The Korean Journal of Applied Statistics (2020) DOI: http://dx.doi.org/10.5351/KJAS.2020.33.3.347
33(3), 347–355

Effect of an unsampled population on the estimation of a population size

Yujin Chung[a,1]

[a]Department of Applied Statistics, Kyonggi University

(Received April 19, 2020; Revised April 30, 2020; Accepted April 30, 2020)

Abstract

An Isolation-with-Migration (IM) model is used to estimate extant population sizes, the splitting time of populations split away from their common ancestral populations, and migration rates between the extant populations. An evolutionary model such as IM models is estimated by analyzing DNA sequences sampled from the extant populations in the model. When a true model includes an unsampled 'ghost' population without data, the unsampled population is often ignored from the evolutionary model to infer. In this paper, we conduct a simulation study to investigate the effect of an unsampled population on the estimation of the size of the sampled population. When there exists an unsampled population that shares migrations with the sampled population, the size estimation of the sampled population was biased. However, the size estimation was improved if an evolutionary model, including the unsampled population, was estimated.

Keywords: Coalescent process, isolation-with-migration model, unsampled population, ghost population, MIST

1. 서론

Isolation-with-Migration (IM) 모형은 여러 집단(populations)의 진화를 설명하는 모형으로 널리 사용되고 있다. 이 진화 모형의 간단한 경우인 이집단 IM 모형(2-population IM model)은 Figure 1.1(c)에서 보여 주듯이 두 집단(population 1과 population 2)이 과거 T_S라는 시간 전에 하나의 공통 조상 집단(common ancestral population)으로부터 분화되어 진화되어 왔고, 분화 이후에 두 집단 간 이주(migrations)가 있어 온 과정을 설명한다 (Chung, 2019). 그리고 두 개의 현존하는 집단과 이들의 공통 조상 집단의 크기(effective population size)는 각각 θ_1, θ_2, θ_a으로 표기하고, 집단의 크기는 모두 시간에 따라 변하지 않는 상수라고 가정한다. 두 집단 사이에 양방향으로 발생하는 이주 사건(migration events)은 이주율(migration rate) m_1과 m_2에 따라 발생하고 이주율 또한 시간이 따라 변하지 않는 상수이다. 따라서 이집단 IM 모형은 여섯개의 모수로 설명되고 이를 $\Psi_2 = (\theta_1, \theta_2, \theta_a, m_1, m_2, T_S)$로 표기한다. 만일 하나의 현존하는 집단만 고려한다면 진화 모형은 Figure 1.1(b)와 같은 단일 집단 모형(single population model)이 되고 이때 모수는 $\Psi_1 = (\theta_1)$이다.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2018R1C1B5044541).

[1]Department of Applied Statistics, Kyonggi University, Kwanggyosan-ro 154-42, Suwon, Gyeonggi-do 16227, South Korea. E-mail: yujinchung@kgu.ac.kr

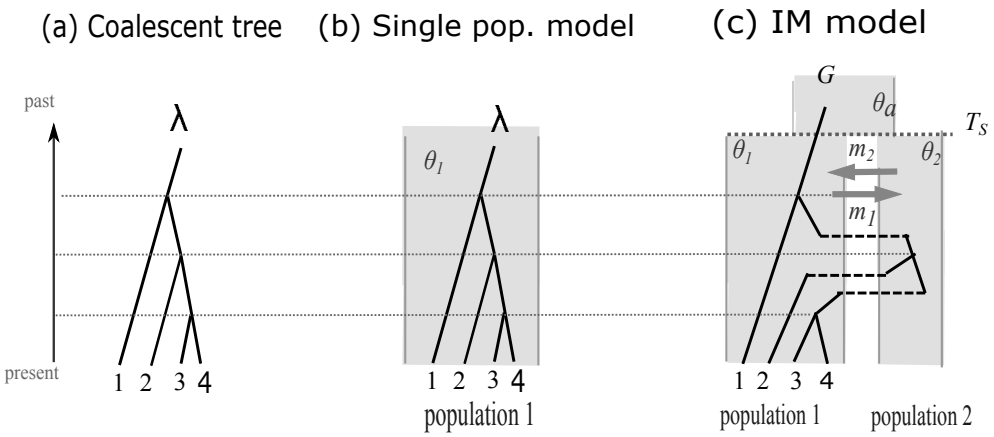

Figure 1.1. (a) An example of coalescent tree (λ) of four sequences. (b) The coalescent tree given a single population model with population size θ_1. That is, $\Psi_1 = (\theta_1)$. The four sequences were sampled from population 1. (c) An example of genealogy (G) given a 2-population IM model, where the four sequences sampled from population 1. Population 2 is an unsampled population. The coalescent tree of the genealogy is in (a), and the IM model has six parameters $\Psi_2 = (\theta_1, \theta_2, \theta_a, m_1, m_2, T_S)$.

진화 모형을 추정하기 위해, 보통 현존하는 집단에서 DNA 염기서열(DNA sequences)을 추출하여 분석한다. 염기서열은 추출한 뒤 정렬하는 과정을 거치는데, 이렇게 정렬된 DNA 염기서열(aligned DNA sequences or DNA alignments)을 분석하여 단일 집단 모형이나 IM 모형을 설명하는 모수를 추정한다. 이 때 하나의 정렬된 DNA 염기서열은 개체마다 염색체의 같은 위치에서 추출된 것이기 때문에 흔히 정렬된 DNA 염기서열을 *locus*라고 부른다. 그리고 DNA 염기서열을 정렬하는 것은 DNA 염기서열이 과거 동일한 개체의 DNA 염기서열로부터 유전된 것이 되도록 DNA alignment를 추정한 것이다. 또한 재조합(recombination)과 같은 생물학적 혹은 진화적 과정에 의해서 loci마다 다른 진화 역사(evolutionary history)를 가질 수 있다. 따라서 loci는 흔히 서로 독립이라 가정한다 (Chung, 2019).

DNA alignment마다 진화 경로(evolutionary path)가 다를 수 있고, 이런 진화 경로를 계통(genealogy, G)이라 부른다. Figure 1.1(c)는 이집단 IM 모형에서 발생할 수 있는 계통의 한 예를 그래프로 표현해 보여주고 있다. 이집단 IM 모형하에서 계통 그래프는 대각선으로 자주 표현되기 하지만 유전자가 한 세대에서 다음 세대로 전달되는 수직적 경로(vertical path)와 한 집단에서 다른 집단으로의 이주 사건, 즉 이주하는 경로를 표현하는 수평적 경로(horizontal path)로 구성된다. 계통 그래프를 시간 역순으로 관찰했을 때, 수직적 경로는 두 개의 lineages가 하나의 lineage로 합쳐지는 사건으로 볼 수 있고 이 사건을 합류 사건(coalescent event)이라고 부른다. 만약 Figure 1.1(c)에서 이주 사건을 제외하고 합류 사건만으로 나무 모양의 그래프를 표현하면 Figure 1.1(a)와 같은 그래프를 얻을 수 있다. 이렇게 합류 사건만으로 구성된 그래프를 합류 나무(coalescent tree, λ)라고 부르겠다. 만약 단일 집단 진화 모형을 가정하면 발생할 수 있는 계통은 Figure 1.1(b)와 같이 합류 사건만으로 구성된 그래프이다. 이 경우에는 계통과 계통의 합류 나무가 항상 동일하다.

진화 모형을 추정하기 위해 n개의 DNA alignments, $D_1, \ldots, D_n$를 관측했다고 하자. 이 때 D_i의 계통을 G_i이고 이 계통의 합류 나무를 λ_i이라고 하자($i = 1, \ldots, n$). 일반적으로 계통과 합류 나무는 관측할 수 없는 잠재 변수(latent variable)이고 진화 모형과 데이터를 연결하는 중요한 역할을 한다. DNA alignments를 분석하여 IM 모형을 추정하는 여러 가지 베이지안 방법들이 개발 되었는데 (Chung, 2019), 이 중 `MIST` (Chung과 Hey, 2017)라는 프로그램은 많은 loci를 분석할 수 있고, 베이지안 추

론 뿐만 아니라 모수의 최대우도추정량을 구할 때도 이용할 수 있다 (Chung, 2020). 프로그램 MIST를 이용한 최대우도추정법 (Chung, 2020)은 이단계 분석(2-step analysis)법이다. 첫 번째 단계에서는 각 DNA alignments를 분석하여 합류 나무 $\lambda_1, \ldots, \lambda_n$를 추정하고, 두 번째 단계에서는 추정된 합류 나무를 분석하여 단일 집단 모형 혹은 이집단 IM 모형을 추정하는 방법이다. 일반적으로 MIST를 이용한 최대우도추정법은 베이지안 방법보다 추정량의 정확도가 떨어지나 계산 속도가 더 빠르다.

진화 모형을 추정할 때, 흔히 관심있는 집단에서 DNA 염기서열을 추출한다. 예를 들면, 어느 population 1이라는 집단의 크기를 추정하고 싶은 경우 이 집단에서 DNA 염기서열을 추출하고, 단일 집단 모형 Ψ_1을 추정한다. 만약 두 집단 population 1과 population 2의 진화모형에 관심있다면 두 집단으로부터 DNA 염기서열을 추출하고, 이집단 IM 모형 Ψ_2을 추정한다. 하지만 이런 분석 방법의 문제점은 참인 진화 모형을 모른다는 것이다. 만약 population 1과 population 2의 참 진화 모형이 Figure 1.1(c)와 같은 이집단 IM 모형이고 population 2로 부터의 이주 사건들이 population 1의 크기(θ_1)에 영향을 미친다고 하자. 이 때 우리의 관심은 θ_1을 추정하는 것이기 때문에, 일반적으로 population 1에서만 데이터를 추출하고 분석하여 Figure 1.1(a)와 같은 단일 집단 모형을 추정한다. 이 처럼 데이터를 추출하지는 않지만 이주 사건을 통해 표본 집단(sampled population)에 영향을 줄 수 있는 population 2와 같은 집단을 미표본 집단(unsampled population) 혹은 ghost population이라고 부른다. 미표본 집단은 현존하는 집단일 수도 있고 멸종한 집단일 수도 있다. 그리고 일반적으로 population 2와 같은 미표본 집단이 존재하는지에 대해 알 수 없다. 만약 미표본 집단이 실제로 존재하나 이를 고려하지 않고 표본 집단에 대한 진화 모형을 추정할 경우, 진화 모형의 모수에 대한 추정량은 편의(bias)를 가질 수 있다고 알려져 있다 (Beerli, 2004; Hey 등, 2018). 그리고 이 편의는 미표본 집단을 포함하여 확장된 진화 모형을 추정함으로써 어느 정도 개선이 될 수 있다. 미표본 집단은 어느 생물에 대해서도 존재 할 수 있기 때문에, 실제 데이터 분석에서도 미표본 집단을 포함하여 확장된 IM 모형을 추정하는 경우가 있다 (Hey 등, 2018). 하지만 집단을 하나 더 추가하여 확장된 IM 모형을 추정하면 계산량이 굉장히 많이 증가할 수 있기 때문에 사용하는 프로그램에 따라 항상 미표본 집단을 추가한 모형을 추정할 수 있는 것은 아니다 (Chung, 2019).

지금까지 개발된 프로그램 중 최대우도추정법은 프로그램 MIST을 이용한 방법이 유일하다. 본 논문에서는 최대우도추정법을 이용하여 진화모형을 추정할 때, 미표본 집단이 미치는 영향을 모의 실험을 통해 알아보고자 한다. 이를 위하여 2장에서는 일반적으로 사용하는 통계적 모형 및 MIST를 이용한 최대우도추정법에 대하여 설명한다. 3장에서는 모의 실험을 통해 집단의 크기를 추정할 때 미표본 집단의 영향에 대해 살펴보겠다. 마지막으로 4장에서 본 논문의 전체적인 내용을 정리하겠다.

2. 표준 통계적 모형과 최대우도추정법

2.1. 표준 통계적 모형

DNA alignments를 분석하여 진화 모형 Ψ를 추정하기 위해 두 가지 종류의 불확실성(uncertainty)에 대한 확률 모형을 고려한다 (Felsenstein, 1988). 첫 번째는 계통이나 합류 나무가 조건으로 주어졌을 때 DNA alignment의 확률 $p(D_i|G_i)$로, 흔히 사용하는 확률 모형은 간단한 mutation 모형 (Kimura, 1969)부터 다양한 substitution 모형들 (Jukes와 Cantor, 1969; Hasegawa 등, 1985; Tavaré, 1986)이 있다. 여기서 λ_i를 G_i의 합류 나무라고 했을 때, λ_i가 조건으로 주어졌을 때 D_i와 이주 사건들은 조건부 독립이어서 $p(D_i|G_i) = P(D_i|\lambda_i)$이 된다 (Chung과 Hey, 2017). 두 번째는 진화 모형이 조건으로 주어졌을 때 잠재 변수의 확률 $p(G_i|\Psi)$로는 확률과정(stochastic process) 중 하나인 합류 과정(coalescent process)을 흔히 사용한다 (Kingman, 1982; Wakeley, 2009). 따라서 두 가지 종류의 확률 모형을 이용

하여, 진화모형 Ψ가 주어졌을 때 i번째 locus의 확률은 다음과 같이 구할 수 있다.

$$p(D_i|\Psi) = \int p(D_i|G_i)p(G_i|\Psi)dG_i = \int p(D_i|\lambda_i)p(\lambda_i|\Psi)d\lambda_i. \tag{2.1}$$

여기서 $\mathcal{M}$을 모든 가능한 이주 사건의 집합이라고 하면 $p(\lambda_i|\Psi) = \int p(G|\Psi)d\mathcal{M} = \int p(\lambda, \mathcal{M}|\Psi)d\mathcal{M}$ 이다 (Chung과 Hey, 2017). 그러면 n개의 loci를 $\mathbf{D} = (D_1, \ldots, D_n)$라고 할 때, Ψ의 가능도 함수(likelihood function)은 다음과 같다.

$$L_1(\Psi|\mathbf{D}) = \prod_{i=1}^{n} p(D_i|\Psi) = \prod_{i=1}^{n} \int p(D_i|\lambda_i)p(\lambda_i|\Psi)d\lambda_i, \tag{2.2}$$

이 가능도 함수는 Felsenstein의 방정식(Felsenstein's equation)이라고 불리고, 단일 모형($\Psi = \Psi_1$)이나 이집단 IM 모형($\Psi = \Psi_2$)처럼 어떤 진화모형에 대해서도 같은 방법으로 가능도 함수를 구할 수 있다.

가능도 함수 $L_1(\Psi)$의 닫힌 형식(closed-form)은 없고, 합류 나무의 공간이 매우 방대하기 때문에 수치적으로 계산하는 것도 어렵다 (Chung, 2019). 따라서 가능도 함수 $L_1(\Psi)$을 최대화하는 최대우도추정량은 매우 제한적인 경우에 대해서만 구할 수 있어 (Zhu와 Yang, 2012), 대신 마르코프 체인 몬테 카를로(Markov chain Monte Carlo; MCMC)를 이용하여 베이지안 추론방법이 많이 개발되었다 (Chung과 Hey, 2017; Hey 등, 2018; Chung, 2020).

2.2. 진화 모형 추정을 위한 최대우도추정법

프로그램 `MIST`을 이용하는 최대우도추정법 (Chung, 2020)은 이단계 분석방법이다. 첫 번째 단계는 식 (2.2)처럼 잠재 변수인 합류 나무 λ_i에 대해 적분하는 대신, $p(D_i|\lambda_i)$을 λ_i의 가능도함수로 고려하고 아래와 같이 λ_i의 최대우도추정량을 구하는 것이다.

$$\widetilde{\lambda}_i = \arg\max_{\lambda_i} p(D_i|\lambda_i). \tag{2.3}$$

합류 나무의 최대우도추정량은 `PAUP*` (Swofford, 2002) 프로그램을 사용하여 구하였다. 두 번째 단계에서는 추정된 합류 나무 $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n$을 데이터로 고려하여 Ψ의 가능도 함수를 아래와 같이 구하였다.

$$L_2\left(\Psi|\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n\right) = \prod_{i=1}^{n} p\left(\widetilde{\lambda}_i|\Psi\right). \tag{2.4}$$

그리고 $L_2(\Psi)$ 가능도 함수를 최대화하는 값을 아래와 같이 Ψ의 추정량으로 사용하였다.

$$\widetilde{\Psi} = \arg\max_{\Psi} L_2\left(\Psi|\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n\right). \tag{2.5}$$

이 단계에서 $\widetilde{\Psi}$를 구하기 위해 `MIST` 프로그램를 이용하였다.

이 최대우도추정법은 잠재 변수 λ_i를 추정하여 두 번째 단계에서 이 추정값을 데이터처럼 사용하였기 때문에, λ_i의 추정값의 추정오차가 $\widetilde{\Psi}$ 에도 영향을 미친다. 따라서 MCMC 시뮬레이션을 통해 λ_i를 생성하는 베이지안 추정법보다 정확도는 떨어진다 (Chung, 2020). 하지만 일반적으로 베이지안 방법보다 계산이 빠르다 (Chung, 2020). 그리고 여러 진화 시나리오가 있는 추정하는 경우 효율적으로 분석할 수 있다. 첫 번째 단계에서 $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n$를 한번 구해두면, 이를 반복적으로 사용할 수 있어 두 번째에서 다른 진화 모형을 추정할 수 있어, 여러가지 진화 모형을 고려할 때 분석을 처음부터 다시 시작하지 않아

도 된다 (Chung, 2020). 예를 들어, 집단 population 1에서 데이터를 추출하였을 때, 이 데이터를 분석하여 $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n$를 구하였다고 하자. 만약 Figure 1.1(b)와 같은 단일집단 모형과 미표본 집단을 고려한 Figure 1.1(c)와 같은 이집단 IM 모형을 둘 다 추정한다면, 같은 $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n$을 사용하고 두 번째 단계만 실행하여 두 진화모형을 추정할 수 있다.

3. 모의 실험

미표본 집단의 영향을 알아보기 위해 모의 실험을 수행하였다. 세 가지 진화 모형을 고려하였다. 첫 번째 모형은 단일 집단 모형으로 집단 크기 $\theta_1 = 1$로 가정하였다. 두 번째 모형은 이집단 IM 모형으로, 집단1에서만 데이터를 채집한 경우이다. 즉, 비록 표본을 집단1에서만 채집하고 집단1의 크기를 추정하는 것이 목표이지만, 미표본집단인 집단2가 집단1의 크기에 영향을 미칠 수 있는 경우이다. 두 가지의 이집단 모형의 모수값을 고려하였는데, 하나는 $\theta_1 = \theta_2 = \theta_a = 3$, $m_1 = m_2 = 2$, $T_S = 4$로 세 집단의 크기는 1로 같고 이주율 또한 2로 같은 경우이고, 다른 하나는 $\theta_1 = 1$, $\theta_2 = 5$, $\theta_a = 3$, $m_1 = 2$, $m_2 = 0.4$, $T_S = 4$로 집단 크기와 이주율이 모두 다른 경우이다. IM모형 중 모수값이 모든 다른 모형은 Chung과 Hey (2017)에서 미표본 집단의 효과를 조사하기 위해 가정한 모형과 동일하다. 우선 집단유전학에서 시뮬레이션을 할 때 많이 사용되는 소프트웨어인 `ms` (Hudson, 2002)을 이용하여 각 모형마다 1,000 loci를 시뮬레이션하였다. 여기서 loci의 개수는 합류 나무의 수와 같다. 각각의 생성한 합류 나무로부터 DNA alignment를 생성하는데 소프트웨어 `seq-gen` (Rambaut와 Grassly, 1997)을 사용하였다. 각 alignment의 길이는 1,000 sites이고 Jukes-Cantor substitution 모형 (Jukes와 Cantor, 1969)을 가정하였다. 각 진화 모형마다 100개의 데이터셋을 생성하였다.

각 데이터셋마다 최대우도추정법을 이용하여 모수를 추정하였다. 우선 첫 번째 단계에서 DNA alignments를 분석하여 합류 나무를 추정하였다. 합류 나무는 `PAUP*` (Swofford, 2002) 소프트웨어를 사용하여 추정하였고, 데이터 생성 모형과 같은 Jukes-Cantor substitution 모형을 가정하였다. 추정된 합류 나무는 두 번째 단계에서 분석되는데, `MIST` 프로그램을 이용하여 가정한 진화모형의 모수를 추정하였다. 각 데이터셋마다 단일집단 모형 (Figure 1.1(B))과 이집단 IM모형 (Figure 1.1(C))을 모두 추정하였다. 두 가지 다른 진화 모형을 추정하기 위해서 첫 번째 분석단계에서 추정된 합류 나무를 다시 추정하지 않고, 두 번째 단계만 다시 실행하여 다른 진화모형을 추정하였다.

또한 데이터 생성에 사용한 합류나무의 참값을 분석하였다. 분석의 첫 번째 단계는 생략하고 두 번째 단계에서 참 합류 나무를 분석하여 단일집단 모형과 이집단 IM 모형을 추정하였다. 이 분석을 통하여 이 단계 분석방법에서 두 번째 단계만의 분석 성능을 평가할 수 있다.

3.1. 거짓 미표본 집단의 영향

참모형은 단일집단모형으로 다른 집단으로부터 영향을 받지 않는 진화모형이다. Population 1 집단의 크기 ($\theta_1 = 1$)를 추정하기 위해 이 집단으로부터 데이터를 생성하였다. 단일집단모형을 가정하였을 때, 즉 참인 진화모형을 가정하였을 때, θ_1의 최대우도추정값은 2.3044이고 표본오차는 0.6598로 과대추정(overestimation)되었다 (Table 3.1). 하지만 참인 합류나무를 분석하였을 때는 추정값이 0.9995(표본오차: 0.0181)로 매우 정확한 추정값을 얻을 수 있었다. 두 편의의 차이는 이단계 분석방법이 합류나무 추정값의 오차를 고려하지 않고 진화모형의 모수를 추정하였기 때문에 발생한 편의로 볼 수 있다.

참인 단일집단 모형이 아닌 이집단 IM 모형 (Figure 1.1)을 가정하고, θ_1의 추정값도 구하였다. 참 합류 나무를 분석하였을 경우 추정값의 평균이 1.0115(표준오차: 0.4024)로 정확한 추정값을 얻을 수 있었다. 이집단 IM 모형의 다른 모수들도 동시에 추정하였기 때문에, 참인 진화모형인 단일집단 모형을 가정했을 때 보다 표준오차가 더 커진 것으로 볼 수 있다. DNA alignments로 부터 추정한 합류 나무를

Table 3.1. Average of MLEs of the population size (θ_1). The true model is a single population model with $\theta_1 = 1$. The MLEs were obtained by analyzing either the true simulated coalescent trees or the coalescent trees estimated from the simulated DNA alignments. The numbers in parenthesis are standard errors.

Inference model	True coalescent trees	Estimated coalescent tree
Single pop. model	0.9995 (0.0181)	2.3044 (0.6598)
2-pop. IM model	1.0115 (0.4024)	0.4625 (0.1891)

Table 3.2. Average of MLEs of the population size (θ_1). The true model is the 2-population IM model with parameters $\theta_1 = \theta_2 = \theta_a = 3$, $m_1 = m_2 = 2$, $T_S = 4$. The MLEs were obtained by analyzing either the true simulated coalescent trees or the coalescent trees estimated from the simulated DNA alignments. The numbers in parenthesis are standard errors.

Inference model	True coalescent trees	Estimated coalescent tree
Single pop. model	1.8320 (0.0393)	2.7800 (0.6257)
2-pop. IM model	1.0098 (0.0541)	0.8585 (1.6528)

Table 3.3. Average of MLEs of parameters in a 2-population IM model. The true model is the 2-population IM model with parameters $\theta_1 = \theta_2 = \theta_a = 3$, $m_1 = m_2 = 2$, $T_S = 4$. The MLEs were obtained by analyzing either the true simulated coalescent trees or the coalescent trees estimated from the simulated DNA alignments. The numbers in parenthesis are standard errors.

Parameters	θ_2	θ_a	m_1	m_2	T_S
True coal. trees	1.0371 (0.1698)	0.9171(0.1951)	1.897 (0.3929)	1.8836 (0.9258)	4.089 (0.5847)
Estimated coal. tree	5.791 (4.4765)	17.14 (6.5030)	9.2121 (2.6865)	8.9397 (2.5481)	5.1072 (1.6921)

분석한 경우, θ_1의 최대우도추정값은 평균 0.4625로 과소추정(underestimation)되었다.

따라서 미표본 집단이 있다고 잘못 가정하였을 때, 참 합류 나무를 분석한 경우에는 집단 크기 추정의 정확도에 미치는 영향이 적었다. 하지만 프로그램 `MIST`를 이용하여 구한 집단 크기의 최대우도추정량은 가정한 진화 모형이 참이든 거짓이든 편의를 가졌다. 따라서 첫 번째 단계에서 더 정확한 합류나무를 추정하거나 두 번째 단계에서 추정오차를 반영하여 분석하는 방법을 개발할 필요가 있다.

3.2. 미표본 집단의 영향

한 집단에서 생성된 데이터를 분석하여 단일집단 모형과 이집단 IM 모형을 추정하였다. 참 진화 모형은 이집단 IM 모형으로 두 개의 현존하는 집단 중 population 1에서 데이터를 추출하였고, 다른 집단인 population 2에서는 데이터를 추출하지 않은 미표본 집단이지만, migrations을 통해 표본집단인 population 1의 크기(θ_1)에 영향을 미칠 수 있다.

참 진화 모형의 모수가 $\theta_1 = \theta_2 = \theta_a = 3$, $m_1 = m_2 = 2$, $T_S = 4$인 경우, Table 3.2에서 θ_1의 추정값과 표준오차를 비교하였다. 참 진화모형은 이집단 IM 모형이지만, 미표번 집단의 존재를 고려하지 못하고 단일 집단 모형을 가정하였을 때, 집단크기의 추정값은 참값 $\theta_1 = 1$보다 과대추정되었다. 참 합류 나무를 분석했을 때는 추정값은 1.832(표준오차: 0.0393)이고 추정된 합류 나무를 분석하였을 때 추정값은 2.78(표준오차: 0.6257)이었다 (Table 3.2). 이집단 IM 모형을 가정하여 모수를 추정하였을 때, 참 합류 나무를 분석하거나 추정된 합류 나무를 분석하는 두 경우 모두 θ_1을 참값에 가까운 값으로 추정하였다 (Table 3.2). 이집단 IM 모형의 다른 모수들은 참 합류 나무를 분석했을 경우, 참값에 가까웠으나 추정된 합류 나무를 분석하였을 때는 추정값들이 참값과 차이가 달랐고 표준오차 또한 매우 컸다 (Table 3.3). 특히, 미표본 집단(unsampled population)에서 직접 채집한 데이터가 없기 때문에 θ_2의 표준오차가 0.1698로 표본 집단의 크기(θ_1)의 표준오차 0.0393보다 컸다.

Table 3.4. Average of MLEs of the population size (θ_1). The true model is the 2-population IM model with parameters $\theta_1 = 1$, $\theta_2 = 5$, $\theta_a = 3$, $m_1 = 2$, $m_2 = 0.4$, $T_S = 4$. The MLEs were obtained by analyzing either the true simulated coalescent trees or the coalescent trees estimated from the simulated DNA alignments. The numbers in parenthesis are standard errors.

Inference model	True coalescent trees	Estimated coalescent tree
Single pop. model	3.7340 (0.0792)	4.5680 (0.6992)
2-pop. IM model	0.9957 (0.0388)	1.1725 (1.5809)

Table 3.5. Average of MLEs of parameters in a 2-population IM model. The true model is the 2-population IM model with parameters $\theta_1 = 1$, $\theta_2 = 5$, $\theta_a = 3$, $m_1 = 2$, $m_2 = 0.4$, $T_S = 4$. The MLEs were obtained by analyzing either the true simulated coalescent trees or the coalescent trees estimated from the simulated DNA alignments. The numbers in parenthesis are standard errors.

Parameters	θ_2	θ_a	m_1	m_2	T_S
True coal. trees	5.339 (0.6308)	2.984 (0.1537)	2.0680 (0.1911)	0.52514 (0.3015)	4.006 (0.0331)
Estimated coal. tree	8.065 (4.6716)	17.219 (5.8805)	2.8162 (1.0812)	2.02500 (1.1690)	8.095 (2.8817)

마찬가지로 참 진화 모형의 모수가 $\theta_1 = 1$, $\theta_2 = 5$, $\theta_a = 3$, $m_1 = 2$, $m_2 = 0.4$, $T_S = 4$인 경우, Table 3.4에서 θ_1의 추정값과 표준오차를 비교하였다. 단일 집단 모형을 가정했을 때, 집단 크기는 참 값인 1보다 큰 값으로 추정되었다. 참 합류나무를 추정하였을 경우는 추정값은 3.734(표본오차: 0.0792)이고 추정된 합류나무를 추정하였을 때는 추정값이 4.568(표본오차: 0.6992)였다 (Table 3.4). 이집단 IM 모형을 추정하였을 때는 추정된 집단 크기가의 값이 0.9957(참 합류나무를 분석한 경우)와 1.1725(추정된 합류나무를 분석한 경우)로 참값에 훨씬 더 가까웠다. 이집단 IM 모형의 다른 모수들의 추정값은 참 합류 나무를 분석했을 경우, 참값에 가까웠다 (Table 3.5). 마찬가지로 미표본 집단(unsampled population)에서 직접 채집한 데이터가 없기 때문에 θ_2의 표준오차가 0.6308로 표본 집단의 크기(θ_1)의 표준오차 0.0388보다 컸다. 추정된 합류 나무를 분석하였을 때는 추정값들이 참값과 차이가 달랐고 표준오차 또한 매우 컸다 (Table 3.5).

표본 집단에 영향을 미치는 미표본 집단이 존재하는 경우, 이를 고려하지 않았을 때 집단 크기의 추정은 심각한 편의(bias)가 발생함을 모의 실험을 통해 확인하였다. 따라서 미표본 집단을 진화 모형에 포함시켜 IM 모형으로 확장된 모형을 추정할 필요가 있다. 그리고 이 때 데이터가 추출된 집단들과 관련된 모수의 추정값만 사용하고, 다른 모수의 추정값을 사용하고 해석하는 것은 주의가 필요하다.

4. 결론

많은 경우 관심있는 집단에서 데이터를 추출하고 분석하여 표본집단만이 현존하는 집단이 되는 진화 모형을 추정한다. 따라서 표본집단과 미표본집단 사이에 이주 사건들이 존재하는 경우에도, 미표본집단은 추정하려고 하는 진화 모형에서 종종 제외된다. 본 연구는 미표본 집단이 진화 모형에서 제외되었을 때 표본집단의 크기 추정에 미치는 영향에 대해 연구하였다. 특히 최대우도추정법 (Chung, 2020)으로 진화 모형을 추정하였을 때 미표본 집단의 영향에 대해 알아보았다.

모의실험을 통해, 참인 진화 모형이 단일모형인 경우, 즉 미표본 집단이 없는 경우에 대해서는 최대우도 추정법은 편향된 추정값을 구하였다. 하지만 참 합류나무를 분석하였을 때는 추정하는 진화모형이 미표본 집단을 포함여부에 상관없이 정확한 표본 집단의 크기를 구하였다. 따라서 편향된 최대우도추정량은 합류 나무의 추정오차에 의해 발생한 것으로 보인다. 미표본 집단이 있지만 이를 진화모형에 포함하지 않고 분석한 경우, 표본집단의 크기는 편향을 보였다. 하지만 미표본 집단을 진화모형에 포함시켰을 때는, 표본집단의 크기를 매우 정확하게 추정할 수 있었다. 참 합류나무를 분석한 경우에는 표본집단 크

기 외 다른 모수들도 표준오차가 크지만 정확하게 추정하였다. 추정된 합류나무를 분석하는 최대우도추정법을 사용하면, 표본집단 크기 외 다른 모수들의 표준오차가 굉장히 커서 추정값을 해석하는데 주의가 많이 필요하다. 이것 또한 추정된 합류 나무의 오차의 영향으로 보인다.

실제 데이터를 분석할 때는 미표본 집단의 존재 여부를 알 수 없으므로, 미표본 집단을 포함한 진화모형을 추정하는 것이 바람직하다. 하지만 미표본 집단을 포함한 진화 모형을 추정하는 것은 여전히 여러가지 어려움이 있다. 첫 번째로 미표본 집단을 포함한 진화 모형을 참 합류나무로 추정하였을 때, 표본집단의 크기는 항상 정확하게 추정하는 것을 모의실험을 통해 확인하였으나, DNA alignments를 분석하였을 때는 편의가 발생하는 경우가 있었다. 따라서 앞으로 첫 번째 단계에서 더 정확한 합류나무를 추정하거나 두 번째 단계에서 추정오차를 반영하여 분석하는 방법에 개발할 필요가 있다. 두 번째로 세 개 혹은 그 이상의 현존하는 집단에 대한 IM 모형을 추정하는 것은 여전히 매우 어려운 문제이다 (Chung, 2019). 하지만 본 연구에서 사용한 최대우도추정법을 여러 집단에 대하여 확장하는 것은 간단하기 때문에, 미표본 집단을 포함한 많은 수십개의 집단에 관한 진화모형을 추정할 수 있는 통계적 방법이 될 것이라 기대한다.

References

Beerli, P. (2004). Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations, *Molecular Ecology*, **13**, 827–836.

Chung, Y. (2019). Recent advances in Bayesian inference of isolation-with-migration models, *Genomics & Informatics*, **17**, e37.

Chung, Y. (2020). A maximum likelihood approach to infer demographic models, *Communications for Statistical Applications and Methods*, **27**, 385–395.

Chung, Y. and Hey, J. (2017). Bayesian analysis of evolutionary divergence with genomic data under diverse demographic models. *Molecular Biology and Evolution*, **34**, 1517–1528.

Felsenstein, J. (1988). Phylogenies from molecular sequences: Inference and reliability, *Annual Review of Genetics*, **22**, 521–565.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *Journal of Molecular Evolution*, **22**, 160–174.

Hey, J., Chung, Y., Sethuraman, A., Lachance, J., Tishkoff, S., Sousa, V. C., and Wang, Y. (2018). Phylogeny estimation by integration over isolation with migration models, *Molecular Biology and Evolution*, **35** 2805–2818.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation, *Bioinformatics*, **18**, 337–338.

Jukes, T. H. and Cantor, C. R. (1969). *Evolution of Protein Molecules*, Academy Press.

Kimura M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, **61**, 893–903.

Kingman J. F. C. (1982). The Coalescent, *Stochastic Processes and their Applications*, **13**, 235–248.

Rambaut, A. and Grassly, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees, *Bioinformatics*, **13**, 235–238.

Swofford, D. (2002). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0*. Sinauer Associates.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *American Mathematical Society: Lectures on Mathematics in the Life Sciences*, **17**, 57–86.

Wakeley, J. (2009) *Coalescent Theory: An Introduction*. Roberts and Company Publishers.

Zhu, T. and Yang Z. (2012). Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Molecular Biology and Evolution*, **29**, 3131–3142.

집단 크기 추정에 대한 미표본 집단의 영향

정유진[a,1]

[a]경기대학교 응용통계학과

(2020년 4월 19일 접수, 2020년 4월 30일 수정, 2020년 4월 30일 채택)

요 약

IM 모형(Isolation-with-Migration model; IM model)은 현존하는 집단들의 크기, 그 집단들이 공통 조상 집단으로부터 분리 된 분화 시간, 그리고 현존 집단 간의 이주율을 추정하는 데 널리 사용되는 진화 모형이다. IM 모형과 같은 진화 모형은 그 진화 모형 내 현존 집단으로부터 추출 된 DNA 염기서열을 분석하여 추정할 수 있다. 참인 진화 모형이 데이터가 추출되지 않은 미표본 집단(unsampled population) 혹은 소위 ghost라 불리는 집단을 포함할 때, 종종 이 미표본 집단을 제외한 진화 모델이 추론된다. 본 논문에서는 미표본 집단이 표본집단의 크기 추정에 미치는 영향을 조사하기 위해 모의실험을 수행하였다. 표본집단과 미표본집단 사이에 이주 사건들이 존재하는 경우, 표본집단의 크기의 추정량은 편향되었다. 그러나 미표본집단을 포함한 진화 모델이 추정되면 표본집단의 크기의 추정량은 많은 경우 개선되었다.

주요용어: 합류 과정, isolation-with-migration 모형, 미표본 집단, ghost 집단, MIST

본 연구는 한국 연구재단의 지원을 받아 수행한 연구임 (No. NRF-2018R1C1B5044541).

[1](16227) 경기도 수원시 영통구 광교산로 154-42, 경기대학교 응용통계학과. E-mail: yujinchung@kgu.ac.kr