

계수과정의 우도함수 유도

오창혁¹

¹영남대학교 통계학과

접수 2013년 12월 19일, 수정 2014년 1월 6일, 게재확정 2014년 1월 12일

요약

계수과정은 다양한 분야에서 활용되고 있으며 그 성질은 강도함수에 의해 결정된다. 일정 구간에서 연속적으로 과정이 관측될 때, 우도함수를 이용하여 강도함수의 모수를 추정할 수 있다. 그러나 기존의 연구는 직관적인 방법에 의한 우도함수 유도이며, 여러 명의 저자에 의해 얻은 우도함수가 일치하지 않아 우도함수를 이용한 최우추정치를 구하는 문제 등의 적용에 어려움이 발생하고 있다. 따라서 이 단신연구에서는 계수과정의 우도함수를 엄밀한 방법으로 유도하여 기존의 문제점을 해결한다.

주요용어: 계수과정, 우도함수, 통계물리적 개념.

1. 서론

지난 한 세기 동안 전염병의 전염 현상을 수리적으로 모형화하기 위한 많은 연구가 이루어졌다. 이에 관한 자세한 역사적 고찰은 Bailey (1975)와 Anderson과 May (1991)에 나타나 있으며, 계수과정의 하나인 확률적 전염모형은 McKendrick (1926)에 의해 처음으로 제시되었다. 전염병의 확률모형은 기본적으로 모집단의 구성원을 질병 감염 상태에 따라 분류하며, 시간의 흐름에 따라서 각 상태가 발생하는 횟수를 계수하는 계수과정이다. 전염병모형 중 많은 연구와 적용이 이루어진 일반화 확률전염병모형 (the general stochastic epidemic model; 확률 SIR 모형)은 민감 (susceptible), 감염 (infective), 회복 (recovered)의 세 가지 상태를 차례로 겪는 모형으로 후천성 면역결핍증, 홍역, 구제역 등의 질병 외에도 많은 질병에 관한 모형화에 사용된다. 확률전염병모형의 질병 전염의 실제 데이터에 대한 적용의 예는 Diekmann과 Heesterbeek (2000)에서 찾아 볼 수 있다.

확률 SIR 모형은 비선형성에 의하여 특정 시점에서의 감염 개체의 수에 대한 확률이 매우 복잡한 반복 계산 (recursive) 식의 형태로 나타나며 (Daley와 Gani, 1999, pp. 66-73 참조), 조건부 확률이나 평균과 분산 등의 폐수식이 얻어지지 않는다. 이러한 문제점에 의해 특정 시점에서의 감염개체 수에 대한 확률을 기본으로 하는 최우추정법은 가능하지 않다. 주어진 시구간 $(0, u]$ 에서 감염과정이 연속적으로 관측될 때, 즉, 모든 감염과 회복의 시점이 관측될 때, 확률 SIR 모형의 모수에 관한 우도함수의 유도를 위해 많은 시도가 이루어졌으며, 이산 시점에서 관측되었을 때, 모수에 관한 추정은 Cauchemez와 Ferguson (2008)과 Oh (2013) 등에 의해 이루어졌다. 우도함수를 이용한 추정 방법은 여러 분야에서 많은 연구가 이루어져 왔다 (Kang 등, 2012; Ha와 Cho, 2012).

확률 SIR 모형에 대한 우도함수를 살펴보기 위해 먼저 확률 SIR 모형을 소개한다. 확률 SIR 모형을 위하여 시점 $t = 0$ 에서 초기 감염 개체 수가 a , 초기 민감 개체 수가 μ 인 전체 크기가 $\mu + a$ 인 폐쇄 모집단을 생각하자. 이과정을 $(0, u]$ 에서 연속적으로 관측한다고 하자. $S(t)$ 는 시점 t 에서 민감 개체의 수,

¹ (712-749) 경북 경산시 대학로 280, 영남대학교 통계학과, 교수. E-mail: choh@yu.ac.kr

$X(t)$ 는 시점 t 까지 감염된 개체의 총수, $I(t)$ 는 시점 t 에서의 감염 개체의 수, $R(t)$ 시점 t 까지의 회복된 개체의 총수, β 는 감염율, γ 는 회복율, \mathcal{G}_t 는 역사 $\{S(\tau), I(\tau); 0 \leq \tau \leq t\}$ 에 의해 생성되는 시그마 알지브라이며, $n_I \leq \nu$ 와 $n_R \leq \nu$ 는 각각 시점 u 까지 감염된 개체 수와 회복된 개체 수를 나타낸다고 하자. Rida (1991)에 따르면 $S(t)$ 와 $I(t)$ 가 주어졌을 때, 미세구간 $(t, t+h]$ 에서 하나의 민감 개체가 감염될 확률을 다음과 같이 표현된다.

$$\begin{aligned} \Pr\{dX(t) = 1, dR(t) = 0 \mid \mathcal{G}_t\} &= \frac{\beta}{\mu} S(t)I(t)h + o(h), \\ \Pr\{dX(t) = 0, dR(t) = 1 \mid \mathcal{G}_t\} &= \gamma I(t)h + o(h), \\ \Pr\{dX(t) = 0, dR(t) = 0 \mid \mathcal{G}_t\} &= 1 - \frac{\beta}{\mu} S(t)I(t)h - \gamma I(t)h + o(h). \end{aligned} \quad (1.1)$$

여기서, 수정항 $o(h)$ 은 $h \rightarrow 0$ 이면 $o(h)/h \rightarrow 0$ 를 만족한다.

확률 SIR 모형에서 주어진 기간 $(0, u]$ 동안 전염 현상의 연속적 관측을 통해 모든 감염사건과 회복사건에 관한 자료를 얻은 경우에 모형의 모수인 감염율과 회복률의 추정엔 최우추정법으로 이루어 질 수 있으며, 이를 위한 우도함수와 최우추정치들 얻기 위한 다양한 시도가 이루어졌다. 이러한 이유는 SIR 모형에서의 확률의 계산이 매우 복잡하다는 것이다 (Bailey, 1975; Daley와 Gani, 1999).

구간 $(0, u]$ 에서 관측된 회복 시점을 $\tau = (\tau_1, \dots, \tau_{n_R})$, 감염 시점을 $\phi = (\phi_1, \phi_2, \dots, \phi_{n_I})$ 라고 할 때, Becker와 Britton (1999)은 Andersen 등 (1993), 402쪽의 계수과정을 이용하여 SIR 모형에서의 감염율과 회복율에 관한 다음의 우도함수를 제시하였다.

$$\prod_{j=1}^{n_R} (\gamma I(\tau_j)) \prod_{k=1}^{n_I} \left(\frac{\beta}{\mu} S(\phi_k) I(\phi_k) \right) \times \exp \left\{ -\frac{\beta}{\mu} \int_0^u S(t)I(t)dt - \gamma \int_0^u I(t)dt \right\}. \quad (1.2)$$

한편, Andersson과 Britton (2000)은 통계물리학에서의 방법을 이용하여 직관적이며 개략적인 방법으로 일반적인 다변량 계수과정의 모수에 관한 우도함수를 유도한 다음 이를 이용하여 SIR 모형의 우도함수를 구하였다.

$$\prod_{j=1}^{n_R} (\gamma I(\tau_j^-)) \prod_{k=1}^{n_I} \left(\frac{\beta}{\mu} S(\phi_k^-) I(\phi_k^-) \right) \times \exp \left\{ -\frac{\beta}{\mu} \int_0^u S(t)I(t)dt - \gamma \int_0^u I(t)dt \right\}. \quad (1.3)$$

식 (1.3)의 식 (1.2)과의 차이점은 앞의 두 곱항에서 관측시점의 좌극한을 사용한다는 것이며, 이는 확률 SIR 모형이 계수과정임을 고려하면 좌극한의 사용 여부는 매우 큰 차이를 나타낸다고 할 수 있다 한편, Kypraios (2009)는 주어진 구간 $[0, u]$ 내에서 전염병 현상이 종료된다는 가정 하에서 즉, 관측 구간동안 모든 감염 개체가 회복되어 더 이상의 감염이 진행되지 않는다는 조건 하에서 다음의 우도함수를 제시하였다.

$$\prod_{j=1}^{n_I} (\gamma I(\tau_j^-)) \prod_{k=2}^{n_I} \left(\frac{\beta}{\mu} S(\phi_k^-) I(\phi_k^-) \right) \times \exp \left\{ -\frac{\beta}{\mu} \int_{\phi_1}^u S(t)I(t)dt - \gamma \int_{\phi_1}^u I(t)dt \right\}. \quad (1.4)$$

우도함수 (1.4)과 우도함수 (1.3)의 차이점은 두 번째의 곱의 항이 $k = 2$ 에서 시작한다는 것과, 적분항의 하한이 0이 아닌 처음 감염 시점 ϕ_1 이라는 것이다. 관측 시작 시점이 0인 점을 고려하면 정보의 손실이 발생함을 알 수 있다. 또한, 감염의 총수와 회복의 총수가 같다. 즉, $n_R = n_I$ 이다. Kypraios (2009)는 우도함수 (1.4)를 이용하여 초기 민감 개체의 수를 미지의 값으로 하여 우도함수를 수치적인 방법으로 최대화하는 값을 찾아 최우추정치를 구하였다. 초기 민감 개체 수는 Huggins 등 (2004)이 마팅계일 방법을 적용하여 추정되었으나, 씨물레이션 결과에 의하면 특정한 모수값을 제외한 부분에서 안정적이지 못한 단점이 발견되었고, Kypraios (2009)은 자신의 최우추정량이 마팅계일 추정량보다 우월함을 보였으나, 이 또한 불안정한 씨물레이션 결과를 가지고 있다. O'Neill과 Roberts (1999)은 구간 $[0, u]$ 에 전염 현상을 관측할 때, 이 구간에서 처음으로 감염이 일어난 시점을 관측의 시작점으로 한 우도

함수를 소개하였는데 이는 Kypraios (2009)의 (1.4)의 우도함수와 같고, Höhle과 Jørgensen (2002)은 생존함수의 개념을 사용하여 이를 증명하였다. O'Neill과 Roberts (1999)의 우도함수는 구간 $[0, u]$ 에서 관측을 하지만 우도함수를 구할 때는 이 구간에서 처음으로 발생하는 감염 시점을 시작점으로 하여 우도함수를 구한다는 단점이 있다.

따라서, 본 단신연구에서는 먼저 Andersson과 Britton (2000)의 방법을 따라서 다변량 계수과정의 우도함수를 엄밀히 유도하여 이를 확률 SIR 모형의 우도함수를 구하는데 사용한다. Andersson과 Britton (2000)의 증명 과정에서 생략된 몇 가지 단계를 구체적으로 명시하고 주어진 구간을 세분화하는 데에 대한 효율적 방법을 제시하였다. 유도에는 기초적 수학적 방법을 사용하였다.

본 논문의 구성은 다음과 같다. 제2절에서는 다변량 계수과정을 주어진 구간에서 관측할 때의 우도함수를 유도한다. 제3절에서는 이를 이용하여 확률 SIR 모형의 우도함수를 유도하고 토의한다.

2. 다변량 계수과정과 우도함수

차수가 k 인 다변량 계수과정 $\{N(t)\}$ 를 생각한다. 정의에 의하여 성분 $\{N_i(t)\}$, $i \in \{1, \dots, k\}$ 와 합 $\{\sum_{i=1}^k N_i(t)\}$ 은 계수과정이고, 각 성분 과정 $N_i(t)$ 은 시점 t 까지 발생하는 해당 사건의 발생횟수로 해석될 수 있다. 시점 t 까지 전체 다변량과정의 역사 \mathcal{H}_t 가 주어졌을 때, 미세구간 $(t, t+h]$ 에서 이런 사건이 발생할 미세확률은 다음을 만족한다.

$$\begin{aligned} P(N_i(t+h) - N_i(t) = 1 | \mathcal{H}_t) &= h\lambda_i(t) + o(h), i = 1, \dots, k, \\ P(N(t+h) - N(t) = 0 | \mathcal{H}_t) &= 1 - h \sum_i \lambda_i(t) + o(h). \end{aligned} \quad (2.1)$$

단, $\lambda_i(t)$ 는 성분 계수과정 $N_i(t)$ 의 음이 아닌 함수인 강도함수이며, 시점 t 까지의 역사에 의존한다. 식 (2.1)은 짧은 시구간에서는 단지 k 개의 서로 다른 사건 중에 단 한가지만 발생할 수 있다는 즉, 해당되는 성분과정 한 개만 1만큼 증가한다는 것을 의미한다.

기존 연구에서 미세확률을 위한 미세구간의 형태가 다양하다. 예를 들면 Rida (1991)과 Andersson과 Britton (2000)은 $(t, t+h]$ 형태를 사용하며, Kypraios (2009)는 $[t, t+h)$ 의 형태를, 그리고 Huggins 등 (2004)는 $(t, t+h)$ 의 형태를 사용한다.

구간 $(0, u]$ 에서 계수과정을 연속적으로 관측한다고 가정하자. 여기서 시작 점 0은 구간에 포함되지 않고, 마지막 점 u 는 포함된다. 참고로 관측을 위해 Andersson과 Britton (2000), 그리고 Kypraios (2009)은 구간 $[0, u]$ 을, Schutt (2010)은 구간 $(0, u)$ 을 사용한다. 그러나, 시작점 0를 포함시키거나 혹은 마지막 점 u 를 제외시키게 되면 우도함수 유도에서 이들 점에 따른 수식을 조절하여야 한다.

구간 $(0, u]$ 를 ν 개의 등구간으로 나누고 각 세분 구간의 길이를 $h = u/\nu$ 라고 하자.

$$(0, u] = \cup_{j=0}^{\nu-1} (jh, (j+1)h].$$

각 세분 구간에 대하여 이 세분 구간 이전의 역사에 의존하여, 이 세분 구간에서 발생하는 사건들의 확률을 계산한다. Andersson과 Britton (2000)와 Schutt (2010)에서의 세분 구간의 형태는 $[t, t+h)$ 이고 Kypraios (2009)는 세분 구간의 형태를 특정하지 않고 우도함수의 결과만 제시하였다.

먼저 기호를 도입한다. $j = 0, 1, \dots, \nu - 1$ 에 대하여

$$\begin{aligned} \Delta N^\nu &= (\Delta N(0), \Delta N(h), \dots, \Delta N((\nu-1)h)), \\ \Delta N(jh) &= (\Delta N_1(jh), \Delta N_2(jh), \dots, \Delta N_k(jh)) \end{aligned}$$

라고 하자. 여기에서 $\Delta_{jh} = [jh, jh + h]$ 대신에 $\Delta_{jh} = (jh, jh + h]$ 을 생각하는 데 그 이유는, Andersson과 Britton (2000)에서처럼 $\Delta_{jh} = [jh, jh + h)$ 라고 두면 $\Delta N_i(jh)$ 는 구간 $(jh, jh + h]$ 내에서 발생한 사건을 계수하지 않기 때문이다.

주어진 관측에 대한 확률 즉, 우도는 다음과 같이 얻어진다.

$$\begin{aligned} & \mathcal{L}(\Delta N^\nu) \\ &= \prod_{j=0}^{\nu} \left(\prod_{i=1}^k \left((h\lambda_i(jh))^{\Delta N_i(jh)} \right) \left(1 - h \sum_{i=1}^k \lambda_i(jh) \right)^{1 - \sum_{i=1}^k \Delta N_i(jh)} + o(h) \right) \\ &= \prod_{j=0}^{\nu} \left(h^{\sum_{i=1}^k \Delta N_i(jh)} \prod_{i=1}^k \left((\lambda_i(jh))^{\Delta N_i(jh)} \right) \left(1 - h \sum_{i=1}^k \lambda_i(jh) \right)^{1 - \sum_{i=1}^k \Delta N_i(jh)} + o(h) \right). \end{aligned}$$

두 개의 큰 괄호 안의 곱은 Δ_{jh} 에서 관측된 것에 의존하여, 어느 확률을 선택할 것인가에 관한 것이다. 참고로 $k = 2$ 인 경우에 $(\Delta N_1(jh), \Delta N_2(jh)) = (1, 0)$, 혹은 $(0, 1)$, $(0, 0)$ 이 된다. 세분 구간의 길이 h 가 작아질수록, 세분 구간의 개수 ν 는 커지며 사건을 포함하게 되는 세분 구간을 제외한 거의 모든 세분 구간이 위 식의 두번째 큰 괄호에 포함되게 된다.

여기서 $h \approx 0$ 이면, 테일러 전개에 의해 $\left(1 - h \sum_{i=1}^k \lambda_i(jh) \right) \approx \exp\left(-h \sum_{i=1}^k \lambda_i(jh)\right)$ 가 되므로 다음이 성립한다.

$$\begin{aligned} & \mathcal{L}(\Delta N^\nu) \\ &= \prod_{j=0}^{\nu-1} \left(h^{\sum_{i=1}^k \Delta N_i(jh)} \prod_{i=1}^k \left((\lambda_i(jh))^{\Delta N_i(jh)} \right) \exp\left(-h \sum_{i=1}^k \lambda_i(jh)\right) + o(h) \right) \end{aligned} \quad (2.2)$$

$$= \prod_{j=0}^{\nu-1} \left(h^{\sum_{i=1}^k \Delta N_i(jh)} \prod_{i=1}^k \left((\lambda_i(jh))^{\Delta N_i(jh)} + o(h) \right) \exp\left(-h \sum_{i=1}^k \lambda_i(jh)\right) \right). \quad (2.3)$$

식 (2.3)은 Andersson과 Britton (2000)의 88쪽의 식과 일치하며, 식 (2.3)은 수렴의 증명을 위해 변형된 식이다. 사건의 발생하지 않는 경우인 $\Delta N_1(jh) = \dots = \Delta N_k(jh) = 0$ 를 제외하면 이 식은 ν 가 증가함에 따라 값이 감소하는데, 그 이유는 h 가 내부의 곱에 나타나기 때문이다. 이는 당연한 결과로서, 한 점에서 사건이 발생할 확률이 0이기 때문이다.

우도 계산을 위해, 각 성분 계수과정 N_i , $i = 1, \dots, k$, 이 r_i 개의 점프를 $t_{i,1} < \dots < t_{i,r_i}$ 에서 가진다고 하고, $r_1 + \dots + r_k = r$ 라고 하자.

$$\begin{aligned} & \mathcal{L}(\Delta N^\nu) \\ &= \prod_{j=0}^{\nu-1} h^{\sum_{i=1}^k \Delta N_i(jh)} \prod_{i=1}^k \left((\lambda_i(jh))^{\Delta N_i(jh)} + o(h) \right) \exp\left(-h \sum_{i=1}^k \lambda_i(jh)\right) \end{aligned}$$

$$\begin{aligned}
&= h^{\sum_{i=1}^k \Delta N_i(0 \cdot h)} \prod_{i=1}^k \left((\lambda_i(0 \cdot h))^{\Delta N_i(0 \cdot h)} + o(h) \right) \exp \left(-h \sum_{i=1}^k \lambda_i(0 \cdot h) \right)^{1 - \sum_{i=1}^k \Delta N_i(0 \cdot h)} \\
&\quad \times h^{\sum_{i=1}^k \Delta N_i(1 \cdot h)} \prod_{i=1}^k \left((\lambda_i(1 \cdot h))^{\Delta N_i(1 \cdot h)} + o(h) \right) \exp \left(-h \sum_{i=1}^k \lambda_i(1 \cdot h) \right)^{1 - \sum_{i=1}^k \Delta N_i(1 \cdot h)} \\
&\quad \times \cdots \\
&\quad \times h^{\sum_{i=1}^k \Delta N_i((\nu-1)h)} \prod_{i=1}^k \left((\lambda_i((\nu-1)h))^{\Delta N_i((\nu-1)h)} + o(h) \right) \\
&\quad \quad \quad \times \exp \left(-h \sum_{i=1}^k \lambda_i((\nu-1)h) \right)^{1 - \sum_{i=1}^k \Delta N_i((\nu-1)h)}.
\end{aligned}$$

표현의 편의를 위하여 $\mathcal{J}_i^\nu = \{j \mid t_{i,m} \in (jh, (j+1)h], j = 0, \dots, \nu-1; m = 1, \dots, r_i\}$, $i = 1, \dots, k$ 에 대하여 $\mathcal{J}^\nu = \cup_{i=1}^k \mathcal{J}_i^\nu$, $\mathcal{E} = \{0, \dots, \nu-1\}$ 라고 나타내자. 그러면 위의 식은 다음의 표현으로 간단히 할 수 있다.

$$\mathcal{L}(\Delta N^\nu) = \prod_{j \in \mathcal{E} \setminus \mathcal{J}^\nu} \exp \left(-h \sum_{i=1}^k \lambda_i(jh) \right) \left(\prod_{i=1}^k \prod_{j \in \mathcal{J}_i^\nu} (h\lambda_i(jh) + o(h)) \right).$$

이 식을 h 의 관측된 점프의 개수의 승

$$h^{\sum_{j=0}^{\nu-1} \sum_{i=1}^k \Delta N_i(jh)}$$

으로 나누어 다음을 얻는다.

$$\mathcal{L}(\Delta N^\nu) = \prod_{j \in \mathcal{E} \setminus \mathcal{J}^\nu} \exp \left(-h \sum_{i=1}^k \lambda_i(jh) \right) \prod_{i=1}^k \prod_{j \in \mathcal{J}_i^\nu} \left(\lambda_i(jh) + \frac{o(h)}{h} \right).$$

충분히 큰 ν 에 대하여, 과정의 각 점프 시점 t_0 는 $j_0 h < t_0 \leq (j_0 + 1)h$ 가 되는 j_0 를 가지며 모든 이와 같은 세분 구간은 겹치지 않는다. 따라서 $\nu \rightarrow \infty$ 인 경우에, $h \rightarrow 0$ 이므로 $\sum_{j \in \mathcal{E} \setminus \mathcal{J}^\nu} h\lambda_i(jh) \approx \sum_{j \in \mathcal{E}} h\lambda_i(jh)$ 이고 $j_0 h$ 는 t_0^- 한다. 따라서 다음이 성립한다.

$$\begin{aligned}
\lim_{\nu \rightarrow \infty} \mathcal{L}(\Delta N^\nu) &= \lim_{\nu \rightarrow \infty} \exp \left(-\sum_{i=1}^k \sum_{j \in \mathcal{E} \setminus \mathcal{J}^\nu} h\lambda_i(jh) \right) \prod_{i=1}^k \prod_{j \in \mathcal{J}_i^\nu} \left(\lambda_i(jh) + \frac{o(h)}{h} \right) \\
&= \exp \left(-\sum_{i=1}^k \int_0^u \lambda_i(t) dt \right) \prod_{i=1}^k \prod_{m=1}^{r_i} (\lambda_i(t_{i,m}^-)).
\end{aligned} \tag{2.4}$$

따라서 이 표현의 로그함수는 다음과 같이 나타낼 수 있다.

$$\ell_u(N(t); 0 \leq t \leq u) = \sum_{i=1}^k \int_0^u (\log[\lambda_i(t)] dN_i(t) - \lambda_i(t_{i,m}^-) dt). \tag{2.5}$$

여기서 $\lambda_i(t^-)$ 는 시점 t 직전의 강도를 나타낸다. 또한 $dN_i(t)$ 항은 계수과정의 점프에 따른 합을 나타낸다.

3. 토의 및 결론

이 연구에서는 다변량 계수과정의 우도함수를 엄격한 방법으로 구하였다. 고려한 세분 구간의 형태는 $(t, t+h]$ 로서 왜 이러한 구간의 형태가 필요한지를 보였다. 식 (2.2)에서 식 (2.3)로 넘어가는 단계와 관측구간 $(0, u]$ 를 ν 개의 세분 구간으로 나누고 이를 바탕으로 회복 시점의 좌극한으로 수렴하는 것을 보인 부분은 증명의 중요한 단계이며, Andersson과 Britton (2000)이나 Schutt (2010)에서는 모두 생략되어 있다.

식 (1.1)로 주어지는 일반화 확률전염병모형에 대한 로그 우도함수는 다변량계수과정의 우도함수인 식 (2.5)를 이용하여 얻을 수 있다:

$$n_R \log \gamma + n_I (\log \beta - \log \mu) + \sum_{j=1}^{n_I} \log S(\phi_j^-) - \frac{\beta}{\mu} \int_0^u S(t)I(t)dt - \gamma \int_0^u I(t)dt \quad (3.1)$$

여기에서 $S(\phi_j^-)$ 은 초기 민감 개체 수 μ 에 의존한다.

식 (3.1)의 우도함수는 Andersson과 Britton (2000)과 같으며 Kypraios (2009)의 우도함수와는 일치하지 않는다. Kypraios (2009)의 우도함수는 구간 $(0, u]$ 에서의 관측에 대한 우도함수라기 보다는 관측구간 $(0, \infty)$ 내에서 처음 발생한 감염의 시간 ϕ_1 부터 전염 상태가 종료될 때까지 즉, 모든 감염 개체가 회복되고 나서 더 이상 감염이 진행되지 않는 시간까지의 구간 (ϕ_1, ∞) 에서의 관측에 대한 우도함수라고 할 수 있다.

어떤 집단에 하나의 전염병이 감염될 때, 그 집단의 구성원 중 일부는 그 질병에 대한 면역성을 이미 보유하고 있거나, 혹은 동질혼합성에서 제외되거나 할 수 있다. 따라서 집단의 구성원 중 그 질병에 감염될 수 있는 구성원의 수 즉, 민감 개체 수는 전염율의 추정 및 예방접종 등의 방역 정책에 큰 영향을 미치게 된다. Kypraios (2009)는 우도함수 (1.4)를 사용하여 시점 0에서의 초기 민감 개체 수의 최대추정치를 구하는 문제를 다루었다. 이때 폐형식 해가 존재하지 않으므로 수치적 방법을 사용하여야 한다. 시뮬레이션을 통하여 Kypraios (2009)는 자신의 최우추정치가 Huggins 등 (2004)의 마팅계일 추정치보다 더 좋음을 보였으나, 최우추정치는 마팅계일 추정치와 마찬가지로 특정 모수의 값에서는 좋은 결과를 가지나 전체적으로 좋은 결과를 가지지 못한다.

전체 구간에서의 관측에 대한 우도함수 (1.3)를 이용한 초기 민감 개체 수의 최우추정치는 저자의 시뮬레이션 결과 (여기에 결과를 제시되지 않음)에 의하면 Kypraios (2009)의 결과 보다는 좋지 않은 성질을 보이거나, Huggins 등 (2004)의 결과와는 유사한 성질을 보였다. 따라서 추후 연구에서 이에 대한 추가적 고찰과 검증이 필요하다고 본다.

우도함수 (1.3)를 이용한 초기민감 개체 수의 최우추정치는 회복시점에 민감한 반응을 보이는 것으로 나타나므로, 이의 조절을 통한 추정방법, 예컨대 벌칙 (penalized) 최우추정법에 관한 추가적 연구가 진행되어야 할 것이다.

References

- Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical models based on counting processes*, Springer, New York.
- Anderson, R. M. and May, R. M. (1991). *Infectious diseases of humans: Dynamics and control*, Clarendon Press, Oxford.
- Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*, Springer-Verlag, New York.
- Bailey, N. T. J. (1975) *The mathematical theory of infectious diseases and its applications*, Griffin, London.

- Becker, N. G. and Britton, T. (1999). Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society B*, **61**, 287–307.
- Cauchemez, S. and Ferguson, N. M. (2008). Likelihood based estimation of continuous-time epidemic models from time-series data: Application to measles transmission in London. *Journal of the Royal Society Interface*, **5**, 885–897.
- Daley, D. J., Gani, J. (1999). *Epidemic modelling: An introduction*, Cambridge University Press, New York.
- Diekmann, O. and Heesterbeek, J. A. P. (2000). *Mathematical epidemiology of infectious diseases: Model building, analysis and interpretation*, Wiley, New York.
- Ha, I. and Cho, G. (2012). H-likelihood approach for variable selection in gamma frailty models. *Journal of the Korean Data & Information Science Society*, **23**, 199–207.
- Höhle, M. and Jørgensen, E. (2002). *Estimating parameters for stochastic epidemics*, Dina research report No. 102, Danish Institute of Agricultural Sciences, Research Centre Foulum, Denmark.
- Huggins, M., Yip, P. S. F. and Lau, E. H. Y. (2004). A note on the estimation of the initial number of susceptible individuals in the general epidemic model. *Statistics & Probability Letters*, **67**, 321–330.
- Kang, S. G., Lee, J. H. and Lee, W. D. (2012). Likelihood based inference for the ratio of parameters in two Maxwell distributions. *Journal of the Korean Data & Information Science Society*, **23**, 89–98.
- Kypriaios, T. (2009). A note on maximum likelihood estimation of the initial number of susceptibles in the general stochastic epidemic model. *Statistics & Probability Letters*, **79**, 1972–1976.
- McKendrick, A. G. (1926). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, **44**, 98–130.
- Oh, C. (2013). Approximation of estimates in the susceptible-infectious-removed epidemic model. *Journal of Mathematics and Statistics*, **9**, 38–42.
- O’Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics, *Journal of the Royal Statistical Society A*, **162**, 121–129.
- Rida, W. N. (1991). Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. *Journal of the Royal Statistical Society B*, **53**, 269–283.
- Schutt, R. (2010). *Topics in model-based population inference*, Ph.D. thesis, Columbia University, New York.

Derivation of the likelihood function for the counting process

Changhyuck Oh¹

¹Department of Statistics, Yeungnam University

Received 19 December 2013, revised 6 January 2014, accepted 12 January 2014

Abstract

Counting processes are widely used in many fields, whose properties are determined by the intensity function. For estimation of the parameters of the intensity functions when the process is observed continuously over a fixed interval, the likelihood function is of interest. However in the literature there are only heuristic derivations and some results are not coincident. We thus in this note derive the likelihood function of the counting process in a rigorous way. So this note fill up a hole in derivation of the likelihood function.

Keywords: Counting process, likelihood function, statistical physics concept.

¹ Professor, Department of Statistics, Gyeongsan 712-749, Korea. E-mail: choh@yu.ac.kr