碩 士 學 位 論 文

불완전 자료에 대한 여러 가지 알고리즘에 관한 연구

-MHEM 알고리즘과 KOSPI 200의 수익률을 중심으로-

高麗大學校 大學院

經濟統計學科

李 義 贊

2011年 12月

全 秀 榮 敎 授 指 導

碩 士 學 位 論 文

불완전 자료에 대한 여러 가지 알고리즘에 관한 연구

-MHEM 알고리즘과 KOSPI 200의 수익률을 중심으로-

이 論文을 經濟學 碩士學位 論文으로 提出함.

2011年 12月

高 麗 大 學 校 大 學 院

經 濟 統 計 學 科

李 義 贊 (印)

李 義 贊의 經濟學 碩士學位論文
審査를 完了함.

2011年　12月

委員長 (印)

委　員 (印)

委　員 (印)

요 약 문

결측자료(missing data), 절단분포(truncated distribution), 중도절단자료(censored data) 등 불완전한 자료(incomplete data)하의 추론문제(incomplete problems)는 통계학에서 자주 발생되는 현상이다. 이런 문제의 해결방법으로 Expectation Maximization(EM), Monte Carlo Expectation Maximization(MCEM), Stochastic Expectation Maximization(SEM) 알고리즘 등이 있지만, 본 연구에서는 정형화된 분포의 가정이 없는 경우에 사용할 수 있는 Metropolis Hastings Expectation Maximization(MHEM) 알고리즘을 제안한다.

본 논문에서는 각 알고리즘들의 개념과 특성을 살펴보고, 각 알고리즘들의 효율성을 고찰하여, 어느 상황에서 어떠한 알고리즘을 적용하는 것이 효율적인지 살펴본다. 또한 본 논문에서는 중도절단자료(censored data)를 이용한 모의실험을 통하여 각 알고리즘의 특성 및 효율성을 비교 분석하고, KOSPI 200 자료를 가지고 MHEM 알고리즘을 이용한 실증분석을 진행하였다.

목 차

표 목 차

그림 목차

제 1장 서 론

통계 분석 대상이 되는 자료에 결측치가 있을 경우, 이러한 불완전한 자료를 가지고 분석을 진행하게 되면 많은 문제점을 갖게 된다. 이때 결측치를 가진 자료를 삭제하는 방법으로부터 여러 가지 다른 값을 가지고 결측치를 대체하는 방법을 고려하게 되는데, 이런 것들을 포함한 여러 가지 방법이 존재한다.

가장 먼저 생각해 볼 수 있는 방법으로는 결측치의 분포에서 최대우도를 갖는 값을 실제 계산을 통해서 구하는 방법이다. 그러나 결측치의 우도함수가 계산을 통해 최대값이 계산되어지지 않는 경우라면, 근사식을 이용해야 하는 경우도 발생한다. 이런 경우 수치해석적인 방법을 통해 최대값을 구할 수 있는데, 이 방법은 프로그래밍이 복잡해지고, 일봉(one mountaintop)의 경우가 아닌 경우에 여러 가지 문제점들이 발생한다. 이런 문제점을 보완하기 위해서 결측치의 충분통계량을 계산하여 최대화하는 과정을 반복적으로 수행하는 시뮬레이션 방법이 제안될 수 있으며, 그 중 가장 널리 알려져 있는 대체 방법으로 Expectation Maximization(EM) 알고리즘(Dempster et al., 1977)이 있다. 이 반복적 알고리즘의 가장 큰 장점은 기존의 근사접근이나, 수치해석적인 방법에 비해 상대적으로 프로그래밍 하기 쉽고, 우도함수 또는 로그우도함수를 최대화시키는 m.l.e로의 단조수렴 추정치를 생성해 낸다는 것이다.

EM 알고리즘은 매우 효과적인 알고리즘이지만, 몇 가지 한계를 가지고 있다. 이런 EM 알고리즘의 한계를 개선하기 위해 지금까지 여러 가지 알고리즘 등이 제안되어 왔다. 하지만 실제 자료들은 대부분이 불완전한 자료들인 경우가 많기 때문에 많은 분야에서 EM 알고리즘을 변형하여 사용하고 있다. 예를 들면, 김승구(2003, 2004, 2005)는 자기공명영상의 올바른 분할을 위해서 효과적인 바이어스 필드보정에 EM 알고리즘을 사용하였다. 또한 이 뿐만 아니라 강만기(2000)는 신뢰성 분석에 있어서 Weibull 분포에 대한 모수 추정 시 변형된 EM 알고리즘을 제안하여 사용하였다. 이런 방식 등으로 제안된 여러 가지 알고리즘 중에서 대표적인 것으로 Monte Carlo Expectation Maximization 알고리즘(wei et al., 1990), Stochastic Expectation Maximization 알고리즘(Celeux et al., 1985) 등이 있다.

본 논문에서는 Metorpolis Hastings Expectation Maximization(MHEM) 알고리즘이라는 변형된 EM 알고리즘을 제안한다. 본문에서는 EM 알고리즘과 변형된 EM 알고리즘 등을 정의하고 각 알고리즘들의 차이점과 특징을 비교 분석하여 어느 상황에서 어떠한 알고리즘을 쓰는 것이 효율적인지 살펴본다. 또한 각각의 알고리즘들을 이용하여 중도절단자료(cencored data)에 대한 모의실험 및 실증분석을 시행해 보았다.

본 논문은 2장에서 불완전 자료의 경우에 기존에 사용하는 알고리즘들의 특징 및 프로세스를 소개하고, 3장에서 본 논문에서 제안하는 MHEM 알고리즘을 소개한다. 그리고 4장에서는 모의실험, 5장에서는 실증분석의 결과를 살펴본다. 마지막으로 6장에서 본 논문의 결론을 정리하였다.

제 2장 알고리즘 소개

2.1 Expectation Maximization 알고리즘

현대 사회에서의 통계는 우리 생활과 아주 밀접한 관계를 가지기 시작했다. 통계가 우리 생활과 가까워질수록 통계 자료 및 통계 분석 결과를 많은 분야에서 활용을 하게 되었다. 이러한 통계 자료 및 통계 분석 결과를 활용하기 위해서는 통계 자료들을 수집해야 하는 것이 선행되어야 한다. 하지만 사회에서 자료들을 수집하는 것은 여러 가지 상황 및 제약 때문에 완전한 자료들을 구한다는 것은 매우 어려운 일이다. 그래서 통계 전반에 걸쳐 결측치와 불완전 자료들에 관한 많은 문제들이 존재하며, 이런 결측치가 존재할 경우, 원 자료를 사용하는데 많은 문제들이 발생된다.

이런 문제점을 해결하기 위해 생각해 볼 수 있는 방법으로는 결측치의 분포에서 최대우도를 갖는 값을 실제 계산을 통해서 구하는 방법이다. 그러나 결측치의 우도함수가 계산을 통해 최대값이 계산되어지지 않는 경우라면, 근사식을 이용해야 하는 경우도 발생한다. 이런 경우 수치해석적인 방법을 통해 최대값을 구할 수 있는데, 이 방법은 프로그래밍이 복잡해지고, 일봉의 경우가 아닌 경우 여러 가지 문제점들이 발생한다. 이런 문제점을 보완하기 위해서 결측치의 충분통계량을 계산하여, 최대화하는 과정을 반복적으로 수행하는 시뮬레이션 방법이 제안될 수 있으며, Dempster et al., 1977에 의해 제안된 Expectation Maximization(EM) 알고리즘은 다양한 불완전한 자료(incomplete data)로부터 최대 우도추정치(m.l.e)를 반복적인 기법을 통해 구할 수 있는 방법으로 위와 같은 문제점들을 다루는데 널리 사용되는 도구가 되었다. 이 반복적 알고리즘의 가장 큰 장점은 기존의 근사접근이나, 수치해석적인 방법에 비해 상대적으로 프로그래밍하기 쉽고, 우도함수 또는 로그우도함수를 최대화시키는 m.l.e로의 단조수렴 추정치를 생성해 낸다는 것이다. 알고리즘의 각 반복은 Expectation 단계(E-step)과 Maximization 단계(M-step)로 구성되어 있기에 이것을 EM 알고리즘이라고 부르며, 관련이론의 단순성과 일반성을 가지고 있으며, 다양한 분야에 대해 적용이 가능하기 때문에 주목 받아왔다. 특히 M.L.E가 쉽게 계산되어지는 지수 족에서의 완전자료인 경우, EM 알고리즘의 M-step의 계산은 마찬가지로 쉽게 계산되어진다.

다음 <표 2.1>은 EM 알고리즘의 E-step과 M-step을 설명하기위한 기호를 설명과 함께 표로 정리한 것이다. 이것을 바탕으로 E-step과 M-step이 어떻게 유도되어 EM 알고리즘을 구성하는지 살펴보겠다.

<표 2.1> EM 알고리즘의 용어 설명

기 호	설 명
$\boldsymbol{x}' = (x_1, x_2, \cdots, x_{n1})$	관측된 자료
$\boldsymbol{z}' = (z_1, z_2, \cdots, z_{n2})$	관측되지 않은 자료
$g(\boldsymbol{x}\|\theta)$	x의 결합 p.d.f
$h(\boldsymbol{x}, \boldsymbol{z}\|\theta)$	관측된 자료와 관측되지 않은 자료의 결합 p.d.f
$k(\boldsymbol{z}\|\theta, \boldsymbol{x})$	관측된 자료가 주어졌을 때 관측되지 않은 자료의 조건부 p.d.f
$L(\theta\|\boldsymbol{x})$	관측된 자료의 우도함수
$L^c(\theta\|\boldsymbol{x}, \boldsymbol{z})$	완전한 자료의 우도함수

EM 알고리즘의 목표는 완전한 우도함수 $L^c(\theta|\boldsymbol{x}, \boldsymbol{z})$를 이용하여 $L(\theta|\boldsymbol{x})$를 최대화 하는 것이다. 이 목표를 위해 특정 값으로 지정되진 않았으나 일정한 $\theta_0 \in \Omega$에 대해 다음과 같은 항등식을 유도 할 수 있다.

$$\log L(\theta|\boldsymbol{x}) = \int \log L(\theta|\boldsymbol{x}) \cdot k(\boldsymbol{z}|\theta_0, \boldsymbol{x})\, dz \quad (\because \int k(\boldsymbol{z}|\theta_0, \boldsymbol{x})\, dz = \int \frac{h(\boldsymbol{x}, \boldsymbol{z}|\theta_0)}{g(\boldsymbol{x}|\theta_0)}\, dz$$
$$= \frac{1}{g(\boldsymbol{x}|\theta_0)} \cdot \int h(\boldsymbol{x}, \boldsymbol{z}|\theta_0)\, dz$$
$$= \frac{1}{g(\boldsymbol{x}|\theta_0)} \cdot g(\boldsymbol{x}|\theta_0) = 1)$$

$$= \int \log g(\boldsymbol{x}|\theta) \cdot k(\boldsymbol{z}|\theta_0, \boldsymbol{x})\, dz$$
$$= \int \log \left(\frac{h(\boldsymbol{x}, \boldsymbol{z}|\theta)}{k(\boldsymbol{z}|\theta, \boldsymbol{x})}\right) \cdot k(\boldsymbol{z}|\theta_0, \boldsymbol{x})\, dz \quad (\because k(\boldsymbol{z}|\theta, \boldsymbol{x}) = \frac{h(\boldsymbol{x}, \boldsymbol{z}|\theta)}{g(\boldsymbol{x}|\theta)} \Rightarrow g(\boldsymbol{x}|\theta) = \frac{h(\boldsymbol{x}, \boldsymbol{z}|\theta)}{k(\boldsymbol{z}|\theta, \boldsymbol{x})})$$
$$= \int [\log h(\boldsymbol{x}, \boldsymbol{z}|\theta) - \log k(\boldsymbol{z}|\theta, \boldsymbol{x})] \cdot k(\boldsymbol{z}|\theta_0, \boldsymbol{x})\, dz$$
$$= \int [\log h(\boldsymbol{x}, \boldsymbol{z}|\theta)] \cdot k(\boldsymbol{z}|\theta_0, \boldsymbol{x})\, dz - \int [\log k(\boldsymbol{z}|\theta, \boldsymbol{x})] \cdot k(\boldsymbol{z}|\theta_0, \boldsymbol{x})\, dz$$
$$= E_{\theta_0}\left[\log L^c(\theta|\boldsymbol{x}, \boldsymbol{z})|\theta_0, \boldsymbol{x}\right] - E_{\theta_0}\left[\log k(\boldsymbol{z}|\theta, \boldsymbol{x})|\theta_0, \boldsymbol{x}\right]$$

($\because$ $u(x_2)$가 x_2의 함수일 경우, $X_1 = x_1$이 주어졌을 때 $u(x_2)$의 조건부 기댓값 이 다음과 같이 존재한다. $E[u(x_2)|x_1] = \int u(x_2) \cdot f_{2|1}(x_2|x_1)\, dx_2$)

기댓값은 조건부 p.d.f $k(\boldsymbol{z}|\theta_0,\boldsymbol{x})$ 하에서 계산되기 때문에 $E_{\theta_0}\left[\log k(\boldsymbol{z}|\theta,\boldsymbol{x})|\theta_0,\boldsymbol{x}\right]$를 무시할 수 있다. 그러면

$$Q(\theta|\theta_0,\boldsymbol{x}) = E_{\theta_0}\left[\log L^c(\theta|\boldsymbol{x},\boldsymbol{z})|\theta_0,\boldsymbol{x}\right]$$

와 같이 Q를 정의하는 기댓값을 EM 알고리즘의 E-step이라고 한다.

EM 알고리즘의 목적은 $L(\theta|\boldsymbol{x})$, 즉 $\log L(\theta|\boldsymbol{x})$를 최대화하는 것이다. 이것은 $Q(\theta|\theta_0,\boldsymbol{x})$를 극대화하는 것과 같다. $Q(\theta|\theta_0,\boldsymbol{x})$의 극대화 과정은 다음과 같다.

1단계 : $\widehat{\theta_{(0)}} = \theta$를 초기 추정값(관측된 우도에 근거한 값)이라고 한다.

2단계 : $m=1,\cdots,N$ 이라고 하자. $\widehat{\theta_{(m)}} = Q(\theta|\widehat{\theta_{(m-1)}},\boldsymbol{x})$를 최대화 하는 값을 θ의 m단계 추정값이라고 한다.

3단계 : 2단계 과정을 통해 추정값 $\widehat{\theta_{(m)}}$을 구한다.

여기서 $m\to\infty$ 일 때 $\widehat{\theta_{(m)}}$가 최대우도 추정값으로 확률 수렴한다.

위에서 $\widehat{\theta_{(m+1)}}$는 $\widehat{\theta_{(m)}}$에 비해 우도함수를 항상 증가시킨다. 이것을 수식으로 표현하면 $L(\widehat{\theta_{(m+1)}}|\boldsymbol{x}) \geq L(\widehat{\theta_{(m)}}|\boldsymbol{x})$와 같다. 이것은 다음과 같은 과정을 통해서 증명 할 수 있다.

$Q(\widehat{\theta_{(m+1)}}|\widehat{\theta_{(m)}},\boldsymbol{x}) - Q(\widehat{\theta_{(m)}}|\widehat{\theta_{(m)}},\boldsymbol{x})$ ($\because$ $\widehat{\theta_{(m+1)}}$이 $Q(\theta|\widehat{\theta_{(m)}},\boldsymbol{x})$를 최대화)

$= E_{\widehat{\theta_{(m)}}}\left[\log L^c(\widehat{\theta_{(m+1)}}|\boldsymbol{x},\boldsymbol{z})\right] - E_{\widehat{\theta_{(m)}}}\left[\log L^c(\widehat{\theta_{(m)}}|\boldsymbol{x},\boldsymbol{z})\right]$

(여기서 기댓값은 조건부 p.d.f $k(\boldsymbol{z}|\widehat{\theta_{(m)}},\boldsymbol{x})$하에서 구해진다.)

$= E_{\widehat{\theta_{(m)}}}\left[\log k(\boldsymbol{z}|\widehat{\theta_{(m+1)}},\boldsymbol{x})\right] - E_{\widehat{\theta_{(m)}}}\left[\log k(\boldsymbol{z}|\widehat{\theta_{(m)}},\boldsymbol{x})\right]$

(기댓값들은 $\widehat{\theta_{(m)}}$과 $\boldsymbol{x}$가 주어졌을 때 $\boldsymbol{z}$의 조건부 p.d.f하에서 구해진다.)

(젠슨 부등식(Jensen's inequality)을 이용)

$= E_{\widehat{\theta_{(m)}}}\left[\log\left[\frac{k(\boldsymbol{z}|\widehat{\theta_{(m+1)}},\boldsymbol{x})}{k(\boldsymbol{z}|\widehat{\theta_{(m)}},\boldsymbol{x})}\right]\right] \geq \log E_{\widehat{\theta_{(m)}}}\left[\frac{k(\boldsymbol{z}|\widehat{\theta_{(m+1)}},\boldsymbol{x})}{k(\boldsymbol{z}|\widehat{\theta_{(m)}},\boldsymbol{x})}\right]$

$\left(= \log\int \frac{k(\boldsymbol{z}|\widehat{\theta_{(m+1)}},\boldsymbol{x})}{k(\boldsymbol{z}|\widehat{\theta_{(m)}},\boldsymbol{x})} \cdot k(\boldsymbol{z}|\widehat{\theta_{(m)}},\boldsymbol{x})dz = \log(1) = 0\right)$

$= E_{\widehat{\theta_{(m)}}}\left[\log\left[\frac{k(\boldsymbol{z}|\widehat{\theta_{(m+1)}},\boldsymbol{x})}{k(\boldsymbol{z}|\widehat{\theta_{(m)}},\boldsymbol{x})}\right]\right] \geq 0$

이러한 극대화 과정을 EM 알고리즘의 M-step이라고 한다.
EM 알고리즘의 E-step과 M-step을 정리하면 다음 <표 2.2>와 같다.

<표 2.2> EM 알고리즘

$\widehat{\theta_{(m)}}$ 가 m번째 단계에서의 추정값을 나타낸다고 하자. (m+1)번째 단계의 추정값을 계산하기 위해 다음의 E-step과 M-step을 반복 시행한다.

E-step : $Q(\theta|\widehat{\theta_{(m)}}, \boldsymbol{x}) = E_{\widehat{\theta_{(m)}}}[\log L^c(\theta|\boldsymbol{x}, \boldsymbol{z})|\widehat{\theta_{(m)}}, \boldsymbol{x}]$를 구한다.

여기서 기댓값은 조건부 pdf $k(\boldsymbol{z}|\widehat{\theta_{(m)}}, \boldsymbol{x})$ 하에서 구해진다.

M-step : $\widehat{\theta_{(m+1)}} = Arg\max Q(\theta|\widehat{\theta_{(m)}}, \boldsymbol{x})$

EM 알고리즘은 많은 장점을 가지고 있지만, 한계점도 가지고 있다. 지금까지 설명한 EM 알고리즘은 기댓값이 쉽게 계산된다는 가정 하에 설명하였다. 하지만 결측 자료가 조건부로 들어간 우도함수의 기댓값을 구하는 E-step은 고차의 적분을 수반하므로 쉽게 계산되지 않는 경우가 발생하게 된다는 단점을 가지고 있다.

2.2 Stochastic Expectation Maximization 알고리즘

Stochastic Expectation Maximization(SEM) 알고리즘은 EM 알고리즘을 적용하기에 어렵고 복잡했던 많은 문제들, 특히 EM 알고리즘의 E-step에서 다차원의 수치적분을 포함하고 있어 해결이 쉽지 않은 경우에 대해 해결하고자 Celeus et al., 1985 가 제안한 알고리즘 이다. SEM 알고리즘의 주된 내용은 각 반복 m에서 $\theta_{(m)}$이 θ에 대한 현재 추정치인 경우, 식 (2.1)의 $k(\boldsymbol{z}|\theta_{(m)}, \boldsymbol{x})$로부터 하나의 표본을 추출하여 결측치 $\boldsymbol{z}$를 채워 넣는다는 것이다.

$$k(\boldsymbol{z}|\theta_{(m)}, \boldsymbol{x}) = \frac{h(\boldsymbol{x}, \boldsymbol{z}|\theta_{(m)})}{\int h(\boldsymbol{x}, \boldsymbol{z}'|\theta_{(m)})dz'} = \frac{h(\boldsymbol{x}, \boldsymbol{z}|\theta_{(m)})}{g(\boldsymbol{x}|\theta_{(m)})} \quad (2.1)$$

결측치 $\boldsymbol{z}$를 이런 방식으로 대체하면 현재 θ에 대해 가지고 있는 모든 정보에 의존하며, 그렇기에 우리는 적절한 의완전자료(pseudo complete data)를 갖게 된다. 일단

의완전자료를 갖게 되면 의완전자료 로그우도함수를 최대화함으로써 갱신되어지는 m.l.e, $\widehat{\theta_{(m+1)}}$을 얻게 된다. 이러한 과정을 반복하면서 SEM 알고리즘이 구성된다.

SEM 알고리즘을 정리하면 다음 <표 2.3>과 같다.

<표 2.3> SEM 알고리즘

$\widehat{\theta_{(m)}}$ 가 m 번째 단계에서의 추정값을 나타낸다고 하자. (m+1)번째 단계의 추정값을 계산하기 위해 다음을 반복 시행한다.

S-step : $k(\boldsymbol{z}|\theta_{(m)}, \boldsymbol{x})$로부터 표본을 추출하여 결측치 $\boldsymbol{z}$로 대체하여 의완전자료 생성

$$\widehat{\theta_{(m)}} = \frac{1}{(T-n_0)} \sum_{n=n_0+1}^{T} \theta^{(n)} \quad (n_0 : \text{소각하는 초기 반복횟수})$$

M-step : $\widehat{\theta_{(m+1)}} = Arg\max \widehat{\theta_{(m)}}$

의완전자료를 제공하는 S-step과 최대화를 시행하는 M-step의 과정은 가벼운 조건(Ip, E. H. S 1994a)하에서 정상분포 π로 수렴해가는 마코브 연쇄(Markov Chain) $\{\theta_{(m)}\}$를 생성한다. 그 정상분포 π는 근사적으로 θ의 m.l.e를 중앙에 두며, SEM 알고리즘의 반복 안에서의 $\theta_{(m)}$의 변화율에 의존하는 분산을 가지고 있다. 또, SEM 알고리즘을 사용하는 대부분의 상황에서 $\theta_{(m)}$의 수렴은 충분히 빠른 것으로 알려져 있다.(Celeux et al., 1985)

2.3 Monte Carlo Expectation Maximization 알고리즘

Monte Carlo Expectation Maximization(MCEM) 알고리즘은 SEM 알고리즘처럼 EM 알고리즘의 단점인 E-step이 고차의 적분을 수반하게 되면 계산되지 않는 경우가 발생한다는 것을 보완하기 위해 Wei et al., 1990가 제안한 알고리즘이다. MCEM 알고리즘은 EM 알고리즘의 E-step의 기댓값 계산에 필요한 적분과정을 몬테카를로(Monte Carlo) 방법으로 해결하여 결측 자료를 구하고 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 찾

을 수 있는데 이것이 몬테카를로 방법이다. 몬테카를로 방법의 장점 중의 하나는 계산이 다른 수학적 방법에 비해 간단하다는 것을 들 수 있다.

MCEM 알고리즘에서 사용한 몬테카를로 적분법은 확률변수들을 발생시켜 어려운 적분 계산을 쉽게 해결하기 위해 개발되었다. 다음과 같은 적분을 생각해 보자.

$$E_{\pi}[h(\boldsymbol{x})] = \int h(\boldsymbol{x}) \cdot \pi(\boldsymbol{x}) dx \tag{2.2}$$

식 (2.2)에서 적분 값을 구하는 것은 힘들지만 많은 수의 확률변수들($x_1, x_2, \cdots, x_n$)을 발생시킬 수 있다면, 대수의 법칙(law of large numbers)에 의해서 근사적으로 구할 수 있다.

$$E_{\pi}[h(\boldsymbol{x})] = \lim_{n\to\infty} \frac{1}{n} \cdot \sum_{i=1}^{n} h(x_i) \tag{2.3}$$

이와 같이 적분을 기댓값의 형태로 바꾸어 적분 값을 추정하는 방식인 몬테카를로 적분은 베이지안 분석에서 요구되는 사후확률 분포함수에 대한 점근적 추론으로도 이용될 수 있다. 실제 추출과정에서 독립적인 표본을 생성해낸다는 것이 쉽지 않다. 이때 마코브 연쇄를 발생시켜 몬테카를로 적분을 적용하는데 이를 마코브 연쇄 몬테 카를로(Markov Chain Monte Carlo) 기법이라 하며 식 (2.3)는 여전히 유효하게 된다.

MCEM 알고리즘을 정리하면 다음 <표 2.4>과 같다.

<표 2.4> MCEM 알고리즘

$\widehat{\theta_{(m)}}$ 가 m번째 단계에서의 추정 값을 나타낸다고 하자. (m+1)번째 단계의 추정 값을 계산하기 위해 다음의 E-step과 M-step을 반복 시행한다.

E-step : $Q(\theta|\widehat{\theta_{(m)}}, \boldsymbol{x}) = \frac{1}{k}\sum_{i=1}^{k} \log L^{c}(\theta|\boldsymbol{x}, \boldsymbol{z})$ 를 계산 (k : 반복횟수)

M-step : $\widehat{\theta_{(m+1)}} = Arg\max Q(\theta|\widehat{\theta_{(m)}}, \boldsymbol{x})$

제 3장 Metropolis Hastings Expectation Maximization 알고리즘

앞에서 설명한 기존의 알고리즘들은 모두 분포의 가정이 필요하거나 분포를 알고 있어야 한다는 특징을 가지고 있다. 본 논문에서는 확률분포가 알려지지 않을 때 모수 추정이 가능하도록 하는 방법으로 메트로폴리스 헤스팅스(Metropolis Hastings, MH) 알고리즘을 이용하여 새롭게 변형된 EM 알고리즘을 제안한다. MH 알고리즘은 관심의 대상이 되는 확률분포가 주어지지 않거나 가정이 되지 않았을 때, 직접 난수를 생성할 수 없으므로, 확률분포가 극한분포를 갖는 마코브 연쇄로부터 난수를 발생시키는 알고리즘이다.

3.1 Markov Chain

임의의 집합 A가 어떤 연속적인 확률변수의 집합 $A = \{X_0, X_1, \cdots\}$일 때 매 시간 $t \geq 0$을 만족하고 다음 단계(state)인 X_{t+1}은 단지 바로 그 이전 단계 X_t에만 의존할 뿐 나머지 $X_0, X_1, \cdots, X_{t-1}$에 대해서는 의존하지 않는 어떤 분포 $P(X_{t+1}|X_t)$에서 추출된다고 할 때 이 확률변수 열을 마코브 연쇄라 하고 다음과 같은 식 (3.1)이 성립한다.

$$P(X_{t+1}|X_0, X_1, \cdots, X_t) = P(X_{t+1}|X_t) \tag{3.1}$$

마코브 연쇄는 초기 상태인 X_0에 영향을 받지 않으며, 몇 가지 조건을 만족하면 $P(\bullet\ |X_0)$는 안정적인 분포함수 π으로 수렴하는 것으로 알려져 있다. 실제로 마코브 연쇄를 발생시켜 임의의 함수 $u(X)$에 대해 기댓값 $E(u(X))$를 추정할 때에는 n번의 반복 중에서 m번의 버림(burn-in)을 생각하여 그 추정치를 에르고딕 평균(ergodic average)인 다음 식 (3.2)와 같이 계산할 수 있다.

$$\hat{E}(u(X)) = \frac{1}{n-m} \sum_{t=m+1}^{n} f(X^{(t)}) \tag{3.2}$$

3.2 Metropolis Hastings 알고리즘

Metropolis 외(1953)에 의해서 제안된 메트로 폴리스(Metropolis, M) 알고리즘(Metropolis et al., 1953)은 확률과정의 마코브 연쇄를 이용한 샘플링 방법 중 한가지로써, 물리학에서의 입자들의 평형상태(equilibrium) 분포를 생성하기 위한 방법으로 제안된 것으로 주로 확률과정으로 설명되는 물리통계에서 자주 사용되고 있다.

시간 t에 따라 변화하는 현상을 확률과정 $\{S_t, t=1,2,\cdots\}$로 나타낼 때, 하나의 단계 마코브 연쇄에 대하여 생각해보자. 시점 t 에서의 상태변수 S_t가 $\{1,2,\cdots,m\}$ 중의 한 값을 가지며, 전이확률(transition probability)이 시점 t와는 무관하게 $P_{ij} = P(S_{t+1} = j | S_t = i)$로 주어져 있다. 전이확률 행렬(transition probability matrix)이 $P = (P_{ij})_{m \times m}$ 로 주어져 있을 때, S_{t+1}의 분포는 S_t의 분포와 조건부 분포인 전치행렬(transpose matrix) P에 의하여 결정된다. 다시 표현하면,

$$P(S_{t+1} = j) = \sum_{i=1}^{m} P(S_{t+1} = j | S_t = i) \bullet P(S_t = i) \tag{3.3}$$

가 되며, 만약 마코브 연쇄가 진행되는 과정 중에

$$\boldsymbol{\pi} \bullet P = \boldsymbol{\pi} \tag{3.4}$$

를 만족하는 분포 $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_m)$가 존재하면 이를 평형분포(equilibrium distribution)라고 한다. 일단 평형상태에 들어가면 분포는 전이확률 행렬에 영향을 받지 않고 평형상태가 계속되기 때문이다. 즉, 임의의 자연수 n에 대하여, $\boldsymbol{\pi} \bullet P^n = \boldsymbol{\pi}$이다. 주어진 전치행렬 P를 사용하여 평형분포 $\boldsymbol{\pi}$를 구하는 것이 일반적인 확률과정의 문제이지만, M 알고리즘은 반대로 평형분포 $\boldsymbol{\pi}$가 주어져 있을 때 이 분포에 따르는 값을 생성하기 위하여 전치행렬 P를 구하여 사용하는 반복 알고리즘이다.

모든 상태들 사이에 다음 식 (3.5)와 같은 등식이 성립할 때, Reversible 마코브 연쇄라고 말한다.

$$\pi_i \bullet P_{ij} = \pi_j \bullet P_{ji}, i \neq j \tag{3.5}$$

다음 식 (3.6)과 같은 등식에서 Reversible(가역)이란 의미를 알 수 있다.

$$\begin{aligned} P(S_t = i, S_{t+1} = j) &= P(S_t = i)(S_{t+1} = j | S_t = i) \\ &= P(S_t = j)(S_{t+1} = i | S_t = j) \\ &= P(S_t = j, S_{t+1} = i) \end{aligned} \tag{3.6}$$

마코브 연쇄가 가역적이면,

$$\boldsymbol{\pi} \cdot P = \left(\sum_{i=1}^{m} \pi_i \cdot P_{ij} \right)_{1 \times m} = \left(\sum_{i=1}^{m} \pi_j \cdot P_{ji} \right) = \left(\pi_j \cdot \sum_{i=1}^{m} P_{ji} \right) = (\pi_j)_{1 \times m} = \boldsymbol{\pi} \tag{3.7}$$

가 되어 평형상태를 만족한다. M 알고리즘은 평형분포 $\boldsymbol{\pi}$에 따르는 값을 생성하기 위하여 $\pi_i \cdot P_{ij} = \pi_j \cdot P_{ji}, i \neq j$,를 만족하는 전치행렬을 수행하는 반복 알고리즘이다. 상태변수가 이산형인 경우와 연속형인 경우로 나누어 M 알고리즘에 대하여 좀 더 상세히 알아보자. 우선 이산형인 경우에 대한 M 알고리즘의 개념은 다음과 같다.

임의의 대칭인 전이행렬(symmetric transition matrix) $Q = (Q_{ij})$가 주어져 있다고 하자. 전치행렬은 시점 t와 무관하므로 임의의 시점 t에서 생각해도 된다. $S_t = i$인 경우, 주어진 전치행렬 Q에 의하여 상태 j가 선택되었을 때, 확률에 의하여 S_{t+1} 값이 상태 i에 머무를 수도 있고, 상태 j로 움직일 수도 있도록 선택이 주어지는 마코브 연쇄를 정의한다. 이때 상태 i에서 다른 상태 j로 움직일 확률을 α_{ij}라고 하면,

$$P_{ij} = \alpha_{ij} \cdot Q_{ij} \quad , \quad \text{for } i \neq j \quad , \quad P_{ii} = 1 - \sum_{i \neq j} P_{ij} \tag{3.8}$$

로 전치행렬 P가 구성된다. $\boldsymbol{\pi} \cdot P = \boldsymbol{\pi}$에서 살펴보면, 임의의 대칭전치행렬 Q에 의하여 상태 j가 선택되었으므로, $\pi_i < \pi_j$ 라면 상태 j로 움직이도록 하는 것은 당연하며, $\pi_i \geq \pi_j$ 인 경우는 상태 i가 상태 j보다 머무는 확률이 크므로 머무는(움직이는 것을 거부할) 확률로 $1 - \alpha_{ij}$를 주는 것이 합리적이다. 따라서 M 알고리즘도 거절법(Rejection Method, J. von Neuman, 1951)의 형태를 가지고 있는 셈이 된다. 위의 생각으로 Metropolis 외(1953)는 다음 식 (3.9)와 같이 α_{ij}를 정의하였다.

$$\alpha_{ij} = \begin{cases} 1, & \pi_j > \pi_i \\ \dfrac{\pi_j}{\pi_i}, & \pi_j \leq \pi_i \end{cases} \tag{3.9}$$

대칭전치행렬 Q와 α_{ij}의 정의를 이용하면,

$$\begin{aligned} \pi_i \bullet P_{ij} &= \pi_i \bullet \alpha_{ij} \bullet Q_{ij} = \pi_i \bullet \min\left\{1, \frac{\pi_j}{\pi_i}\right\} \bullet Q_{ij} = \min\{\pi_i, \pi_j\} \bullet Q_{ij} \\ &= \min\{\pi_i, \pi_j\} \bullet Q_{ji} = \pi_j \bullet \min\left\{\frac{\pi_i}{\pi_j}, 1\right\} \bullet Q_{ji} \\ &= \pi_j \bullet P_{ji} \end{aligned} \tag{3.10}$$

가 되어 전치행렬 P로 이루어진 마코브 연쇄는 가역적이다. 따라서 $\boldsymbol{\pi} \bullet P = \boldsymbol{\pi}$가 성립되며 마코브 연쇄를 수행하면서 분포 π에 따르는 값을 생성할 수 있는 것이다.

상태변수가 연속형인 경우에도 마찬가지의 방법으로 하면 된다. 상태변수 X의 확률밀도함수가 $\pi(\boldsymbol{x})$로 주어져 있으며, $\pi(\boldsymbol{x})$를 따르는 값들을 생성하는 것이 목적이다.

요약하면, M 알고리즘은

1. 대칭인 전치행렬을 임의로 하나 선정한다.
2. 가역성을 만족하기 위한 적당한 α_{ij}를 정의한다.
3. 주어진 분포 π를 사용하여 α_{ij}들을 구한다.
4. $\boldsymbol{\pi} \bullet P = \boldsymbol{\pi}$를 평형상태를 만든 후, 마코브 연쇄를 수행하면서 주어진 분포 $\boldsymbol{\pi}$에 따르는 값을 생성하는 기법이다.

메트로 폴리스 헤스팅스(Metropolis Hastings, MH, Hastings, W. 1970)알고리즘은 Hastings(1970)가 M 알고리즘에서 대칭인 경우뿐만 아니라 비대칭인 경우도 고려하여 M 알고리즘을 개선한 알고리즘 이다.

MH 알고리즘을 정리하면 다음 <표 3.1>과 같다.

<표 3.1> MH 알고리즘

$x^{(t)}$가 주어졌을 때,

Step 1 : $Y_t \sim T(\boldsymbol{y}|x^{(t)})$를 생성한다. (T : proposal distribution)

Step 2 :

$$X^{(t+1)} = \begin{cases} Y_t & with\ probability\ \rho(x^{(t)}, Y_t) \\ x^{(t)} & with\ probability\ 1-\rho(x^{(t)}, Y_t) \end{cases} \quad where\ \rho(\boldsymbol{x}, \boldsymbol{y}) = \min\left\{ \frac{f(\boldsymbol{y})}{f(\boldsymbol{x})} \frac{T(\boldsymbol{x}|\boldsymbol{y})}{T(\boldsymbol{y}|\boldsymbol{x})}, 1 \right\}$$

을 실행한다.

3.3 Metropolis Hastings Expectation Maximization 알고리즘

Metropolis Hastings Expectation Maximization(MHEM) 알고리즘은 본 논문에서 제안하는 알고리즘으로 기존의 EM 알고리즘과 MH 알고리즘을 결합하여 적용한 방법이다. MHEM 알고리즘은 EM 알고리즘에서 E-step의 기댓값 계산에 필요한 적분과정을 MH 알고리즘으로 해결하여 결측 자료를 구하고 이로부터 M-step을 통하여 모수를 추정하는 방법이다. 기존의 알고리즘들과 MHEM 알고리즘의 가장 큰 차이점은 정형화된 분포의 가정이 필요 없다는 장점을 가지고 있다는 것이다. 반면에 MHEM 알고리즘이 기존의 알고리즘들에 비해 모수 추정치의 정확도가 떨어진다는 단점을 가지고 있다. 하지만 현실에서의 대부분의 불완전 자료들은 어떤 정형화된 분포를 따르지 않는 자료들이 더 많이 존재한다. 이럴 경우 기존의 알고리즘들을 이용하기 쉽지 않지만 MHEM 알고리즘을 이용하면 쉽게 모수를 추정할 수 있다.

MHEM 알고리즘을 정리하면 다음 <표 3.2>와 같다.

<표 3.2> MHEM 알고리즘

$\widehat{\theta_{(m)}}$ 가 m번째 단계에서의 추정값을 나타낸다고 하자. (m+1)번째 단계의 추정값을 계산하기 위해 다음의 E-step과 M-step을 반복 시행한다.

E-step : MH(Metropolis Hastings) 알고리즘을 이용하여 $Q(\theta|\widehat{\theta_{(m)}}, \boldsymbol{x})$ 를 구한다.

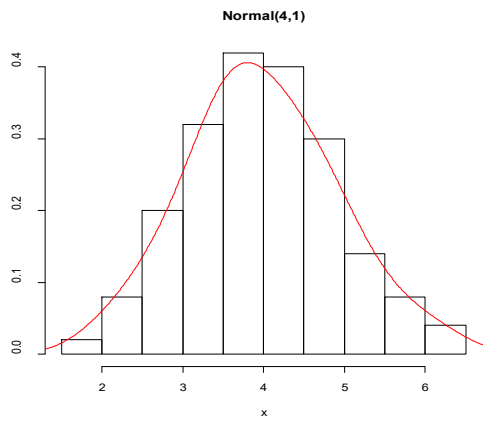
M-step : $\widehat{\theta_{(m+1)}} = Arg\max Q(\theta|\widehat{\theta_{(m)}}, \boldsymbol{x})$

제 4장 모의실험

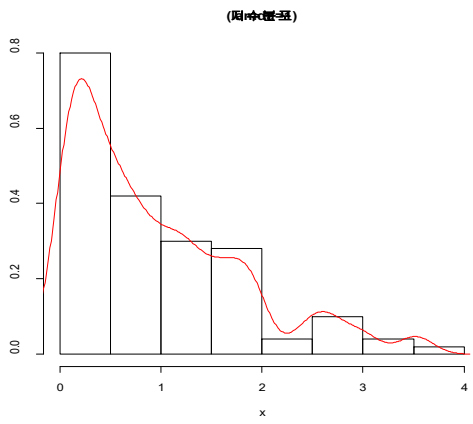
4장에서는 2장과 3장에서 설명한 알고리즘들을 좀 더 쉽게 비교 분석하기 위해 각 알고리즘들을 이용하여 모의실험을 진행하였다. 모의실험을 위한 통계 패키지로는 R-software(2.10.1)를 사용하였다.

4.1 자료 소개

본 논문의 모의실험에서는 평균이 4이고 분산이 1인 정규분포와 λ가 1인 지수분포를 따르는 자료를 각각 100개씩 생성하였다. 생성한 자료의 그래프는 <그림 4.1>과 <그림 4.2>와 같다.



<그림 4.1> 정규분포 자료의 그래프



<그림 4.2> 지수분포 자료의 그래프

이렇게 생성한 자료 중 정규분포는 4.5를, 지수분포는 1.5를 기준으로 오른쪽으로 절단하여 오른쪽으로 중도 절단된 자료(censored data)를 설정하였다. 즉, 정규분포에서 4.5보다 작거나 같은 값은 관측된 자료로 설정하고, 4.5보다 큰 경우는 결측 자료로 설정하여 관측된 자료를 바탕으로 결측된 자료를 포함한 완전자료의 평균을 추정하기 위해 설정하였다. 지수분포 또한 정규분포와 똑같은 방식으로 1.5를 기준으로 관측된 자료와 결측된 자료로 구분하여 설정하였다.

생성된 자료들의 평균과 관측된 자료들로 설정된 자료의 평균은 <표 4.1>과 같다.

<표 4.1> 모의실험 자료

	완전한 자료의 평균	관측된 자료의 평균
정규분포	3.9740	3.5346
지수분포	0.9556	0.5568

<표 4.1>과 같이 결측된 자료를 제외하고 관측된 자료를 가지고 평균을 추정하면 실제 값과 많은 차이가 있다. 그래서 관측된 자료만 가지고 추정한 평균을 이용하면, 많은 문제를 발생시킬 수 있다. 4장의 모의실험에서는 앞에서 설명한 알고리즘들을 이용하여 결측된 자료를 보완하여 완전한 자료의 평균을 추정한다.

4.2 정규분포

4.1절에서 설명한 정규분포를 따르는 자료를 바탕으로 각각의 알고리즘들을 이용하여 모의실험을 진행하였다.

4.2.1 정규분포-EM 알고리즘

EM 알고리즘에 적용하기 위해서는 E-step의 $Q(\theta|\widehat{\theta_{(m)}}, \boldsymbol{x})$ 을 계산해야 한다. 이 함수의 계산과정은 다음과 같다.

Y를 평균이 θ이고 분산이 1인 정규분포를 따르는 확률변수라고 가정하자. 그러면 완전한 자료의 로그 우도함수는 식 (4.1)과 같다.

$$Y_i \sim N(\theta, 1) \quad < Y : \text{관측치} >$$

$$
\begin{aligned}
L^c(\theta|\boldsymbol{y},\boldsymbol{z}) &= \prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{(y_i-\theta)^2}{2}}\cdot\prod_{i=m+1}^{n}\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{(z_i-\theta)^2}{2}} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n\cdot\prod_{i=1}^{m}e^{-\frac{(y_i-\theta)^2}{2}}\cdot\prod_{i=m+1}^{n}e^{-\frac{(z_i-\theta)^2}{2}} \\
&\propto \prod_{i=1}^{m}e^{-\frac{(y_i-\theta)^2}{2}}\cdot\prod_{i=m+1}^{n}e^{-\frac{(z_i-\theta)^2}{2}}
\end{aligned}
\tag{4.1}
$$

$\boldsymbol{z}$를 결측치라 하고, $\boldsymbol{z}$가 정규분포를 따른다고 가정하자. 그러면 관측된 자료가 주어졌을 때 관측되지 않은 자료의 조건부 p.d.f는 식 (4.2)와 같다.

$$
\begin{aligned}
&\boldsymbol{z} = (z_{n-m+1},\cdots,z_n) \qquad <\boldsymbol{z}: \text{결측치}> \\
&\boldsymbol{z} \sim k(\boldsymbol{z}|\theta,\boldsymbol{y}) = \frac{1}{(2\pi)^{(n-m)/2}}\cdot e^{\sum_{i=m+1}^{n}\frac{(z_i-\theta)^2}{2}}
\end{aligned}
\tag{4.2}
$$

완전한 자료의 로그우도함수는 식 (4.3)과 같다.

$$
\ell = Log\,L^c(\theta|\boldsymbol{y},\boldsymbol{z}) = -\sum_{i=1}^{m}\frac{(y_i-\theta)^2}{2}-\sum_{i=m+1}^{n}\frac{(z_i-\theta)^2}{2} \tag{4.3}
$$

식 (4.3)을 이용하여 완전한 자료의 로그우도함수 $\log L^c(\theta)$의 조건부 기댓값은 다음 식 (4.4)와 같다.

$$
E(\ell) = -\sum_{i=1}^{m}\frac{(y_i-\theta)^2}{2}-\frac{1}{2}\cdot\sum_{i=m+1}^{n}E\left[(z_i-\theta)^2\right] \tag{4.4}
$$

따라서 식 (4.4)를 최대화 시키는 모수 θ를 찾기 위해 θ에 관하여 미분하면 식 (4.5)와 같이 전개된다.

$$
\frac{\partial E(\ell)}{\partial\theta} = \sum_{i=1}^{m}(y_i-\theta)-\frac{1}{2}\cdot\sum_{i=m+1}^{n}\left[\frac{\partial}{\partial\theta}\int(z_i-\theta)^2\cdot\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{(z_i-\theta)^2}{2}}dz_i\right]
$$

$$= m \cdot \bar{y} - m \cdot \theta - \frac{1}{2} \cdot \sum_{i=m+1}^{n} \left[\int \frac{1}{\sqrt{2\pi}} \cdot \frac{\partial}{\partial \theta} (z_i - \theta)^2 \cdot e^{-\frac{(z_i-\theta)^2}{2}} dz_i \right]$$

$$-2(z_i - \theta) \cdot e^{-\frac{(z_i-\theta)^2}{2}} + (z_i - \theta)^2 \cdot (z_i - \theta) \cdot e^{-\frac{(z_i-\theta)^2}{2}}$$

$$= m \cdot \bar{y} - m \cdot \theta - \frac{1}{2} \cdot \sum_{i=m+1}^{n} \left[-2 \cdot E(z_i) + 2 \cdot \theta + E\left[(z_i - \theta)^3 \right] \right]$$

$$= m \cdot \bar{y} - m \cdot \theta - \frac{1}{2} \cdot (n-m) \cdot (-2 \cdot E(z_1) + 2 \cdot \theta) = 0 \qquad (4.5)$$

θ에 관하여 전개하면 다음 식 (4.8)과 같은 값을 얻는다. 식 (4.5)를 전개하면 식 (4.6)과 같이 전개된다.

$$m \cdot \bar{y} - m \cdot \theta - (n-m) \cdot E(z_1) - (n-m) \cdot \theta = 0 \qquad (4.6)$$

식 (4.6)에서 가로를 전개하여 좌변과 우변으로 정리하면 다음 식 (4.7)과 같다.

$$m \cdot \bar{y} - (n-m) \cdot E(z_1) = n \cdot \theta \qquad (4.7)$$

식 (4.7)을 θ에 관하여 정리하면 식 (4.8) 과 같이 나타난다.

$$\hat{\theta} = \frac{m \cdot \bar{y} + (n-m) \cdot E(z_1)}{n} \quad , where \quad E(z_1) = \frac{\phi(a-\hat{\theta})}{1 - \Phi(a-\hat{\theta})} \qquad (4.8)$$

식 (4.8)에서의 $\phi(a-\hat{\theta})$ 과 $\Phi(a-\hat{\theta})$은 다음 식 (4.9)와 식 (4.10)과 같다.

$$\phi(a-\hat{\theta}) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(a-\hat{\theta})^2}{2}} \qquad (4.9)$$

$$\Phi(a-\hat{\theta}) = \frac{1}{2}\left(1 + erf\left(\frac{a-\hat{\theta}}{\sqrt{2}}\right)\right) \qquad (4.10)$$

0

여기서 $erf(x)$는 오차함수(error function)로 식 (4.11)과 같다.

$$erf(x) = \frac{2}{\sqrt{\pi}} \cdot \int_0^x e^{-t^2} dt \ : \text{오차함수} \tag{4.11}$$

위와 같은 과정에 의해서 E-step의 $Q(\theta|\widehat{\theta_{(m)}}, \boldsymbol{x})$ 을 구할 수 있다. 이 후 M-step에 의해서 최대화 과정을 수행하면 EM 알고리즘에 의해서 정규분포의 형태를 따르는 불완전한 자료의 모수를 추정할 수 있다.

4.2.2 정규분포-SEM, MCEM, MHEM 알고리즘

앞에서 설명한 다른 알고리즘들은 앞에서 살펴본 EM 알고리즘의 계산과정 중 $E(z_1)$을 각각 몬테카를로 방법이나 MH 알고리즘 등으로 대체하여 적용하면 된다. 예를 들어 MCEM 알고리즘인 경우 $E(z_1)$을 몬테카를로 방법에의 해서 정규분포에서 충분히 많은 샘플을 추출하여, 추출한 값들의 평균으로 대체한다. 이렇게 생성된 평균값을 $E(z_1)$에 적용하면 MCEM 알고리즘을 이용하여 불완전한 자료의 모수를 추정할 수 있다.

4.3 지수분포

4.1절에서 설명한 지수분포를 따르는 자료를 바탕으로 각각의 알고리즘들을 이용하여 모의실험을 진행하였다.

4.3.1 지수분포-EM 알고리즘

EM 알고리즘에 적용하기 위해서는 E-step의 $Q(\theta|\widehat{\theta_{(m)}}, \boldsymbol{x})$ 을 계산해야 한다. 이 함수의 계산과정은 다음과 같다.

W를 평균 θ인 지수분포를 따르는 확률변수라고 가정하자. 그러면 확률밀도함수는 식 (4.12)와 같다.

$$f(w;\theta) = \theta^{-1} \cdot \exp(-w/\theta) \cdot I_{(0,\infty)}(w) \ , \ \theta > 0 \tag{4.12}$$

식 (4.12)에서 지수함수는 $I_{(0,\infty)}(w) = \begin{cases} 1 & , \ w > 0 \\ 0 & , \ otherwise \end{cases}$ 이다. 따라서 분포함수는 다음 식 (4.13)과 같이 주어진다.

$$F(w;\theta) = \{1 - \exp(-w/\theta)\} \bullet I_{(0,\infty)}(w) \tag{4.13}$$

$\boldsymbol{y} = \left(y_1^T, \cdots\cdots, y_n^T\right)^T$는 관측된 자료를 나타내고, 여기서 $y_j = \left(c_j, \delta_j\right)^T$로 두면 관측치 w_j가 중도절단 될 경우 $\delta_j = 0$, w_j가 관측된 경우 $\delta_j = 1$로 한다. 즉, 만약 관측치 w_j가 절단되지 않았을 경우 실제적인 값은 $w_j = c_j$, $j = 1, \cdots\cdots, n$ 이고, 반면에 만약 c_j에서 중도절단 되었다면 $w_j > c_j$, $j = 1, \cdots\cdots, n$가 된다.

r개의 중도 절단되지 않는 관측치를 $w_1, \cdots\cdots, w_r$이고, n-r개의 중도 절단된 결측치를 $w_{r+1}, \cdots\cdots, w_n$이라 하자.

완전한 자료 벡터 $\boldsymbol{x}$를 다음 식 (4.14)와 같이 나타낸다.

$$\boldsymbol{x} = \left(w_1, \cdots\cdots, w_n\right)^T = \left(w_1, \cdots\cdots, w_r, z^T\right)^T \tag{4.14}$$

여기서 $z^T = \left(w_{r+1}, \cdots\cdots, w_n\right)^T$은 n-r개 중도 절단된 자료를 포함하고 있다.

완전한 자료의 로그우도함수는 다음 식 (4.15)와 같다.

$$\log L^c(\theta) = \sum_{j=1}^{n} \log g_c\left(w_j;\theta\right) = -n \bullet \log\theta - \theta^{-1} \sum_{j=1}^{n} w_j \tag{4.15}$$

$L^c(\theta)$는 관측되지 않은 자료 $w_{r+1}, \cdots\cdots, w_n$에서 선형으로 보일 수 있고, E-step에서 $Q(\theta|\widehat{\theta_{(k)}}, \boldsymbol{y})$ 함수는 관측된 자료 $\boldsymbol{y}, \theta_0$에 대한 현재의 고정 값 $\theta_{(k)}$가 주어진 조건부 기대치로 구할 수 있다. 지수분포의 무기억성(lack of memory)에 의해서, $W_j > c_j$ 조건하에서 $W_j - c_j$의 조건부 분포는 평균이 θ인 지수분포를 따른다. $W_j(> c_j)$의 조건부 확률밀도함수는 다음 식 (4.16)과 같다.

$$f(\boldsymbol{w}, \theta) = \theta^{-1} \bullet \exp\{-(w_j - c_j)/\theta\} \bullet I_{(c,\infty)}(w_j) \ , \ \theta > 0 \tag{4.16}$$

식 (4.16)으로부터 기댓값은 식 (4.17)과 같이 구해진다.

$$E_\theta^{(k)}(W_j|\boldsymbol{y}) = E_\theta^{(k)}(W_j|\, W_j > c_j) = c_j + E_\theta^{(k)}(W_j) = c_j + \theta^{(k)}\ ,\ j = r+1, \cdots, n \quad (4.17)$$

식 (4.17)을 이용한 완전한 자료의 로그우도함수 $\log L^c(\theta)$의 조건부 기댓값은 다음 식 (4.18)과 같다.

$$\begin{aligned} Q(\theta|\widehat{\theta_{(k)}}, \boldsymbol{y}) &= E\left[-n \bullet \log\theta - \theta^{-1}\left\{\sum_{j=1}^{n} w_j\right\}\right] \\ &= E\left[-n \bullet \log\theta - \theta^{-1}\left\{\sum_{j=1}^{r} c_j + \sum_{j=r+1}^{n} (w_j|w_j > c_j)\right\}\right] \\ &= -n \bullet \log\theta - \theta^{-1}\left\{\sum_{j=1}^{r} c_j + \sum_{j=r+1}^{n} E(w_j|w_j > c_j)\right\} \\ &= -n \bullet \log\theta - \theta^{-1}\left\{\sum_{j=1}^{r} c_j + \sum_{j=r+1}^{n} (c_j + E(w_j))\right\} \\ &= -n \bullet \log\theta - \theta^{-1}\left\{\sum_{j=1}^{r} c_j + (n-r)\theta^{(k)}\right\} \end{aligned} \quad (4.18)$$

따라서 식 (4.18)의 Q함수를 최대화 시키는 모수 θ를 찾기 위해 θ에 관하여 미분하면 다음 식 (4.19)와 같이 된다.

$$\begin{aligned} \frac{\partial Q(\theta|\widehat{\theta_{(k)}}, \boldsymbol{y})}{\partial\theta}\Big|_{\hat{\theta}} &= \frac{\partial}{\partial\theta}\left[-n \bullet \log\theta - \theta^{-1}\left\{\sum_{j=1}^{r} c_j + \sum_{j=r+1}^{n} \left(c_j + \theta^{(k)}\right)\right\}\right] \\ &= \frac{\partial}{\partial\theta}\left[-n \bullet \log\theta - \theta^{-1}\left\{\sum_{j=1}^{r} c_j + (n-r)\theta^{(k)}\right\}\right] \\ &= \frac{-n}{\theta} + \frac{1}{\theta^2}\left\{\sum_{j=1}^{n} c_j + (n-r)\theta^{(k)}\right\} = 0 \end{aligned} \quad (4.19)$$

식 (4.19)를 θ에 관하여 전개하면 아래 식 (4.20)과 같은 값을 얻는다.

$$\begin{aligned}\theta^{(k+1)} &= \left\{\sum_{j=1}^{r} c_j + \sum_{j=r+1}^{n} E_\theta^{(k)}(W_j|y)\right\}/n \\ &= \left\{\sum_{j=1}^{r} c_j + \sum_{j=r+1}^{n} \left(c_j + \theta^{(k)}\right)\right\}/n \\ &= \left\{\sum_{j=1}^{r} c_j + (n-r)\theta^{(k)}\right\}/n\end{aligned} \tag{4.20}$$

위와 같은 과정에 의해서 E-step의 $Q(\theta|\widehat{\theta_{(m)}}, \boldsymbol{x})$ 을 구할 수 있다. 이 후 M-step에 의해서 최대화 과정을 수행하면 EM 알고리즘에 의해서 지수분포의 형태를 따르는 불완전한 자료의 모수를 추정할 수 있다.

4.3.2 지수분포-SEM, MCEM, MHEM 알고리즘

앞에서 설명한 다른 알고리즘들은 앞에서 살펴본 EM 알고리즘의 계산과정 중 $\sum_{j=1}^{r} c_j$ 를 각각 몬테카를로 방법이나 MH 알고리즘 등으로 대체하여 적용하면 된다. 예를 들어 MCEM 알고리즘인 경우 $\sum_{j=1}^{r} c_j$을 몬테카를로 방법에의 해서 지수분포에서 충분히 많은 샘플을 추출하여, 추출한 값들의 평균으로 대체한다. 이렇게 생성된 평균값을 $\sum_{j=1}^{r} c_j$에 적용하면 MCEM 알고리즘을 이용하여 불완전한 자료의 모수를 추정할 수 있다.

4.4 모의실험 결과

모의실험에서 각 알고리즘들에 대한 θ의 초기치를 정규분포는 0으로 하고, 지수분포는 0.5로 하여 알고리즘들마다 각각 10000번씩 반복을 진행 하였다. 그 결과 다음 <표 4.2>와 같은 모의 실험한 결과가 나왔다.

<표 4.2> 모의실험 결과

		정규분포	지수분포
real value		3.9740	0.9556
EM	평균	3.9831	1.0305
	표준편차	0.0126	0.0017
MCEM	평균	3.9832	1.0304
	표준편차	0.0192	0.0257
SEM	평균	3.9832	1.0306
	표준편차	0.0016	0.0034
MHEM	평균	3.8278	0.8901
	표준편차	0.0121	0.0020

모의실험의 결과 자료들의 실제 평균값과 EM, SEM, MHEM 알고리즘들을 이용하여 추정한 평균이 결측된 자료를 제외하고 관측된 자료를 가지고 평균을 추정한 값보다 더욱 좋은 결과를 보여준다는 것을 알 수 있다. 물론 각각의 알고리즘들마다 장·단점이 다르기 때문에 모의실험 결과에는 약간의 차이가 있다. 그 중 MHEM 알고리즘을 사용한 결과가 다른 알고리즘들을 사용한 결과보다 정확도가 떨어지는 것을 볼 수 있다. 하지만 대체적으로 실제 평균값에 가깝게 추정되는 것을 확인 할 수 있다. 또한 MHEM 알고리즘은 본 모의실험처럼 정형화된 분포가 아닌 정형화되지 않은 분포일 때, 다른 알고리즘들과 달리 쉽게 사용할 수 있다는 장점을 가지고 있기 때문에 추정치의 정확도가 약간 떨어지는 것을 상쇄할 수 있다.

제 5장 실증분석

5.1 배 경

우리나라의 금융기관과 기업은 1997년에서 1998년 사이에 IMF 위기를 겪으면서 유동성 부족으로 인하여 '위험관리능력결핍'을 지적 받게 되었다. 또한 오늘날 국가 간 자본이동의 속도가 가속화되고 그 방법 또한 다양화되어 가면서 낙후된 국내 금융 시장의 위험 관리 시스템 대신에 체계적이고 유동적인 위험 관리 시스템이 요청되고 있는 실정이다. 이와 같은 현실은 우리에게 위험 관리의 중요성을 일깨워 주었고 이전의 전통적인 위험 관리 기법에서 벗어나 변화하는 금융 환경 속에서 시장 위험을 측정하고, 예상되는 손실 가능성을 제공하는 새로운 위험관리기법을 요구하게 되었다. 이러한 요구에 부응하여 최근 많이 사용하는 위험관리기법이 Value-at-Risk(VaR) 기법이다.

VaR 분석기법이란 주어진 신뢰수준에서 포트폴리오의 목표보유기간동안 기대되는 최대손실, 즉 향후 불리한 시장가격변동이 특정 신뢰 구간 내에서 발생하는 경우 입을 수 있는 포트폴리오의 최대 손실 규모를 산출하는 기법을 말한다.

이러한 VaR 분석기법을 조금 더 적극적으로 활용하기 위해서 최근까지의 자료들을 관측된 자료들로 정하고, 미래의 자료를 결측자료로 가정하여, 중도 절단된 자료로 설정한다. 이와 같이 설정하여 본 논문에서는 MHEM 알고리즘을 통해서 미래의 자료를 포함한 VaR을 추정하여 위험 관리에서 조금 더 능동적으로 대처하고자 한다.

5.2 자료 소개

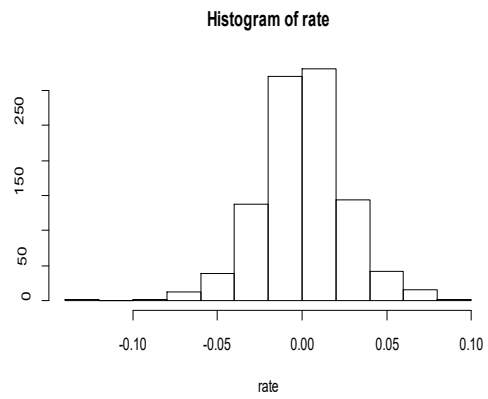
본 논문의 실증분석에 사용되는 자료는 증권거래소에서 제공하는 1999년 1월 4일부터 2003년 4월 7일까지의 KOSPI 200의 자료와 2005년 1월 3일부터 2010년 12월 30일까지의 KOSPI 200의 자료이다. 이와 같이 구간을 설정한 이유는 외환위기 이후의 KOSPI 200 자료들을 바탕으로 MHEM 알고리즘을 이용하여 VaR을 추정하여 MHEM 알고리즘으로 추정한 VaR이 정확하게 모형을 반영하고 있는지 모의실험 하여 살펴본 후, 최근 자료들을 이용하여 실증분석을 진행하였다. 또한 1999년 1월 4일부터 2003년 4월 7일까지의 KOSPI 200자료를 설정한 이유는 외환 위기에는 금융 경색, 리스크 프리미엄의 폭등 등으로 인하여 VaR 추정 결과에 대하여 명확하고 확실한 의미를 부여할 수 없다는 점에서 외환위기 이후의 자료를 사용하였으며, 그 구간 동안 신용카드 와 신용대출의 무분별한 사용으로 경제적 위기로 인하여 자료의 변동 폭이 컸으며, 그 구간을 MHEM 알고리즘을 이용하여 VaR 추정 결과를 잘 보여준다면, 정확하게 모형을 반영하고 있다고 판단하였다.

VaR 측정 시 고려해야 할 중요 요소인 보유기간은 1일 종가 지수를 기준으로 분석

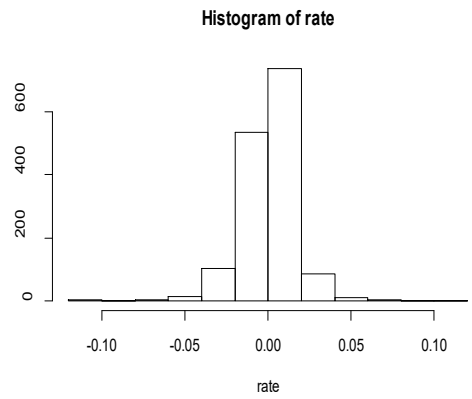
하였다. 사실 보유 기간은 필요에 따라 달라질 수 있는데 하루 단위로 선정하는 이유는 변동성이 관찰되면서 BIS에서 요구하는 것이 2주 VaR임에도 불구하고 대부분의 금융회사들은 내부위험제어의 목적으로 하루 동안 손실을 막을 수 있는 VaR을 적용하기 때문이다. 실제 VaR을 추정할 때 분석의 대상인 주가 지수의 일별 수익률은 연속 복리 수익률(continuously compounded returns)로 이는 다음의 식 (5.1)과 같다.

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \tag{5.1}$$

실무적으로 가격보다는 수익률이 보다 더 통계적으로 의미를 갖게 되며, 또한 주어진 가격에 대한 상대적인 변화를 측정하기 위하여 절대 수익률($D_t = P_t - P_{t-1}$)보다는 상대수익률($R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$)과 연속 복리 수익률이 선호된다. 본 논문에서 연속 복리 수익률을 사용하는 이유는 보유기간이 1일 초과 시에 연속 복리 수익률의 경우 단순히 일일 연속 복리 수익률을 합하면 되는 단순함이 있기 때문이다.



<그림 5.1> 수익률(1999~2003) 자료의 그래프



<그림 5.2> 수익률(2005~2010) 자료의 그래프

<그림 5.1>과 <그림 5.2>는 1999년 1월 4일부터 2003년 4월 7일과 2005년 1월 3일부터 2010년 12월 30일까지의 KOSPI 200 수익률의 히스토그램이다. KOSPI 200 수익률의 히스토그램을 보여주는 <그림 5.1>과 <그림 5.2>를 살펴보면 수익률 분포의 평균이 거의 0에 가깝고, 좌우비대칭인 모양을 나타낸다. 이는 주가수익률의 자료가 정규

분포와 유사하지만, 실제로 주가수익률의 분포는 극치 부분이 정규분포보다 두껍게 나타난 결과로 보인다. 이 사실을 좀 더 체계적으로 살펴보기 위해 <표 5.1>과 같은 통계량들을 살펴보았다.

<표 5.1> KOSPI 200 수익률의 기초 통계량

기 간	평균	표준편차	왜도	첨도	Shapiro-Wilk normality test	P-value
99.01.04~03.04.07	9.02×10^{-5}	0.025	-2.2×10^{-1}	4.54	0.9856	1.215×10^{-8}
05.01.03~10.12.30	0.000573	0.0156	-1.3×10^{-1}	2.13	0.9267	2.2×10^{-16}
표준정규분포	0	1	0	3	-	-

<표 5.1>에 KOSPI 200 수익률의 실제 기간에 따른 평균, 표준편차, 왜도, 첨도, Shapiro-Wilk 통계량과 정규분포 검정에 대한 유의 수준 P값이 정규 분포의 기초 통계 값과 비교되어 정리되어있다. 평균을 통해서는 KOSPI 200 주가지수의 수익률이 0을 중심으로 하는 분포인가를 보게 되고, 표준편차를 통해서는 데이터의 기간에 따른 변동성의 변화를 보고자한다. 왜도는 평균 근방의 비대칭 정도를 나타내는 값으로 정규분포를 따른다면 0의 값을 가지게 될 것이다. 첨도는 정규분포에 대비한 상대적인 고도(peakness)와 편평도(flatness)를 측정하는 것으로 정규분포의 첨도 값 3보다 상대적으로 높은 값을 갖는다는 것은 꼬리가 두터운 분포임을 의미하고 낮은 값을 갖는다면 정규분포에 비하여 좁은 영역에 분포되어 있음을 의미하는 것이다.

<표 5.1>에서의 통계량들을 보면 알 수 있듯이 실제 자료들이 왜도가 0이 아니고 첨도가 3이 아닌 것을 알 수 있다. 이것은 실제 자료들이 정규분포를 따르지 않는 다는 것을 보여준다. 그렇다면 실제 자료들의 분포가 정규분포를 따르는가에 대한 더 엄밀한 검증을 위해서 Shapiro-Wilk normality test를 이용하여 검정하였다. Shapiro-Wilk normality test는 귀무가설을 '자료들이 정규분포를 따른다.'하고 대립가설을 '자료들이 정규분포를 따르지 않는다.'라고 하고 test하는 것이다. <표 5.1>을 살펴보면 Shapiro-Wilk normaility test에 대한 P-value가 모두 매우 작으므로 귀무가설을 기각한다. 즉, 실제 자료들은 정규분포를 따르지 않는다는 결과를 보여주고 있다.

이와 같이 실제 자료들은 대부분 우리가 알고 있는 특정한 분포를 따르고 있는 것보다는 대부분 우리들이 모르는 분포를 따르고 있다. MHEM 알고리즘은 정형화된 분포를 따르지 않는 자료들을 이용하여 모수를 추정할 수 있다는 장점을 가지고 있다. 본 실증분석에서는 이러한 MHEM 알고리즘의 장점을 이용하여 실증분석을 진행하였다.

5.3 Kernel function

5.1절에서 실제 자료들은 정규분포가 아닌 다른 분포를 따르는 자료들이라는 것을 알 수 있었다. 그러면 실제 자료들은 어떠한 분포를 따르는지를 알아보기 위해 비모수적 분석 방법에서 많이 사용되고 있는 커널 분석 방법론(Kernel analysis methods)을 이용하여 실제 자료들이 어떠한 분포를 따르고 있는지 알아보도록 하겠다.

커널 분석 방법론은 아래 식 (5.2)을 만족시키는 연속밀도함수인 커널함수(kernel function) $K(\bullet)$가 있다고 가정한다.

$$\int_{-\infty}^{\infty} K(\phi)\, d\phi = 1 \tag{5.2}$$

이때 커널함수의 추정치는 식 (5.3)와 같게 된다.

$$\hat{f}(x) = \frac{1}{nh} \bullet \sum_{i=1}^{n} K\left(\frac{\chi_i - x}{h}\right) = \frac{1}{nh} \bullet \sum_{i=1}^{n} K(\phi_i) \;\; where \;\; \phi_i = \frac{\chi_i - x}{h} \tag{5.3}$$

여기서 h는 대역값(window width, bandwidth)이고, n은 표본의 크기이다. 또한 분포를 도출할 때 h와 $K(\bullet)$의 선택이 필요한데 일반적으로 사용되는 기준은 아래 식 (5.4)의 MISE(the Mean Intergrated Squared Error)이다.

$$MISE = E\left(\int \left[\hat{f}(x) - f(x)\right]^2 dx\right) = \int \left[Bias(\hat{f})^2 + V(\hat{f})\right] dx \tag{5.4}$$

그러나 MISE를 얻을 수 없으므로 MISE의 근사치를 이용하는데, 편의(Bias)와 분산(Variance)의 근사치를 이용하여 AMISE를 얻을 수 있다. 정확한 $\hat{f}$의 편의와 분산은 다음 식(5.5), (5.6)와 같다.

$$Bias(\hat{f}) = E(\hat{f}) - f = \int K(\phi) \bullet \left[f(h\phi + x) - f(x)\right] d\phi \tag{5.5}$$

$$V(\hat{f}) = \frac{1}{nh} \bullet \int K^2(\phi) \bullet f(h\phi + x) d\phi - \frac{1}{n} \bullet \left[\int K(\phi) \bullet f(h\phi + x) d\phi\right]^2$$

(5.6)

그리고 위의 식 (5.5)와 (5.6)를 Taylor 급수로 전개하여 정리하면 각각의 근사치를 구할 수 있다. 이렇게 구한 근사치를 식 (5.4)에 대입하여 AMISE를 구할 수 있다.

$$AMISE = \frac{1}{4} \cdot \lambda_1 \cdot h^4 + \lambda_2 \cdot \frac{1}{h} \ , \ where \ \lambda_1 = \mu_2^2 \cdot \int (f^{(2)}(x))^2 dx \qquad (5.7)$$
$$\lambda_2 = \int K^2(\phi) d\phi$$

여기서 μ_2는 $\mu_2 = \int \phi^2 \cdot K(\phi) d\phi$ 이고, $f^{(2)}(x)$는 $f(x)$의 2차 도함수 이다.

최적의 h는 편의와 분산의 상충관계를 잘 조절해 줄 수 있는 값이어야 하고, 이것이 의미하는 것은 AMISE를 최소화하는 h를 말한다. AMISE를 h에 대해 미분하여 그 식을 0으로 놓고 방정식을 풀면 AMISE를 최소화하는 h를 다음의 식 (5.8)과 같이 구할 수 있다.

$$h = cn^{-\frac{1}{5}}, \ where \ c = \left(\frac{\lambda_2}{\lambda_1}\right)^{\frac{1}{5}} \qquad (5.8)$$

그리고 MISE는 대역값 h뿐만 아니라, $\int K^2(\phi) d\phi$를 통해 커널함수의 선택에 의해서도 영향을 받는다.

<표 5.2> 대표적인 커널 함수

종 류	식	효율성
Epanechnikov	$\frac{3}{4}\left(1-\frac{1}{5}\phi^2\right)/\sqrt{5}$ for $\lvert\phi\rvert < \sqrt{5}$ 0 *otherwise*	1
Biweight	$\frac{15}{16}(1-\phi^2)^2$ for $\lvert\phi\rvert < 1$ 0 *otherwise*	0.9939
Triangular	$1-\lvert\phi\rvert$ for $\lvert\phi\rvert < 1$ 0 *otherwise*	0.9859
Gaussian	$\frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}\phi^2\right]$	0.9512
Rectangular	$\frac{1}{2}$ for $\lvert\phi\rvert < 1$ 0 *otherwise*	0.9295

자료원 : Silverman(1986), p43

<표 5.2>에서 제시된 커널함수에서 최적 커널 Epanechnikov 커널 함수와 다른 커널 함수들을 이용하여 MISE를 비교해 보면 효율성에 큰 차이가 없음을 알 수 있

다.(Silverman., 1986, p43) 본 논문에서는 2차 미분이 가능한 확률밀도함수인 Gaussian 커널 함수를 이용하여 분포를 추정하였다.

Gaussian 커널 함수를 선택하면 최적의 h는 다음의 식 (5.9)과 같다.

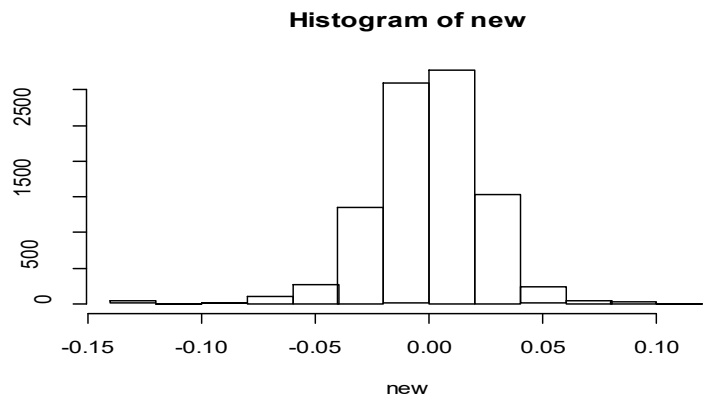
$$h = cn^{-\frac{1}{5}}, \quad where \quad c = 1.06 \cdot \hat{\sigma} \tag{5.9}$$

이것을 바탕으로 분포를 추정하면,

$$\begin{aligned}\hat{f}(x) &= \frac{1}{nh} \cdot \sum_{i=1}^{n} K\left(\frac{\chi_i - x}{h}\right), \ where\ h = 1.06 \cdot \hat{\sigma} \cdot n^{-\frac{1}{5}}\ ,\ \chi_i : \text{실제 자료} \\ &= \frac{1}{n \cdot 1.06 \cdot \hat{\sigma} \cdot n^{-\frac{1}{5}}} \cdot \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}\left(\frac{\chi_i - x}{1.06 \cdot \hat{\sigma} \cdot n^{-\frac{1}{5}}}\right)^2\right] \\ &= \frac{1}{\sqrt{2\pi} \cdot 1.06 \cdot \hat{\sigma} \cdot n^{\frac{4}{5}}} \cdot \sum_{i=1}^{n} \exp\left[-\frac{1}{2}\left(\frac{\chi_i - x}{1.06 \cdot \hat{\sigma} \cdot n^{-\frac{1}{5}}}\right)^2\right]\end{aligned} \tag{5.10}$$

식 (5.10)과 같은 분포를 추정할 수 있다.

5.4 분포의 검증



<그림 5.3> 수익률(1999~2003) 생성 자료의 그래프

5.2에서 커널 함수를 이용하여 추정한 분포가 실제 자료를 잘 반영하고 있는지 확인해 보았다. <그림 5.3>은 추정한 분포를 이용하여 수익률 자료를 생성한 것의 히스토그램이다. 그림에서 보면 알 수 있듯이 5.1절의 <그림 5.1>과 유사한 형태를 보여 주고 있다. 조금 더 정확히 살펴보기 위하여 <표 5.3>과 같이 실제 수익률 자료의 통계량과 추정한 분포를 이용하여 생성한 수익률 자료의 통계량을 비교하였다. 표를 보면 알 수 있듯이 커널 함수를 이용하여 추정한 분포가 실제 수익률 분포를 잘 반영하고 있는 것을 알 수 있다.

<표 5.3> 수익률 생성 자료(2005.01.03~2010.12.30) 통계량

기 간	평 균	표준편차
1999.01.04~2003.04.07	9.020963×10^{-5}	0.0250011
커널함수로 추정한 분포로 생성한 자료	8.678718×10^{-5}	0.02417639

5.5 모의실험(1999.01.04~2003.04.07)

1999년 1월 4일부터 2003년 4월 7일까지의 자료를 가지고 1999년 1월 4일부터 2001년 11월 21일까지의 자료는 관측된 자료하고, 2001년 11월 22일부터 2003년 4월 7일까지의 자료는 관측되지 않은 자료로 설정하고, 관측된 자료들을 바탕으로 관측되지 않은 자료들을 MHEM 알고리즘을 이용하여 추정하였다.

<표 5.4> 모의실험 결과(1999.01.04~2003.04.07)

기 간	평 균
1999.01.04~2001.11.21	0.0001911282
2001.11.22~2003.04.07	−0.0001230409
추정결과(2001.11.22.~2003.04.07.)	−0.0001552036
1999.01.04~2003.04.07	9.020963×10^{-5}
추정결과(1999.01.04.~2003.04.07.)	7.987821×10^{-5}

<표 5.4>를 보면 알 수 있듯이 모의 실험한 결과가 실제 수익률 자료를 잘 반영 하고 있다고 할 수 있다. 이 기간 동안 우리나라 경제는 신용카드 와 신용대출의 무분별한 사용으로 경제적 위기로 인하여 주가에 대한 수익률의 변동 폭이 컸으며, 그 구간을 MHEM 알고리즘을 이용하여 VaR의 추정 결과를 잘 보여주고 있다. 또한 수치상으로는 매우 작은 값이지만 KOSPI 200의 자본금이 매우 큰 금액이므로 이것을 고려하였을 때 수익률 값은 의미를 가진다고 생각하며, 약간의 오차가 있지만 잘 반영하였다고 볼 수 있다.

5.6 실증분석(2005.01.03~2010.12.30)

지금까지의 분포추정, 모의실험 등을 바탕으로 정형화되지 않고, 추정한 분포를 사용하여 MHEM 알고리즘을 이용하여 추정한 결과 값이 실제 수익률 자료를 잘 추정하고 있다는 것을 알 수 있었다. 이것들을 바탕으로 최근 수익률 자료를 바탕으로 관측되지 않은 앞으로의 수익률을 예측했다. 2005년부터 2010년까지의 KOSPI 200의 수익률 자료를 가지고 2011년의 수익률을 예측하였다. 즉, 관측된 자료인 2005~2010년 자료를 가지고 관측되지 않은 2011년 자료의 평균 수익률을 예측 하였다.

<표 5.5> 실증분석 결과(2005.01.03~2010.12.30)

기 간	평 균
2005.01.03.~2010.12.30	0.0005725131
MH 알고리즘을 이용한 추정 결과	0.0001994397

<표 5.5>는 실증분석 결과이다. 2005~2010년까지의 KOSPI 200의 평균 수익률은 0.0005725131이다. 이것을 바탕으로 추정한 2011년의 KOSPI 200의 평균 수익률은 0.0001994397로 추정하였다. 2005년부터 2010년까지의 우리나라 경제의 변동성은 매우 높았다. 2008년 미국 발 세계 금융위기로 인해서 변동성이 확대 되었다. 이와 같은 이유 때문에 주가의 수익률 또한 매우 높은 변동성을 보였다. 이 구간 동안의 주가의 수익률을 반영한 커널함수는 앞으로의 수익률을 추정할 때 좋은 결과를 반영할 것이라고 생각한다. 위 <표 5.5>는 이러한 결과를 반영하여 추정된 결과이다. 하지만 우리나라 주식 시장은 급격한 경제 성장 및 주식 시장의 짧은 역사 그리고 해외 자금 의존도 등으로 인하여 해외 경제의 상황에 따라 급변하는 특징을 가지고 있다. 예를 들어 우리나라 기업은 안정적인 구조를 가지고 있음에도 해외에서 커다란 이슈 및 투자 심리를 위축하는 사건이 발생한다면, 해외 자금의 투자금 회수 및 외국인 투자자들의 투자 심리 위축 등으로 많은 영향을 받아서 주가의 수익률에 큰 영향을 미친다. 이러한 특성 때문에 외부에서 강한 충격이 온다면 본 논문에서 예측한 KOSPI 200의 수익률과 다른 방향으로 진행 될 수 있다. 그러므로 본 논문에서 예측한 KOSPI 200의 수익률을 절대적인 지표로 삼기보다는 주가의 수익률을 예측함에 있어서 참고사항으로 하여 포트폴리오를 구성하면 더 좋을 것이다.

제 6장 결 론

불완전한 자료에 대하여 모수를 추정하고자 할 때 많은 방법들이 있지만, 일반적으로 반복적인 방법에 의해 최우추정량을 구하는 EM 알고리즘이 많이 사용된다. 하지만 EM 알고리즘은 결측된 자료가 조건부로 들어간 우도함수의 기댓값을 구하는 E-step은 고차의 적분을 수반하는 경우가 많으므로 계산이 용이하지 않은 문제점이 발생하게 된다. 이러한 문제점을 보완하기 위해서 MCEM 알고리즘이나 SEM 알고리즘 등이 개발되었다. 그러나 EM, MCEM, SEM 알고리즘 등은 각각의 장·단점을 가지고 있지만, 모두 정형화된 분포의 가정이 필요한 알고리즘이다. 그래서 본 논문에서는 정형화된 분포의 가정이 필요하지 않은 MH 알고리즘과 EM 알고리즘을 결합하여 적용한 MHEM 알고리즘을 제안하였다.

MHEM 알고리즘은 기존의 EM, MCEM, SEM 알고리즘 보다 모수 추정치의 정확도가 약간 떨어진다는 단점을 가지고 있지만, 정형화된 분포의 가정이 없는 경우 EM, MCEM, SEM 알고리즘은 사용할 수 없지만 MHEM 알고리즘을 사용하여 불완전 자료에 대해 모수를 추정할 수 있다는 장점이 단점을 보완하고 있다. 현대사회에서는 대부분의 실제 자료들이 특정 정형화된 분포를 따르고 있다기보다는 정형화되지 않은 분포를 따르고 있다. 이러한 현대사회에서 MHEM 알고리즘은 정형화되지 않은 분포를 따르는 불완전한 자료라도 알고리즘을 사용하여 모수를 추정할 수 있다는 장점을 가지고 있다.

본 논문에서는 오른쪽 중도 절단된 자료를 생성하여 모의실험을 통해 모수를 추정하였다. 이것을 위한 방법으로 EM, SEM, MCEM, MHEM 알고리즘을 사용하였다. 그 결과 MHEM 알고리즘을 이용하여 모수를 추정한 결과의 정확도가 약간 떨어졌지만 대체적으로 모수를 잘 추정하였다. 모의실험은 정형화된 분포를 가정하고 그 분포에서 자료를 생성하였기 때문에 MHEM 알고리즘의 효율성이 조금은 떨어진 것으로 보인다. 하지만 모의실험에서처럼 정형화된 분포가 아니라 정형화되지 않은 분포를 따르는 자료에서는 다른 알고리즘들 사용할 수 없지만, MHEM 알고리즘은 쉽게 사용할 수 있다는 효율성이 있다. 따라서 본 논문에서는 정형화되지 않은 분포를 가지고 있는 오른쪽 중도 절단된 자료에 대해서 MHEM 알고리즘을 이용하여 실증분석을 하였다. 그 결과 본 논문에서 제안한 MHEM 알고리즘의 효율성을 잘 보여주고 있다.

참고문헌

[1] Celeux, G. , Diebolt, J. (1985). The SEM algorithm : A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, *Comp. Statist. Quart., 2, 73-82.*

[2] Dempster, A. P. , Laird, N. M. and Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm, *Journal of the Royal Statistical Society, B 39. 1-38.*

[3] Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika, 57 : 97-109.*

[4] Ip, E. H. S (1994a) A stochastic EM estimator in the presence of missing data theory and applications. *Technical report, Department of Statictics, Stanford University.*

[5] J. von Neuman(1951), Various Techniques used in connection with Random digits, *National Bureau of Standards, Applied Math. Series, vol. 12, pp 36*

[6] Robert, C. P. , Casella, G. (2004). *Monte Carlo Statistical Methods-second edition* : Springer.

[7] Madras, N (2002). *Lectures on Monte Carlo methods* : American Mathematical Society.

[8] Metropolis, N. , Rosenbluth, A. W. , Rosenbluth, M. N. , Teller, A. H. and Teller, E. (1953). Equations of state Calculation by fast Computing Machines, *J. Chem. Physics, Vol. 21, pp. 1087.*

[9] 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.*

[10] 강만기 (2000). Weibull 분포에서 MEM 알고리즘에 의한 모수 추정, *Journal of the Korean Data Analysis Society, Vol. 2, No. 2, June 2000, pp. 229.*

[11] 김승구 (2003). 자기공명영상분할에서 바이어스 필드 보정을 위한 재귀적 EM 알고리즘, *Journal of the Korean Data Society, Vol. 5, no. e, June 2003, pp. 323-336.*

[12] 김승구 (2004). 정규혼합모형의 대용량자료 적합을 위한 일반화 Incremental EM 알고리즘에 대한 연구, *Jounal of the Korean Data Analysis Society, Vol. 6, No. 4, pp. 1031-1041.*

[13] 김승구 (2005). 인자분석자 혼합모형을 위한 Incremental EM 알고리즘, *Journal of*

the Korean Data Analysis Society, Vol. 7, No. 5, pp. 1605-1614.

[14] 김행선 (2003). 위험관리수단으로서 VaR(Value at Risk)의 추정 방법의 비교 및 분석 : 서강대학교 *대학원 석사학위 논문.*

[15] 남준우 (2000). 비모수커널추정법에 의한 확률밀도함수의 추정, *계량경제학보 11권 4호 pp. 105~121*

[16] 박찬호 (1997). 중도절단된 지수분포의 모수추정에서 효율적인 MCEM 알고리즘의 이용 : *동국대학교 대학원 석사학위 논문.*

[17] 손건태 (2005). *전산통계개론 제 4판* : 자유아카데미

[18] 이혁종 (1998). EM 알고리즘과 그 실증분석 : *고려대학교 대학원 석사학위 논문.*