

몬테 카를로 expectation maximization 방법을 이용한 확률적 질병 확산 모형 추정 연구[†]

최보승¹ · 윤용화²

¹고려대학교 세종캠퍼스 국가통계전공 · ²대구대학교 전산통계학과

접수 2017년 12월 26일, 수정 2018년 1월 11일, 게재확정 2018년 1월 16일

요약

본 논문에서는 질병 확산 모형을 구축하고 추정하기 위한 통계적 방법을 제안한다. 질병의 확산 과정을 모형화 하기 위하여 전통적으로 미분 방정식을 이용한 방법이 제안되어 왔다. 미분 방정식을 이용한 모형 구축은 질병의 확산 과정이 결정적 추세를 따른다는 가정을 한다. 본 연구에서는 질병의 확산 과정이 확률적 추세를 따른다는 가정에서 확률적 화학 반응 모형을 이용하여 질병 확산 모형을 구축하고자 하였다. 특히 확률적 화학 모형의 반응 상수를 추정하기 위하여 몬테 카를로 EM (Monte Carlo expectation maximization; MCEM) 방법을 이용하였다. 제안된 MCEM 방법은 대표적인 질병 확산 모형 가운데 하나인 SIRS (susceptible - infected - recovered - susceptible) 모형에 적용하였고 그 결과를 베이지안 추론을 기반으로 하는 MCMC (Markov Chain Monte Carlo) 방법과 비교하였다. MCEM의 결과는 상대적으로 안정적이고 빠른 수렴의 결과를 제공하였다. 또한 2009년 미국에서 발생한 신종 플루의 초기 확산 자료에 적합하여 모형의 추정에 적용하였다. 본 연구에서 제안한 MCEM 방법은 베이지안 추정 방법의 하나의 대안으로 활용될 수 있을 것이다.

주요용어: 몬테 카를로 EM 알고리즘, 에스아이알에스 모형, 질병 확산 모형, 확률적 화학 반응 모형.

1. 서론

돼지 독감 혹은 신종 플루라고 불리우는 H1N1에 의한 유행병은 2009년 우리나라를 비롯한 전세계를 강타 하였다. 특히 우리나라는 잘못된 질병 확산의 예측에 따라 초기 백신 확보를 실패 함으로서 매우 큰 사회적 혼란을 야기하였다. 우리나라의 겸역 당국과 보건 당국 또한 전염병의 확산을 막기 위해 다양한 노력을 진행하였다. 정책적인 방법을 가지고 질병의 유행을 막는 것도 매우 중요한 문제이다. 이와 더불어 질병의 유행에 대한 대응의 일환으로 질병의 유행을 설명 할 수 있는 통계적 모형의 개발은 예방의 정도를 측정하고 질병 통제를 확립하는 있어서 효과적인 대책을 수립하는데 이용될 수 있다. 본 연구에서는 역학 조사를 통해 수집된 자료를 이용하여 질병의 확산을 설명하고 통계적 방법에 기반을 둔 모형의 구축을 수행하고자 한다.

질병의 확산 과정을 설명하기 위한 모형의 구축은 질병 확산 초기 시점에서 관찰된 데이터 또는 과거에 창궐하였던 유사한 전염병의 데이터를 사용하여 수행 할 수 있을 것이다. 연구를 통하여 수리적인 모형을 구축하고 구축된 모형을 이용하여 질병의 확산에 대한 실제 상황을 적절하게 기술 할 수 있다면 미

[†] 이 논문은 대구대학교 교내 연구비로 지원받아 수행된 연구임 (No.20150340).

¹ (30019) 세종특별자치시 세종로 2011, 고려대학교 세종캠퍼스 국가통계전공, 조교수.

² 교신저자: (38453) 경상북도 경산시 대구대로 201, 대구대학교 전산통계학과, 교수.

E-mail: yhyoon@daegu.ac.kr

래의 질병 확산의 정도를 예측할 수 있을 것이다. 모형에 기초한 예측은 질병과 관련된 방제와 대책의 정책 수립을 위한 기초 자료로 사용될 수 있다.

역사적으로 사람 혹은 가축과 같이 접촉을 통하여 전염되는 질병에 대한 수학적 접근이 시도되어왔다. 질병의 확산 과정을 설명하기 위한 모형의 구축은 초기에 미분 방정식 (ordinary differential equations; ODEs)을 가지고 구축되었다. 미분 방정식 모형은 전염병의 이동이 결정적 (deterministic)인 움직임을 따른다는 가정 하에 전염병의 추세를 파악할 수 있는 모형이다. 결정적 움직임을 가정한 모형은 전체 모형을 근사적으로 설명할 수 있도록 단순화 시킨다. 결과적으로 대수의 법칙 (law of large number)에 근거하여 전체 추세를 평균값으로 수렴하게 모형을 구축하게 된다 (Andersson과 Britton, 2000, 5장). 그러나 미분 방정식을 이용한 근사 모형은 질병의 확산 과정을 다소 단순화시키는 문제점을 갖는다. 또한 질병 감염 대상군이 대 단위이고 초기 감염자가 상대적으로 매우 적은 상태에서는 질병 확산의 초기 단계에서 질병의 확산 과정을 제대로 식별하지 못하는 문제를 가질 수 있다 (Schwartz 등, 2015).

실제로 질병의 확산 과정은 결정적인 움직임을 따른다는 가정 보다는 확률적 (stochastic) 움직임을 따른다고 볼 수 있으며 확률적 모형을 이용하여 전염병의 전파를 설명하는 것이 더 합리적일 수 있다 (Keeling 등, 2001). 확률적 모형은 이산 상태 공간에서 연속 시간 마르코프 연쇄 (continuous time Markov Chain)를 따르는 모형으로 표현 될 수 있다. 확률적 모형을 통하여 질병 혹은 전염병의 확산 과정이 고유하게 가질 수 있는 변동성을 함께 설명할 수 있게 된다. 하지만 결정적인 모형을 사용하는 대신에 확률적 모형을 구축하고자 하는데 있어서는 더 많은 시간과 자료의 수집이 요구된다.

질병 혹은 전염병의 확산 과정을 설명하고자 하는 모형 구축과 관련된 국내외 연구를 살펴보자. Ryu와 Choi (2015)는 미분 방정식을 이용하여 결정적 추세 가정에서 고전적 SIR (susceptible - infected - recovered) 모형의 모수 추정을 수행하였다. 이 때 최소제곱법으로 모수를 추정하였으며 이를 국내 말라리아 발병자 수 자료에 적합하였다. Seo와 Choi (2015)는 질병의 확산과정이 확률적 움직임을 따른다는 가정에서 모형을 구축하고자 하였고 그 방법을 2009년 우리나라를 강타한 신종플로 확산의 초기 과정에 적용하였다. 이 때 정확하게 관찰되지 않은 회복군의 숫자를 확률적으로 추출하는 방법을 함께 적용하였다. Lim 등 (2016)과 Do 등 (2017)은 2015년 5월부터 2016년 1월까지 조사된 메르스 코로나바이러스 (middle east respiratory syndrome coronavirus; MERS-CoV)에 의한 호흡기 질환 환자 자료를 이용한 모형 구축 연구를 진행하였다. 두 연구는 모두 확률적 확산 과정을 가정하여 모형 적합을 진행하였다. Lim 등 (2016)의 연구에서는 질병관리본부에서 공식적으로 유행병의 종식을 선언한 2015년 7월 28일 이후 퇴원을 하거나 사망을 하지 않은 13명의 환자 자료를 결측 자료로 고려하여 이를 회복군으로 이동하는 과정을 추가로 추정하는 일종의 결측치 대체를 모형 추정 전체에 포함하여 모형 추정을 진행하였다. Do 등 (2017)의 연구에서는 Lim 등 (2016)의 연구와는 다르게 SIR 모형에 노출기 혹은 잠복기로 불리우는 단계를 추가하여 SEIR (susceptible - exposed - infected - recovered) 모형을 적용하였다. SIR 모형에 SEIR 모형에서 추정된 모수 가운데 감염율에 해당하는 모수의 추정치에는 차이를 보였으나 질병의 확산 정도를 나타내는 기초감염재생산수 (basic reproduction number)는 큰 차이를 보이지 않았으며 각각 1.128과 1.182로 추정되었다. Eom 등 (2017)은 전염병의 감염 확률을 직접 추정하는 연구를 진행하였다. 1928년 Lowell Reed와 Wade Frost가 제안한 방법으로 기본적으로 조건부 이항분포를 따르는 확률 모형을 구축하여 관찰된 자료로부터 감염 확률을 추정하는 방법이 제시되었다. 이를 Reed-Frost 모형이라 부른다 (Deijfen, 2011). Eom 등 (2017)은 2016년 6월부터 약 두 달간 수집된 서아프리카 카메룬 공화국에 위치한 Maroua 지역의 콜레라 질병 자료를 이용하여 Reed-Frost 모형을 적용한 후 감염 확률을 추정하는 연구를 진행하였다.

그러나 이상의 방법들을 적용하기 위해서는 모형이 간단하여야 하며 자료가 충분히 확보 되어야 한다. 특히 연속 시간 마르코프 연쇄를 가정하는데 있어서 모든 시점에서의 자료가 온전하게 관찰되어야 한다. 그러나 현실적인 상황에서 이는 불가능한 가정이다. 본 논문에서는 부분적으로 분리되어 수집된

질병 자료에 대하여 확률적 화학 반응 모형 (stochastic reaction model or stochastic kinetic network model)을 적용하여 모형 구축을 시도한 후 보다 효율적인 추정 방법을 소개하고 질병 확산 모형의 추정 및 구축을 수행하고자 하였다. 본 연구에서는 확률적 화학 반응 모형을 구축하기 위하여 EM 알고리즘 (Dempster 등, 1977)의 몬테 카를로 (Monte Carlo) 버전 (Wei와 Tanner, 1990)을 이용하였다. 확률적 화학 반응 모형을 추정하기 위한 몇 가지 베이지안 (Bayesian) 접근법이 제안되었다. Boys 등 (2008)과 Wilkinson (2012)는 Gillespie 알고리즘 (Gillespie, 1997)에 기초한 베이지안 추론 접근법을 제안하였고 MCMC (Markov Chain Monte Carlo) 기법을 사용하여 이산 시점에서 관측된 종 (species)의 자료를 조건부로 관측되지 않은 종의 자료를 근사적인 방법으로 추출한 후 확률적 화학 반응 모형의 모수를 추출하는 MCMC 방법을 제안하였다. Choi와 Rempala (2012) 또한 베이지안 접근법에 기반을 둔 MCMC 방법을 제안하였다. 그들은 확률적 화학 반응 모형에서 관찰되지 않은 종의 변화와 시간을 일괄적으로 추출하여 MCMC 과정에서 개선하는 방법을 제안하였다. 이 때 uniformization 방법 (Rodrigue 등, 2006)을 적용하였고 그 결과를 2009년 미국에서 창궐한 신종 플루의 초기 자료에 적용하였다. Choi (2015)는 확률적 화학 반응 모형의 모수를 추정하는데 있어서 기본적으로 Boys 등 (2008)의 방법으로부터 출발하여 추가적을 평활 기법을 적용하여 추정 결과를 안정화 시키는 연구를 진행하였다.

본 연구의 구성은 다음과 같다. 2 절에서는 확률적 화학 반응 모형을 이용한 질병 확산 모형 구축과 모형 추정을 위하여 본 연구에서 제안하고 있는 방법을 소개한다. 3 절에서는 모의 실험을 통하여 생성된 인공 자료를 이용하여 모형 추정 결과를 소개한다. 이 때 질병 확산 모형 가운데 하나인 SIRS 모형을 이용하였다. 그리고 기존의 다른 MCMC 방법과의 비교를 함께 수행하였다. 또한 미국에서 발생하였던 신종 인플루엔자 (H1N1)의 초기 유행성 자료를 사용하여 실제 자료 분석을 수행하였다. 마지막 4 절에서는 이 연구에 대한 정리와 요약을 서술한다.

2. 통계적 모형 추정 방법

2.1. 확률적 질병 확산 모형

일반적으로 잘 알려진 확률적 질병 확산 모형은 William Kermack와 Anderson McKendrick에 의해 소개된 SIR 모형이다 (Andersson과 Britton, 2000). SIR 모형은 질병의 전파 과정에 따라 전체 모집단을 크게 3개의 부분 집단으로 구분하였다. 먼저 S (susceptible)는 질병에 감염될 가능성이 있는 감염 대상군을 나타낸다. I (Infected)는 질병에 감염되어 감염 대상군의 개체에 질병을 전파 시킬 수 있는 감염군을 나타낸다. 마지막으로 R (recovered 또는 removed)은 감염군에 있던 한 개체가 일정 시간이 흐른 후에 회복되거나 혹은 사망하여 더 이상 질병을 감염시키지 않는 집단으로 회복군을 나타낸다. 감염자와의 접촉에 따라 감염 대상군에 있는 한 개체는 감염군으로 이동한 후 다시 일정 시간이 흐른 후에 회복군으로 이동하게 된다. 이들 세 단계의 앞 글자를 가지고 SIR 모형이라고 부르게 된다.

SIR 모형은 여러 파생 모형을 가지고 있다. 감염된 개체가 감염군으로 이동하기 전에 일정한 잠복기를 가지게 되는 경우 잠복기에 해당하는 E (exposed) 단계를 추가한 모형이 SEIR 모형이다 (Do 등, 2017). SIS 모형은 질병에서 회복되는 순간부터 다시 질병에 감염될 수 있는 질병에서 사용될 수 있는 모형이고 SIRS는 질병에서 회복되어 회복군에 있던 한 개체가 일정 시간이 지난 후에 다시 질병에 감염될 가능성이 있는 질병에 적용될 수 있는 모형으로 콜레라와 같은 질병이 대표적이라 할 수 있다 (Koepke 등, 2016).

확률적 질병 확산 모형을 설명하기 위하여 SIRS 모형을 예를 들어 보자. SIRS 모형에서 $S(t)$, $I(t)$, $R(t)$ 는 각각 시점 t 에서 감염 대상군, 감염군, 회복군의 상태 (state) 혹은 관찰된 집단의 수를 나타낸다. SIRS 모형에서 전체 모집단의 수는 M 으로 고정되어 있다고 가정하고 또한 모든 시점에서 $S(t) +$

$I(t) + R(t) = M$ 을 만족한다고 가정한다. SIRS 모형은 다음의 확률적 화학 반응 모형으로 표현할 수 있다.

$$\begin{aligned} S + I &\xrightarrow{h_1} 2I \quad ; \quad h_1(S, I, R, \theta_1) = \theta_1 SI, \\ I &\xrightarrow{h_2} R \quad ; \quad h_2(S, I, R, \theta_2) = \theta_2 I, \\ R &\xrightarrow{h_3} S \quad ; \quad h_3(S, I, R, \theta_3) = \theta_3 R. \end{aligned} \quad (2.1)$$

이 모형에서 첫 번째 반응식은 감염 대상군의 한 개체가 감염군의 한 개체와 접촉하여 감염군으로의 이동을 나타내는 식이고 두 번째 반응식은 감염군의 한 개체가 일정 시간이 지난 후로 회복되어 회복군으로의 이동을 나타나는 식이다. 마지막 세 번째 식은 회복군의 한 개체가 일정 시간이 흐른 후에 면역력이 떨어져서 다시 감염 대상군으로의 이동을 나타내는 식이다. 각각의 반응 식에는 반응 상수 $\theta_1, \theta_2, \theta_3$ 이 할당되어 있으며 실제 자료가 주어졌을 때 이 세 반응 상수는 추정하여하는 모수가 된다. 각각의 반응식은 모두 각 반응식의 발생 정도를 나타내는 위험 함수 (hazard function 또는 rate law), $h_1(S, I, R, \theta_1)$, $h_2(S, I, R, \theta_2)$, $h_3(S, I, R, \theta_3)$ 가 할당되어 있다. 각 반응식의 발생이 연속 시간 마르코브 연쇄를 따르고 그 때의 발생은 이산적이라 할 때 발생은 포아송 확률 과정을 따르게 되며 이 때 각 위험 함수는 포아송 확률 과정의 모수가 된다 (Choi과 Rempala, 2012).

최초시점 $t = 0$ 에서의 각 종의 값 $S(0), I(0), R(0)$ 과 반응 상수 $\theta_1, \theta_2, \theta_3$ 가 주어졌을 때 Gillespie 알고리즘 (Gillespie, 1977)을 이용하여 SIRS 모형을 모의실험을 통하여 구현 할 수 있다. 다음 Figure 2.1은 모의실험을 통하여 구현된 SIR 모형의 한 예이다. 초기치와 반응 상수는 각각 $S(0) = 100$, $I(0) = 1$, $R(0) = 0$ 이고 $\theta_1 = 0.01$, $\theta_2 = 0.2$, $\theta_3 = 0.1$ 로 주어졌다. 파란색 실선은 감염 대상군, 가운데 녹색 파선은 감염군, 후반부 위쪽 붉은 점선은 회복군을 나타낸다.

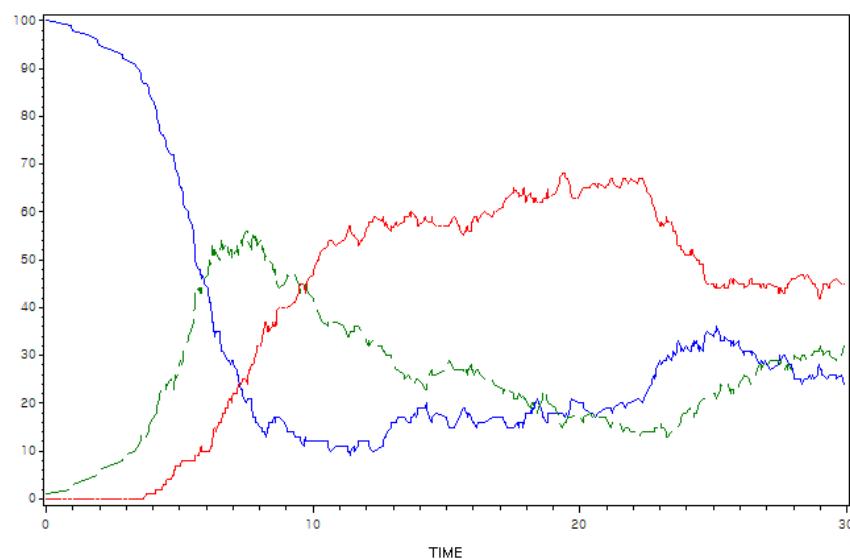


Figure 2.1 SIRS model trajectory for susceptible, infective, recovered respectively with total population = 100

2.2. 반응 상수의 통계적 추론

관찰된 시구간 $(0, T]$ 에서 확률적 화학 반응 모형을 따르는 전체 확률 과정 \mathbf{X} 라 하선. 그리고 이 화학 반응 모형은 x_1, x_2, \dots, x_u 개의 종 (species)를 가지고 R_1, R_2, \dots, R_v 의 반응 (reaction)으로 구성되어 있다고 하자. SIRS 모형을 예로 들자면 $u = 3$ 이고 $v = 3$ 이다. 각각의 반응식에는 반응식의 정도를 설명할 수 있는 위험 함수 $h_1(\mathbf{x}, \theta_1), \dots, h_v(\mathbf{x}, \theta_v)$ 가 할당되어 있으며 여기서 $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_v)$ 는 반응 상수로 추정하여야 할 모수이다. \mathbf{X} 가 모든 연속 시간에서 모두 관찰되었다고 하였을 때 Gillespie 방법에 의하여 우도 함수는 다음과 같이 구할 수 있다 (Boyes 등, 2008).

$$L(\boldsymbol{\theta} | \mathbf{X}) = \prod_{i=1}^T \prod_{j=1}^{n_i} h_{k_{ij}}(x(t_{ij}), \theta_{k_{ij}}) \exp \left(\int_0^T h_0(x(t), \theta) dt \right) \quad (2.2)$$

여기서 $h_0(x(t), \theta) = \sum_{k=1}^v h_k(x, \theta_k)$ 이다. 우도 함수 구축의 자세한 설명은 Wilkinson (2012)를 참조 할 수 있다. 식 (2.2)에서 위험 함수 $h_k(x, \theta_k)$ 는 반응 상수와 반응 상수와 독동인 식의 곱의 형태로 분리 될 수 있다. 즉 $h_k(x, \theta_k) = \theta_k g_k(x)$ 이 성립하고 이에 따라 식 (2.2)는 다음과 같이 정리된다.

$$\begin{aligned} L(\boldsymbol{\theta} | \mathbf{X}) &= \left(\prod_{i=1}^T \prod_{j=1}^{n_i} \theta_{k_{ij}} g_{k_{ij}}(x(t_{ij})) \right) \exp \left(- \int_0^T \sum_{k=1}^v \theta_k g_k(x(t)) dt \right) \\ &\propto \left(\prod_{k=1}^v \theta_k^{\sum_{i=1}^T r_{ki}} \right) \exp \left(- \sum_{k=1}^v \int_0^T \theta_k g_k(x(t)) dt \right) \\ &= \prod_{k=1}^v \theta_k^{\sum_{i=1}^T r_{ki}} \exp \left(- \theta_k \int_0^T g_k(x(t)) dt \right) \\ &= \prod_{k=1}^v L_k(\theta_k | \mathbf{X}). \end{aligned} \quad (2.3)$$

식 (2.3)에서 r_{ki} , $k = 1, \dots, v$, $i = 1, \dots, T$ 는 구간 $(i, i+1]$ 에서 발생한 반응 k 의 수를 나타낸다. 결국 $r_k = \sum_{i=1}^T r_{ki}$ 는 전체 관찰 구간에서 k 번째 반응이 발생한 총 수를 나타낸다. 이제 식 (2.3)으로부터 개별적인 반응 상수 θ 에 대한 추론이 가능하다. 결과적으로 각 θ 에 대한 우도함수는 감마 분포의 형태를 가지며 이로부터 각 θ 에 대한 최대 우도 추정치는 다음과 같이 주어진다.

$$\hat{\theta}_k = \frac{\sum_{i=1}^T r_{ki}}{\int_0^T g_k(x(t)) dt} = \frac{r_k}{\int_0^T g_k(x(t)) dt}, \quad k = 1, \dots, v. \quad (2.4)$$

베이지안 추론을 위하여 각 반응 상수에 공액 사전분포를 할당할 수 있다. 각 반응 상수에 대하여 다음과 같이 독립적인 감마분포를 사전분포로 할당한다.

$$\theta_k \sim \Gamma(a_k, b_k), \quad k = 1, \dots, v. \quad (2.5)$$

이제 우도 함수 (2.3)와 사전 함수 (2.5)로 부터 다음과 같은 사후 분포 함수를 구축할 수 있으며 이 분포로부터 베이지안 추론이 가능하다.

$$\theta_k | \mathbf{x} \sim \Gamma \left(a_k + r_k, b_k + \int_0^T g_k(x(t)) dt \right), \quad k = 1, \dots, v. \quad (2.6)$$

실제로 \mathbf{X} 의 모든 값들이 관찰되는 것은 거의 불가능 하며 일정한 시간 간격을 두고 (예를 들면 매초, 매일, 매월) 특정 시점 혹은 단위 시점에 관찰된 종의 숫자만이 자료로 주어질 수 있다. 단위 시간 사이

에서 관찰되지 않은 종들의 변화는 결측 정보라 할 수 있다. 완전히 관찰된 자료와 결측 자료를 구분하기 위하여 결측이 발생한 경우를 \mathbf{X}_{mis} 로 나타낸다. 이제 결측 정보 \mathbf{X}_{mis} 는 모수의 추정 과정에서 관찰되지 않은 자료에 해당하기 때문에 이에 대한 적절한 대체 또는 추정이 포함되어야 한다. 모형의 추정 과정에서 관찰되지 않은 자료는 추가적인 모수에 해당한다. 관찰되지 않은 종들에 대한 적절한 추정을 포함하여 반응 상수에 대한 추정을 수행하기 위한 방법으로 베이지안 방법이 제안되어 왔다. 베이지안 방법에서는 $p(\mathbf{X}_{mis}, \theta | \mathbf{X})$ 와 같이 결합 사후분포로부터 관찰되지 않은 종의 정보 \mathbf{X}_{mis} 와 반응 상수 θ 를 각각의 조건부 사후 분포로부터 추출하는 Gibbs sampler 방법이 제안되어 왔다. Boys 등 (2008)과 Choi와 Rempala (2012)는 \mathbf{X}_{mis} 의 조건부 사후분포로부터 추출을 위하여 Metropolis-Hastings 알고리즘을 이용하였다. 그러나 식 (2.6)에서와 같이 모든 \mathbf{X}_{mis} 를 추출하는 대신에 방응의 수인 r_k 의 정보만 주어져도 반응 상수에 대한 추정이 가능하다. 본 연구에서는 이러한 점을 착안하여 r_k 에 대하여 몬테 카를로 EM (Monte Carlo expectation maximization; MCEM) 방법을 적용하여 방응 상수에 대한 추론에 대한 대안을 제시하고자 한다.

2.3. 몬테 카를로 EM (MCEM) 알고리즘

관찰되지 않은 각 반응의 수 r_{ik} , $i = 1, \dots, T$, $k = 1, \dots, v$ 를 결측치로 고려 한 후 EM 알고리즘 (Dempster 등, 1977)을 이용한다. E-step에서는 관찰되지 않은 r_{ik} 의 기대값을 계산하고 M-step에서는 우도함수를 최대화 시키는 MLE $\hat{\theta}$ 를 구한다. 우도 함수 2.3로부터 각 반응 상수 θ_k 에 대한 우도 함수에 로그를 취하면 다음과 같다.

$$\log L(\theta_k | \mathbf{X}) = \sum_{i=1}^T r_{ki} \log \theta_k - \theta_k \int_0^T g_k(x(t)) dt, \quad k = 1, \dots, v. \quad (2.7)$$

E-step에서는 $\hat{\theta}_k$ 의 값이 주어졌을 때 관찰되지 않은 r_{ki} 의 기대값을 계산함으로서 로그 우도 함수 (2.7)의 기대값을 구한다.

$$E_{old}[\log L(\theta_k | \mathbf{X})] = \log \theta_k E_{old} \left[\sum_{i=1}^T r_{ki} \right] - \theta_k \int_0^T g_k(x(t)) dt, \quad k = 1, \dots, v. \quad (2.8)$$

M=step에서는 식 (2.8)를 최대화 시키는 θ_k , $k = 1, \dots, v$ 를 계산한다.

$$\hat{\theta}_k^{new} = \frac{E_{old} \left[\sum_{i=1}^T r_{ki} \right]}{\int_0^T g_k(x(t)) dt}, \quad k = 1, \dots, v.$$

그러나 r_{ik} 의 조건부 분포가 정확히 알려져 있지 않기 때문에 이에 대한 기대값을 구하는 것은 또 다른 문제가 될 수 있다. 이에 대한 대안방법으로 본 연구에서는 E-step에서 기대값을 바로 계산하지 않고 몬테 카를로 방법을 이용하여 기대값을 계산하는 몬테 카를로 EM 방법을 이용하고자 하였다 (Wei와 Tanner, 1990). 이제 EM 알고리즘에서의 E-step은 다음과 같이 몬테 카를로 방법으로 대체될 수 있다.

- a. Draw $r_{ki}^1, r_{ki}^2, \dots, r_{ki}^m \stackrel{iid}{\sim} p(r_{ki} | \theta_k, X)$, $k = 1, \dots, v$, $i = 1, \dots, T$.
- b. $\hat{E}_{old}[\log L(\theta_k | X)] = \frac{1}{m} \sum_{j=1}^m \log L(\theta_k | X, r_{ki}^j)$, $k = 1, \dots, v$, $i = 1, \dots, T$.

이제 몬테 카를로 E-step에서 각 시구간 $[i, i+1]$ 에서 관찰되지 않은 r_{ki} 의 추출 문제를 다루어 보자. r_{ki} 는 각 반응의 발생 숫자이고 그로부터 생성되는 전체 프로세스 \mathbf{X} 는 연속시간 마르코프 체인을 따르는 이산적인 과정이라 할 수 있다. 즉 r_{ki} 는 counting process를 따른다 할 수 있다. counting

process의 가장 대표적인 예는 포아송 과정 (Poisson process)이며 결과적으로 각 반응의 발생정도는 포아송 과정을 따른다고 볼 수 있다. 이와 같은 접근은 확률적 질병 확산 과정을 모의실험 하기 위한 방법인 Gillespie 알고리즘에서 적용된다. Gillespie 알고리즘을 조금 더 효율적으로 적용하기 위한 방법으로 $\tau - leaf$ 방법이 제안 되었다 (Gillespie, 2001). 본 연구에서도 $\tau - leaf$ 방법과 유사한 과정을 이용하여 r_{ki} 의 추출을 수행하고자 한다. $\tau - leaf$ 방법은 시 구간 $[i, i+1)$ 에서 시작 시점을 정보만을 가지고 추출을 진행한다. 즉 $X(i)$ 의 정보만을 가지고 추출을 진행한다. 본 연구에서는 이를 확장하여 $X(i)$ 와 함께 $X(i+1)$ 의 정보도 함께 이용하여 추출을 진행한다. $\tau - leaf$ 방법에서는 r_{ki} 의 추출을 위하여 i 시점에서의 위험함수 값인 $h_k(x(i), \theta_k)$ 를 모수로 하는 포아송 분포로부터 표본을 추출한다. 시 구간의 종료 시점인 $i+1$ 에서의 정보를 함께 사용하여 r_{ki} 의 추출은 다음의 포아송 분포로부터 먼저 표본을 추출한다.

$$r_{ki}^* \sim Poi((h_k(x(i), \theta_k) + h_k(x(i+1), \theta_k))/2).$$

추출한 후 다음과 식과 같이 최종적으로 갱신한다. 여기서 s_{ki} 는 구간 $[i, i+1)$ 에서 최소한으로 발생되어야 하는 반응의 수이다.

$$r_{ki}^{(j)} = \begin{cases} r_{ki}^* & \text{if } r_{ki}^* > s_{ki}, \\ s_{ki} & \text{if } r_{ki}^* \leq s_{ki}. \end{cases} \quad (2.9)$$

3. 자료 분석

본 절에서는 가상 자료와 실제 자료를 이용하여 2절에서 소개한 MCEM 방법에 의한 확률적 질병 확산 모형의 추정 결과를 제시한다. 또한 소개된 방법의 성능을 비교하기 위하여 Boyes 등 (2008)과 Choi와 Rempala (2012)가 제안한 MCMC 방법을 이용한 베이지안 추정 결과를 함께 제시한다.

3.1. 모의 실험 자료

MCEM 방법을 이용한 모형 추정을 위하여 확률적 질병 확산 모형을 따르는 자료를 모의 실험을 통하여 생성한 후 생성된 자료에 대한 모형 적합을 진행하였다. 분석에 이용된 질병 확산 모형은 2.1절에서 소개된 SIRS 모형이다. SIRS 모형에서는 R로 표현되는 회복군의 경우 사망 혹은 회복을 의미하고 더 이상 질병에 감염되지 않는다. 하지만 SIRS 모형에서는 일정 시간이 흐른 후에 회복군에 머물러 있던 개체는 다시 감염 대상군은 S로 이동할 수 있다. 식 (2.1)의 확률적 화학 반응 모형으로 표현된 SIRS 모형으로부터 Gillespie 알고리즘을 이용하여 모형 자료를 생성할 수 있다. Gillespie 알고리즘을 이용한 모형의 생성 과정은 Ryu와 Choi (2015)를 참조하면 된다. 3개의 반응식에 대한 반응 상수는 $\theta = (0.02, 0.2, 0.1)$ 로 설정하였다. 전체 모집단의 크기는 $M = 25$ 로 하였고 감염 대상군, 감염군, 회복군의 초기값은 각각 $S(0) = 24$, $I(0) = 1$, $R(0) = 0$ 로 설정하였다. 전체 시구간은 $T = 30$ 으로 하였다. 이와 같이 초기값을 지정하고 반응 상수를 결정한 후 Gillespie 알고리즘을 이용하여 전체 경로를 생성할 수 있다. 유사한 결과가 Figure 2.1에 제시되어 있다. 하지만 실제 상황과 유사하게 하기 위하여 생성된 전체 경로자료를 이용하지 않고 $[0, T]$ 의 단위시간에서 관찰된 자료만을 저장하여 분석을 위한 자료를 생성하였다. 생성된 결과는 Figure 3.1에서 확인할 수 있다. 검은색 실선으로 표시된 것이 감염 대상군의 전체 경로이고 붉은 색 점선으로 표시된 것이 감염군의 전체 경로이다. 파란색 파선으로 표시된 것이 회복군을 나타낸다. 대략 시점이 25에 도달하면 안정화 되는 모습을 확인할 수 있다.

2.2절에서 소개된 MCEM 방법을 적용하기 위해서는 나누어진 단위 시 구간 $[i, I+1)$ 에서 r_{ki} 의 하한값을 미리 계산해 두어야 한다. 이는 다음의 식과 같이 결정된다.

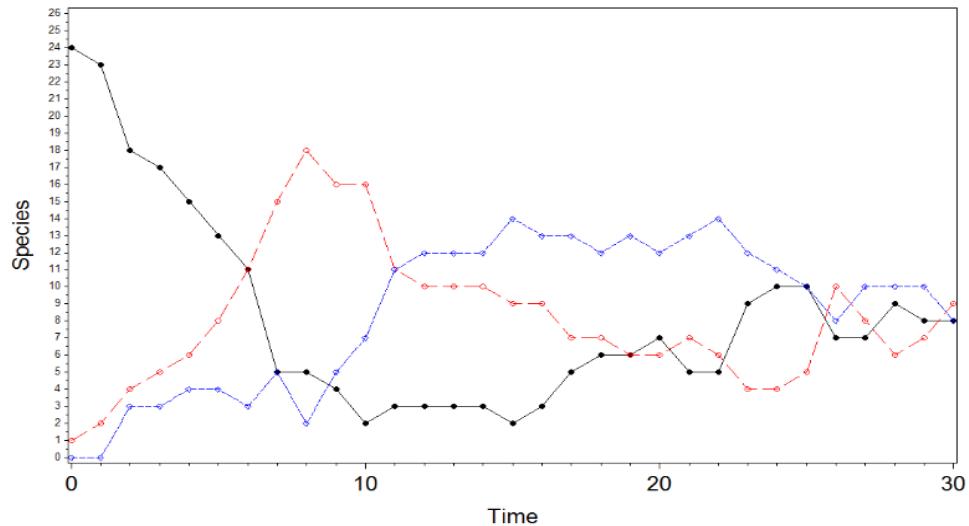


Figure 3.1 Simulated SIRS model; solid line: S, dotted line: I, dashed line: R

if $S(i+1) - S(i) > 0$ then

$$\begin{aligned} s_{3i} &= S(i+1) - S(i), \\ s_{2i} &= s_{3i} - (R(i+1) - R(i)), \\ s_{1i} &= s_{2i} + (I(i+1) - I(i)), \end{aligned}$$

if $I(i+1) - I(i) > 0$ then

$$\begin{aligned} s_{1i} &= I(i+1) - I(i), \\ s_{3i} &= s_{1i} - \max(S(i) - S(i+1), 0), \\ s_{2i} &= s_{3i} + (R(i+1) - R(i)), \end{aligned}$$

if $R(i+1) - R(i) > 0$ then

$$\begin{aligned} s_{2i} &= (R(i+1) - R(i)) + \max(S(i+1) - S(i), 0), \\ s_{1i} &= s_{2i}(I(i+1) - I(i)), \\ s_{3i} &= s_{1i} + (S(i+1) - S(i)). \end{aligned}$$

이 식에서 $S(i)$, $I(i)$, $R(i)$ 는 각 시점 i 에서 관찰된 감염 대상군, 감염군, 회복군의 수를 나타낸다. Figure 3.1에서 점이나 원으로 표시된 숫자를 의미한다. 그리고 s_{ki} , $k = 1, 2, 3$ 은 관찰 시 구간에서 세 반응에 대한 발생 건수 r_{ki} , $k = 1, 2, 3$ 의 하한 값을 나타낸다. 이 s_{ki} , $k = 1, 2, 3$ 가 식 (2.9)에 적용되게 된다.

MCEM 방법을 이용한 추정의 성능을 비교하기 위하여 MCMC를 통한 베이지안 추정과 비교를 수행하였다. 두 가지 방법을 고려하였는데 첫 번째 방법은 Boys 등 (2008)이 제안한 방법이고 두 번째 방법은 Choi와 Rempala (2012)가 제안한 방법이다. Boys 등 (2008)의 방법은 r_{ki} 를 추출하기 위하-

여 approximation block updating 방법을 이용한 것이다. 이에 반하여 Choi와 Rempala (2012)는 uniformization 방법을 이용하여 관찰되지 않은 전체 과정을 모두 추출하는 방법이다. 세 방법은 모두 5,000번의 반복 수행을 진행하였고 이 가운데 초기 1,000개의 반복은 burning set으로 하여 제외하고 추정 결과에 이용하였다. MCEM 방법에서 몬테 카를로 E-step의 반복 숫자는 100으로 하였다. 모든 분석은 SAS/IML V9.4로 수행되었다.

다음 Table 3.1는 MCMC 방법에 의한 추정 결과를 정리한 표이다. 첫 번째 줄의 Uniform은 Choi와 Rempala (2012)의 추정 결과이고 두 번째 줄의 MH는 Boys 등 (2008)의 추정 결과이다. 마지막 줄의 MCEM은 본 연구에서 제안한 방법이다. 각 칸의 숫자들은 반응 상수 $\theta_1, \theta_2, \theta_3$ 의 4,000 표본 평균을 나타내고 괄호안은 표준편차를 계산한 것이다. Uniform과 MH는 베이지안 방법이기 때문에 사후 표본 평균에 해당한다. 실제값이 각각 0.02, 0.2, 0.1임을 고려하였을 때 세 방법은 모두 비슷하고 비교적 정확한 결과를 주고 있다. MCEM의 방법이 미세하지만 상대적으로 조금 큰 편향을 보이고 있다. 이에 반하여 MCEM 방법은 표준편차가 매우 작은 것을 확인할 수 있다. MCEM 방법에서는 E-step에서 100개의 표본을 추출하여 몬테카를로 평균을 계산하였기 때문에 상대적으로 매우 작은 표준편차가 계산된 것을 어느 정도 예견할 수 있다.

Table 3.1 Posterior means and standard deviation (in parenthesis) for MCMC using Uniformization method (Uniform), MCMC using Metropolis-Hastings algorithm (MH), and MC-EM simulation (MCEM)

	θ_1	θ_2	θ_3
Uniform	0.0197(0.0039)	0.1902(0.0392)	0.1059(0.0322)
MH	0.0199(0.0039)	0.1885(0.0394)	0.1059(0.0318)
MCEM	0.0205(0.0001)	0.1835(0.0076)	0.1248(0.0006)

다음 Figure 3.2는 각 방법에 따른 5,000번의 반복수행 결과를 시도표로 표시한 것이다. 세 반응 상수 가운데 첫 번째 반응 상수인 θ_1 의 결과만을 정리한 것이다. 기존의 두 베이지안 방법은 큰 차이를 보이지 않으며 5,000번의 반복수행에서 안정적으로 수렴한 결과를 보여 주고 있다. 마지막 줄은 MCEM 방법의 수렴 결과를 보여준다. 분석에 이용된 컴퓨터는 quad core CPU 4.0GHz, windows 10 x64-based processor에서 수행되었다. 기존의 MCMC 방법의 경우 약 3초 내외가 소요되었다. MCEM 방법의 경우 상대적으로 매우 긴 약 46초의 시간이 소요되었다. 이는 E-step에서 반복의 수를 100으로 하였기 때문에 실제적으로 MCEM의 방법이 훨씬 많은 반복을 수행하였기 때문인 것으로 기인한다. 그러나 MCEM의 방법은 상대적으로 매우 작은 (약 1-200번) 반복의 수를 가지고 수렴된 결과를 보여주고 있기 때문에 수렴이 이르는 속도는 충분히 출일 수 있다고 할 수 있다.

3.2. 실증 자료 분석

본 연구에서 제안하고 있는 MCEM 방법을 실제 자료 분석이 적용하여 보자. 분석에 이용된 자료는 2009년 미국에서 발생한 A/H1N1 인플루엔자 전염병 자료이다. 우리나라에서는 처음에 돼지독감이라고 불리다 이후 신종 플루라는 이름이 사용되었다. 미국의 질병 관리 본부 (center for disease control and prevention; CDC)에서 전염병 발병 초기시점에 수집된 자료로 2009년 4월 26일부터 5월 14일까지 총 19일 동안 매일 환자수를 기록한 자료이다 (Choi와 Rempala, 2012). 다음 Table 3.2는 수집된 자료를 정리한 표이다.

Table 3.2의 자료를 가지고 SIRS 모형을 적합 하는 데는 제약이 있다. 수집된 자료는 SIRS 모형에서 오직 I에 해당하는 감염군의 자료만이 존재한다. 따라서 축약된 모형이 필요하다. 신종 플루의 감염 대상군은 모집단의 모든 사람이 되고 주어진 자료는 발병 초기의 자료이다. 감염자 수의 증가에 따른 감염 대상군의 감소의 비율은 전체 모집단에 차지하는 비중이 매우 작기 때문에 감염 대상군은 모형에서 제외

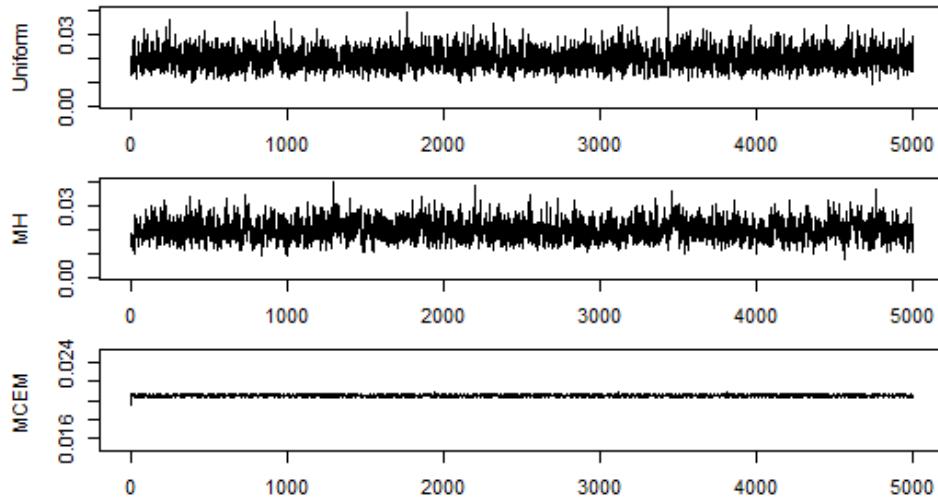


Figure 3.2 Trace plot for the parameter θ_1 . First panel is Uniformization method, second panel is MH method, and the last panel is MCEM method

Table 3.2 Daily counts of the number of H1N1 cases in the US from 26, Apr, 2009 to 14, May, 2009 (19 days)

Day	1	2	3	4	5	6	7	8	9
Count	20	40	64	91	109	141	160	226	279
Day	10	11	12	13	14	15	16	17	18
Count	403	642	896	1639	2254	2532	2600	3009	3352
Day	19								
Count	4298								

시킬 수 있을 것이다. 그리고 모형을 추정하는데 있어서도 회복군의 숫자는 필요가 없다. 이러한 특성을 반영하여 감염군 하나만을 고려하는 모형을 이용할 수 있을 것이다. 이에 Table 3.2을 적용하기 위한 모형으로 다음과 같은 *birth-and-death* 과정을 적용하였다.

$$\begin{aligned} Y &\xrightarrow{h_1} 2Y; \quad h_1 = h_1(Y) = \theta_1 Y, \\ Y &\xrightarrow{h_2} \emptyset; \quad h_2 = h_2(Y) = \theta_2 Y. \end{aligned} \quad (3.1)$$

식 (3.1)에서 첫 번째 반응식은 질병의 감염을 나타내고 두 번째 반응식은 질병으로부터의 회복을 나타낸다. 주어진 자료가 전염병 전파 과정의 초기 시점의 자료이기 때문에 감염자의 수가 끊임없이 증가하고 있다. 감염자의 숫자가 회복자의 숫자보다 월등히 클 것이고 반응의 속도를 나타내는 위험 함수는 $h_1 \gg h_2$ 의 관계가 성립한다. 이는 다시 $\theta_1 \gg \theta_2$ 가 성립하게 된다. 한편으로는 꾸준한 증가로 두 번째 반응의 발생 (질병의 회복)은 측정이 힘들지만 매우 작을 것으로 예측할 수 있다. 이러한 특성을 고려하고 감염률에 해당하는 θ_1 의 추정에 초점을 맞추기 위하여 θ_2 의 값은 0.001로 고정하였다 (Choi와 Rempala, 2012). 총 3,000번의 반복 수행을 하였으며 이 가운데 1,000번의 초기 수행을 burning set으로 제거하고 후반부 2,000번의 결과만을 정리하였다. 추정치는 $\hat{\theta}_1 = 0.275$ 이고 표준편차는 0.0041로 계산되었다. Uniformization 방법을 이용한 경우 추정치는 $\hat{\theta}_1 = 0.207$ 이고 표준편차는 0.0036이다. 분석에 걸린 시간은 MCEM의 경우 약 8초가 소요되었으나 uniformization 방법은 1분 이상의 시간이 소요되었다. Uniformization 방법은 계산 과정에서 행렬의 지수를 계산하여야 하는 단점을 가지고 있다.

이 때문에 이 자료에서는 실제적으로 매우 긴 분석 시간이 소요되었다.

모형의 적합 정도를 검토하기 위하여 추정된 반응 상수값 0.275와 θ_2 의 가정한 값 0.001을 가지고 Gillespie 알고리즘을 이용하여 모형을 구축하였다. 그 결과는 다음 Figure 3.3에 표시되어 있다. 그림에서 붉은색 점으로 표시된 것이 Table 3.2의 자료를 표시한 것이고 붉은색 실선으로 표시된 부분이 적합된 결과이다. 전반적인 추세를 잘 따라가고 있는 것을 확인할 수 있다.

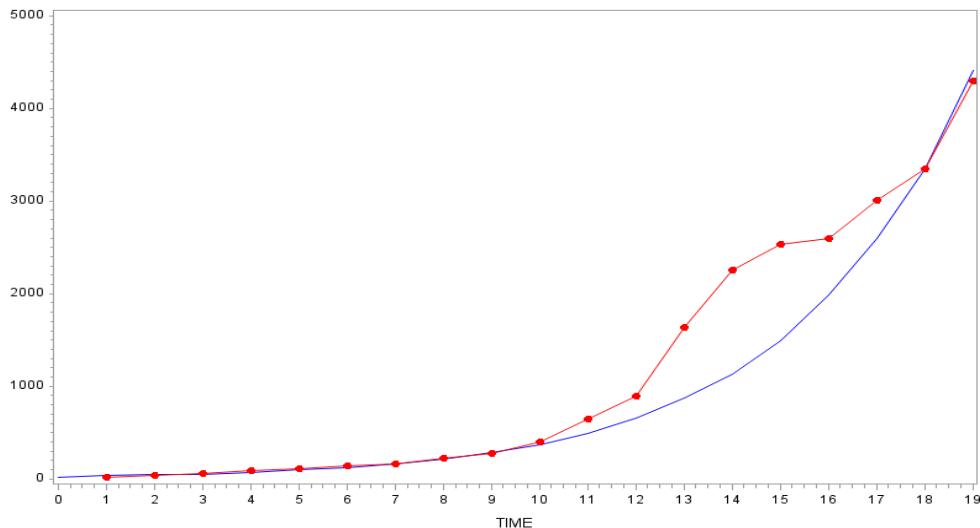


Figure 3.3 The model fitted (solid line) and observed (dotted line) counts of H1N1 data in the US during the oneset of the epidemic

4. 결론

이 논문의 목적은 질병의 확산 과정을 설명하고 전염병 확산 모형을 구축하기 위한 통계적 모형을 구축하는 것이다. 전통적으로 결정적 추세를 가정한 미분 방정식 모형이 많이 사용되었는데 본 연구에서는 확률적 추세를 가정하는 확률적 화학 반응 모형으로 질병 확산 모형의 구축을 수행하였다. 확률적 질병 확산 과정을 설명하는 대표적인 방법 가운데 하나인 SIRS 모형을 가지고 모형의 추정 및 구축을 수행하였다. 모형의 추정을 위하여 관측된 자료의 한계를 결측 자료로 고려하여 모수의 추정의 과정에서 결측 자료의 추정을 함께 수행하는 MCEM 방법을 제안하였다. 제안된 방법은 모의 실험을 통하여 생성된 SIR식 모형의 적합에 적용 되었다. 이때 본 연구에서 소개된 방법의 성능을 비교하기 위하여 두 가지 베이지안 추론에 기반을 둔 MCMC 방법을 이용한 추정 결과와 비교를 진행하였다. 마지막으로 미국에서 발생한 신종 플루 자료를 이용한 모형 적합을 수행하였다. 모의 실험과 실제 자료를 이용한 적합 결과 안정적인 결과를 얻을 수 있었다. 하지만 분석에 이용된 자료의 구조가 상대적으로 복잡하지 않고 규모가 크지 않은 자료에 적합을 수행한 한계를 가지고 있다. 보다 복잡하고 대용량 자료에 대한 분석이 요구된다 할 수 있다.

References

- Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*, Springer, New York.
- Boys, R. J., Wilkinson, D. J. and Kirkwood, T. B. B. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, **18**, 125-135.
- Choi, B. (2015). An estimation method for stochastic reaction model. *Journal of the Korean Data & Information Science Society*, **26**, 813-826.
- Choi, B. and Rempala, G. A. (2012). Inference for discretely observed stochastic kinetic networks with applications to epidemic modeling. *Biostatistics*, **13**, 153-165.
- Deijfen, M. (2011). Epidemics and vaccination on weighted graphs. *Mathematical biosciences*, **232**, 57-65.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of Royal Statistical Society Series B*, **39**, 1-38.
- Do, M., Kim, J. and Choi, B. (2017). A study of epidemic model using SEIR model. *Journal of the Korean Data & Information Science Society*, **28**, 296-307.
- Eom, E., Hwang, J. and Choi, B. (2017). An estimation method of probability of infection using Reed - Frost model. *Journal of the Korean Data & Information Science Society*, **28**, 57-66.
- Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, **81**, 2340-2361.
- Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, **155**, 1716-1733.
- Keeling, M. J., Woolhouse, M. E. J., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J. and Grenfell, B. T. (2001). Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science*, **294**, 813-817.
- Koepke, A. A., Longini Jr, I. M., Halloran, M. E., Wakefield, J. and Minin, V. N. (2016). Predictive modeling of cholera outbreaks in Bangladesh. *The Annals of Applied Statistics*, **10**, 575-595.
- Lim, Y., Do, M. and Choi, B. (2016). A construction of susceptible - infected - removed model using Korean MERS pandemic data. *Journal of the Korean Data Analysis Society*, **18**, 105-115.
- Rodrigue, N., Philippe, H. and Lartillot, N. (2008). Uniformization for sampling realizations of Markov processes: Applications to Bayesian implementations of codon substitution models. *Bioinformatics*, **24**, 56-67.
- Ryu, S. and Choi, B. (2015). Development of epidemic model using the stochastic method. *Journal of the Korean Data & Information Science Society*, **26**, 301-312.
- Schwartz, E. J., Choi, B. and Rempala, G. A. (2015). Estimating epidemic parameters: Application to H1N1 pandemic data. *Mathematical Biosciences*, **270**, 198-203.
- Seo, M. and Choi, B. (2015). An estimation method for stochastic epidemic model. *Journal of the Korean Data Analysis Society*, **17**, 1247-1259.
- Wilkinson, D. J. (2012). *Stochastic modelling for systems biology*, 2nd Eds., CRC Press, Boca Raton.
- 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**, 699-704.

Stochastic epidemic model estimation using Monte Carlo expectation maximization algorithm[†]

Boseung Choi¹ · Yong Hwa Yoon²

¹Department of National Statistics, Korea University Sejong campus

²Department of Statistics and Computer Science, Daegu University

Received 26 December 2017, revised 11 January 2018, accepted 16 January 2018

Abstract

In this paper, we introduce a statistical method for modeling the spread of disease. Historically, the ordinary differential equation is proposed to construct the epidemic model. However, the deterministic approach for the epidemic model is too simplified to capture the stochastic behavior of spread of disease. We consider the stochastic kinetic networks model for the epidemic modeling and we proposed MCEM (Monte Carlo expectation maximization) method to perform the statistical inference for reaction constants of the stochastic epidemic model. We applied our MCEM method to a synthetic data from the representative stochastic epidemic model, named SIRS (susceptible - infected - recovered - susceptible) model and we compared proposed MCEM method with two Bayesian MCMC methods. The MCEM result gives stable and faster convergence results. We also MCEM method to the data from the onset of early pandemic of H1N1 in the US. The proposed MCEM method can be an alternative to the estimation method for the stochastic epidemic model.

Keywords: Epidemic model, MCEM algorithm, SIRS model, stochastic chemical reaction model.

[†] This research is supported by Daegu University research grant in 2014 (No.20150340).

¹ Assistant professor, Department of National Statistics, Korea University Sejong campus, Sejong 30019, Korea.

² Corresponding author: Professor, Department of Statistics and Computer Science, Daegu University, Gyeongbuk 38453, Korea. E-mail: yhyoon@daegu.ac.kr