응용통계연구 (2012)
25(1), 183–196
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.1.183

불완전 자료에 대한 Metropolis-Hastings Expectation Maximization 알고리즘 연구

전수영[1] · 이희찬[2]

[1]고려대학교 정보통계학과, [2]고려대학교 경제통계학과

(2011년 11월 18일 접수, 2011년 12월 2일 수정, 2011년 12월 2일 채택)

요 약

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, Monte Carlo Expectation Maximization, Stochastic Expectation Maximization 알고리즘 등을 이용하는 방법이 있지만, 정형화된 분포의 가정이 필요하다는 단점을 가지고 있다. 본 연구에서는 정형화된 분포의 가정이 없는 경우에 사용할 수 있는 Metropolis-Hastings Expectation Maximization(MHEM) 알고리즘을 제안하고자 한다. MHEM 알고리즘의 효율성은 중도절단자료(censored data)를 이용한 모의실험과 KOSPI 200 수익률의 실증자료분석를 통해 알수 있었다.

주요용어: 불완전한 자료, Expectation Maximization, Monte Carlo Expectation Maximization, Stochastic Expectation Maximization, Metropolis-Hastings Expectation Maximization.

1. 서론

현대 사회에서의 통계는 우리 생활과 아주 밀접한 관계를 가지고 있기 때문에, 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 일이어서 통계 전반에 걸쳐 결측치와 불완전 자료들에 관한 많은 문제들이 존재한다.

통계 분석 대상이 되는 자료에 결측치가 있을 경우, 이러한 불완전한 자료를 가지고 분석을 진행하게 되면 많은 문제점을 갖게 된다. 이러한 문제를 해결하는 방법으로 결측치를 가진 자료를 삭제하는 방법으로부터 여러 가지 다른 값을 가지고 결측치를 대체하는 방법 등 여러 방법 등을 고려할 수 있는데, 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 모의실험 방법이 제안될 수 있으며, 그 중 가장 널리 알려져 있는 대체 방법으

[1]교신저자: (339-700) 충남 연기군 조치원읍 세종로 2511, 고려대학교 과학기술대학 정보통계학과, 조교수.
E-mail: scheon@korea.ac.kr

표 **1.1.** EM 알고리즘

$\hat{\theta}_{(t)}$를 t번째 단계에서의 추정값이라 하면, $(t+1)$번째 단계의 추정값 $\hat{\theta}_{(t+1)}$은 다음의 E-step과 M-step을 반복 시행함으로써 얻어진다.

E-step: $Q\left(\theta|\hat{\theta}_{(t)}, \boldsymbol{x}\right) = E_{\hat{\theta}_{(t)}}\left[\log L^c(\theta|\boldsymbol{x}, \boldsymbol{z})|\hat{\theta}_{(t)}, \boldsymbol{x}\right]$를 계산한다.
여기서 기대값은 조건부 확률분포함수 $k\left(\boldsymbol{z}|\hat{\theta}_{(t)}, \boldsymbol{x}\right)$하에서 구해진다.

M-step: $\hat{\theta}_{(t+1)} = \text{Arg max } Q\left(\theta|\hat{\theta}_{(t)}, \boldsymbol{x}\right)$.

로 Expectation Maximization(EM) 알고리즘 (Dempster 등, 1977)이 있다. 예로 김승구 (2003, 2004, 2005)는 자기공명영상의 올바른 분할을 위해서 효과적인 바이어스 필드보정에 EM 알고리즘을 사용하였고, 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, EM 알고리즘을 보안한 Stochastic Expectation Maximization(SEM) 알고리즘 (Celeux와 Diebolt, 1985)과 Monte Carlo Expectation Maximization(MCEM) 알고리즘 (Wei와 Tanner, 1990) 등이 있다.

Dempster 등 (1977)에 의해 제안된 Expectation Maximization(EM) 알고리즘은 다양한 불완전한 자료(incomplete data)로부터 최대 우도추정치를 반복적인 기법을 통해 구할 수 있는 방법으로 위와 같은 문제점들을 다루는데 널리 사용되는 도구가 되었다. 이러한 반복적 알고리즘의 가장 큰 장점은 기존의 근사접근이나, 수치해석적인 방법에 비해 상대적으로 프로그래밍 하기 쉽고, 우도함수 또는 로그우도함수를 최대화시키는 최대 우도추정치로의 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 최대 우도추정치가 쉽게 계산되어지는 지수족에서의 완전자료인 경우, EM 알고리즘의 M-step의 계산은 마찬가지로 쉽게 계산되어진다.

먼저 $L(\theta|\boldsymbol{x})$는 관측된 자료 $\boldsymbol{x}$의 우도함수라 하고, $k(\boldsymbol{z}|\theta, \boldsymbol{x})$를 관측된 자료 $\boldsymbol{x}$가 주어졌을 때 관측되지 않은 자료 $\boldsymbol{z}$의 조건부 확률분포함수라 하자. 그리고 $L^c(\theta|\boldsymbol{x}, \boldsymbol{z})$는 완전한 자료 $(\boldsymbol{x}, \boldsymbol{z})$의 우도함수라 하자. EM 알고리즘은 다음 표 1.1과 같다.

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.

Stochastic Expectation Maximization(SEM; Celeux와 Diebolt, 1985) 알고리즘은 EM 알고리즘을 적용하기에 어렵고 복잡했던 많은 문제들, 특히 EM 알고리즘의 E-step에서 다차원의 수치적분을 포함하고 있어 해결이 쉽지 않은 경우에 대해 해결하기 위해 제안된 알고리즘이다. SEM 알고리즘은 각 반복 t에서 $\hat{\theta}_{(t)}$이 θ에 대한 현재 추정치인 경우, 표 1.1의 $k(\boldsymbol{z}|\hat{\theta}_{(t)}, \boldsymbol{x})$로부터 하나의 표본을 추출하여 결측치 $\boldsymbol{z}$를 채워 넣는 알고리즘이다.

$$k\left(\boldsymbol{z}|\hat{\theta}_{(t)}, \boldsymbol{x}\right) = \frac{h\left(\boldsymbol{x}, \boldsymbol{z}|\hat{\theta}_{(t)}\right)}{\int h\left(\boldsymbol{x}, \boldsymbol{z}'|\hat{\theta}_{(t)}\right) dz'} = \frac{h\left(\boldsymbol{x}, \boldsymbol{z}|\hat{\theta}_{(t)}\right)}{g\left(\boldsymbol{x}|\hat{\theta}_{(t)}\right)}, \tag{1.1}$$

여기서 $h(\boldsymbol{x}, \boldsymbol{z}|\theta)$는 관측된 자료와 관측되지 않은 자료의 결합 확률분포함수이다. 결측치 $\boldsymbol{z}$를 이런 방식으로 대체하면 현재 θ에 대해 가지고 있는 모든 정보를 포함하게 되어 적절한 의완전자료(pseudo

표 **1.2.** SEM 알고리즘

E-step: $k\left(\boldsymbol{z}|\hat{\theta}_{(t)},\boldsymbol{x}\right)$로부터 표본을 추출하여 결측치 $\hat{\boldsymbol{z}}$로 대체하여 의완전자료를 생성한다.

$$\theta = \frac{1}{(T-n_0)}\sum_{n=n_0+1}^{T}\theta^{(n)}.$$

여기서 $\theta^{(n)}$은 $L(\hat{\theta}_{(t)}|\boldsymbol{x},\hat{\boldsymbol{z}})$의 최대우도추정치이며, n_0는소각되는 초기 반복횟수이다.

M-step: $\hat{\theta}_{(t+1)} = \text{Arg max}\, Q\left(\theta|\hat{\theta}_{(t)},\boldsymbol{x}\right)$.

표 **1.3.** MCEM 알고리즘

E-step: $Q\left(\theta|\hat{\theta}_{(t)},\boldsymbol{x}\right) = \frac{1}{k}\sum_{i=1}^{k}\log L^c(\theta|\boldsymbol{x},\boldsymbol{z})$를 계산한다. 여기서 k는 반복횟수이다.

M-step: $\hat{\theta}_{(t+1)} = \text{Arg max}\, Q\left(\theta|\hat{\theta}_{(t)},\boldsymbol{x}\right)$.

complete data)를 갖게 된다. 일단 의완전자료를 갖게 되면 의완전자료 로그우도함수를 최대화함으로써 갱신되어지는 최대우도추정치 $\hat{\theta}_{(t+1)}$을 얻게 된다. 이러한 과정을 반복하면서 SEM 알고리즘이 구성된다 (표 1.2).

의완전자료를 제공하고 최대화를 시행하는 SEM 알고리즘은 가벼운 조건 (Ip, 1994)하에서 정상분포 $\boldsymbol{\pi}$로 수렴해가는 마코브 연쇄(Markov Chain) $\hat{\theta}_{(t)}$를 생성시키며, 그 정상분포 $\boldsymbol{\pi}$는 근사적으로 θ의 최대우도 추정치를 중앙에 두며, 알고리즘의 반복 안에서의 $\hat{\theta}_{(t)}$의 변화율에 의존하는 분산을 가지고 있다. 또한, SEM 알고리즘을 사용하는 대부분의 상황에서 $\hat{\theta}_{(t)}$의 수렴은 충분히 빠른 것으로 알려져 있다 (Celeux와 Diebolt, 1985).

Monte Carlo Expectation Maximization(MCEM) 알고리즘은 SEM 알고리즘처럼 EM 알고리즘의 단점인 E-step이 고차의 적분을 수반하게 되면 계산되지 않는 경우가 발생한다는 것을 보완하기 위해 Wei와 Tanner (1990)가 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 찾을 수 있는데 이것이 몬테카를로 방법이다. 몬테카를로 방법의 장점 중의 하나는 계산이 다른 수학적 방법에 비해 간단하다는 것을 들 수 있다 (표 1.3).

이제까지 살펴본 알고리즘들은 확률분포가 알려져 있지 않을때 이용하지 못하는 단점을 가지고 있다. 따라서 본 논문에서는 확률분포가 알려지지 않을 때 유용한 Metropolis-Hastings Expectation Maximization(MHEM) 알고리즘이라는 변형된 EM 알고리즘을 제안하고자 한다.

2장에서 본 논문이 제안하는 MHEM 알고리즘을 소개하고, 3장에서는 모의실험을 통해 각 알고리즘의 성능을 비교한다. 4장에서는 KOSPI 200 수익률 자료에 대해 MHEM 알고리즘을 이용한 실증분석을 살펴보고, 5장에서 본 논문의 결론을 정리한다.

표 **2.1.** MHEM 알고리즘

E-step : MH 알고리즘을 이용하여 $Q\left(\theta|\hat{\theta}_{(t)}, \boldsymbol{x}\right)$를 계산한다.

Step 1: $y \sim T\left(\cdot|\hat{\theta}_{(t)}\right)$를 생성한다 ($T$: proposal distribution).

Step 2: $\theta = \begin{cases} y, & \text{with probability } \rho\left(\hat{\theta}_{(t)}, y\right), \\ \hat{\theta}_{(t)}, & \text{with probability } 1-\rho\left(\hat{\theta}_{(t)}, y\right), \end{cases}$ where $\rho(\boldsymbol{x}, \boldsymbol{y}) = \min\left\{\frac{Q(\boldsymbol{y})}{Q(\boldsymbol{x})}\frac{T(\boldsymbol{x}|\boldsymbol{y})}{T(\boldsymbol{y}|\boldsymbol{x})}, 1\right\}$.

M-step: $\hat{\theta}_{(t+1)} = \text{Arg max } Q\left(\theta|\hat{\theta}_{(t)}, \boldsymbol{x}\right)$.

2. 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, 간접적으로 확률분포가 극한분포를 갖는 마코브 연쇄로부터 난수를 발생시켜 추론하는 알고리즘이다.

Metropolis 등 (1953)에 의해 제안된 메트로 폴리스(Metropolis) 알고리즘은 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. 메트로 폴리스-헤스팅스(Metropolis-Hastings; MH) 알고리즘 (Hastings, 1970)은 메트로 폴리스 알고리즘에서 대칭인 경우뿐만 아니라 비대칭인 경우도 고려하여 메트로 폴리스 알고리즘을 개선한 알고리즘이다.

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 2.1과 같다.

3. 모의실험

본 모의실험은 Robert와 Casella (2004, p.178)에 의해 제시된 예를 이용하여 정규분포를 따르는 자료를 생성시켜, 1장과 2장에서 설명한 각 알고리즘들을 비교 분석하고 MHEM 알고리즘의 효율성을 알아보았다. 모의실험을 위한 통계 패키지로는 R-software(version 2.10.1)를 사용하였다.

3.1. 자료소개

본 논문의 모의실험을 위해 평균이 4이고 분산이 1인 정규분포를 따르는 자료를 100개 생성하였다 (그림 3.1).

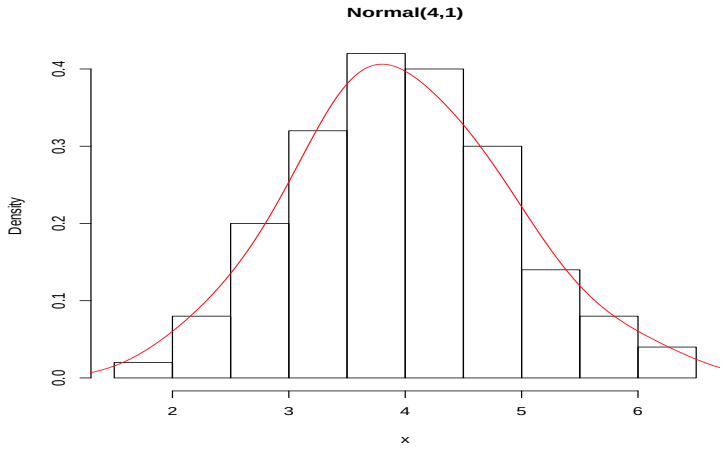

그림 3.1. 정규분포 자료의 그래프

표 3.1. 모의실험 자료

완전한 자료의 평균	관측된 자료의 평균
3.9740	3.5346

이렇게 생성한 자료 중 4.5를 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. 생성된 자료들의 평균과 관측된 자료들로 설정된 자료의 평균은 표 3.1과 같다.

표 3.1과 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. 3장의 모의실험에서는 앞에서 설명한 알고리즘들을 이용하여 결측된 자료를 보완하여 완전한 자료의 평균을 추정한다.

3.2. 알고리즘 비교

3.1절에서 설명한 정규분포를 따르는 자료를 바탕으로 각각의 알고리즘들을 이용하여 모의실험을 진행하였다.

3.2.1. EM 알고리즘 EM 알고리즘에 적용하기 위해서는 E-step의 $Q(\theta|\hat{\theta}_{(t)}, \boldsymbol{x})$을 계산해야 한다. 이 함수의 계산과정은 다음과 같다.

Y를 평균이 θ이고 분산이 1인 정규분포를 따르는 확률변수라고 가정하자. 이때 m개 자료만 관찰되고 $n-m$개의 자료는 결측되었다고 하자. 그러면 완전한 자료의 로그 우도함수는 식 (3.1)과 같다.

$$\ell = \mathrm{Log}\, L^c(\theta|\boldsymbol{y}, \boldsymbol{z}) = -\sum_{i=1}^{m} \frac{(y_i-\theta)^2}{2} - \sum_{i=m+1}^{n} \frac{(z_i-\theta)^2}{2}. \tag{3.1}$$

$\boldsymbol{z}$를 결측치라 하고, $\boldsymbol{z}$가 정규분포를 따른다고 가정하면 관측된 자료가 주어졌을 때 관측되지 않은 자료

표 **3.2.** 모의실험 결과

	완전 자료	관찰 자료	EM	MCEM	SEM	MHEM
평균	3.9740	3.5346	3.9831	3.9832	3.9832	3.8278
표준편차			0.0126	0.0192	0.0016	0.0121

$\boldsymbol{z}$의 조건부 확률분포함수는 식 (3.2)와 같다.

$$k(\boldsymbol{z}|\theta,\boldsymbol{y}) = \frac{1}{(2\pi)^{\left(n-\frac{m}{2}\right)}} \cdot e^{-\sum_{i=m+1}^{n}\frac{(z_i-\theta)^2}{2}}. \tag{3.2}$$

식 (3.1)을 이용한 완전한 자료의 로그우도함수 $\log L^c(\theta)$의 조건부 기댓값은 다음 식 (3.3)과 같다.

$$E(\ell) = -\sum_{i=1}^{m}\frac{(y_i-\theta)^2}{2} - \frac{1}{2}\cdot\sum_{i=m+1}^{n} E\left[(z_i-\theta)^2\right]. \tag{3.3}$$

따라서 식 (3.3)을 최대화 시키는 모수 θ를 찾기 위해 θ에 관하여 미분한뒤 정리하면 식 (3.4)와 같이 나타난다.

$$\hat{\theta} = \frac{m\cdot\bar{y} + (n-m)E(\theta^*)}{n}, \quad \text{where} \;\; E(\theta^*) = \frac{\phi\big(a-\hat{\theta}\big)}{1-\Phi\big(a-\hat{\theta}\big)}. \tag{3.4}$$

여기서 $\phi(\cdot)$와 $\Phi(\cdot)$는 각각 표준정규분포의 확률분포함수와 누적분포함수이다.

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

3.2.2. SEM, MCEM, MHEM 알고리즘 SEM, MCEM과 MHEM 알고리즘들은 앞절에서 살펴본 EM 알고리즘의 계산과정 중 $E(\theta^*)$을 각각 몬테카를로 방법이나 MH 알고리즘 등으로 대체하여 적용하면 된다. 예를 들어 MCEM 알고리즘인 경우 $E(\theta^*)$을 몬테카를로 방법에 의해서 정규분포에서 충분히 많은 샘플을 추출하여, 추출한 값들의 평균으로 대체한다. 이렇게 생성된 평균값을 $E(\theta^*)$에 적용하면 MCEM 알고리즘을 이용하여 불완전한 자료의 모수를 추정할 수 있다.

3.3. 모의실험 결과

모의실험에서 각 알고리즘들에 대한 θ의 초기치를 0으로 하여 알고리즘들마다 각각 반복을 10000번씩 진행 하였다.

모의실험 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 (표 3.2). 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.

4. 실증분석

4.1. 배경

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 VaR를 추정하여 위험 관리에서 조금 더 능동적으로 대처하고자 한다.

4.2. 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 VaR를 추정하여 MHEM 알고리즘으로 추정한 VaR가 정확하게 모형을 반영하고 있는지 보기 위함이다. 이를 위해 모의실험을 살펴보았고, 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 자료가 필요하였으며, 그 구간 동안 신용카드와 신용대출의 무분별한 사용으로 경제적 위기로 인하여 자료의 변동 폭이 커 그 구간을 MHEM 알고리즘을 이용하여 VaR 추정 결과를 잘 보여준다면, 정확하게 모형을 반영하고 있다고 판단하였기 때문이다.

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 BIS에서 요구하는 것이 2주 VaR임에도 불구하고 대부분의 금융회사들은 내부위험제어의 목적으로 하루 동안 손실을 막을 수 있는 VaR를 적용하기 때문이다. 실제 VaR를 추정할 때 분석의 대상인 주가 지수의 일별 수익률은 연속 복리 수익률(continuously compounded returns)로 이는 다음의 식 (4.1)과 같다.

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right). \tag{4.1}$$

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 ($R_t = (P_t - P_{t-1})/P_{t-1}$)과 연속 복리 수익률이 선호된다. 연속 복리 수익률을 사용하는 이유는 보유기간

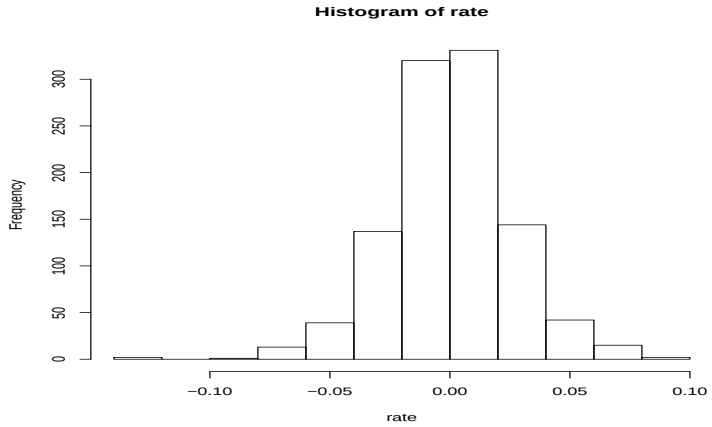

그림 4.1. 수익률(1999∼2003) 자료의 그래프

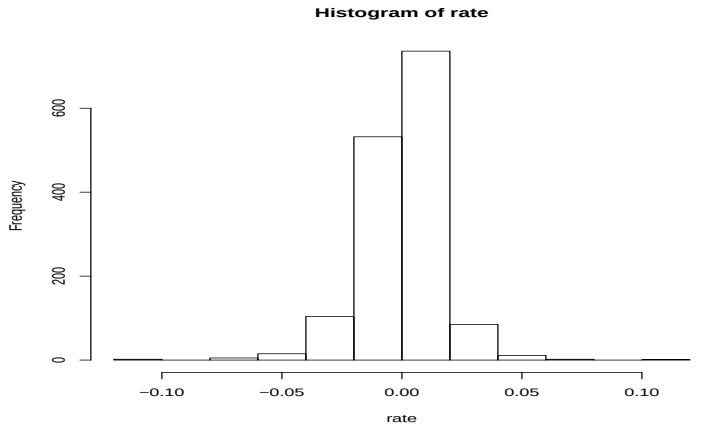

그림 4.2. 수익률(2005∼2010) 자료의 그래프

이 1일 초과 시에 연속 복리 수익률의 경우 단순히 일일 연속 복리 수익률을 합하면 되는 단순함이 있기 때문이다 (김행선, 2003).

그림 4.1과 그림 4.2는 1999년 1월 4일부터 2003년 4월 7일과 2005년 1월 3일부터 2010년 12월 30일까지의 KOSPI 200 수익률의 히스토그램이다. KOSPI 200 수익률의 히스토그램을 보여주는 그림 4.1과 그림 4.2를 살펴보면 수익률 분포의 평균이 거의 0에 가깝고, 좌우비대칭인 모양을 나타낸다. 이는 주가수익률의 자료가 정규분포와 유사하지만, 실제로 주가수익률의 분포는 극치 부분이 정규분포보다 두껍게 나타난 결과로 보인다. 이 사실을 좀 더 체계적으로 살펴보기 위해 표 4.1과 같은 통계량들을 보았다.

표 4.1에 KOSPI 200 수익률의 실제 기간에 따른 평균, 표준편차, 왜도, 첨도, SW(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

표 **4.1.** 모의실험 자료

기간	평균	표준편차	왜도	첨도	SW 통계량	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	–	–

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 4.1에서의 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. 또한 실제 자료들의 분포가 정규분포를 따르는가에 대한 더 엄밀한 검증을 위해서 Shapiro-Wilk 정규성 검정을 해 본 결과 실제 자료들은 정규분포를 따르지 않는다는 결과를 보여주고 있다.

이와 같이 실제 자료들은 특정한 분포를 따르지 않는 경우가 대부분이다. MHEM 알고리즘은 정형화된 분포를 따르지 않는 자료들을 이용하여 모수를 추정할 수 있다는 장점을 가지고 있다. 본 실증분석에서는 이러한 MHEM 알고리즘의 장점을 이용하여 실증분석을 진행하였다.

4.3. Kernel function

4.2절에서 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 방법(Kernel analysis method)을 이용하여 실제 자료들이 어떠한 분포를 따르고 있는지 알아보도록 하겠다.

커널 분석 방법은 아래 식 (4.2)를 만족시키는 연속밀도함수인 커널함수(kernel function) $K(\cdot)$가 있다고 가정한다.

$$\int_{-\infty}^{\infty} K(\phi)d\phi = 1. \tag{4.2}$$

이때 커널함수의 추정치는 식 (4.3)과 같게 된다.

$$\hat{f}(x) = \frac{1}{nh} \cdot \sum_{i=1}^{n} K\left(\frac{\chi_i - x}{h}\right) = \frac{1}{nh} \cdot \sum_{i=1}^{n} K(\phi_i), \tag{4.3}$$

여기서 χ_i는 실제자료, $\phi_i = (\chi_i - x)/h$, h는 대역값(window width, bandwidth), 그리고 n은 표본의 크기이다. 또한 분포를 도출할 때 h와 $K(\cdot)$의 선택이 필요한데 일반적으로 사용되는 기준은 아래 식 (4.4)의 MISE(Mean Intergrated Squared Error)이다.

$$\text{MISE} = E\left(\int \left[\hat{f}(x) - f(x)\right]^2 dx\right) = \int \left[\text{Bias}\left(\hat{f}\right)^2 + V\left(\hat{f}\right)\right] dx. \tag{4.4}$$

그러나 MISE를 얻기가 어려우므로 MISE의 근사치를 이용하는데, 편의(Bias)와 분산(Variance)의 근사치를 이용하여 AMISE를 얻을 수 있다. 정확한 $\hat{f}$의 편의와 분산은 다음 식 (4.5), (4.6)과 같다.

$$\text{Bias}\left(\hat{f}\right) = E\left(\hat{f}\right) - f = \int K(\phi) \cdot [f(h\phi + x) - f(x)]d\phi, \tag{4.5}$$

$$V\left(\hat{f}\right) = \frac{1}{nh} \cdot \int K^2(\phi) \cdot f(h\phi + x)d\phi - \frac{1}{n} \cdot \left[\int K(\phi) \cdot f(h\phi + x)d\phi\right]^2. \tag{4.6}$$

표 **4.2.** 대표적인 커널 함수 (Silverman, 1986, p.43)

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

그리고 위의 식 (4.5)와 (4.6)을 Taylor 급수로 전개하여 정리하면 각각의 근사치를 구할 수 있다. 이렇게 구한 근사치를 식 (4.4)에 대입하여 AMISE를 구할 수 있다.

$$\text{AMISE} = \frac{1}{4}\cdot\lambda_1\cdot h^4 + \lambda_2\cdot\frac{1}{h}, \tag{4.7}$$

여기서 $\lambda_1 = \mu_2^2\cdot\int(f^{(2)}(x))^2dx$, $\lambda_2 = \int K^2(\phi)d\phi$, $\mu_2 = \int\phi^2\cdot K(\phi)d\phi$이고, $f^{(2)}(x)$는 $f(x)$의 2차 도함수이다.

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 (4.8)과 같이 구할 수 있다.

$$h = \left(\frac{\lambda_2}{\lambda_1}\right)^{\frac{1}{5}}\cdot n^{-\frac{1}{5}}. \tag{4.8}$$

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

표 4.2에서 제시된 커널함수에서 최적 커널인 Epanechnikov 커널함수와 다른 커널함수들을 이용하여 MISE를 비교해 보면 효율성에 큰 차이가 없음을 알 수 있다 (Silverman, 1986, p.43). 따라서 본 논문에서는 2차 미분이 가능한 확률밀도함수인 Gaussian 커널 함수를 이용하여 분포를 추정하였다.

Gaussian 커널 함수를 선택하면 최적의 h는 다음의 식 (4.9)와 같다.

$$h = 1.06\cdot\hat{\sigma}\cdot n^{-\frac{1}{5}}. \tag{4.9}$$

이것을 바탕으로 다음과 같이 커널함수를 추정할 수 있다 (김행선, 2003).

$$\begin{aligned}\hat{f}(x) &= \frac{1}{nh}\cdot\sum_{i=1}^{n}K\left(\frac{\chi_i - x}{h}\right), \quad \text{where } \ h = 1.06\cdot\hat{\sigma}\cdot n^{-\frac{1}{5}}\\ &= \frac{1}{n\cdot 1.06\cdot\hat{\sigma}\cdot n^{-\frac{1}{5}}}\cdot\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\cdot\exp\left[-\frac{1}{2}\left(\frac{\chi_i - x}{1.06\cdot\hat{\sigma}\cdot n^{-\frac{1}{5}}}\right)^2\right]\\ &= \frac{1}{\sqrt{2\pi}\cdot 1.06\cdot\hat{\sigma}\cdot n^{\frac{4}{5}}}\cdot\sum_{i=1}^{n}\exp\left[-\frac{1}{2}\left(\frac{\chi_i - x}{1.06\cdot\hat{\sigma}\cdot n^{-\frac{1}{5}}}\right)^2\right].\end{aligned} \tag{4.10}$$

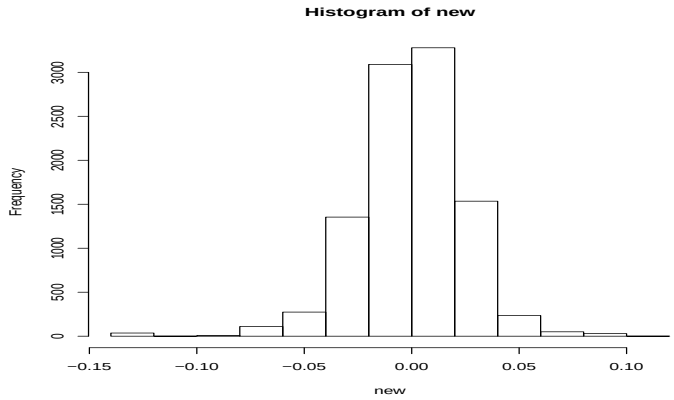

그림 4.3. 수익률(1999∼2003) 생성 자료의 그래프

표 4.3. 수익률 생성자료(2005.01.03∼2010.12.30) 통계량

기간	평균	표준편차
1999.01.04∼2003.04.07	9.020963×10^{-5}	0.0250011
커널함수로 추정한 분포로 생성한 자료	8.678718×10^{-5}	0.02417639

표 4.4. 모의실험 결과(1999.01.04∼2003.04.07)

기간	평균
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}

4.4. 분포의 검증

4.4절에서는 커널 함수를 이용하여 추정한 분포가 실제 자료를 잘 반영하고 있는지 확인해 보고자 한다. 그림 4.3은 추정한 분포를 이용하여 수익률 자료를 생성한 것의 히스토그램이다. 그림에서 보면 알 수 있듯이 4.1절의 그림 4.1과 유사한 형태를 보여 주고 있다. 조금 더 정확히 살펴보기 위하여 표 4.3과 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.

4.5. 모의실험(1999.01.04∼2003.05.07)

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

표 4.4를 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

표 **4.5.** 실증분석 결과(2005.01.03∼2010.12.30)

기간	평균
2005.01.03∼2010.12.30	0.0005725131
MHEM 알고리즘을 이용한 추정 결과	0.0001994397

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.

4.6. 실증분석(2005.01.03∼2010.12.30)

지금까지의 분포추정, 모의실험 등을 바탕으로 정형화되지 않은 추정된 분포에 MHEM 알고리즘을 이용하여 추정한 결과, 그 결과값이 실제 수익률 자료를 잘 추정하고 있다는 것을 알 수 있었다. 이것들을 바탕으로 본 실증분석에서는 최근 수익률 자료를 바탕으로 관측되지 않은 앞으로의 수익률을 예측해 보고자 한다. 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 4.5는 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 4.5는 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.

5. 결론

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-step에서 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 떨어진다

는 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. 이를 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 결과는 본 논문에서 제안한 MHEM 알고리즘의 효율성을 잘 보여주고 있다.

참고문헌

강만기 (2000). Weibull 분포에서 MEM 알고리즘에 의한 모수 추정, *Journal of the Korean Data Analysis Society*, **2**, 299–305.

김승구 (2003). 자기공명영상분할에서 바이어스 필드 보정을 위한 재귀적 EM 알고리즘, *Journal of the Korean Data Society*, **5**, 323–336.

김승구 (2004). 정규혼합모형의 대용량자료 적합을 위한 일반화 Incremental EM 알고리즘에 대한 연구, *Journal of the Korean Data Analysis Society*, **6**, 1031–1041.

김승구 (2005). 인자분석자 혼합모형을 위한 Incremental EM 알고리즘, *Journal of the Korean Data Analysis Society*, **7**, 1605–1614.

김행선 (2003). 위험관리수단으로서 VaR(Value at Risk)의 추정 방법의 비교 및 분석, 서강대학교 대학원 석사학위 논문.

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, *Computational Statistics*, **2**, 73–82.

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.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their application, *Biometrika*, **57**, 97–109.

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

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, Second edition, Springer.

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**, 1087–1091.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Champman and Hall.

Wei, G. C. G. and Tanner, M. A. (1990). A monte carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, *Journal of the American Statistical Association*, **85**, 699–714.

Metropolis-Hastings Expectation Maximization Algorithm for Incomplete Data

Sooyoung Cheon[1] · Heechan Lee[2]

[1]Department of Informational Statistics, Korea University
[2]Department of Economics and Statistics, Korea University

(Received November 18, 2011; Revised December 2, 2011; Accepted December 2, 2011)

Abstract

The inference for incomplete data such as missing data, truncated distribution and censored data is a phenomenon that occurs frequently in statistics. To solve this problem, Expectation Maximization(EM), Monte Carlo Expectation Maximization(MCEM) and Stochastic Expectation Maximization(SEM) algorithm have been used for a long time; however, they generally assume known distributions. In this paper, we propose the Metropolis-Hastings Expectation Maximization(MHEM) algorithm for unknown distributions. The performance of our proposed algorithm has been investigated on simulated and real dataset, KOSPI 200.

Keywords: Incomplete data, Expectation Maximization, Monte Carlo Expectation Maximization, Stochastic Expectation Maximization, Metropolis-Hastings Expectation Maximization.

[1]Corresponding author: Assistant Professor, Department of Informational Statistics, Korea University, Jochiwon 339-700, Korea. E-mail: scheon@korea.ac.kr