Journal of the Korean Data & Information Science Society
2015, **26**(4), 813–826

http://dx.doi.org/10.7465/jkdi.2015.26.4.813
한국데이터정보과학회지

확률적 방법에 기반한 화학 반응 모형의 모수 추정 방법[†]

최보승[1]

[1]대구대학교 전산통계학과

접수 2015년 4월 18일, 수정 2015년 5월 11일, 게재확정 2015년 6월 3일

요 약

본 연구는 화학 반응 모형의 추정 문제를 다루고 있다. 화학 반응 모형이란 생화학 분야에서 종(species) 들 간의 상호작용을 통한 변화 과정을 설명하기 위한 모형으로 생화학 분야 뿐 만 아니라 질병의 확산과정을 설명하는데 적용하는 모형이다. 본 연구에서는 화학 반응 모형 안에서 종들의 움직임이 확률적이라는 가정하에 Gillespie 알고리즘을 이용하여 모형 추정을 위한 우도함수를 구축하였다. 제한적인 자료구조 하에서 베이지안 접근법에 기반하여 MCMC (Markov chain Monte Carlo) 방법에 기반한 모수의 추정 방법을 제안하였다. 제안된 방법들은 생태계 포식자-피식자 관계를 설명하기 위한 Lotka-Volterra 모형과 유전자 전사 (gene transcription) 과정을 설명하기 위한 L1 retrotransposition 모형에 적용하였다. 그 결과 우수한 추정 결과를 보였다.

주요용어: 로트카-볼테라 모형, 마르코프 연쇄 몬테 카를로 방법, 유전자 전사, 질레스피 알고리즘, 화학 반응 모형.

1. 서론

화학 반응 모형 (chemical reaction model) 이란 주로 생화학 분야에서 종의 합성, 상호작용 들을 통하여 변형, 확산, 소멸되어 가는 과정을 설명하기 위하여 사용되는 모형으로 chemical kinetics 모형이라고도 불리 운다. 유전학 분야에서는 DNA 전사, 유전자 조절 (gene regulation), 메신저 RNA의 소멸 등을 모형화하기 위하여 사용된다. 또한 생태계에서 상호 경쟁관계에 놓여 있는 종족들 사이에서 발생하는 적자생존, 포식자-피식자 관계 등을 설명하기 위하여 사용된다. 이와 같이 생명과학 또는 생화학 분야의 다양한 분야에서 종들의 확산 과정을 설명하기 위하여 구축된 화학 반응 모형은 여러 컴퓨터 알고리즘을 이용하여 구현할 수 있으며 모의실험을 통하여 그 결과를 시각적으로 확인 할 수도 있다.

화학 반응 모형은 자연과학 뿐 만 아니라 사회과학 분야에도 적용할 수 있다. Lee 등 (2002)와 Lee 등 (2003)는 화학 반응 모형을 이용하여 우리나라 주식시장 변화에 대한 동태적 분석을 진행하였다. 코스피와 코스닥의 주식시장의 흐름을 상호 간 경쟁관계에 놓여있다는 가정 하에서 그 흐름을 모형화 하여 추정하는 연구를 진행하였다. 이 때 흐름의 변화는 결정적 (deterministic)이라는 가정 하에 화학 반응 모형을 상미분 방정식 (ordinary differential equation; ODE)에 적용한 후 비선형 모형에 대한 최소제곱법을 이용하여 모수를 추정하였다. Kim (2010)은 세계 출판시장 현황 자료를 이용하여 일반 종이책과 전자책 간의 경쟁관계를 화학 반응 모형을 이용하여 분석하였다. 이 연구 또한 상미분 방정식을 적용하여 최소제곱법을 이용하여 모형 추정을 진행하였다.

[†] 이 논문은 대구대학교 교내연구비로 지원받아 수행된 연구임 (No.20140330).
[1] (712-714) 경북 경산시 대구대로 201, 대구대학교 전산통계학과, 조교수. E-mail: bchoi@daegu.ac.kr

화학 반응 모형이 주로 적용되는 응용 분야 가운데 하나는 질병 확산 모형이다. Kermack와 McKendrick (1927)은 질병의 확산 및 소멸 과정을 설명하기 위하여 SIR (susceptible-infected-recovered) 모형을 제시하였다. 이는 전체 모집단을 세 개의 집단으로 나누어 이를 각각 감염 대상군 (susceptible), 감염군 (infected), 회복군 (recovered 또는 removed)으로 구분한 후 병에 걸리지 않은 사람이 감염 대상군으로부터 출발하여 감염군, 회복군으로 순차적으로 이동하는 과정을 화학 반응 모형을 이용하여 설명하고자 하였다. Hwang 등 (2007)은 SIR 모형을 이용하여 우리나라의 말라리아와 신증후군출혈열, 그리고 홍역 자료에 적용하여 모형 적합을 수행하였다. 또한 자료의 변동은 결정적이라는 가정하에 비선형 회귀식에 대한 최소제곱법을 적용하여 모수를 추정하였다. Ryu와 Choi (2015)는 SIR 모형의 구축 문제를 다루었는데 Gillespie (1977)가 제시한 Gillespie 알고리즘을 이용하여 추정된 모형에 대한 확률적 변동을 가정한 모형 구축을 시도하였고 그 결과를 우리나라 말라리아 자료에 적용하였다. SIR 모형은 질병 확산 과정을 설명하고자 하는 모형 가운데 기본이 되는 모형으로 보다 다양한 형태로 확장이 가능하다. SIRS (susceptible-infected-recovered-susceptible)모형은 회복군으로 넘어온 개체가 완전한 면역 상태에 놓이는 것이 아니라 다시 일정 시간이 지난 이후에 감염대상군으로 이동할 수 있음을 고려하는 모형이고 SEIR (susceptible-exposed-infected-recovered)모형은 감염대상군에서 감염군으로 이동하는 사이에 잠재기 (exposed)를 거쳐서 간다고 가정하는 모형이다. SIS (susceptible-infected-susceptible)모형은 질병으로부터 회복되었다 하더라도 다시 질병에 걸릴 수 있음을 가정하는 모형이다. 이와 같은 변형된 형태의 여러 질병 확산 모형들 모두 화학 반응 모형을 이용하여 설명할 수 있는 모형들이다.

생화학 분야나 생명과학 분야 뿐 만 아니라 다양한 분야에서 활용되고 있는 화학 반응 모형의 추정과 관련된 많은 연구들은 자료의 흐름이 결정적이라는 가정 하에서 상미분 방정식을 이용하여 최소제곱법에 기반한 추정 방법들이 대부분을 차지하고 있다. 화학 반응 모형을 이용하여 표현하고자 하는 종들의 움직임이 결정적이라고 가정하게 되면 그 모형이 단순해지고 모형을 추정하는데 상대적으로 쉬운 방법을 이용하여 구축할 수 있는 장점이 있다. 그러나 생태계나 자연 현상에서 존재하는 종들의 흐름이 결정적으로 움직인다는 가정보다는 확률적 (stochastic)으로 움직인다고 가정하는 것이 보다 적절할 수 있다 (Andersson과 Britton, 2000). 확률적 움직임을 가정한 경우 모형을 추정하는데 있어서 적절한 확률 모형을 가정하고 우도함수에 기반하여 모형을 추정할 수 있다. 그러나 자료의 제약 등에 의하여 최대우도추정이 불가능 한 경우가 발생할 수 있으며 이 때는 베이지안 방법을 적용하여 모수를 추정할 수 있다.

본 연구의 목적은 화학 반응 모형의 추정 방법에 대한 연구이다. 기본적으로 화학 반응 모형의 움직임이 확률적이라는 가정으로부터 출발하여 우도함수에 기반한 모형 추정문제를 다루고자 한다. Gillespie가 제안한 모의실험 방법에 의하여 확률적 화학 반응 모형을 구축하는 경우 (Gillespie, 1977), 관찰된 종들의 움직임은 연속 시간 마르코프 연쇄 (continuous time Markov chain)을 따른다고 가정한다. 그러나 관찰된 데이터들은 이산적인 시간에 따라 관찰되는 경우가 대부분이다. 그렇기 때문에 많은 경우에 데이터들은 불완전한 상태로 관찰된다. 기본적으로 추정하고자 하는 화학 반응 모형이 단순한 형태를 가진다면 우도함수를 이용한 최대우도추정으로 모수를 추정할 수 있다. 그러나 상대적으로 복잡한 모형을 가지는 경우 최대우도추정에는 여러 제약이 따르게 되며 이를 해결하기 위하여 베이지안 방법을 적용하여 MCMC를 이용한 모의실험을 통하여 모수를 추정하는 방법들이 제안되어 왔다. 베이지안 방법 또한 만족스러운 결과를 제공하고 있으나 이산적인 시간에 따라 관찰된 제한적인 데이터에 이용하여 모형을 추정하는 관계로 수렴 속도가 느리고 안정적이지 못한 결과를 보이는 경우가 있다. 본 연구에서는 이를 조금 더 개선하여 상대적으로 안정적으로 모수를 추정하는 방법을 제안하였다.

본 연구의 진행 순서는 다음과 같다. 제 2절에서는 화학 반응 모형의 정의와 함께 구현 방법에 대하여 설명한다. 3절에서는 확률적 가정 하에서 화학 반응 모형의 추정에 대하여 설명한다. 4절에서는 가상 자료를 이용한 모형의 구축 결과에 대하여 설명하면서 기존의 방법과의 비교 분석 결과를 제시하고자

한다. 마지막 5절에서는 본 연구의 한계점에 대하여 논하고자 한다.

2. 화학 반응 모형

2.1. 화학 반응 모형

확률적 화학 반응 모형은 다수의 종들과 종들 간의 상호작용으로 구성된 동태적 시스템을 의미한다. 동태적 시스템을 구성하는 종들의 변화는 연속시간 마르코프 연쇄를 따른다고 가정한다. 특정시점에서 관찰된 종들의 크기를 X라 할 때 시간의 흐름에 따라 벡터 X의 변화는 종들 간의 상호작용을 설명하는 반응(reaction)에 의하여 결정되어 진다. 따라서 화학 반응 모형은 종들과 종들 간의 상호작용을 설명하는 반응들로 구성되어 진다. 모형이 총 K개의 반응으로 구성되어 있다고 할 때 k번째 반응은 이 반응에 의해서 소멸되는 종의 수 ν_k와 생성되는 종의 수 ν'_k에 의해서 결정된다. 또한 각각의 반응은 위험함수 $h_k(x)$에 비례하여 발생한다고 가정한다. 시점 t에서 k번째 반응이 발생함으로써 종들의 변화는 다음의 식으로 표현된다.

$$X(t) = X(t-) + \nu'_k - \nu_k,$$

여기서 $(t-)$은 이전 시점의 극값을 나타낸다. 시점 t까지 번째 k반응이 발생한 총 수를 $R_k(t)$라 할 때

$$R_k(t) = Y_k\left(\int_0^t h_k(X(s))ds\right) \tag{2.1}$$

를 만족하며 여기서 Y_k는 독립적인 포아송 확률과정 (Poison process)를 따른다. 또한 위험함수 $h_k(x)$는 포아송 확률과정의 rate를 나타낸다. 이제 $X(t)$는 다음과 같은 방정식을 만족한다.

$$X(t) = X(0) + \sum_k R_k(t)(\nu'_k - \nu_k) = X(0) + \sum_k Y_k\left(\int_0^t h_k(X(s))ds\right)(\nu'_k - \nu_k) \tag{2.2}$$

식 (2.1)은 화학 반응 모형 안에서 보다 직관적으로 표현된다. 예를 들어 $A + B \xrightarrow{k} C$와 같이 표현될 수 있는데 이는 k번째 반응이 일어남으로써 종 A와 B의 개체가 하나씩 소멸하고 종 C의 개체가 하나 증가함을 의미한다. 이를 일반화하여 표현하여 보면 전체 모형이 v개의 종으로 구성되어 있고 각각의 종들을 $A_1, A_2, \ldots, A_v$이라 할 때

$$\sum_{i=1}^{v} \nu_{ik} A_i \to \sum_{i=1}^{v} \nu'_{ik} A_i, \quad k = 1, 2, \ldots, K \tag{2.3}$$

와 같이 표현된다.

2.2. Lotka - Volterra 모형

Lotka - Volterra 모형은 화학 반응 모형을 설명하는데 있어서 가장 대표적으로 쓰이는 모형이다. Lotka - Volterra 모형은 자연 생태계에서 피식자-포식자간의 먹이사슬과 적자생존 과정을 설명하기 위하여 제시된 모형으로 두 종인 피식자와 포식자 간의 상호작용을 통하여 각 종이 생성과 소멸과정을 설명하기 위한 모형이다. 모형 정의를 위하여 X_1을 먹이사슬관계에서 피식자라 하고 X_2를 포식자라고 하자. 이 모형은 식 (2.3) 형태의 반응 3개로 구성된 모형이다.

$$\begin{aligned} X_1 &\xrightarrow{\theta_1} 2X_1, & h_1(\boldsymbol{X}, \theta_1) &= \theta_1 X_1 \\ X_1 + X_2 &\xrightarrow{\theta_2} 2X_2, & h_2(\boldsymbol{X}, \theta_2) &= \theta_2 X_1 X_2 \\ X_2 &\xrightarrow{\theta_3} \emptyset, & h_3(\boldsymbol{X}, \theta_3) &= \theta_3 X_2 \end{aligned} \tag{2.4}$$

이 모형에서 각 반응식은 반응 상수 θ_1, θ_2, θ_3가 할당되어 있다. 실제 모형의 추정 문제에서 이 반응상수는 추정하여야 하는 모수가 된다. 모형 (2.4)에서 첫 번째식은 피식자의 생성 작용을 설명하는 반응이고, 두 번째 식은 피식자의 소멸과 포식자의 생성을 나타내는 반응이고, 마지막 식은 포식자의 소멸을 나타내는 반응이다. 예를 들어 두 번째 반응이 발생하게 되면 피식자의 개체 하나와 포식자의 개체 하나가 소멸되어 포식자의 개체 2개가 생성된다. 결과적으로 피식자는 한 개체가 감소하고 포식자는 한 개체가 증가한다. Figure 2.1은 Lotka - Volterra 모형에 대하여 Gillespie 알고리즘을 이용하여 생성한 그림이다. X_1과 X_2의 초기치는 각각 100과 50으로 하였으며 전체 관찰 시간은 50으로 하였다. 그리고 각 반응상수의 값은 $\theta_1 = 1$, $\theta_2 = 0.005$, $\theta_3 = 0.6$으로 할당 하였다. 윗쪽의 (a)가 두 종 X_1과 X_2의 전체 경로 (trajectory)를 나타낸다. 확률적 화학 반응 모형에서는 종들의 변화는 연속 시간 마르코프 연쇄를 따른다고 가정한다. 식 (2.1)과 같이 각 반응의 발생 건수는 포아송 확률과정을 따른다고 하였기 때문에 식 (2.4)의 각 반응식들은 위험함수 $h_1(\boldsymbol{X}, \theta_1)$, $h_2(\boldsymbol{X}, \theta_2)$, $h_3(\boldsymbol{X}, \theta_3)$에 비례하여 발생하게 된다. 확률적 화학 반응 모형을 구현하기 위한 Gillespie 알고리즘은 다음과 같다.

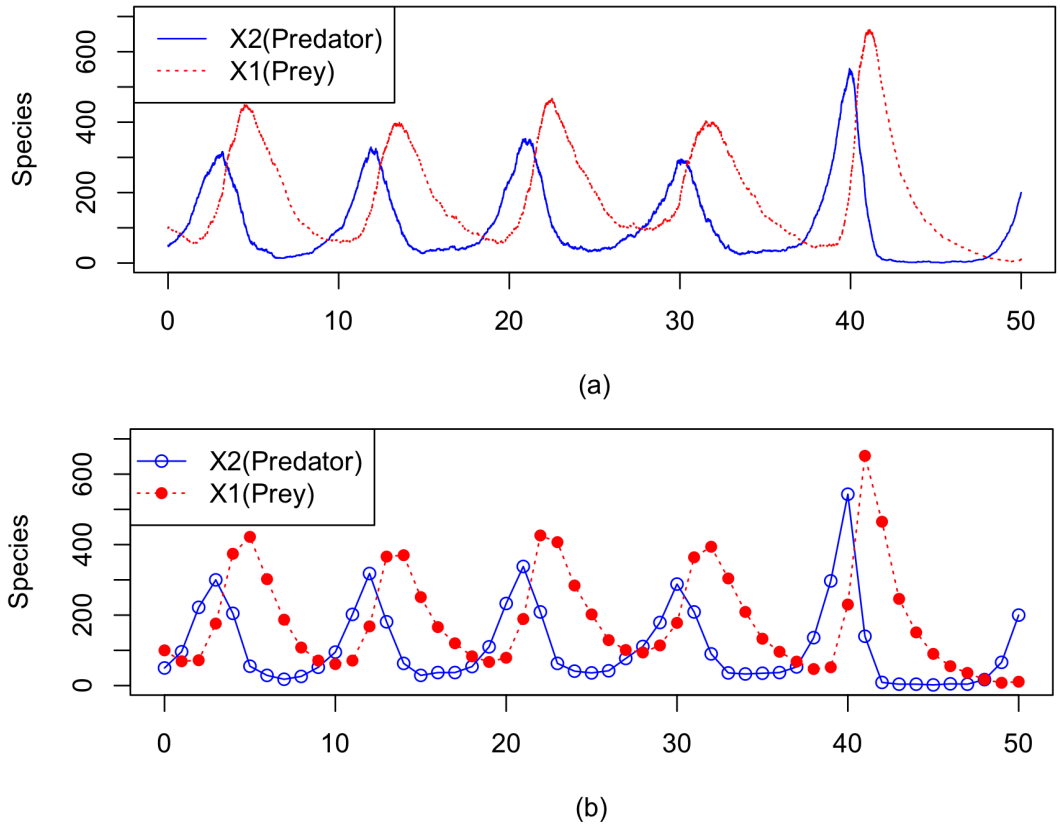

Figure 2.1 Stochastic chemical reaction trajectory for Lotka-Volterra model. (X_1 (dotted line) and X_2 respectively) with initial population are X_1 of 100 and X_2 of 50 and $\boldsymbol{\theta} = (1, 0.005, 0.6)$. Pannel (a) depicts the true series of species occur in the continues time points and (b) The actual sequence of reaction species can be observed on the discrete time points.

Gillespie 알고리즘

Step 1. 특정 시점에서 관찰된 종을 $\boldsymbol{X}(t)$라 할 때 이때의 위험함수 $h_k(\boldsymbol{x})$, $k = 1, \ldots, K$를 계산한다.

Step 2. 위험함수의 합 $h_0(\boldsymbol{x}) = \sum_k h_k(\boldsymbol{x})$를 모수로 하는 지수분포 $\exp(h_0)$로 부터 특정 반응이 일어날때 까지의 시간 t를 추출한다.

Step 3. 모형의 v개의 반응 가운데 어떠한 반응이 발생하였지를 추출한다. 이때 발생 확률은 위험함수에 비례하도록 $h_k(\boldsymbol{x})/h_0(\boldsymbol{x})$로 한다.

Step 4. Step 2에서 추출된 시간의 총 합이 최종 시간 T에 도달할때 까지 Step 2와 Step 3을 반복한다.

3. 모형 추정 방법

3.1. 우도함수의 구축

먼저 Gillespie 알고리즘에 근거한 우도함수의 구축방법부터 설명한다. 확률적 화학 반응 모형의 우도함수는 두 가지 부분으로 구분된다. 첫 번째는 반응이 일어날때까지의 시간이고 두 번째는 어떠한 반응이 일어났는가에 대한 부분이다. 반응의 발생은 포아송 과정을 따른다고 하였기 때문에 특정 반응이 일어날때 까지의 시간은 포아송 과정으로부터 지수 분포를 따르게 된다. 이 때의 모수는 화학 반응 모형의 모든 위험 함수의 합 $h_0(\boldsymbol{X}, \boldsymbol{\theta}) = \sum_k h_k(\boldsymbol{X}, \theta_k)$를 가진다. 또한 k번째 반응은 확률 $h_k(\boldsymbol{X}, \theta_k)/h_0(\boldsymbol{X}, \boldsymbol{\theta})$을 가지는 이산 확률 분포를 가지게 된다. 전체 관찰 시구간이 $[0, T]$라 할 때 확률적 화학 반응 모형에 대한 우도함수는 다음과 같이 주어진다 (Boys 등, 2008; Choi와 Rempala, 2012; Seo와 Choi, 2015).

$$L(\theta|\boldsymbol{X}) = \prod_{i=1}^{n} h_{k_i}(x(t_{i-1}), \theta_{k_i}) \exp\Big(- \int_0^T h_0(x(s), \theta) ds \Big), \tag{3.1}$$

여기서 n은 전체 모형에서 발생한 반응의 총 합이 된다. 또한 모형 (2.4)에서 본 바와 같이 각각의 위험함수는 다음과 같이 추정해야 하는 반응상수와 이와 독립인 함수 $g_k(\cdot)$의 곱의 형태로 정리된다.

$$h_k(x, \theta_k) = \theta_k g_k(x) \quad k = 1 \ldots, K$$

이로부터 우도함수 (3.1)은 다음과 같이 반응 상수 θ_k에 따른 곱의 형태로 정리될 수 있다.

$$\begin{aligned} L(\theta|\boldsymbol{X}) &= \Big(\prod_{i=1}^{n} \theta_{k_i} g_{k_i}(x(t_{i-1})) \Big) \exp\Big(- \int_0^T \sum_{k=1}^{K} \theta_k g_k(x(s)) ds \Big) \\ &\propto \prod_{k=1}^{v} \theta_k^{r_k} \exp\Big(- \theta_k \int_0^T g_k(x(s)) ds \Big) \\ &= \prod_{k=1}^{v} L_k(\theta_k|\boldsymbol{X}), \end{aligned} \tag{3.2}$$

이 식에서 r_k는 관찰된 k번째 반응의 총 수를 나타낸다. 이로부터 각 반응상수에 대한 통계적 추론은 독립적인 우도함수 $L_k(\theta_k|\boldsymbol{X})$를 가지고 수행이 가능해 진다. 결과적으로 이 우도함수는 감마분포의 형태를 가지게 되며 이로부터 다음과 같은 반응 상수에 대한 최대우도 추정치를 구할 수 있다.

$$\hat{\theta}_k = \frac{r_k}{\int_0^T g_k(x(s)) ds}, \quad k = 1, \ldots, K. \tag{3.3}$$

또한 베이지안 방법에 의한 모수 추론도 가능하다. 이를 위해서 적절한 사전분포를 할당하여야 하는데 개별 반응 상수에 대한 우도함수가 감마분포 이기 때문에 공액 사전분포로 감마분포 $\Gamma(a_k, b_k),\ k = 1, \ldots, K$를 할당할 수 있다. 우도함수 (3.2)과 감마 사전분포로부터 사후분포는 다음과 같은 감마분포를 따르게 된다.

$$\theta_s|\boldsymbol{x} \sim \Gamma\left(a_k + r_k, b_k + \int_0^T g_k(x(s))ds\right), \quad k = 1, \ldots, K. \tag{3.4}$$

식 (3.3)를 계산하거나 사후분포 (3.4)로 부터 최대 사후 (maximum posterrior estimator) 또는 사후분포로 표본을 추출하여 베이지안 추정량을 계산하기 위해서는 추가적인 계산이 필요하다.

$\int_0^T g_k(x(s))ds$의 경우 사다리꼴 공식과 같은 수치적분 방법을 이용하여 근사적으로 계산이 가능하다. 하지만 r_k는 전체 관찰구간에서 발생한 k번째 반응의 총 수를 나타내는데 실제 자료로부터 관찰하는데 한계가 있다. 구축한 화학 반응 모형이 상대적으로 간단한 경우 r_k가 쉽게 계산되는 경우가 있다. Seo와 Choi (2015)는 우리나라의 신종플루 감염자 자료에 대하여 SIR 모형을 적용하였다. 약간의 가정과 제약 조건을 통하여 r_k를 직접적으로 구하여 사후분포 (3.4)로부터 직접적으로 반응상수의 추정치를 계산 하였다. 그러나 복잡한 화학 반응 모형에서는 정확한 r_k를 구하는 것이 불가능하다. Figure 2.2의 아랫쪽 그림 (b)를 살펴보자. 이론적으로 화학 반응 모형은 연속시간 마르코프 연쇄를 따르기 때문의 윗쪽 그림 (a)의 형태로 구현이 가능하다. 그러나 실제로 관찰이 되는 경우에는 그림 (b)와 같이 이산적인 시간에서만 관찰이 가능하다. 그림 (b)는 Gillespie 알고리즘을 통해 구현된 Lotka-Volterra 모형의 전체 경로 가운데 단위 시간에서 종의 변화 값들을 찾아 점과 원으로 표시한 것이다. 따라서 관찰된 시점 사이에서 X_1과 X_2가 정확하게 어떻게 변화하였지 알 수 없다. 즉 사후분포로부터 반응 상수를 추출하는 베이지안 방법에서 추가적으로 정확하게 관찰되지 않은 r_k의 값을 계산하는 부분이 추가 되어야 한다. 즉 관찰되지 않은 일종의 결측값을 대체하는 부분이 추가되어야 한다. 본 연구에서는 이를 위하여 Boys 등 (2008)에 의하여 제시된 방법을 기반으로 하여 상대적으로 안정적인 결과를 주는 개선 방법을 이용하였다. Boys 등 (2008)은 Lokka - Volterra 모형을 이용하여 이 문제를 베이지안 방법을 제안하였다. 그들이 제안한 방법은 일종의 block updating method (Shephard와 Pitt, 1997; Liechty와 Roberts, 2001)를 이용하는 방법이다.

3.2. 베이지안 방법을 이용한 모수 추정

베이지안 방법을 이용한 모수 추정 방법에 대하여 알아보자. Figure 2.1의 아랫쪽 그림 (b)와 같이 자료가 관찰되어 있다고 가정할 때 전체 시구간이 자료가 관찰된 시점에 따라 단위시간별로 나뉜다. 그리고 이 나누어진 시구간에 따라 독립적으로 각 반응을 수를 추출한다. 예를 들어 구분된 시구간 가운데 하나를 $[j, j+1]$이라 하면 이 구간의 시작 시점인 j와 끝 시점의 $j+1$에서는 각 종의 크기가 주어져 있다. Lotka - Volterra 모형의 경우 이 구간내에서 첫 번째 반응의 발생 숫자만 결정 되면 나머지 두 반응의 숫자는 자동적으로 결정된다. 이와 같은 과정을 나누어진 모든 시 구간에 독립적으로 시행하여 생성된 반응의 수를 합쳐 전체 반응의 수가 결정된다. j번째 구간에서 생성하여야 하는 첫 번째 반응의 수를 r_{1j}라 하자. r_{1j}를 추출하기 위한 제안분포로 random walk chain 방법을 이용하였다. MCMC simulation 과정에서 현 시점의 값을 $r_{1j}^{(m)}$이라 할 때 random walk chain 방법을 이용한 Metropolis-Hastings 알고리즘은 다음과 같다.

알고리즘 1

Step 1. $\lambda = 1 + r_{1j}^{(m)}/c$를 모수로 하는 포아송분포로부터 독립적으로 두개의 난수를 추출하고 그 차를 구한다. 이를 y라 할 때 후보 값 $r_{1j}^{(*)} = r_{1j}^{(m)} + y$를 계산한다. 여기서 c는 Metropolis - Hastings

알고리즘의 조정상수 (tunning constant) 역할을 한다. 제안 분포 역할을 하는 y의 분포는

$$p(y) = \exp(-2\lambda)I_y(2\lambda)$$

이며 여기서 $I_y(\cdot)$은 Bessel 함수이다 (Johnson와 Kotz, 1969).

Step 2. 추출된 각 반응의 수는 포아송 분포를 따른다. 이 때 포아송 분포의 모수는 $\mu = \{h_1(\boldsymbol{x}(j), \theta_1)+ h_1(\boldsymbol{x}(j+1), \theta_1)\}$로 계산된다. 그리고 Step 2에서 제시된 제안 분포를 이용하여 다음의 식을 추출된 난수에 대한 채택확률 $\min(1, A)$을 계산한다.

$$A = \frac{\exp(-2\lambda^{(*)})I_y(2\lambda^{(*)}) \cdot \mu^{r_{1j}^{(*)}}/(r_{1j}^{(*)}!)}{\exp(-2\lambda^{(m)})I_y(2\lambda^{(m)}) \cdot \mu^{r_{1j}^{(m)}}/(r_{1j}^{(m)}!)}.$$

여기서 $\lambda^{(m)} = 1 + r_{1j}^{(m)}/c$, $\lambda^{(*)} = 1 + r_{1j}^{(*)}/c$이 된다.

Step 3. 균일분포 $uniform(0, 1)$으로부터 표본을 추출하여 이 값이 Step 2에서 계산된 확률보다 작을 경우 Step 1에서 추출된 $r_{1j}^{(*)}$를 새로운 표본으로 대체하고 그렇지 않을 경우 $r_{1j}^{(m)}$을 다시 이용한다.

이와 같은 방법을 나누어진 모든 시구간 $[j, j+1]$, $j = 0, 1, \ldots, T-1$에서 수행한다. 이 작업이 끝나게 되면 관찰되지 않은 전체 반응수를 추출할 수 있게 된다.

이제 베이지안 방법을 이용하여 반응 상수를 추출하기 위한 전체적인 방법은 다음과 같은 Gibbs 표본 추출 알고리즘을 이용한다.

알고리즘 2

Step 1. 반응상수 θ_k, $k = 1, 2, \ldots, K$에 대한 적절한 초기치를 추출한다.

Step 2. Algorithm 1을 이용하여 r_k, $k = 1, 2, \ldots, K$를 추출한다.

Step 3. 식 (3.4)의 감마분포로부터 반응상수를 추출한다.

Step 4. 반응상수가 수렴할 때까지 Step 2, Step 3을 반복 수행한다.

3.3. 평활 방법을 이용한 모수 추정 방법

3.2절에서 소개한 방법은 복잡하지 않은 화학 반응 모형에서는 상대적으로 좋은 결과를 제시한다. 하지만 관찰되지 않은 종들의 경로를 직접 추출하지 않고 반응의 수를 추출하는 근사적인 방법이라 할 수 있다. Choi와 Rempala (2012)는 각 반응의 발생 수를 추출하는데 있어서 Metropolis - Hastings 알고리즘 대신에 uniformaization 방법 (Hobolth와 Stone, 2009; Rodrigue 등, 2008)을 이용하여 Algorithm 2의 Step 2에서 제시된 방법 대신에 관찰되지 않은 종들의 변화와 종들의 발생 시간을 직접 추출하는 방법을 제안하였다. 이들이 제시한 방법은 Boys 등 (2008)의 방법보다 복잡하지만 종들의 경로를 직접 추출하므로 보다 복잡한 화학 반응 모형에서 모형 적합이 잘 되는 경향을 보이고 있다. 또한 모의실험을 통하여 제안하고 있는 방법에 우수함을 보였다. 하지만 이들의 연구는 전체적인 계산시간이 오래 걸리고 매우 거대한 행렬 연산이 요구되는 단점이 있다.

본 연구에서는 Boys 등 (2008)이 제안한 방법을 조금더 보완하여 상대적으로 안정적인 결과를 제공하는 방법을 제안하고자 한다. 식 (3.3)의 최대우도 추정량은 또한 다음과 같은 추정 함수의 적률 추정량이 된다.

$$\int_0^T dR_k(t) - \theta_k \int_0^T g_k(\boldsymbol{x}(s))ds = 0, \ k = 1, \ldots, K$$

이로부터 추정량을 계산하기 위해서는 마찬가지로 $R_k(T)$와 $\int_0^T g_k(\boldsymbol{x}(s))ds$의 추정량을 구해야 하는데 $\int_0^T g_k(\boldsymbol{x}(s))ds$은 최대우도추량을 구할 때 사용하였던 방법을 그대로 이용하여 사다리꼴 공식을 이용하여 같은 값을 계산한다. $R_k(T)$를 계산하기 위해서 MCMC 과정에서 추출된 값들을 이용하여 다음과 같은 가중 평균값을 계산한다.

$$R_k(T)^{(l)} = \begin{cases} r_k^{(0)} & \text{if } l = 0 \\ \alpha r_k^{(l)} + (1-\alpha) R_k(T)^{(l-1)} & \text{if } l > 0 \end{cases}$$

MCMC 과정 중에 생성되는 $r_k^{(l)}$을 일종의 시계열 자료처럼 고려하여 지수평활법에서 적용하는 평활 상수를 이용한 후 가중평균을 계산하였다. 현 시점에서 추출된 $r_k^{(l)}$과 직전 시점에서 계산 평활된 통계량 $R_k(T)^{(l-1)}$의 가중 평균을 계산하여 $R_k(T)^{(l)}$의 추정량을 최신 값으로 대체한다. 평활 상수 α를 조절함으로서 평활의 정도를 결정할 수 있다. 이 값이 1에 가까우면 최근 생성한 값에 더 많은 가중치를 주게 되며 0에 가까울 수록 평활의 효과가 없어지게 된다.

4. 모의실험

4.1. Lotka-Volterra 모형

이제 모의실험 자료를 이용하여 본 연구에서 소개하고 있는 방법들 간에 비교 분석을 수행하여 보자. 우선 Gillespie 알고리즘 방법을 이용하여 Lotka-Volterra 모형 (2.4)를 생성하였다. 3개의 반응에 반응상수는 각각 $\theta_1 = 1$, $\theta_2 = 0.005$, $\theta_3 = 0.6$으로 하였고 X_1과 X_2의 초기치는 각각 100과 50으로 하였다. 즉 $X_1(0) = 100$, $X_2(0) = 50$가 된다. 전체 관찰 시간은 $T = 50$으로 하였다. 기본적으로 Gillespie 알고리즘을 수행하여 데이터를 생성하면 Figure 2.1의 윗쪽 그림 (a)와 같은 결과를 얻게 된다. 그러나 실제 상황에서는 이와 같은 연속적인 시간에 자료가 관찰될 수 없다. 이를 위하여 단위시간의 관찰값을 취하면 아랫쪽 그림 (b)와 같은 자료를 얻을 수 있으며 이 자료가 관찰되었다고 가정하고 이 모의 실험 자료를 이용하여 모형 적합을 수행하였다.

모형 적합을 위하여 3장에서 소개한 MCMC 방법과 지수평활법을 추가한 방법을 적용하여 모형 적합을 시도 하였다. 반응 상수의 사전분포에 대한 초 모수로 각각 $a_1 = a_2 = a_3 = 0.001$, $b_1 = b_2 = b_3 = 0.001$을 할당 하였다. Matropolis-Hastings 알고리즘을 위한 조절 상수로 $c = 120$로 하였다. 총 1,000번의 반복 수행을 하였으며 이 가운데 처음 500번을 제외하고 후반부 500개만을 가지고 모수의 추정치를 구하였다.

다음 Table 4.1은 MCMC 방법에 의한 1,000번의 반복 수행 결과 가운데 후반부 500번의 결과를 정리한 표이다. 첫 번째 줄은 기존의 Boys 등 (2008)의 MCMC 방법에 따른 결과이고 두 번째 줄은 본 연구에서 제시한 방법에 의한 결과이다. 각각 MCMC1과 MCMC2로 표시하였다. 사후분포로부터 추출된 표본의 평균과 표준편차를 계산한 것이다. 평균값을 보면 모두 모수와 큰 차이가 없음을 볼 수 있다. 매우 정확한 추정 결과를 제공한다. 표준편차를 비교하였을 때 두 번째 방법이 상대적으로 작은 표준편차를 보여 주고 있다. 보다 안정적인 추정 결과를 제공한다고 할 수 있다.

Table 4.1 Posterior means and standard deviation (in parenthesis) of MCMC simulation for Lotka-Volterra model

	θ_1	θ_2	θ_3
MCMC1	1.043 (0.0206)	0.005 (0.0001)	0.621 (0.012)
MCMC2	1.038 (0.0065)	0.005 (0.00003)	0.612 (0.003)

다음 Figure 4.1은 1000번의 반복에 따른 시도표를 나타낸다. Table 4.1과 마찬가지고 윗 쪽 그림은 MCMC1, 아래 쪽 그림은 MCMC2를 나타낸다. 각 그림에서 맨 윗 쪽은 선부터 차례로 θ_1, θ_3, θ_2를 나타낸다. 1000번의 반복에서 모두 200번 이후 안정적으로 수렴하고 있는 모습을 보이고 있다. 두 방법간에 비교를 하였을 때의 Table 4.1과 마찬가지로 MCMC2의 방법이 보다 안정적인 결과를 제공하고 있다.

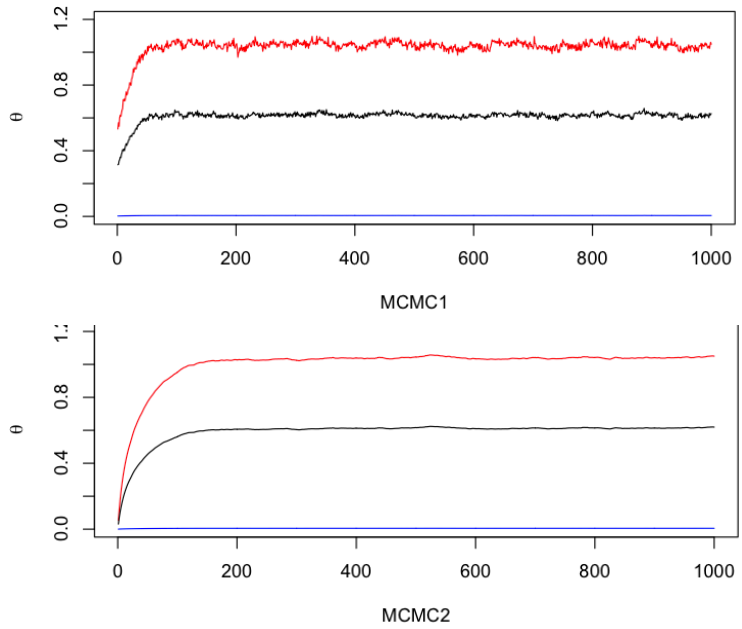

Figure 4.1 Plots of traces of converged and thinned MCMC output of reaction constants for Lotka - Volterra model

4.2. L1 retrotransposition 모형

L1 retrotransposition 모형은 유전자 전사 과정을 설명하기 위한 모형으로 다음과 같은 4개의 반응으로 구성된 모형이다 (Rempala 등, 2006).

$$\begin{aligned} NN &\xrightarrow{\theta_1 TProt} mRNA, \\ mRNA &\xrightarrow{\theta_2} \emptyset \\ \emptyset &\xrightarrow{\theta_3} TProt \\ TProt &\xrightarrow{\theta_4} \emptyset \end{aligned} \tag{4.1}$$

이 모형 (4.1)은 3개의 종으로 구성되어 있다. 모형에서 NN으로 표시한 유전체 DNA (gDNA), 촉매 단백질 (TProt), 메신저 RNA (mRNA)로 이 가운데 gDNA는 모형 내에서 종의 크기가 변하지 않는다고 가정한다. 따라서 실제 이 모형은 2개의 종과 4개의 반응으로 구성되어 있다고 볼 수 있다.

모형 (4.1)에서 첫 번째 반응은 mRNA의 생성을 나타내는 반응으로 TProt를 촉매제로 하여 mRNA가 생성된다. 세 번째 반응은 PTrot의 자가 생성을 나타내는 반응이다. 두 번째와 네 번째 반응은 각각

mRNA와 TProt의 분해 (degraduation)를 나타내는 반응이다. 각각의 반응에는 반응 상수 θ_1, θ_2, θ_3, θ_4가 할당되어 있다. 시점 t에서의 종의 상태 $X(t) = (mRNA(t), TProt(t))$는 $t \to \infty$에 따라 다음의 값을 수렴한다.

$$X(\infty) = (\frac{\theta_1\theta_3}{\theta_2\theta_4}, \frac{\theta_3}{\theta_4}) \tag{4.2}$$

다음 Figure 4.2은 Gillespie 알고리즘을 이용하여 모형 (4.1)을 구현한 것이다. mRNA와 TProt의 초기값은 모두 1로 하였으면 반응상수값은 $\theta_1 = 0.575$, $\theta_2 = 0.25$, $\theta_3 = 6.25$, $\theta_4 = 0.25$로 설정하였다. 마지막으로 전체 시구간은 $T = 1,000$으로 하였다. 그림 위쪽의 선이 mRNA의 경로를 나타내며 아래쪽의 선이 TProt의 경로를 나타낸다. Lotka-Volterra 모형과 달리 매우 빠른 변동을 가지고 있음을 볼 수 있다. 각 시도표의 가운데를 지나는 직선은 mRNA와 TProt의 수렴값을 나타내는데 각각 57.5와 25이다.

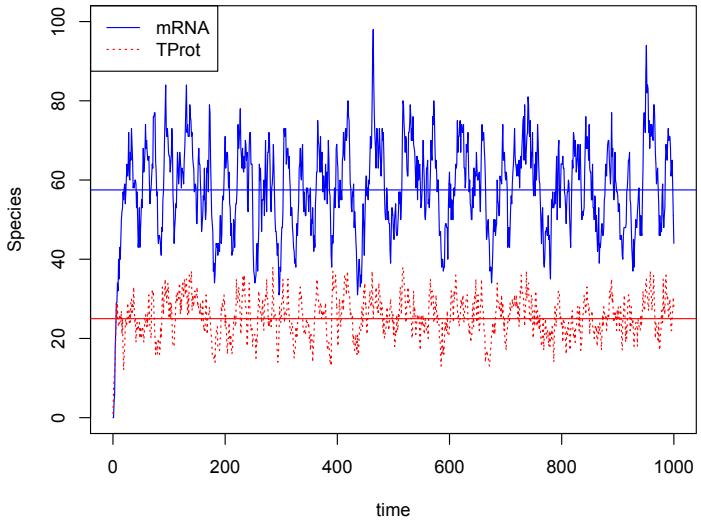

Figure 4.2 Simulated path for gene transcription model using Gillespie algorithm with reaction constants of $\theta_1 = 0.575$, $\theta_2 = 0.25$, $\theta_3 = 6.25$, $\theta_4 = 0.25$ and initial values of mRNA=1, TProt=1.

MCMC 방법을 통한 모수 추정 결과를 살펴보자. 총 20,000번의 반복 실행을 하였고 이 가운데 처음 10,000번을 burning set으로 제외하고 후반부 10,000번을 이용하여 모수 추정을 수행하였다. 다음 Table 4.2는 사후평균과 표준편차를 정리한 것이다. Lotka-Volterra 모형과 마찬가지로 사후평균에는 큰 차이가 없으나 표준편차에서 본 연구에서 제시된 방법이 상대적으로 작은 표준편차를 보이고 있다. Figure 4.2에서 살펴볼 수 있는 바와 같이 전반적으로 종의 흐름이 더 빠르고 불 안정한 것으로 볼 수 있다. 또한 반응상수에 대한 추정 문제에서도 2개의 종의 가지고 4개의 반응상수를 추정함으로 전반적으로 모수의 추정이 쉽지 않은 것을 볼 수 있다. 그러나 mRNA와 TProt의 수렴값을 계산해 보면 MCMC1의 경우 58.49, 25.50이고 MCMC2의 경우 58.53, 25.52로 참값과 큰 차이가 없는 것을 볼 수 있다.

추정결과를 그림을 통하여 살펴보자. Figure 4.3은 MCMC 과정의 20,000 반복에 대한 반응상수의 시도표를 나타낸다. 위쪽의 그림은 기존의 MCMC 방법에 의한 결과이고 아래 쪽 그림이 본 연구에

서 제시하고 있는 방법의 결과이다. 본 연구의 결과가 상대적으로 안정적인 결과를 제시 하고 있다. MCMC 과정에서의 수렴여부를 확인하기 위하여 potential scale reduction factor (Gelman and Rubin, 1992)를 계산하였다. 20,000번의 반복 실행에서 가장 큰 반응상수에 대한 값이 1.04로 계산되어 모두 1.1 이하가 되었다. MCMC 과정에서 수렴에 큰 문제가 없음을 확인할 수 있었다.

Table 4.2 Posterior means and standard deviation (in parenthesis) of MCMC simulation for gene transcription model

	θ_1	θ_2	θ_3	θ_4
MCMC1	0.4778 (0.0168)	0.2083 (0.0074)	5.5137 (0.2417)	0.2162 (0.0095)
MCMC2	0.4719 (0.0035)	0.2057 (0.0015)	5.4706 (0.0346)	0.2144 (0.0013)

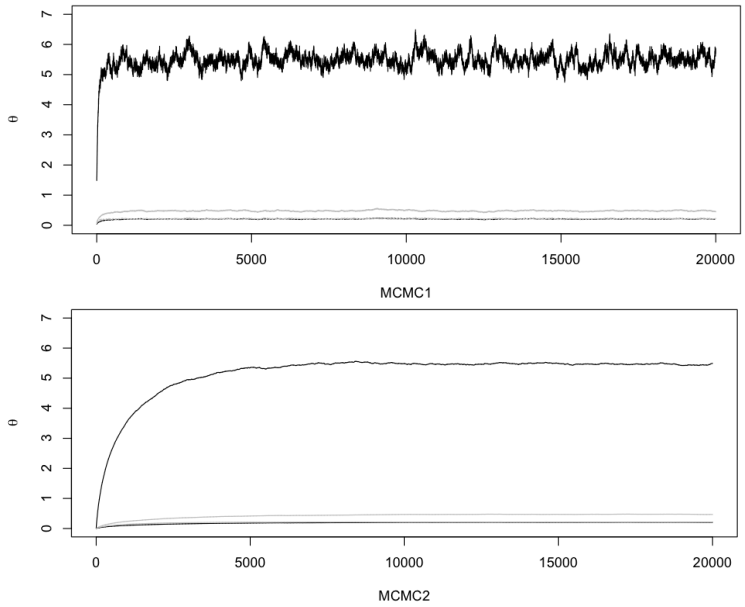

Figure 4.3 Plots of traces of converged and thinned MCMC output of parameter for gene trnascription model

5. 결론

본 연구에서는 확률적 화학 반응 모형에 대한 소개와 함께 우도함수에 기반한 추정방법을 소개하였다. 지금까지의 화학 반응 모형의 구축과 관련된 연구들이 대부분 종들의 흐름이 결정적인 방법에 따라 움직인다라고 가정을 한 후 상미분 방정식을 이용하여 모형을 구축하고자 하는 방법이 주로 제안되어 왔다. 이 때 모수의 추정은 최소제곱법을 사용하여 수행되었다. 본 연구에서는 이와는 다르게 화학 반응 모형의 움직임이 결정적이지 않고 확률적으로 움직인다고 가정 한 후에 확률적 화학 반응 모형을 구축하고자 하였다. 먼저 모형의 구축을 위하여 Gillespie 알고리즘을 이용하여 모형을 구축하고자 하였다. 화학 반응 모형에서 모형의 추정 문제는 곧 모수의 역할을 하는 반응 상수의 추정 문제로 귀결될 수 있는데 본 연구에서는 Gillespie가 제안한 방법에 근거하여 우도 함수를 구축한 후 우도 함수에 기반한 모수 추정을 수행하고자 하였다. 이 때 이산적으로 관찰되는 자료의 한계를 극복하기 위하여 베이지안 접근법에

기반한 MCMC 방법을 이용하여 최종적으로 모수의 추정을 수행하였다. 기존의 Metropolis-Hastings 방법을 이용한 모형의 추정 방법을 소개하였고 기존의 방법을 조금 더 개선하여 상대적으로 안정적인 추정 결과를 제시하는 방법을 제안하였다. 제 안된 방법을 이용하여 모수의 추정 방법을 비교하기 위하여 Lotka-Volterra 모형과 유전자 전사 모형을 본 연구에서 제시한 방법을 적용하여 보았다. 모형에서 모두 본 연구에서 제안하고 있는 방법이 상대적으로 안정적인 결과를 보였다.

기본적으로 본 연구에서 제시하고 있는 방법들은 일종의 근사적인 방법으로 상대적으로 간단한 화학 반응 모형에서는 추정 결과가 좋으나 복잡한 모형에서는 추정 결과가 좋지 않은 단점을 지니고 있다. 이를 보완할 수 있는 방법으로 Uniformization 방법을 이용한 추정 방법이 제시되고 있으나 매우 큰 행렬의 거듭 제곱 연산을 수행하여야 하는 단점이 있다. 향후 연구를 통하여 이러한 단점들을 보완할 수 있는 방법이 제시될 수 있을 것이다.

본 연구에서 소개하고 있는 확률적 화학 반응 모형은 생화학 분야 뿐만 아니라 질병의 확산 과정을 설명하고자 하는 모형에 많이 적용되고 있다. 이 뿐 만 아니라 사회과학 분야나 경영학 분야에서 상호 경쟁 또는 협력 관계에 있는 기업 또는 집단의 문제에도 적용할 수 있을 것이다. 그 확장 가능성이 매우 큰 모형이라 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. L. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, **18**, 125-135.

Choi, B. and Rempala, G. A. (2012). Inference for discretely observed stochastic kinetic networks with applications to epidemic modeling. *Biostatistics*, **13**, 153-165.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457-472.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, **81**, 2340-2361.

Hobolth, A. and Stone, E. A. (2009). Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *Annals of Applied Statistics*, **3**, 1204-1230.

Hwang, N. A., Jeong, B. Y., Lim, Y. C. and Park, J. S. (2007). Diseases data analysis using SIR nonlinear regression model. *Journal of the Korean Data Analysis Society*, **9**, 49-59.

Johnson, N. L. and Kotz, S. (1969). *Discrete distribution. In: Distributions in Statistics*, Vol. 1, Wiley, New York.

Kermack, W. O. and McKendrkck, A. G. (1991). Contributions to the mathematical theory of epidemic-I (Reprint from the 1927 original). *Bulletin of Mathematical Biology*, **53**, 33-55.

Kim, K. (2010). An analysis on the competition patterns between Paper-book and E-book using the Lotka-Volterra model. *Journal of the Korea Academia-Industrial cooperation Society*, **11**, 4766-4773.

Lee, S. J., Lee, D. J. and Oh, H. (2003). A dynamic analysis on the competition relationships in Korean stock market using Lotka-Volterra model. *Journal of Korean Institute of Industrial engineers*, **29**, 14-20.

Lee, S. J., Oh, H. and Lee, D. J. (2002). An analysis on the competition between KOSPI and KOSDAQ using the Lotka-Volterra model. *Proceedings of Korean Institute of Industrial engineers*, 1052-1058.

Liechty, J. C. and Roberts, G. O. (2001). Markov chain Monte Carlo methods for switching diffusion models. *Biometrika*, **88**, 299-315.

Rempala, G. A., Ramos, K. S. and Kalbfleishe, T. (2006). A stochastic model of gene transcription: An application L1 retrotransposition events. *Journal of Theoretical Biology*, **242**, 101-116.

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. R. and Choi, B. (2015). Development of epidemic model using the stochastic method. *Journal of the Korean Data & Information Science Society*, **26**, 301-302.

Seo, M. O. and Choi, B. (2015). An estimation method for stochastic epidemic model. *Journal of the Korean Data Analysis Society*, Submitted.

Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, **84**, 653-667.

Wilkinson, D. J. (2012). *Stochastic modelling for systems Biology*, 2nd Eds., CRC Press, Boca Raton.

Journal of the Korean Data & Information Science Society
2015, **26**(4), 813–826

http://dx.doi.org/10.7465/jkdi.2015.26.4.813
한국데이터정보과학회지

An estimation method for stochastic reaction model[†]

Boseung Choi[1]

Department of Statistics and Computer Science, Daegu University

Received 18 April 2015, revised 11 May 2015, accepted 3 June 2015

Abstract

This research deals with an estimation method for kinetic reaction model. The kinetic reaction model is a model to explain spread or changing process based on interaction between species on the Biochemical area. This model can be applied to a model for disease spreading as well as a model for system Biology. In the search, we assumed that the spread of species is stochastic and we construct the reaction model based on stochastic movement. We utilized Gillespie algorithm in order to construct likelihood function. We introduced a Bayesian estimation method using Markov chain Monte Carlo methods that produces more stable results. We applied the Bayesian estimation method to the Lotka-Volterra model and gene transcription model and had more stable estimation results.

Keywords: Gene transcription model, Gillespie algorithm, Lotka-Volterra model, MCMC, stochastic kinetic reaction model.

[†] This research is supported by Daegu University Research Grant in 2014 (No.20140330).
[1] Assistant professor, Department of Statistics and Computer Science, Daegu University, Gyeongbuk 712-714, Korea. E-mail: bchoi@daegu.ac.kr