표본기간을 이동할 경우 총 238번의 정규혼합분포를 추정할 수 있다. 이 기간 내의 분포들의 성분은 주로 3개로 예측하였으나 거의 대부분의 분포에서 2개의 성분을 갖고 있는 정규 혼합 분포였다. AIC와 BIC는 2개의 성분을 갖고 있는 분포와 3개의 성분을 갖고 있는 분포 사이에 큰 차이를 보이고 있지는 않다. 그러나 본 논문에서는 조금이라도 더 확실한 2개의 성

분을 갖고 있는 정규혼합분포를 이용하여 논의하였다.

사후검정결과 위반의 수에 대한 기댓값이 제시되어 있고, 조건부 VaR에 대한 검정결과가 제시되어 있다. 즉 각 수치는 위반의 수 $\sum I_t$을 의미하며 괄호 안에는 이에 대한 p값이 제시되어 있다. 특이점으로는 p값이 클수록 더 유의하다고 볼 수 있는데 이는 McNeil and Frey(2000)와 일치시키기 위한 것이다.

사후검정결과를 요약하면 우리가 구한 정규혼합분포에 의한 VaR은 모든 구간에서 유의함을 보여 조건부 VaR를 잘 추정하고 있음을 알 수 있었다.

제 7장. 결 론

본 연구는 한국거래소(Korea Exchange)의 주가지수 200을 대상으로 수익률 분포를 추정하고 이로부터 위험측정값인 VaR을 구할 수 있었다. 수익률의 분포로는 정규혼합분포를 이용하였고 추정방법으로는 MCEM 알고리즘을 사용하였다. 정규혼합분포는 정규분포들의 혼합으로 이루어진 분포이며 두꺼운 꼬리를 갖는 여러 분포에서 많은 활용도가 가능하다. 특히 불완전 자료에 대한 모수를 추정하고자 할 때 사용하는 가장 일반적인 방법인 EM 알고리즘에서 불완전 자료의 분포를 가정하지 않고 결측값(missing) 데이터 자체를 Metropolis-Hastings 알고리즘을 이용하여 생성시키는 방식을 사용한다. 본 논문에서 다루었던 정규혼합분포의 경우는 그 분포가 널리 알려진 정규분포로 사용하였으나 실제 정형화되지 않은 분포의 결측값(missing) 데이터의 생성에 더욱 필요한 알고리즘이다. 또한 위험 측정값인 VaR의 경우에서 보듯 정확하게 분포를 파악하게 되면 우리가 대비할 수 있는 위험을 더 정밀하게 측정할 수 있다.

본 논문은 McNeil and Frey(2000)와 비슷하게 1000개의 수익률로 이루어진 표본을 이용하여 각 표본에 대해 정규혼합분포를 추정하였다. 정확한 분포를 통하여 생성된 조건부 VaR의 추정도 정규혼합분포에서 1%, 5%, 95%, 그리고 99% 모두에서 유의함을 보였으며 극단값 분포를 활용하여 구하는 것보다는 데이터가 따르는 정확한 분포를 통하여 위험 측정값을 구하는 것이 다중최빈값을 갖는 경우에서 더 활용성이 높으며 충분히 리스크 측정도구로서 사용될 수 있음을 보였다.

참 고 문 헌

[1] 서영수, (2012), *금융과 리스크 관리*, 교문사

[2] 윤종인, (2011), 깁스샘플러를 이용한 정규분포혼합 및 VaR의 추정과 사후검정 : *통계연구(2011), 제16권 제1호, 60-81.*

[3] 이희찬, (2012), 불완전 자료에 대한 여러 가지 알고리즘에 관한 연구 : *고려대학교 대학원 석사학위 논문.*

[4] Akaike, H., (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control 19 (6): 716–723*

[5] Christian, P.R. George C. (2004). *Monte Carlo Statistical Methods,* second edition, Springer.

[6] Dempster, A., Laird, N. and Rubin, D. (1977), Maximum likelihood from incomplete data via the EM algorithm : *Journal of the Royal Statistical Society, Series B, 39(1):138.*

[7] Hastings. W.K. (1970) Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika, 57(1):97-109,*

[8] Jarque, C.M., Bera, A.K. (1981), Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. *Economics Letters 7 (4): 313–318.*

[9] McCulloch, C. E., (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models : *Journal of the American Statistical Association.*

[10] McLachlan, G, and Pell, D. (2000), *Finite Mixture Model,* John Wiley & Sons, New York.

[11] McNil, A.J., and Frey, R. (2000), Estimation of Tail-related Risk Measures for Heteroscedastic financial Time Series : An Extreme Value Approach, *Journal of Empirical Finance 7, 271-301.*

[12] McNeil, Frey and Embrechts (2005), *Quantitative Risk Management: Concepts, Techniques, and Tools*, Princeton University Press.

[13] Metropolis N, Rosenbluth. A.W., Rosenbluth. M.N., Teller A.H., and Teller. E. (1953) Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics, 21(6):1087-1092,*

[14] Schwarz, G. E.,(1978), Estimating the dimension of a model, *Annals of Statistics 6 (2): 461-464*

[15] Wei, Greg. C.G. and Tanner, Martin A. (1990), A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms, *Journal of the American Statistical Association, Vol. 85, 699-714.*