

碩 士 學 位 論 文

MH-SAMC 알고리즘을 이용한
빅데이터 소셜네트워크 분석

高麗大學校 大學院

應用統計學科

朴 圓 涇

2020年 02月

全 秀 榮 教 授 指 導

碩 士 學 位 論 文

MH-SAMC 알고리즘을 이용한
빅데이터 소셜네트워크 분석

이 論 文 을 統 計 學 碩 士 學 位 論 文 으 로
提 出 함 .

2020年 02月

高 麗 大 學 校 大 學 院

應 用 統 計 學 科


朴 圓 涇 (印)





朴 圓 涇의 統計學 碩士學位論文

審査를 完了함.

2020年 02月

委員長 김우영 (印) 

委員 홍승만 (印) 

委員 임성수 (印) 

요 약 문

소셜네트워크(Social Network) 자료는 객체(actor) 또는 노드(node)들 사이의 관계를 나타내기 위한 자료로 사회학, 경제학, 경영학을 비롯한 사회과학 분야에서뿐만 아니라 물리학, 의학, 생물학 등의 자연과학과 같이 다양한 분야에서 사용되고 있다. 소셜네트워크에서 노드는 한 사람 또는 가족, 조직과 같이 더 큰 사회적 그룹을 나타내고 연결선(edge, link, tie, arc)은 두 노드 사이의 연결여부 혹은 연결강도를 나타낸다.

소셜네트워크 분석은 이러한 연결선을 통해 노드들 간의 규칙적인 관계패턴을 파악하여 네트워크의 전체적인 구조를 이해하고 분석하는 것으로 분석 목적에 따라 적절한 통계적 모형을 이용해 다양한 의미를 도출해 낼 수 있다. 소셜네트워크 분석에 이용되는 통계적 모형의 우수성은 정확한 모수추정과 분석속도에 의존한다. 소셜네트워크 분석에 사용되는 지수족 랜덤 그래프 모형과 잠재변수를 고려한 모형은 고전적 모형으로 매우 복잡하고 모수를 추정하기에 어려움이 많다.

따라서 이를 개선하기 위해 본 연구는 메트로폴리스 헤스팅스(Metropolis-Hastings, MH; Hastings, 1970) 알고리즘과 확률적 근사 몬테카를로(stochastic approximation Monte Carlo, SAMC; Liang et al., 2007) 알고리즘을 이용한 방법을 제안한다. SAMC 알고리즘은 국소 트랩의 문제점을 해결할 수 있으며 분포의 제한된 범위에서만 샘플링하여 더 빠르고 정확한 모수추정을 할 수 있는 장점이 있다. 제안된 방법은 모의실험과 실자료 분석을 통해 기존의 방법보다 더 효율적인 추정치를 제공하였으며 우수한 분석속도를 보여준다.

핵심어 : 소셜네트워크, Exponential-family Random Graph Model, Latent Position Cluster Model, Markov Chain Monte Carlo, Metropolis-Hastings, Stochastic Approximation Monte Carlo

목 차

요 약 문	
목 차	
표 목 차	
제1장 서 론	1
제2장 소셜네트워크 기존 모형	4
2.1 지수족 랜덤 그래프 모형	4
2.2 잠재변수 모형	6
제3장 SAMC 알고리즘을 이용한 소셜네트워크 분석	13
3.1 SAMC 알고리즘	13
3.2 추론 개선	15
제4장 실증분석	20
4.1 모의실험	20
4.2 Sampson's Monks Data	29
4.3 Physician Data	34
제5장 결 론	41
참 고 문 헌	42

표 목 차

<표 1>	28
<표 2>	28
<표 3>	29
<표 4>	33
<표 5>	34
<표 6>	34
<표 7>	38
<표 8>	39
<표 9>	40
<그림 1>	21
<그림 2>	22
<그림 3>	24
<그림 4>	25
<그림 5>	27
<그림 6>	31
<그림 7>	36

제1장 서론

네트워크는 노드(node) 또는 객체(actor)들 사이의 관계를 나타내는데 널리 사용되며, 현대 사회학의 핵심 기술로 부상하게 되었다. 이러한 네트워크는 인류학, 인구통계학, 커뮤니케이션 연구, 경제학, 지리학, 역사학, 정보학, 조직학, 정치학, 사회 심리학, 사회 언어학 등을 비롯한 사회과학 분야에서뿐만 아니라 물리학, 의학, 생물학, 컴퓨터 과학 등의 개발 연구 분야에서도 사용된다 (Gormley, Murphy, 2010).

소셜네트워크(social networks)에서 각 노드는 개인 또는 사회집단을 나타내고, 각 연결선(edge, link, tie, arc)은 두 노드 사이의 연결 여부 또는 강도를 나타낸다(Handcock et al., 2007). 노드는 가족 또는 조직과 같이 더 큰 사회적 그룹을 나타내거나 공항, 서버 또는 위치와 같은 객체 또는 개념, 문자, 변수 등의 추상적인 개체를 나타내는데 사용할 수 있다(Ryan et al., 2017).

소셜네트워크 분석은 연결되거나 연결되지 않은 많은 노드들 간의 규칙적인 관계패턴을 파악하여 네트워크의 전체적인 구조를 추정하고 분석하는 것으로 분석 목적에 따라 다양한 의미를 도출할 수 있다. 일반적으로 시각화되는 사회 구조의 예로는 소셜 미디어 네트워크, 우정과 지인 네트워크, 협동 그래프, 질병 전파 등이 있으며, 이러한 네트워크는 종종 사회학을 통해 시각화된다. 이와 관련된 연구는 오랫동안 연구되어 왔으며 대표적인 책으로 Wasserman, Faust(1994)가 있다(Hunter et al., 2012).

소셜네트워크 자료는 n 개의 노드와 연결선 $y_{i,j}$ 로 구성된다. $y_{i,j}$ 가 관계의 유무를 나타내는 이항변수라면 친구들 사이의 관계 유무, 회사들 간의 협력관계 유무 등이 될 수 있다. 자료는 $n \times n$ 형태의 행렬 Y 로 정의하며 소시오매트릭스(sociomatrix)라고 부른다. 또한, 네트워크에서의 관계는 노드 i 에서 j 로 가는 연결선이 있는 경우 반드시 노드 j 에서 i 로 가는 연결선이 있는 네트워크를 비방향 네트워크(undirected network)라고 하고, 그렇지 않은 네트워크를 방향성

네트워크(directed network)로 나뉘지며 자기 자신과의 관계(i, i)는 가질 수 없게 제한한다. 소셜네트워크의 노드를 효율적이고 유연하게 군집화(clustering)하여 네트워크의 구조를 파악하는 것이 소셜네트워크 분석의 목표이며 이를 위한 다양한 통계적 방법이 적용되고 있다.

소셜네트워크 분석에 관한 선행연구는 연결선들의 조건부 독립을 가정한 p_1 모형(Holland, Leinhardt, 1981)이 제안되었다. 하지만 모형 절약성 부족과 공변량을 고려하지 않은 문제가 존재해 확률변수 개념을 추가한 p_2 모형(van Duijn et al., 2004)이 제안되었다. p_2 모형은 모형이 복잡하고 해석하기 어려운 문제점이 있어 p_1 모형과 p_2 모형의 한계를 보완한 지수족 랜덤 그래프 모형(exponential-family random graph model, ERGM)이 제안되었다. 지수족 랜덤 그래프 모형은 Frank, Strauss(1986)에 의해 일반화되었으며 지속적으로 연구되고 있지만 계산상의 어려움과 모형 퇴화 문제가 존재한다. 이러한 문제를 해결하기 위해 잠재변수를 고려한 모형이 제안되었으며 그 중 Hoff et al.(2002)은 각 노드가 알려지지 않은 잠재적 위치를 가지고 있다고 가정한 잠재 공간 모형을 제안했다. 잠재 공간모형은 모수추정 시 비식별 문제가 존재한다. Handcock et al.(2007)은 잠재 공간모형을 바탕으로 모형 기반 군집화 개념을 적용한 잠재적 위치 군집모형을 제안했으나 대용량 네트워크에 적용 시 계산적 문제가 존재한다.

본 논문에서는 대용량 네트워크에도 계산적 문제를 보완하여 더 정확하고 빠르게 모수를 추정하기 위해 메트로폴리스 헤스팅스(Metropolis-Hastings, MH; Hastings, 1970) 알고리즘과 확률적 근사 몬테카를로(stochastic approximation Monte Carlo, SAMC; Liang et al., 2007) 알고리즘을 이용하여 소셜네트워크 분석을 수행하고자 한다. 제2장에서는 소셜네트워크 분석을 위해 사용되는 여러 가지 통계적 모형 중 소셜네트워크 분석에 사용된 기존 모형 지수족 랜덤 그래프 모형, 잠재변수 모형, 잠재적 위치 군집모형 세 가지를 소개한다. 제3장에서는 SAMC 알고리즘을 소개하고, 소셜네트워크 분석에서의 추론 방법을 알아본다. 제4장에서는 모의실험 자료와 실자료를 바탕으로 기존 MH 알고리즘,

변형 근사 알고리즘, 본 연구에서 제안하는 MH-SAMC 알고리즘을 비교한다.

제2장 소셜네트워크 기존 모형

소셜네트워크 분석에 사용되는 대표적인 고전적 통계모형으로는 지수족 랜덤 그래프 모형과 잠재변수 모형, 그리고 잠재적 위치 군집모형이 있다.

2.1 지수족 랜덤 그래프 모형

지수족 랜덤 그래프 모형은 네트워크의 구조를 네트워크 내에 존재하는 모든 연결선들의 결합분포로 표현한 대표적인 모형이다. p_1 모형과 p_2 모형의 한계를 보완한 모형으로 초기에는 p^* 모형으로 불렸으나 현재는 ERGM으로 더 많이 불린다.

p_1 모형은 Holland, Leinhardt(1981)에 의해 제안되었으며 연결선들의 조건부 독립을 가정하고 연결선들의 상호성(reciprocity)을 모형화하는 것이 목적이다. p_2 모형은 p_1 모형에서 확장됐으며 van Duijn et al.(2004)이 제안했다. 노드들간의 이질성을 고려할 수 있게 돼 더 정교한 모형이 되었다.

ERGM은 Frank, Strauss(1986)에 의해 일반화되었으며 Lusher et al.(2013)이 ERGM의 기본 이론적 가정을 다음과 같이 정리하였다.

Lusher's ERGM 기본 가정:

- a) 사회 연결망은 국소적(local)으로 나타난다.
- b) 네트워크의 연결은 스스로 구성될 뿐만 아니라 노드의 속성과 다른 외생적 요인들에 의해 영향을 받는다.
- c) 네트워크의 유형은 구조적 과정의 증거로 볼 수 있다.
- d) 여러 과정들이 동시에 일어날 수 있다.
- e) 사회 연결망은 구조적이며 확률적이다.

위와 같은 가정을 전제로 구성된 식은 (1)과 같다.

$$P_{\theta}(Y=y) = \frac{\exp\{\eta(\theta)^{\top} g(y)\}}{\kappa(\theta)}, \quad y \in S \quad (1)$$

θ 는 모수이고 $g(\cdot)$ 는 네트워크의 특성을 나타내는 충분통계량이다. 식 (1)과 같이 노드들의 연결 확률을 지수족의 형태로 표현할 수 있다면 모수에 의존하지 않는 충분통계량과 네트워크의 특성을 쉽게 찾을 수 있다. S 는 y 의 표본공간이고 $\eta(\cdot)$ 는 자연 모수이다. $\kappa(\theta)$ 는 정규화 상수로 식 (2)와 같다.

$$\kappa(\theta) = \sum_y \exp\{\eta(\theta)^{\top} g(y)\} \quad (2)$$

따라서, 식 (1)을 모두 더하면 1이 된다. Frank, Strauss(1986)는 나머지 네트워크가 조건으로 주어지고 두 노드가 적어도 하나의 노드를 공유하고 있을 경우 확률적으로 종속이라는 마코브(Markov)가정 하에서 ERGM의 통계량을 계산했다. 하지만 분모의 정규화 상수 $\kappa(\theta)$ 의 계산이 어려워 추론 시 문제점이 발생한다. 이를 해결하기 위해 유사 우도추정(pseudo-likelihood estimation), 확률적 근사에 의한 최우추정(MLE by stochastic approximation), 몬테카를로 최대화에 의한 최우추정(MLE by Monte Carlo maximization), 베이지안 방법과 같은 방법들이 고안되었고 본 연구에서는 이 방법들의 추론 방법에 대한 설명은 생략한다.

ERGM은 네트워크의 전체적인 특성을 모형화하는데 적절하지만 다음과 같은 문제가 있다. 첫째, ERGM은 이해하기 어렵고 모형 퇴화(model degeneracy)와 같은 불필요한 특성을 가진다. 둘째, ERGM의 우도함수는 통계적으로 계산하기 복잡하고 어렵다. 셋째, 노드들 사이에 관찰되지 않은 이질성(heterogeneity)이나 구조가 있을 수 있다.

이 중 ERGM의 가장 큰 문제점은 네트워크의 크기가 증가함에 따라 네트워

크의 특정 부분에 확률이 집중되는 퇴화현상이다. 퇴화현상은 모형 부적합 문제를 일으킬 수 있으며 이를 개선하기 위한 방법들이 연구되고 있다. 이러한 ERGM의 여러 문제 때문에 잠재변수를 고려한 새로운 모형들이 제안되었다.

2.2 잠재변수 모형

잠재변수 모형은 크게 네 가지로 구분된다. 랜덤효과 모형(random effect model)과 혼합효과 모형(mixed effect model), 확률적 블록모형(stochastic block model), 잠재 공간모형(latent space model)으로 구분되며 이 모형들을 기반으로 근사 모형을 사용한 변형 방법도 있다. 잠재변수 모형 중 Wyatt et al.(2008)과 Koskinen(2009)을 제외하고 모든 잠재변수 모형은 연결 쌍(dyad)들의 조건부 독립을 가정한다. 만약 Z 가 연속형 또는 이산형인 잠재변수를 나타내는 경우에 모형은 잠재변수가 주어졌을 때 연결 쌍(노드들의 고유한 비 순서 쌍, $(Y_{i,j}, Y_{j,i})$)들의 조건부 독립을 가정한다.

$$P_{\theta}(Y=y|Z=z) = \prod_{(i,j) \in Y} P_{\theta}(Y_{i,j}=y_{i,j}, Y_{j,i}=y_{j,i}|Z=z)$$

방향을 고려한 네트워크는 연결 쌍들이 조건부 독립이라고 해서 연결선 $Y_{i,j}$ 와 $Y_{j,i}$ 가 항상 독립임을 의미하지 않는다. 상호성으로 인해 잠재변수가 조건으로 주어져도 $Y_{i,j}$ 와 $Y_{j,i}$ 가 종속이 될 수 있다. 그러나 대부분의 잠재변수 모형에서 연결선 $Y_{i,j}$ 들은 조건부 독립을 강하게 가정한다.

$$P_{\theta}(Y=y|Z=z) = \prod_{(i,j) \in Y} P_{\theta}(Y_{i,j}=y_{i,j}|Z=z) \quad (3)$$

잠재변수 모형의 장점으로서는 첫째, 연결 쌍들의 조건부 독립을 가정하더라도 네트워크의 종속성을 파악할 수 있다. 잠재 공간모형은 잠재변수 모형에서 개

선된 모형으로 잠재변수를 효율적으로 사용해 상호성과 전이성(transitivity)같은 종속적 구조를 파악할 수 있다. 둘째, 연결선들의 조건부 독립가정으로 모형 퇴화의 문제를 해결하여 모형구축이 쉽다. 셋째, 마코브체인 몬테카를로(Markov chain Monte Carlo, MCMC) 방법을 사용할 수 있어 계산상의 이점이 있다.

2.2.1 랜덤효과 모형과 혼합효과 모형

초기 잠재변수 모형은 van Duijn(1995)의 랜덤효과 모형이다. van Duijn의 랜덤효과 모형은 Holland, Leinhardt(1981)의 p_1 모형에서 모형 절약성(parsimony) 부족으로 인해 제안되었다. p_1 모형은 공변량들을 무시했지만 van Duijn은 공변량들을 확률변수로 고려하여 p_2 모형을 제안했다. 모수추정을 위해 van Duijn et al.(2004)은 p_1 모형이 일반화 선형모형(generalized linear model)으로 표현될 수 있는 점과 p_2 모형이 일반화 선형혼합모형(generalized linear mixed model)으로 표현될 수 있는 사실을 활용했다. Zijlstra et al.(2009)은 p_2 모형을 위한 베이저안 MCMC 방법을 개발했다. 계산에 필요한 방법은 R 패키지 eigenmodel을 통해 구현할 수 있다.

2.2.2 확률적 블록모형

확률적 블록모형은 Snijders, Nowicki(1997)에 의해 연구되었고 Nowicki, Snijders(2001)에 의해 확장되었다. 확률적 블록모형은 노드들을 블록(block)이라 불리는 하위집합으로 나눈다. 블록 구성원(block membership)이 조건부로 주어진 경우 연결선들은 독립이다. 블록 간의 연결확률은 노드가 속한 블록에 의존한다. 또한 같은 블록에 속한 노드들의 연결확률은 다른 노드와의 연결 확률보다 큰 값을 가진다. Tallberg(2005)는 블록 구성원을 예측하기 위해 공변량을 포함하였다. Airolti et al.(2008)은 노드의 블록 구성원이 노드의 쌍에 의존

하는 혼합 구성원 모형(mixed membership model)이라는 더 발전된 확률적 블록모형을 제안했다. 즉, 노드가 속한 블록은 노드들의 연결선 여부에 따라 달라진다는 개념이다.

Nowicki, Snijders(2001)가 제안한 확률적 블록모형은 두 가지 가정에 기초한다.

첫째, 노드의 집합은 K 개 블록으로 나뉘지며 K 는 알려진 고정 값이다.

$$Z_i | \pi_1, \dots, \pi_K \sim \text{Multinomial}(1; \pi_1, \dots, \pi_K) \quad (4)$$

둘째, $P_\theta(Y_{i,j} = 1 | Z_i = z_i, Z_j = z_j) = \theta_{z_i, z_j}$ 에서 $\theta_{k,l}$ 은 군집 k 에 있는 노드가 군집 l 의 노드와 연결될 확률을 의미한다.

확률적 블록모형은 원래 이항 네트워크자료를 주로 분석하기 위해 개발되었지만 특정 값이 있는 연결선(Mariadassou et al.,(2010))과 범주형 연결선(Jernite et al.,(2014))과 같은 자료의 분석도 가능하게 되었다(Bouveyron et al., 2016).

그러나 확률적 블록모형은 블록의 라벨(label)에 따라 우도함수가 변하지 않아 모수를 식별할 수 없는 라벨 전환(label-switching) 문제점이 있다. 즉, 각각 다른 블록의 라벨이 같은 우도함수의 값을 가지는 것이다. 문제 개선을 위해 모수에 대한 순서를 제한하거나 두 노드가 같은 블록에 속해 있는지 알려주는 지시함수를 사용하는 방법 등 지속적으로 라벨 전환 문제에 대한 연구가 이루어지고 있다.

2.2.3 잠재 공간모형

잠재 공간모형은 Hoff et al.(2002)에 의해 제안되었다. 네트워크자료의 확률적 모형이며 잠재적 위치는 일반적인 통계이론들로 추정된다. 각 노드는 유클리디언(Euclidean) 공간에서 잠재적 위치(latent position)를 갖는다(Hoff et al.,

2002). 즉, 각 노드는 알려지지 않은 잠재적 위치 Z 를 가진다고 가정한다. 또한 이 모형을 이용해 전이성과 관측된 자료로부터 동질성을 찾을 수 있다. 잠재 공간모형은 거리 공간을 유클리디언 공간으로 가정하는 경우(Hoff et al., 2002)와 울트라메트릭(ultrametric)으로 가정하는 경우(Schweinberger, et al., 2003)가 있다. 두 경우 모두 식 (5)의 형태로 표현된다.

$$\text{logit}(P_{\beta}(Y_{i,j} = 1 | Z = z_i, Z = z_j)) = \beta_0 + x_{i,j}^{\top} \beta + d(z_i, z_j) \quad (5)$$

식 (5)에서 $x_{i,j}$ 는 연결 쌍(i, j)의 공변량을 나타내고 β_0 는 네트워크의 밀도(density)를 조절하는 모수이며 $d(\cdot, \cdot)$ 는 i 와 j 가 가지는 잠재적 위치의 거리 함수이다. 잠재 공간모형은 노드의 특성을 나타내는 발신자와 수신자 효과를 추가하여 Hoff(2005)에 의해 아래와 같은 식으로 확장되었다.

$$\text{logit}(P_{\beta}(Y_{i,j} = 1 | Z = z_i, Z = z_j)) = \beta_0 + x_{i,j}^{\top} \beta + d(z_i, z_j) + \delta_i + \gamma_j$$

δ 은 발신자 효과를 의미하고 γ 은 수신자 효과를 의미하며 두 효과는 정규분포를 가정한다.

2.2.4 잠재위치 군집모형

잠재위치 군집모형(latent position cluster model, LPCM)은 Handcock et al.(2007)에 의해 제안되었다. 잠재위치 군집모형은 Hoff et al.(2002)의 잠재 공간모형을 바탕으로 모형 기반 군집화(Fraley, Raftery, 2002) 개념을 적용한 모형이다. 이 모형은 전이성과 관측된 자료로부터 동질성을 찾을 수 있고 동시에 노드들의 군집화까지 가능하다.

잠재적 위치 군집모형은 d 차원의 유클리디언 잠재공간에서 노드들이 관찰되지 않은 임의의 잠재위치 z_i 를 가지고 있다고 가정한다. 두 노드 사이의 연결

확률은 노드들의 잠재위치가 주어진 경우 다른 연결선들과 독립이라고 가정하고 식 (6)과 같이 표현할 수 있다.

$$P(Y|Z, X, \beta) = \prod_{i \neq j} P(y_{i,j} | z_i, z_j, x_{i,j}, \beta) \quad (6)$$

Z 는 모든 노드들의 잠재위치를 나타내는 $n \times d$ 행렬로 $Z = (z_1, z_2, \dots, z_n)^\top$ 이다. z_i 는 $z_i = (z_{i1}, z_{i2}, \dots, z_{id})$ 인 $1 \times d$ 벡터로 각 노드가 가지는 잠재위치를 나타낸다. β 는 추정해야 할 모수 중 하나로 스칼라 값을 가진다. 로지스틱 회귀모형을 이용하여 식 (6)을 (7)과 같이 모형화할 수 있다.

$$\log[\text{odds}(y_{i,j} = 1 | z_i, z_j, x_{i,j}, \beta)] = \beta x_{i,j} - |z_i - z_j| \quad (7)$$

여기서 어떤 사건 A 의 로그 오즈는 $\log[\text{odds}(A)] = \log[P(A)/\{1 - P(A)\}]$ 이며 식 (7)에서 $|z_i - z_j|$ 는 잠재공간에서 노드 i 와 j 사이의 거리 차이 값이다. 모든 노드들의 확률은 다음과 같다.

$$P(Y|Z, \beta, X) = \prod_{i=1}^n \prod_{j \neq i}^n \left[\frac{\exp(\beta x_{i,j} - |z_i - z_j|)}{1 + \exp(\beta x_{i,j} - |z_i - z_j|)} \right]^{y_{i,j}} \left[\frac{1}{1 + \exp(\beta x_{i,j} - |z_i - z_j|)} \right]^{(1 - y_{i,j})}$$

연결확률은 β 와 z_i 에 의존하므로 정확한 β 와 z_i 의 추정이 필요하다. 잠재적 위치 군집모형이 잠재 공간모형과 다른 점은 첫째, 노드들의 군집화를 위해 잠재위치 z_i 를 G 개의 다변량 정규분포의 유한한 혼합으로부터 뽑는다. 각각의 다변량 정규분포는 군집별로 다른 평균과 공분산 행렬을 가진다.

$$z_i \sim \sum_{g=1}^G \lambda_g MVN_d(\mu_g, \sigma_g^2 I_d) \quad (8)$$

식 (8)에서 λ_g 는 노드가 g 번째 군집에 속할 확률로 $\lambda_g \geq 0$ ($g = 1, \dots, G$)이고 $\sum_{g=1}^G \lambda_g = 1$ 이다. I_d 은 $d \times d$ 단위행렬이다. 식 (8)은 Banfield, Raftery(1993)에 의해 제안되었으며 변수들의 군집화를 위한 모형이다.

둘째, 잠재위치가 너무 큰 값을 가지지 않게 다음과 같은 제약조건을 설정한다.

$$\sqrt{\left(\frac{1}{n} \sum_i |z_i|^2\right)} = 1 \quad (9)$$

잠재적 위치 군집모형의 모수추정에는 두 가지 방법이 있다. 첫 번째는 두 단계에 걸쳐 최우추정치를 계산하는 방법이다. 먼저 군집화되지 않은 잠재 공간 모형의 최우추정치를 구한 뒤 혼합모형에 대한 최우추정치를 계산한다. 이 방법은 상대적으로 빠르고 간단하지만 잠재위치를 추정할 때 군집의 정보를 이용하지 않는다. 두 번째는 MCMC 샘플링을 이용한 완전한 베이지안(fully Bayesian) 방법으로 잠재공간과 군집화모형을 동시에 추정한다. 첫 번째 방법보다 계산 측면이나 수학적으로 더 까다롭지만 군집화에 관한 정보와 잠재 위치의 불확실성의 정보 손실을 막을 수 있다. Handcock et al.(2007)은 두 번째 방법을 사용하여 모수추정을 하였다.

최근의 연구에서는 잠재 위치 군집모형에서 기존의 표본생성 분포에서 군집의 정보를 이용하여 일정한 분산이 아닌 각 군집별 분산을 이용하여 표본생성 분포를 사용하였다(김지용, 2017). 식 (9)에서 잠재위치 z_i 에 대한 제약조건을 두지 않아 더 유연한 모수추정을 하였다. 하지만 R 프로그램을 통한 분석으로 인하여 시간이 오래 걸리기 때문에 충분한 반복 수를 통한 분석에 제한이 있었으며, 대용량 네트워크 데이터에 적용할 수 없다는 한계가 있었다.

본 연구에서는 고전적 모형에 비해 이해하기 쉽고 고전적 모형에서 발생한

문제점들을 보완한 잠재 위치 군집모형을 바탕으로 연구를 진행하였다.

2.2.5 변형방법

베이저안 MCMC 방법은 잠재변수 모형에서 유용하게 사용되지만 속도 측면에서 느리다. 또한, 노드 개수가 많은 대용량 네트워크의 경우엔 적용할 수 없는 문제가 있다. 변형방법은 베이저안 MCMC 방법의 근사적인 대안으로 빠르게 실현가능하다. 이를 바탕으로 확률적 블록모형의 근사 최우추정은 Daudin et al.(2008)에 의해 도입되었고 변형된 기댓값 최대화(expectation-maximization, EM) 알고리즘을 제시하여 600개 이상의 노드에 적용했다. 근사 최우추정의 일치성은 Nowicki, Snijders(2001)의 확률적 블록모형을 고려한 Celisse et al.(2011)에 의해 확립되었다.

다양한 변형방법 중 Salter-Townshend, Murphy(2013)는 잠재 공간모형의 근사 베이저안 추정방법을 제안하였고 80개 이상의 노드에 적용했다. 이 방법은 MCMC 방법을 통한 표본추출 대신 최적화(optimization) 방법을 사용하여 대용량 네트워크에도 적용가능하게 했다(Salter-Townshend, Murphy, 2013). 변형방법은 본 논문에서 비교하는 모형들과 달리 다른 모형에서 표본을 추출하기 때문에 추정된 모수 값을 직접 비교하기가 어려워, 기존의 방법과 제안하는 방법의 적합된 모형의 잠재위치 산점도와 분석속도 비교를 위해서만 변형방법을 사용하였다.

제3장 MH-SAMC 알고리즘을 이용한 소셜네트워크 분석

본 장에서는 기존 샘플링 방법인 MH 알고리즘보다 우수한 성능을 보이는 SAMC 알고리즘에 대한 소개와 소셜네트워크 모형에서의 적용에 대해 설명한다.

3.1 SAMC 알고리즘

표본공간 분할을 통해 국소 트랩(local trap) 문제를 해결하는 SAMC 알고리즘(Liang et al, 2007)을 간단하게 설명한다. 먼저 다음과 같은 형태의 분포로부터 표본을 추출하고자 한다.

$$f(x) = \frac{1}{Z} \psi(x), x \in \mathcal{X} \quad (10)$$

Z 는 정규화 상수이고, \mathcal{X} 는 표본공간, $\psi(x)$ 는 비음(non-negative) 함수이다. 표본공간은 정해진 상수 u_0, \dots, u_{m-1} 에 따라 함수 $U(x)$ 가 $m+1$ 개의 분리된 하위영역으로 분할된다. 즉, $E_0 = \{x : U(x) \leq u_0\}$, $E_1 = \{x : u_0 < U(x) \leq u_1\}$, \dots , $E_{m-1} = \{x : u_{m-2} < U(x) \leq u_{m-1}\}$, $E_m = \{x : U(x) > u_{m-1}\}$ 으로 표현할 수 있다. SAMC는 미리 정해진 빈도를 이용하여 각 영역에서 표본을 추출한다. $m+1$ 개의 하위영역은 비어 있지 않은 영역(non-empty)로 가정한다. 즉, $i = 0, \dots, m$ 에 대해 $w_i = \int_{E_i} \psi(x) dx > 0$ 이다. $\pi = (\pi_0, \pi_1, \dots, \pi_m)$ 를 $0 < \pi_i < 1$ 와 $\sum_{i=0}^m \pi_i = 1$ 를 만족하는 $m+1$ 벡터인 각 부분 영역에서의 원래 표본함수라고 정의한다. $i = 0, \dots, m$ 에 대해 $\theta_i = \log(\int_{E_i} \psi(x) dx / \pi_i) = \log(\frac{w_i}{\pi_i})$, $\theta = (\theta_0, \theta_1, \dots, \theta_m)$, Θ 는 θ 의 공간이라 정

의한다. $\psi(x)$ 의 일반적인 선택을 위해 $\Theta = \mathbb{R}^{m+1}$ 이라고 한다. $\theta^{(t)} = (\theta_0^{(t)}, \theta_1^{(t)}, \dots, \theta_m^{(t)})$ 는 t 번째 반복에서 얻어진 θ 의 추정치이다. 식 (10)번을 식 (11)번 식과 같이 다시 표현할 수 있다.

$$f_{\theta^{(t)}}(x) \propto \sum_{i=0}^m \frac{\psi(x)}{e^{e_i^{(t)}}} I(x \in E_i) \quad (11)$$

$\{\gamma_t\}$ 는 어떤 $\tau \in (1, 2)$ 에 대해

$$(a) \sum_{t=1}^{\infty} \gamma_t = \infty, \quad (b) \sum_{t=1}^{\infty} \gamma_t^{\tau} < \infty \quad (12)$$

를 만족하는 양수이고 단조증가하는 수열이라고 정의한다. 본 연구에서는 $t_0 > 1$ 인 어떤 정해진 값에 대해 $\gamma_t = \frac{t_0}{\max(t_0, t)}$, $t = 1, 2, \dots$ 을 사용한다. 이와 같은 가정을 이용하여 SAMC 알고리즘을 정리하면 다음과 같다.

SAMC 알고리즘:

(a) (샘플링) 목표(target) 분포를 가지고 한 번의 MH 갱신에 의해 표본

$x^{(t+1)}$ 추출한다.

(a.1) 제안(proposal) 분포 $q(x^{(t)}, y)$ 에 따라 y 를 생성시킨다.

(a.2) 채택확률을 다음과 같이 구한다.

$$\alpha_{x^{(t)}, y} = \min \left\{ 1, e^{\theta_{f(x^{(t)})}^{(t)} - \theta_{f(y)}^{(t)}} \frac{\psi(y) q(y, x^{(t)})}{\psi(x^{(t)}) q(x^{(t)}, y)} \right\}$$

(a.3) 채택확률 $\alpha_{x^{(t)}, y}$ 의 확률로 $x^{t+1} = y$ 을 채택하고, $1 - \alpha_{x^{(t)}, y}$ 의 확률로

$x^{t+1} = x^t$ 을 채택한다.

(b) (θ 갱신) $\theta^* = \theta_t + \gamma_{t+1}(e_{t+1} - \pi)$ 라고 정의한다.

$J(x)$ 는 표본 x 가 속해 있는 하위영역의 인덱스이며 $e_{t+1} = (e_{t+1,1}, \dots, e_{t+1,m})$ 이다. 만약 $x^{(t)} \in E_i$ 이면 $e_{t+1,i} = 1$, 그렇지 않으면 $e_{t+1,i} = 0$ 이다. 만약 $\theta^* \in \Theta$ 이면 $\theta_{t+1} = \theta^*$, 그렇지 않으면 $\theta_{t+1} = \theta^* + c^*$ 이다. 즉, y 가 뽑힌 지역에 낮은 가중치를 부여하여 다른 영역에서 y 가 뽑힐 수 있도록 θ 를 갱신한다. 여기서 $c^* = (c^*, \dots, c^*)$ 이고, $\theta^* + c^* \in \Theta$ 를 만족하는 임의의 벡터이다.

3.2 추론 개선

본 절에서는 기존 Handcock et al.(2007)에 의해 제안된 잠재위치 군집모형에서의 추론 방법에 대해 설명한 후 SAMC 알고리즘을 이용한 추론 개선방법을 소개한다.

먼저 식 (6)-(9)와 MCMC 샘플링을 이용해 잠재위치 군집모형의 베이지안 추정을 실시한다. 여기서 i 번째 노드가 어느 군집에 속해 있는지 알려주는 새로운 변수 K_i 를 이용한다. K_i 는 군집구성(group membership) 변수를 의미한다. 모수에 대한 사전분포는 다음과 같다.

$$\begin{aligned}\beta &\sim Normal(\xi, \psi^2) \\ (\lambda_1, \dots, \lambda_G) &\sim Dirichlet(\nu_1, \dots, \nu_G) \\ \mu_g &\sim MVN_d(0, \omega^2 I_d), \quad g = 1, \dots, G \\ \sigma_g^2 &\sim \sigma_0^2 \in \nu \chi_\alpha^2, \quad g = 1, \dots, G\end{aligned}$$

$\xi, \psi^2, (\nu_1, \dots, \nu_G), \sigma_0^2, \alpha, \omega^2$ 은 하이퍼 파라미터(hyperparameters)들로 Handcock et al.(2007)과 같이 $\xi = 0, \psi^2 = 2I, \nu_g = 3, \sigma_0^2 = 0.103, \alpha = 2, \omega^2 = 2$ 로

설정했다. MCMC 알고리즘은 앞에서 정의한 사전분포, 잠재위치 z_i 그리고 군 집 구성변수 K_i 를 이용하여 모수추정을 반복한다. 모수추정에는 깃스 샘플링 (Gibbs sampling)과 MH 알고리즘을 이용한다. z_i 와 β 의 경우 완전 조건부 사후분포를 구할 수 없어 MH 알고리즘을 사용하고 남은 모수들은 깃스 샘플링을 사용해 추정한다. 잠재위치와 군 집 구성변수를 비롯한 수식에 명시되지 않는 모수들은 ‘others’로 표시한다. 완전 조건부 사후분포는 아래와 같다. $\phi_d(\cdot; \mu, \Sigma)$ 는 d 차원 다변량 정규분포의 확률밀도를 나타낸다.

$$z_i | K_i = g, \text{ others} \propto \phi_d(z_i; \mu_g, \sigma_g^2 I_d) P(Y | Z, \beta), \quad i = 1, \dots, n \quad (13)$$

$$\beta | Z, \text{ others} \propto \phi(\beta; \xi, \psi^2) P(Y | Z, \beta), \quad (14)$$

$$\lambda | \text{others} \sim \text{Dirichlet}(m + \nu), \quad (15)$$

$$\mu_g | \text{others} \sim \text{MVN}_d\left(\frac{m_g \bar{z}_g}{m_g + \sigma_g^2 / \omega^2}, \frac{\sigma_g^2}{m_g + \sigma_g^2 / \omega^2} I\right), \quad g = 1, \dots, G \quad (16)$$

$$\sigma_g^2 | \text{others} \sim (\sigma_0^2 + d s_g^2) \in v \chi_{\alpha + m_g d}^2, \quad g = 1, \dots, G \quad (17)$$

$$P(K_i = g | \text{others}) = \frac{\lambda_g \phi_d(z_i; \mu_g, \sigma_g^2 I_d)}{\sum_{r=1}^G \lambda_r \phi_d(z_i; \mu_r, \sigma_r^2 I_d)}, \quad i = 1, \dots, n, g = 1, \dots, G \quad (18)$$

여기서 m_g , s_g^2 과 \bar{z}_g 는 다음과 같이 정의한다.

$$m_g = \sum_{i=1}^n I_{[K_i = g]},$$

$$s_g^2 = \frac{1}{d} \sum_{i=1}^n (z_i - \mu_g)^\top (z_i - \mu_g) I_{[K_i = g]},$$

$$\bar{z}_g = \frac{1}{m} \sum_{i=1}^n z_i I_{[K_i = g]},$$

모수추정을 위해 Handcock et al.(2007)에 의해 제안된 알고리즘은 다음과 같

다.

Handcock's MH 알고리즘(2007):

t 번째 반복에서 얻은 표본을 $Z^t = (z_1^t, z_2^t, \dots, z_n^t)^\top$, $z_i^t = (z_{i1}^t, z_{i2}^t, \dots, z_{id}^t)$, β^t , K_i^t , μ_g^t , $\sigma_g^{2(t)}$, λ_g^t 라고 하자.

(단계 1) MH 알고리즘을 이용해 새로운 $Z^{t+1} = (z_1^{t+1}, \dots, z_n^{t+1})^\top$ 을 생성하고

각 노드는 임의의 순서로 업데이트 한다. $i = 1, \dots, n$ 에 대하여,

(a) 표본생성분포 $MVN_d(z_i^t, \delta_Z^2 I_d)$ 로부터 새로운 z_i^* 를 생성한다.

(b) 채택확률을 다음과 같이 구한다.

$$\alpha_{z_i^t, z_i^*} = \min \left\{ 1, \frac{P(Y|Z^*, \beta^t) \phi_d(z_i^*; \mu_{K_i}, \sigma_{K_i}^2 I_d)}{P(Y|Z^t, \beta^t) \phi_d(z_i^t; \mu_{K_i}, \sigma_{K_i}^2 I_d)} \right\}$$

(c) 채택확률 $\alpha_{z_i^t, z_i^*}$ 의 확률로 $z_i^{t+1} = z_i^*$ 을 채택하고, $1 - \alpha_{z_i^t, z_i^*}$ 의 확률로 $z_i^{t+1} = z_i^t$ 을 채택한다.

(단계 2) MH 알고리즘을 이용하여 새로운 β^{t+1} 을 생성한다.

(a) 표본생성분포 $Normal(\beta^t, \delta_\beta^2 I_d)$ 로부터 새로운 β^* 를 생성한다.

(b) 채택확률을 다음과 같이 구한다.

$$\alpha_{\beta^t, \beta^*} = \min \left\{ 1, \frac{P(Y|Z^{t+1}, \beta^*) \phi_d(\beta^*; \xi, \psi^2)}{P(Y|Z^{t+1}, \beta^t) \phi_d(\beta^t; \xi, \psi^2)} \right\}$$

(c) 채택확률 $\alpha_{\beta^t, \beta^*}$ 의 확률로 $\beta^{t+1} = \beta^*$ 을 채택하고, $1 - \alpha_{\beta^t, \beta^*}$ 의 확률로 $\beta^{t+1} = \beta^t$ 을 채택한다.

(단계 3) 식 (10)-(15)를 이용하여 남은 모수 K_i^t , μ_g^t , $\sigma_g^{2(t)}$, λ_g^t 를

깁스 샘플링(Gibbs sampling)을 이용하여 K_i^{t+1} , μ_g^{t+1} , $\sigma_g^{2(t+1)}$, λ_g^{t+1} 로 업데이트 한다.

잠재위치 군집모형은 z_i 와 β 를 추정할 때 MH 알고리즘을 사용한다. (단계 1)에서는 잠재위치를 추정하기 위해 MH 알고리즘을 사용하였는데 잠재위치는 글자 그대로 잠재적인 변수로 추론에 확실한 근거를 제시할 수 없어 추론 개선에 큰 영향을 미치지 않는다고 판단하여 본 연구에서는 (단계 2)에서 β 를 추정할 때 MH 알고리즘 대신 SAMC 알고리즘을 적용하였고, 남은 모수들은 깁스 샘플링을 통해 추정하였다.

R에서의 잠재위치 군집모형의 모수추정을 위한 알고리즘은 latentnet 패키지에 내장된 ergmm 함수를 사용하며 기본 형식은 다음과 같다.

```
ergmm(formula, response = NULL, family = "Bernoulli", fam.par = NULL,
       control = control.ergmm(), user.start = list(), prior = ergmm.prior(),
       tofit = c("mcmc", "mkl", "mkl.mbc", "procrustes", "klswitch"),
       Z.ref = NULL, Z.K.ref = NULL, seed = NULL, verbose = FALSE)
```

여기서 ‘formula’ 부분은 반드시 필요한 부분으로 모형 식이 들어간다. 나머지는 옵션으로 상황에 따라 사용 여부가 달라지며 별도로 지정하지 않을 경우 명시된 값이 기본값으로 지정된다. ergmm의 기본적인 사용을 위한 코드 설명은 4장에서 설명한다.

본 논문에서는 모수 β 의 개선된 추정을 위해 Handcock et al.(2007)에서 사용한 MH 알고리즘 대신 SAMC 알고리즘을 이용한 MH-SAMC 알고리즘을 다음과 같이 제안한다.

MH-SAMC 알고리즘:

t 번째 반복에서 얻은 표본을 $Z^t = (z_1^t, z_2^t, \dots, z_n^t)^\top$, $z_i^t = (z_{i1}^t, z_{i2}^t, \dots, z_{id}^t)$, β^t , K_i^t , μ_g^t , $\sigma_g^{2(t)}$, λ_g^t 라고 하자.

(단계 1) Handcock's MH 알고리즘(2007)과 동일하다.

(단계 2) (샘플링) 목표분포 $\beta | Z, \text{others} \propto \phi(\beta; \xi, \psi^2) P(Y | Z, \beta)$ 를 가지고

한 번의 MH 갱신으로 표본 $\beta^{(t+1)}$ 추출한다.

(a.1) 제안 분포 $Normal(\beta^t, \delta_\beta^2 I_d)$ 로부터 새로운 β^* 를 생성시킨다.

(a.2) 채택확률을 다음과 같이 구한다.

$$\alpha_{\beta^{(t)}, \beta^*} = \min \left\{ 1, e^{\frac{\theta_{f(\beta^*)}^{(t)} - \theta_{f(\beta^t)}^{(t)}}{\delta_\beta^2}} \frac{\psi(\beta^*) q(\beta^*, \beta^{(t)})}{\psi(\beta^{(t)}) q(\beta^{(t)}, \beta^*)} \right\}$$

(a.3) 채택확률 $\alpha_{\beta^{(t)}, \beta^*}$ 의 확률로 $\beta^{t+1} = \beta^*$ 을 채택하고, $1 - \alpha_{\beta^{(t)}, \beta^*}$ 의 확률로

$\beta^{t+1} = \beta^t$ 을 채택한다.

(b) (θ 갱신) $\theta^* = \theta_t + \gamma_{t+1}(e_{t+1} - \pi)$ 라고 정의한다.

(단계 3) Handcock's MH 알고리즘(2007)과 동일하다.

본 연구에서는 SAMC 알고리즘을 시행하기 위해서 식 (14)로부터 추출할 표본구간을 제한시키기 위한 MH 알고리즘을 10000번 시행한 후 90000번의 SAMC 알고리즘을 시행하였다. β 를 제한된 구간에서만 균일하게 샘플링하면 기존의 표본 추출분포 전 구간에서 샘플링 하는 것보다 훨씬 효율적이며 샘플링 과정에서 발생하는 국소 트랩 문제를 해결할 수 있어 더 정확한 모수추정 개선과 빠른 실행 속도가 나올 것으로 기대된다.

제4장 실증분석

본 장에선 모의실험을 통해 세 가지 모수추정 알고리즘을 비교하고자 한다. 비교에 사용된 알고리즘으로 잠재위치 군집모형에서 쓰인 Handcock's MH 알고리즘(HMH), Salter-Townshend, Murphy(2013)가 제안한 알고리즘(STM) 그리고 본 논문에서 제안한 MH-SAMC 알고리즘이다. STM 알고리즘의 경우 HMH 알고리즘에 사용된 모형의 근사적 변형모형을 사용하여 나머지 2개의 알고리즘과 같은 모수 추정치를 제공하지 않아 직접 비교할 수 없다. 따라서 STM 알고리즘은 적합한 잠재위치 산점도를 통해 비교하였다. 실증분석을 위한 통계 패키지로는 R 버전 3.6.0을 사용하였다.

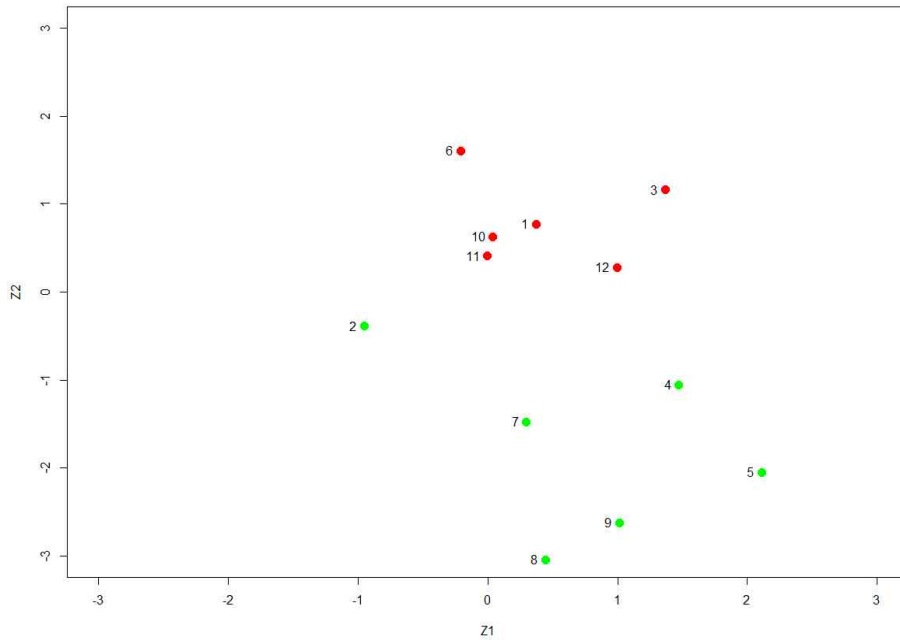
4.1 모의실험

모의실험을 위해서는 소셜네트워크 자료인 행렬 Y 를 생성해야 한다. 우선 노드와 군집의 수가 작은 자료에 먼저 적용하기 위해 2차원 공간을 바탕으로 2개의 군집으로 설정했다. 모수추정 결과의 비교를 위해 모수의 참값은 다음과 같이 설정하고 아래 식으로부터 생성하였다.

$$\begin{aligned}
 \beta &= 1.0 \\
 \mu &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 0 & 0 \end{bmatrix} \\
 \sigma^2 &= \begin{bmatrix} \sigma_1^2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 0.1 & 0.5 \end{bmatrix} \\
 z_i &\sim MVN_2(\sum K_{ig}\mu_g, \sum K_{ig}\sigma_g^2 I_g) \\
 Y_{i,j} &\sim Bernoulli(\text{logit}^{-1}(\beta - |z_i - z_j|)) \\
 \lambda^2 &= \begin{bmatrix} \lambda_1^2 & \lambda_2^2 \end{bmatrix} = \begin{bmatrix} 0.3 & 0.7 \end{bmatrix}
 \end{aligned} \tag{19}$$

여기서 $i, j = 1, \dots, N$ 이며 모의실험 자료는 12개의 노드로 설정하였다. 공변량

X 는 대각원소가 0이며 나머지 값은 모두 1인 12×12 행렬로 지정하였다. 잠재 위치 z_i 를 식 (7)에 따라 랜덤하게 생성한 산점도는 <그림 1>과 같다.



<그림 1> 랜덤 잠재위치 z_i

모든 모수의 설정이 끝난 뒤 식 (19)의 로짓함수의 역함수를 이용해 노드들 간의 연결확률이 0.5 이상이면 1로, 0.5 미만은 0으로 정의하였다. 그 결과 생성된 모의실험 자료는 <그림 2>와 같다. 모의실험 자료 Y 는 12×12 인 행렬로 노드 간 관계유무가 0과 1로 되어있는 이항자료이다. 그리고 자신과의 관계는 허용하지 않아 대각원소는 0이다.

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]
[1,]	0	0	1	0	0	1	0	0	0	1	1	1
[2,]	0	0	0	0	0	0	1	0	0	1	1	0
[3,]	1	0	0	0	0	1	0	0	0	1	0	1
[4,]	0	0	0	0	1	0	1	0	0	0	0	1
[5,]	0	0	0	1	0	0	0	0	1	0	0	0
[6,]	1	0	1	0	0	0	0	0	0	1	1	0
[7,]	1	0	0	1	0	0	0	0	1	0	0	0
[8,]	0	0	0	0	0	0	0	0	1	0	0	0
[9,]	0	0	0	0	1	0	1	1	0	0	0	0
[10,]	1	1	1	0	0	1	0	0	0	0	1	1
[11,]	1	1	0	0	0	1	0	0	0	1	0	1
[12,]	1	0	1	1	0	0	0	0	0	1	1	0

<그림 2> 모의실험 자료 Y

모의실험 자료를 이용해 추정해야 하는 모수는 총 9개다. μ_{11} 과 μ_{12} 는 첫 번째 군집의 평균, μ_{21} 과 μ_{22} 는 두 번째 군집의 평균이며 σ_1^2 과 σ_2^2 은 각 군집의 분산을 나타낸다. 마지막으로 λ_1 과 λ_2 는 각 군집에 속할 확률이다. 제안된 알고리즘을 사용한 모수 β 의 정확한 추론에 초점을 두었다. 다음으로 모의실험 자료를 이용해 세 가지 알고리즘의 모수 추정결과를 비교해 보았다.

4.1.1 Handcock's MH 알고리즘

HMH을 통한 모수추정은 R의 latentnet패키지의 ergmm함수를 사용한다.

```
ergmm(y ~ euclidean(d = 2, G = 2), control = control.ergmm(sample.size =
100000, burnin = 0, interval = 1))
```

모형 적합에는 유클리디언 거리를 사용했다. y 는 앞에서 생성한 모의실험 자료가 들어가고 공간은 2차원($d=2$) 그리고 군집 수는 2개($G=2$)이다. 위 모형에서 'control = control.ergmm' 옵션은 기본값으로 제공되는 값들인데 sample.size는 모형적합에서 최종적으로 저장할 sample 개수를 지정하고, burnin은 burnin sample 개수를, 마지막으로 interval은 sampling 중에서 최종

적으로 저장할 간격을 지정하는 옵션이다. 본 분석에서는 ‘burnin sample = 0’으로 하여 따로 burn sample을 생성하지 않았으며, 각 실행 결과를 모두 저장하기 위해 ‘interval = 1’로, 샘플의 수를 위해 ‘sample.size = 100000’으로 지정하였다.

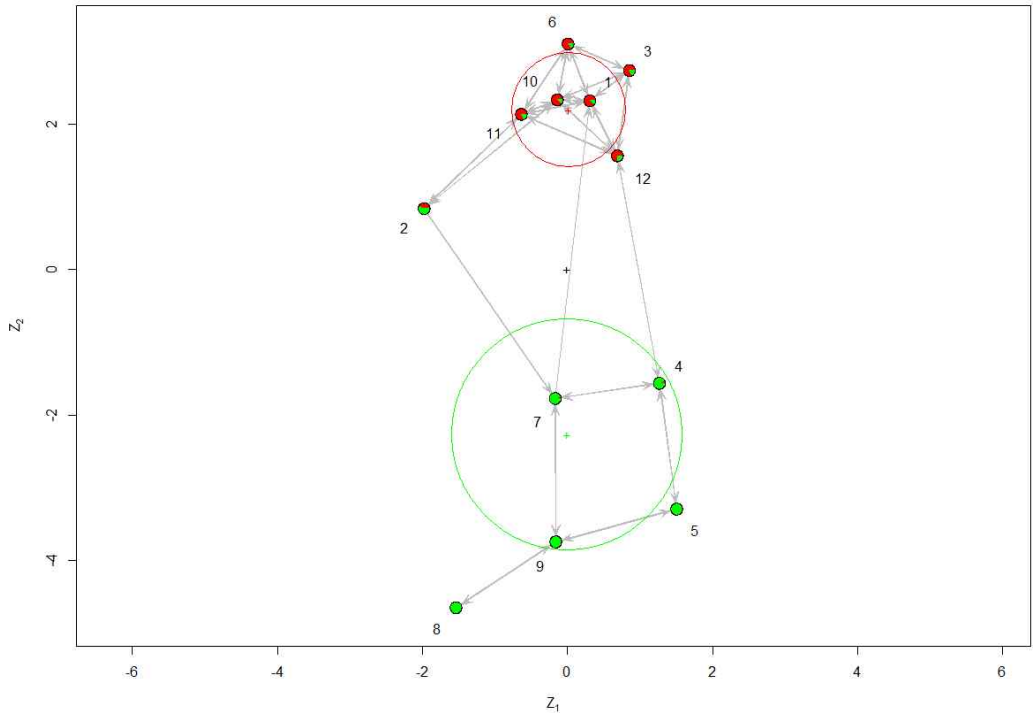
모형적합 후 잠재위치의 산점도를 보고 노드들의 군집화를 확인할 수 있다. 이 때 필요한 명령어는 R에서 plot으로 다음과 같고 <그림 2>를 통해 확인할 수 있다.

```
plot(x , pie = TRUE , vertex.cex = 2.5)
```

x 에는 ergmm을 통해 적합한 결과가 들어간다. ‘pie = TRUE’는 옵션으로 각 노드가 현재의 군집에 속할 사후확률(posterior probabilities)을 제공한다. ‘vertex.cex = 2.5’는 노드의 크기를 조정하는 옵션이다.

<그림 3>를 보면 총 노드의 수는 12개이고 x 축과 y 축은 각각 2차원 공간의 잠재위치 z_1 과 z_2 이다. 노드(1, 3, 6, 10, 11, 12)가 군집 1, 나머지 노드(2, 4, 5, 7, 8, 9)들이 군집 2로 군집화되었다.

노드들이 군집에 속할 사후확률을 제공한다. 각 노드에서 군집과 동일한 색의 부분이 클수록 군집에 속할 사후확률이 큰 것을 의미한다.



<그림 3> HMM를 통해 적합한 잠재위치 산점도

4.1.2 Salter-Townshend, Murphy 알고리즘

변형방법 중 Salter-Townshend, Murphy(2013)가 제안한 알고리즘 STM은 R의 VBLPCM패키지의 `vblpcmfit`함수를 이용한다. STM은 모수추정 시 기존의 잠재 위치 군집모형에 사용한 MCMC 방법이 아닌 근사적인 방법을 통한 최적화 방법을 사용한다. 그로 인해 대용량 네트워크에서도 빠른 속도로 적용 가능하다.

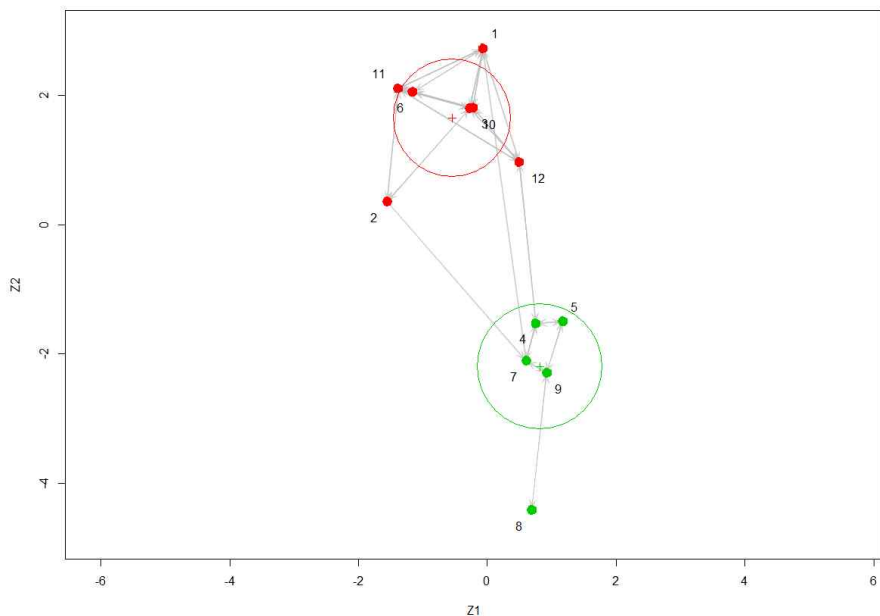
```
vblpcmfit((vblpcmstart(y, d = 2, G = 2)), STEPS = 100000)
```

`vblpcmfit`함수에서 맨 처음 들어가는 문장은 모형적합을 위한 모수들의 초기

값이 들어가야 하기 때문에 `vblpcmstart` 함수가 사용된다. `vblpcmstart` 함수는 빠르게 모수들의 값을 계산하며 안에 들어가는 y 는 소셜네트워크 자료이다. 잠재위치 군집모형과 마찬가지로 2차원($d=2$) 공간과 군집 수는 2개($G=2$)이다. `vblpcmfit`함수에서 'STEPS = 20' 옵션은 최대 반복 횟수를 지정하는 옵션으로 기본값이 20번 반복이며 본 연구에서는 HMM 방법과 동일하게 반복을 지정하였다. 잠재위치의 산점도를 그리기 위한 명령어는 잠재위치 군집모형과 다르게 `plot.vblpcm`을 사용한다.

`plot.vblpcm(x, R2 = 0.5, xlab = 'Z1', ylab = 'Z2')`

x 에는 `vblpcmfit`함수를 통해 적합한 결과가 들어간다. ' $R2 = 0.5$ '는 각 노드의 크기를 지정해주는 옵션이다. '`xlab = 'Z1'`'과 '`ylab = 'Z2'`'는 각각 x 축과 y 축의 이름을 지정한다.

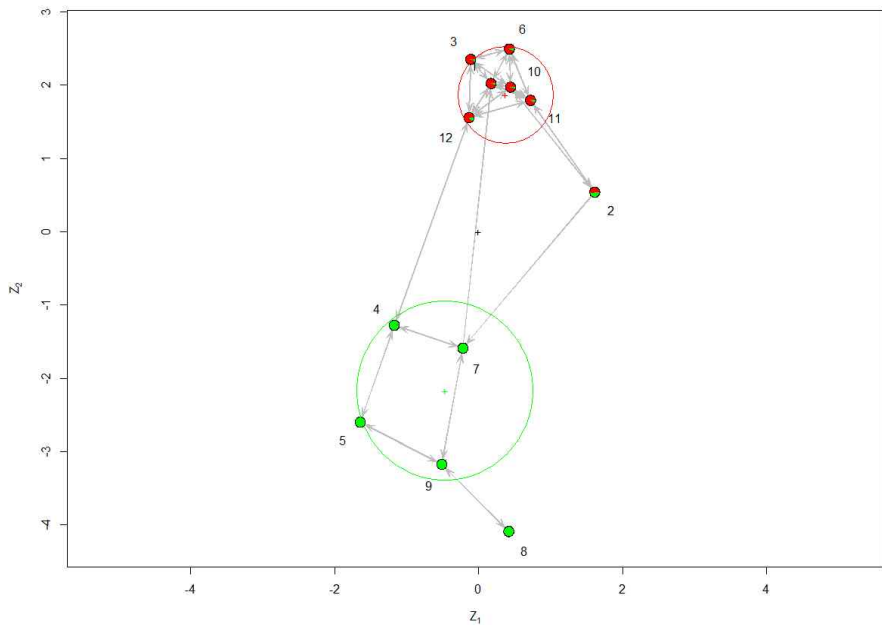


<그림 4> STM을 통해 적합한 잠재위치 산점도

<그림 4>는 <그림 3>와 비슷해 보이지만 각 노드가 현재의 군집에 있을 사후확률을 제공하지 않는다. 또한 빨간색으로 표시된 군집 2의 경우 노드들이 군집의 평균과 분산을 바탕으로 그려진 원 안으로 밀집된 정도가 약하며, 실제 노드 2의 경우는 초록색 군집으로 포함되어야 하지만 STM 방법으로 추정된 결과는 다른 군집에 포함되어 있는 것을 확인할 수 있다. HMM 방법과는 다른 변형된 근사적 모형을 사용하기 때문에 HMM 방법에서 추정한 모수를 직접 비교할 수 없다. 이러한 이유로 STM 방법은 적합 산점도 형태와 분석 시간으로만 비교했다.

4.1.3 MH-SAMC 알고리즘

SAMC알고리즘은 표본생성 분포에서 특정 구간에서만 샘플링 하도록 제한을 두기 때문에 (단계 2)에서 MH 알고리즘을 SAMC 알고리즘으로 대체했을 때 더 개선된 모수추정이 기대된다. 모수추정은 HMM 방법에서 사용했던 R의 'latentnet' 패키지를 구성하고 있는 소스파일을 수정하여 사용했다. 첫 10000번까지 반복은 기존 방법과 동일하게 진행하였으며, 이렇게 뽑힌 10000개의 샘플을 가지고 $\hat{\beta}$ 의 목표(target) 분포에 제한을 지정하였으며, 남은 90000번은 SAMC 알고리즘을 통해 $\hat{\beta}$ 을 추정하였다. 모형적합 후 잠재위치의 산점도 또한 앞서 HMM 방법에서 설명한 방법과 동일하게 적용하였다.



<그림 5> MH-SAMC를 통해 적합한 잠재위치 산점도

<그림 5>는 MH-SAMC 방법으로 적합한 잠재위치 산점도이다. <그림 3>처럼 HMM 방법으로 적합한 그림과 유사하게 나타났지만 차이점은 모든 노드 2와 노드 8을 제외한 노드들은 모두 군집의 표준편차 안에 속해있다는 점이다. HMM 방법에서는 노드 3, 5, 6, 12가 각각 해당하는 군집의 표준편차 범위 밖에 위치해 있다.

HMM 방법과 MH-SAMC 방법 두 알고리즘의 모수추정 결과와 편의(bias), 평균제곱오차(mean squared error, MSE)를 각각 <표 1>과 <표 2>에 정리했다.

<표 1>는 각 모형을 총 10번 적합하여 추정한 모수들의 평균값을, <표 2>는 편이의 절대값 평균과 MSE를 정리한 표다. 그 결과 제안한 방법이 기존의 방법들보다 평균적으로 작은 편의를 보였다. 앞서 모형 적합 후 잠재위치의 산점도 비교에서도 보았지만 MH-SAMC 방법으로 적합했을 때 대부분의 노드들이 군집 안에 위치해 있으면서 각 군집의 표준편차가 더 적게 나왔다.

<표 1> 모의실험 모수 추정값의 평균

모수	참 값	HMH	MH-SAMC
$\hat{\beta}$	1.0	1.1923	0.8105
$\hat{\mu}_{11}$	3.0	-0.5675	1.9972
$\hat{\mu}_{12}$	3.0	0.5385	-0.2269
$\hat{\mu}_{21}$	0.0	0.7026	-1.5066
$\hat{\mu}_{22}$	0.0	-0.3921	0.1719
$\hat{\sigma}_1^2$	0.1	0.3418	0.1641
$\hat{\sigma}_2^2$	0.5	0.1764	0.1084
$\hat{\lambda}_1$	0.3	0.4164	0.4167
$\hat{\lambda}_2$	0.7	0.5836	0.5833

<표 2> 모의실험 모수 추정값의 편의와 MSE

모수	편의		MSE	
	HMH	MH-SAMC	HMH	MH-SAMC
$\hat{\beta}$	0.1923	-0.1895	0.2118	0.0372
$\hat{\mu}_{11}$	-3.5675	-1.0028	13.9941	1.0075
$\hat{\mu}_{12}$	-2.4615	-3.2269	7.7193	10.4131
$\hat{\mu}_{21}$	0.7026	-1.5066	2.1556	2.2709
$\hat{\mu}_{22}$	-0.3921	0.1719	1.6777	0.0297
$\hat{\sigma}_1^2$	0.2418	0.0641	0.1238	0.0042
$\hat{\sigma}_2^2$	-0.3236	-0.3916	0.1342	0.1534
$\hat{\lambda}_1$	0.1164	0.1167	0.0581	0.0136
$\hat{\lambda}_2$	-0.1164	-0.1167	0.0581	0.0136
절대값 평균	0.9016	0.7541	2.9037	1.5492

<표 3> 모의실험 자료 분석시간

	HMH	MH-SAMC	STM
실행시간(초)	22	20	1

<표 3>은 모의실험 자료를 이용하여 세 알고리즘으로 분석한 시간을 정리한 표이다. STM 방법은 최대 반복 수 100000번으로 지정했지만 반복 과정에서 근사 변형 EM 알고리즘이 수렴하면 더 이상 반복을 진행하지 않고 결과를 저장하고 끝내는 방식으로 가장 빠른 분석시간을 보였으며, MH-SAMC 방법이 근소하게 HMH 방법보다 빠른 속도를 보였다.

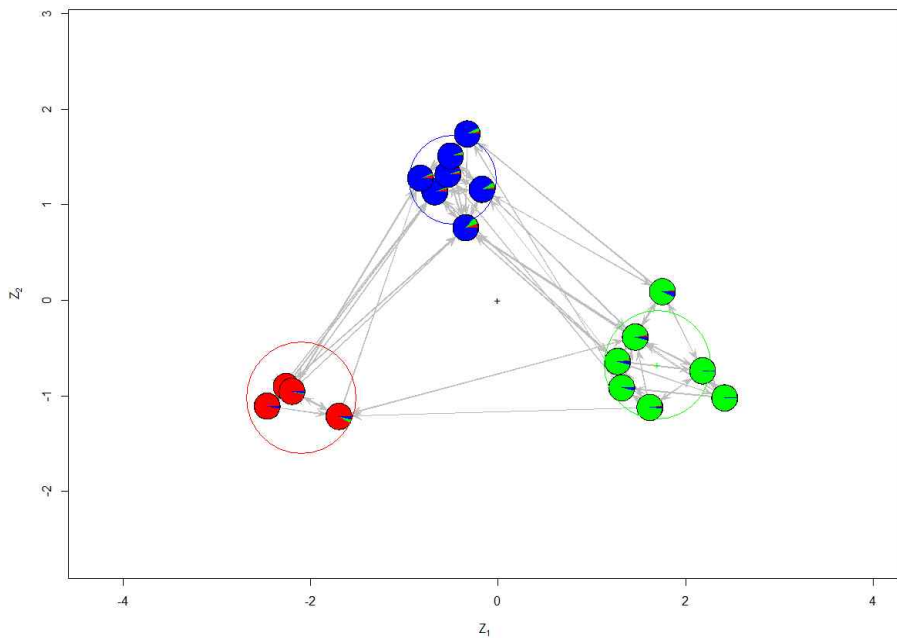
결과적으로 분석결과로부터 본 연구에서 제안한 알고리즘이 기존의 알고리즘보다 더 빠르게, 효율적인 추정치를 제공하는 방법임을 알 수 있다. 이를 토대로 노드의 수가 증가한 실자료에 적용하여 모수 추정결과를 비교해 보았다.

4.2 Sampson's Monks Data

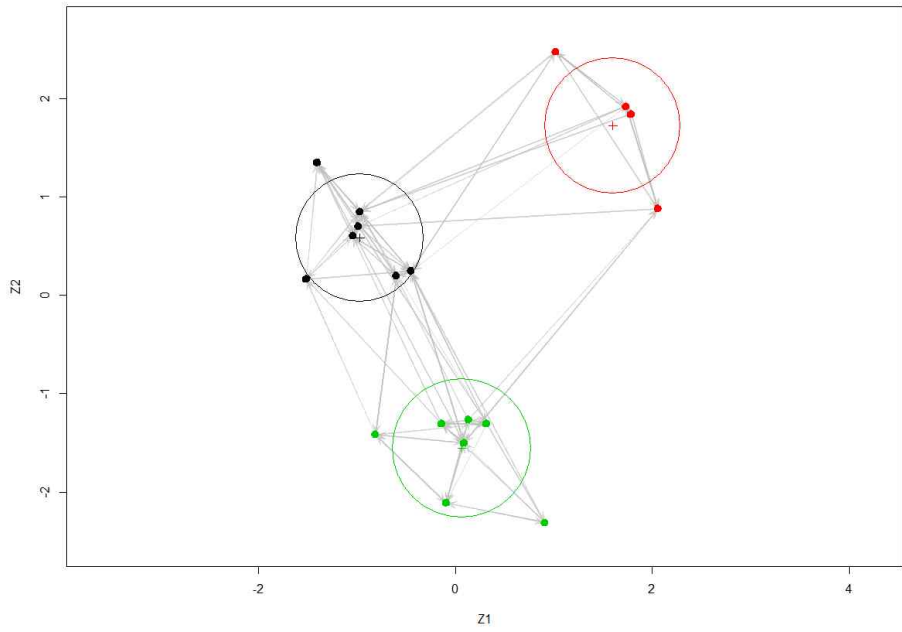
소셜네트워크 분석에서 가장 대표적으로 사용되는 예제는 Sampson(1968)의 수도승 자료이다(Krivitsky, Handcock 2008). Sampson은 고립된 미국 수도원에 서 수도승들과 머무르면서 같이 생활하고 인터뷰 및 관찰을 통해 18명의 수도승 사이의 사회적 관계 자료를 얻었다. Sampson은 사회적 관계 중 다른 수도승에게 가지는 '선호도'에 집중하였다. 18명의 수도승에게 3번에 걸쳐 가장 선호하는 수도승 3명을 조사하였다. 각각의 자료는 'samplk1', 'samplk2', 'samplk3'에 저장되었으며 모두 이항자료이다. 또한 Sampson은 최종 수집한 자료를 바탕으로 3분류로 노드를 군집화하였다. 젊은 터키인(the Young Turks), 충성스러운 야당(the Loyal Opposition) 그리고 추방자들(the Outcasts)로 나누었다.

분석에 사용한 자료는 ‘samplike’로 3번에 걸쳐 조사한 자료들의 합이다. 3개의 자료를 합한 뒤 이항 값을 가지는 소셜네트워크 자료로 전환했다. 그 결과 0과 1로 이루어진 18×18 행렬이 생성되었다. 자기 자신과의 관계는 제한하여 대각원소가 0을 가진다.

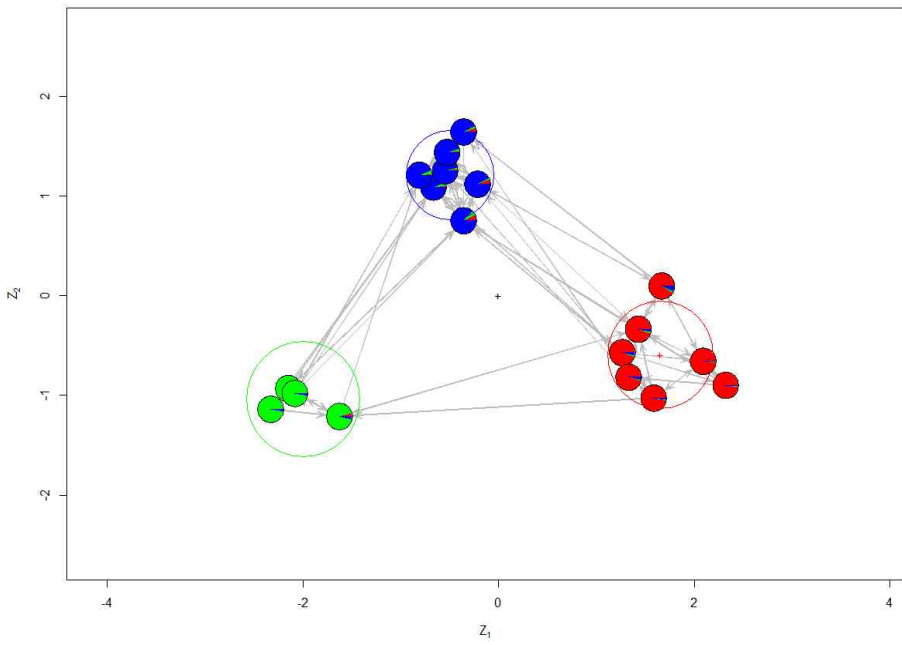
Sampson이 제공한 군집에 관한 정보를 고려하여 2차원 공간($d=2$)과 군집의 수는 3개($G=3$)로 설정하였다.



(a) HMM를 통한 잠재위치 산점도



(b) STM을 통한 잠재위치 산점도



(c) MH-SAMC을 통한 잠재위치 산점도

<그림 6>

각 군집은 세 가지 방법 모두 우수하게 진행되었다. <그림 6>을 보면 HMMH와 MH-SAMC 방법이 STM 방법보다 군집을 나타내는 원 안으로 노드들이 더 밀집되어 있다. 이는 근사 변형방법은 정확한 추론에 있어서 적합하지 않는 분석방법이라는 것을 의미한다.

<표 4>은 실자료의 모수 추정결과를 비교한 것이다. 모수 추정결과, 두 방법의 추정값이 크게 차이가 나지 않지만 각 군집의 분산인 $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, $\hat{\sigma}_3^2$ 의 추정치가 HMMH 결과보다 MH-SAMC 결과가 더 낮다는 것을 보아 MH-SAMC 방법이 더 나은 결과를 보인다는 것을 확인할 수 있다.

<표 5>는 실자료의 모수 참값을 알 수 없기 때문에 편의와 MSE를 구할 수 없다. 따라서 각 알고리즘을 10번 시행한 모수 추정치의 표준편차를 계산하였다. 제안된 방법이 추정한 모든 모수의 표준편차가 0에 가깝게 나왔으며 이는 MH-SAMC 방법이 더 우수한 모수추정이 이루어진 것을 확인할 수 있다.

<표 6>는 수도승 실자료를 세 알고리즘을 통해 분석한 시간을 비교한 표이다. STM 방법의 경우 수도승 자료가 빠른 수렴을 보였고, 가장 빠른 분석속도가 나왔다. 예상대로 MH-SAMC 방법이 HMMH 방법보다 조금 더 빠르게 분석되었다는 것을 확인할 수 있다.

<표 4> 수도권승 자료 모수 추정값의 평균

모수	HMH	MH-SAMC
$\hat{\beta}$	1.3245	0.8969
$\hat{\mu}_{11}$	-0.2102	1.6138
$\hat{\mu}_{12}$	0.6093	-0.6170
$\hat{\mu}_{21}$	-0.2465	-0.3855
$\hat{\mu}_{22}$	0.6425	1.1972
$\hat{\mu}_{31}$	-0.8228	-1.9776
$\hat{\mu}_{32}$	-0.5800	-0.9327
$\hat{\sigma}_1^2$	0.2669	0.1239
$\hat{\sigma}_2^2$	0.2259	0.1092
$\hat{\sigma}_3^2$	0.2275	0.1528
$\hat{\lambda}_1$	0.4402	0.3889
$\hat{\lambda}_2$	0.3434	0.3889
$\hat{\lambda}_3$	0.2164	0.2222

<표 5> 수도권 자료 모수 추정값의 표준편차

모수	HMH	MH-SAMC
$\hat{\beta}$	0.1418	0.0004
$\hat{\mu}_{11}$	1.3525	0.0009
$\hat{\mu}_{12}$	2.2159	0.0005
$\hat{\mu}_{21}$	1.3729	0.0002
$\hat{\mu}_{22}$	2.1904	0.0005
$\hat{\mu}_{31}$	1.2869	0.0005
$\hat{\mu}_{32}$	0.9537	0.0005
$\hat{\sigma}_1^2$	0.0708	0.0001
$\hat{\sigma}_2^2$	0.0657	0.0000
$\hat{\sigma}_3^2$	0.0713	0.0000
$\hat{\lambda}_1$	0.0357	0.0000
$\hat{\lambda}_2$	0.0310	0.0000
$\hat{\lambda}_3$	0.0461	0.0000
평균	0.7565	0.0003

<표 6> 수도권 자료 분석시간

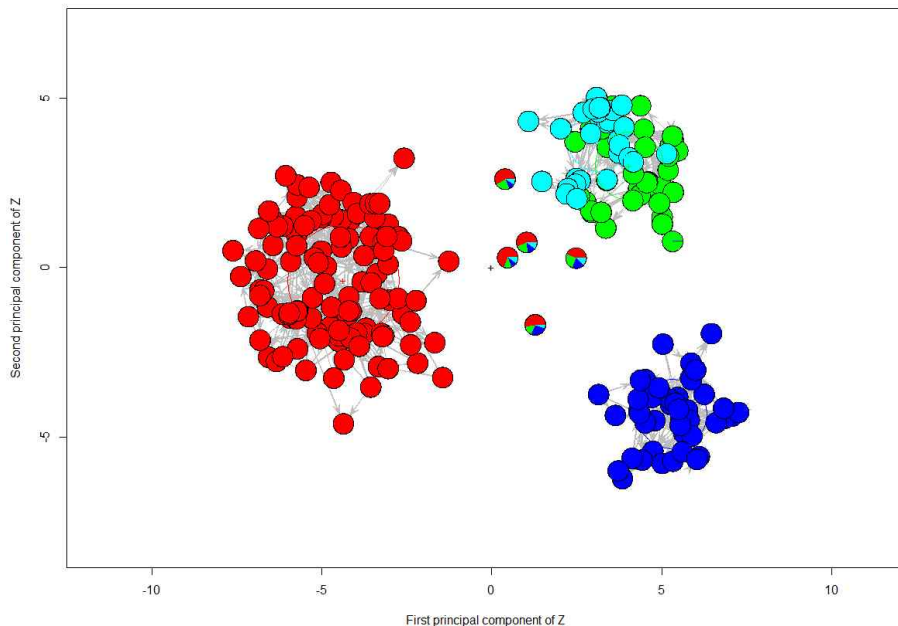
	HMH	MH-SAMC	STM
실행시간(초)	43	36	2

4.3 Physician Data

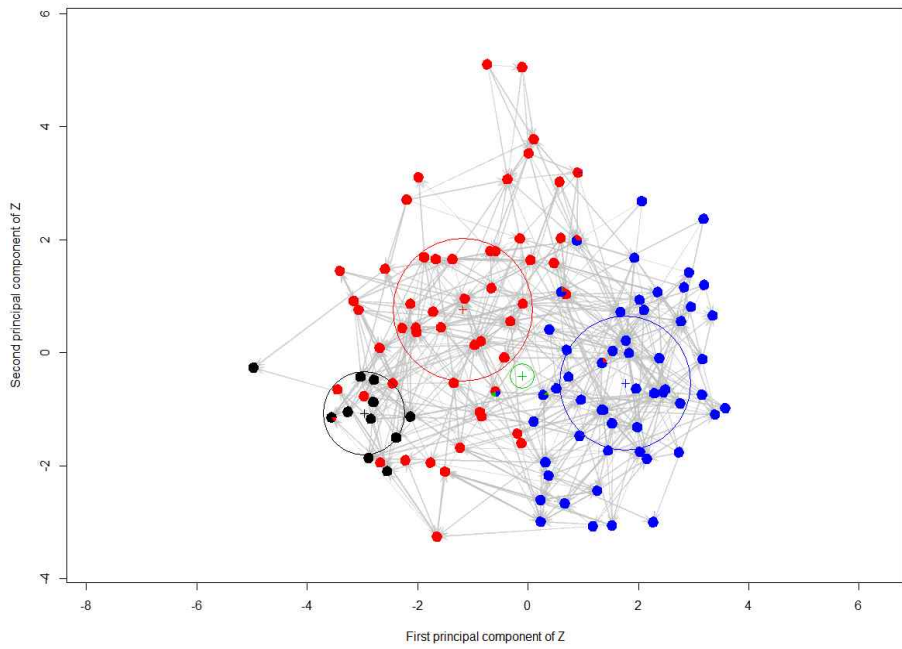
다음 예제는 Ron Burt(1966)의 내과 의사 자료이다(Burt, 1987). Burt는 Coleman, Katz, Menzel이 의료 혁신에 관해 수집한 자료를 이용했다. 일리노이

(Illinois), 피오리아(Peoria), 블루밍턴(Bloomington), 퀸시(Quincy), 게일즈버그(Galesburg) 4개 도시에 있는 의사로부터 수집한 데이터이다(Coleman et al., 1957). 새로운 약물인 ‘tetracycline’에 대한 네트워크 관계의 영향에 관심을 가졌다. 246명의 내과 의사에게 3개의 질문을 하고 그 질문에 해당되는 내과 의사 3명을 조사하였다. 첫째, 치료 요법에 관련된 정보나 조언이 필요할 때 누구에게 질문하는지 둘째, 일주일 동안 치료법을 논의하는 의사는 누구인지 셋째, 가장 자주 사교적으로 만나는 의사가 누구인지 조사하였다. 모두 이항자료이며, 분석에 사용한 자료는 3번에 걸쳐 조사한 자료의 합이다. 자기 자신을 답할 수 없도록 제한하여 대각원소는 0이다.

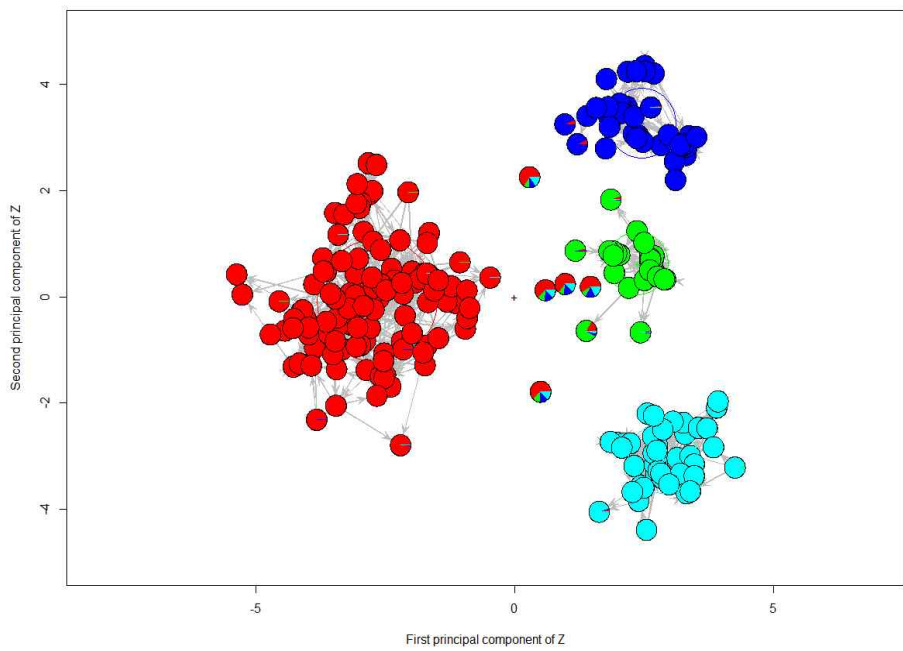
분석에는 3차원 공간($d=3$)으로 지정하였고, Burt가 제공한 군집에 관한 정보를 고려하여 군집의 수는 4개($G=4$)로 설정하였다. 모형적합 이후 산점도는 잠재위치 공간 3차원 적합값을 주성분 점수를 계산하여 제 1주성분을 x축, 제 2주성분을 y축으로 하여 2차원으로 나타내어 비교했다.



(a) HMM를 통한 잠재위치 산점도



(b) STM을 통한 잠재위치 산점도



(c) MH-SAMC를 통한 잠재위치 산점도

<그림 7>

<그림 7>을 보면 HMM과 MH-SAMC 방법에 비해 STM 방법은 지정 군집 수 4개로 했지만 초록색 군집에는 어떠한 노드도 속하지 않았으며 군집이 정확하게 나누어지지 않았다. 적합 잠재위치 산점도 우측 상단 군집에서 HMM 방법은 초록색 군집과 하늘색 군집이 섞여있는 반면에 MH-SAMC 방법은 정확하게 4개의 군집이 나누어져 있음을 알 수 있다. 이는 MH-SAMC 방법이 더 우수하게 모수추정을 하여 군집화가 잘 되었다는 것을 의미한다.

<표 7>은 내과 의사 실자료의 모수 추정결과를 비교한 것이다. 모수 추정결과에서 $\hat{\beta}$ 과 각 군집의 평균은 참값을 알 수 없기 때문에 어떤 방법이 더 우수하다고 판단할 수는 없지만 각 군집의 분산 추정치 $\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\sigma}_3^2, \hat{\sigma}_4^2$ 이 HMM 방법보다 MH-SAMC 방법이 더 낮게 나왔다는 점에서 MH-SAMC 방법이 더 나은 결과를 보여주고 있다. 각 군집에 속할 확률 추정값은 두 방법이 비슷한 값을 보이고 있다.

<표 8>은 내과 의사 실자료를 사용하여 두 알고리즘을 10번 시행한 모수 추정치의 표준편차를 구하였다. HMM 방법보다 MH-SAMC 방법으로 구한 추정치들의 표준편차가 대체적으로 더 낮았으며 추정값들의 표준편차 평균을 보아도 MH-SAMC 방법이 더 낮다는 것을 확인할 수 있다. 이는 MH-SAMC 방법이 더 안정적인 추정을 한다고 판단할 수 있다.

<표 7> 내과 의사 자료 모수 추정값의 평균

모수	HMH	MH-SAMC
$\hat{\beta}$	0.5900	-0.2968
$\hat{\mu}_{11}$	-3.8586	-0.6066
$\hat{\mu}_{12}$	0.2389	-0.6791
$\hat{\mu}_{13}$	0.9467	-0.8754
$\hat{\mu}_{21}$	1.0567	0.1157
$\hat{\mu}_{22}$	-1.3635	0.7926
$\hat{\mu}_{23}$	-0.9310	-0.0614
$\hat{\mu}_{31}$	3.5632	0.9030
$\hat{\mu}_{32}$	0.8213	0.4990
$\hat{\mu}_{33}$	-0.6978	0.1007
$\hat{\mu}_{41}$	1.3169	0.4310
$\hat{\mu}_{42}$	-0.5773	1.4170
$\hat{\mu}_{43}$	2.5896	0.7125
$\hat{\sigma}_1^2$	5.0440	2.7465
$\hat{\sigma}_2^2$	2.7682	1.2474
$\hat{\sigma}_3^2$	2.4409	1.1626
$\hat{\sigma}_4^2$	2.1960	1.0973
$\hat{\lambda}_1$	0.4104	0.4187
$\hat{\lambda}_2$	0.2384	0.2274
$\hat{\lambda}_3$	0.1868	0.1888
$\hat{\lambda}_4$	0.1644	0.1650

<표 8> 내과 의사 자료 모수 추정값의 표준편차

모수	HMH	MH-SAMC
$\hat{\beta}$	0.0951	0.0926
$\hat{\mu}_{11}$	5.8312	2.5390
$\hat{\mu}_{12}$	2.7816	2.0701
$\hat{\mu}_{13}$	3.2717	2.2989
$\hat{\mu}_{21}$	3.6411	3.4226
$\hat{\mu}_{22}$	7.0942	2.1847
$\hat{\mu}_{23}$	4.8942	3.3014
$\hat{\mu}_{31}$	3.8218	3.0256
$\hat{\mu}_{32}$	5.3575	3.4509
$\hat{\mu}_{33}$	4.5042	3.1704
$\hat{\mu}_{41}$	3.6060	3.2974
$\hat{\mu}_{42}$	5.9634	2.6581
$\hat{\mu}_{43}$	6.0703	3.0294
$\hat{\sigma}_1^2$	0.5077	0.4606
$\hat{\sigma}_2^2$	0.8545	0.3780
$\hat{\sigma}_3^2$	0.4941	0.3359
$\hat{\sigma}_4^2$	0.6202	0.2758
$\hat{\lambda}_1$	0.0520	0.0594
$\hat{\lambda}_2$	0.0413	0.0261
$\hat{\lambda}_3$	0.0128	0.0240
$\hat{\lambda}_4$	0.0239	0.0200
평균	2.8352	1.7072

<표 9> 내과 의사 자료 분석 시간

	HMH	MH-SAMC	STM
실행 시간(초)	5092	4972	23694

<표 9>는 본 논문에서 이용한 알고리즘을 이용하여 내과 의사 실자료를 분석하는데 걸린 시간을 정리하였다. MH-SAMC 방법이 HMH 방법보다 조금 더 빠르게 분석을 완료하였고, STM 방법은 수도권 자료에서 수렴이 빠르게 이루어져 오랜 시간이 걸리지 않았지만 대용량 데이터를 분석할 때 수렴이 제대로 이루어지지 않을 경우 더 오래 걸리며, 정확한 추론이 힘들다는 것을 알 수 있다.

제5장 결 론

소셜네트워크 분석에는 이전에 나온 여러 가지 통계적 모형이 사용되고 있다. 소셜네트워크 분석에 많이 사용되고 있는 고전적 모형인 지수족 랜덤 그래프 모형과 잠재변수를 고려한 모형은 모형이 복잡할 뿐만 아니라 모수추정에 계산적인 어려움이 많다. 이러한 이유로 모수 추정보다는 군집화에 초점을 두는 경우가 많다. 하지만 군집화에 초점을 두게 되면 모형이 더 복잡해지며, 빅데이터 소셜네트워크 분석에서는 추정을 위한 시간이 오래 걸릴뿐더러 정확한 모수추정이 더 힘들어진다. 본 논문에서는 제한된 샘플링 공간에서 모수추정을 정확하게 한다면 대용량 소셜네트워크 데이터를 더 빠르게 분석하면서 군집화 역시 효율적으로 진행될 것으로 기대하고 모수추정에 초점을 두었다.

본 논문에선 기존의 알고리즘(HMH)이 목표 분포의 전체구간인 $(-\infty, \infty)$ 구간에서 샘플링하는 것과 다르게 전체 공간에서 샘플링 구간의 자기조절(self-adjusting)능력을 가지고 있는 SAMC 알고리즘을 사용한 방법을 제시하였다. 샘플링 도중에 발생하는 국소 트랩을 방지하여 더 효율적인 모수추정을 하도록 했다. 그 결과 제안된 방법은 모의실험과 실자료를 통해 다른 방법보다 더 개선된 추정치를 바탕으로 우수한 군집화 성능을 보였으며 더 빠른 분석속도를 제공하였다. 이는 노드 수가 더 많은 모형에서 더 큰 효율을 보일 것으로 예상된다.

다만 본 연구에서는 추정하고자 하는 모수 중에서 β 에 대해서만 SAMC 알고리즘을 적용하였다. z_i 는 노드들의 알려지지 않은 잠재위치를 가정하였기 때문에 SAMC 알고리즘 적용효율이 높지 않았다. 또한 이전 연구(김지용, 2017)에서 향후 과제로 언급했던 계산상의 이점을 살려 C 프로그램을 통한 연구를 하였지만 $\hat{\beta}$, \hat{z}_i 를 각각 추정 과정에서 모든 노드들의 잠재공간을 갱신하는 부분에서 많은 시간이 걸리는 문제가 있다. 더 빠르고 정확하게 추정할 수 있는 모형을 개발한다면 빅데이터 시대에 맞춰 효율적인 추정이 될 것으로 기대한다.

참 고 문 헌

- [1] 김지용 (2017). 소셜네트워크자료를 위한 잠재적 위치 군집 분석 개선 연구. 고려대학교 일반 대학원 석사학위논문.
- [2] Airoldi, E., Blei, D., Fienberg, S., Xing, E. (2008). Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, 9, 1981–2014.
- [3] Banfield, J. D., Raftery, A. E. (1993). Model-based Gaussain and non-Gaussian clustering. *Biometrics*, 49(3), 803–821.
- [4] Bouveyron, C., Latouche, P., Zreik, R. (2016). The Stochastic Topic Block Model for the Clustering of Networks with Textual Edges, *Statistics and Computing* (in press).
- [5] Burt, Ronald S. (1987). Social Contagion and Innovation: Cohesion Versus Structural Equivalence, *American Journal of Sociology*, 92, 1287–1335.
- [6] Celisse, A., Daudin, J. J., Pierre, L. (2011). Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model. *Electronic Journal of Statistics*, 6, 1847–1899.
- [7] Coleman, J. S., E. Katz, H. Menzel. (1957). The Diffusion of an Innovation Among Physicians, *Sociometry*, 20, 253–270.
- [8] Daudin, J. J., Picard, F., Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18(2), 173–183.
- [9] Fraley, C., Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 611–631.
- [10] Frank, O., Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395), 832–842.
- [11] Gormley, I. C., Murphy, T. B. (2010). A mixture of experts latent position cluster model for social network data, *Statistical Methodology*, 7(3), 385–405.

- [12] Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks, *Journal of the Royal Statistical Society Series A*, 170, 301–354.
- [13] Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika*, 57(1), 97–109.
- [14] Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent Space Approaches to Social Network Analysis, *Journal of the American Statistical Association*, 97(460), 1090–1098.
- [15] Hoff, P. (2005). Bilinear Mixed-Effects Models for Dyadic Data. *Journal of the American Statistical Association*, 100(469), 286–295.
- [16] Holland, P. W., Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs(with discussion). *Journal of the American Statistical Association*, 76(373), 33–50.
- [17] Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012). Computational Statistical Methods for Social Network Models, *Journal of Computational and Graphical Statistics*, 21(4), 856–882.
- [18] Jernite, Y., Latouche, P., Bouveyron, C., Rivera, P., Jegou, L., Lamassé, S. (2014). The random subgraph model for the analysis of an acclesiastical network in merovingian gaul. *Annals of Applied Statistics*, 8(1), 55–74.
- [19] Koskinen, J. (2009). The Linked Importance Sampler Auxiliary Variable Metropolis Hastings Algorithm for Distributions with Intractable Normalising Constants. *Working paper*.
- [20] Krivitsky, P. N., Handcock, M. S. (2008). Fitting Position Latent Cluster Models for Social Networks with latentnet, *Journal of Statistical Software*, 24(5).
- [21] Liang, F., Liu, C., Carroll, R. (2007). Stochastic approximation in Monte Carlo computation, *Journal of the American Statistical Association*, 102(447), 305–320.
- [22] Lusher, D., Koskinen, J., Robins, G. (2013). *Exponential Random Graph Models for*

Social Networks, Cambridge : Cambridge University Press.

- [23] Mariadassou, M., Robin, S., Vacher, C. (2010). Uncovering latent structure in valued graphs: a variational approach. *Annals of Applied Statistics*, 4(2), 715–742.
- [24] Nowicki, K., Snijders, T. A. B. (2001). Estimation and Prediction for Stochastic Blockstructures. *Journal of the American Statistical Association*, 96(455), 1077–1087.
- [25] Ryan, C., Wyse, J., Friel, N. (2017). Bayesian model selection for the latent position cluster model for social networks, *Network Science*, 5(1), 70–91.
- [26] Salter-Townshend, M., Murphy, T. B. (2013). Variational Bayesian inference for the Latent Position Cluster Model for network data, *Computational Statistics and Data Analysis*, 57(1), 661–671.
- [27] Sampson, S. (1968). A novitiate in a period of change: An experimental and case study of social relationships, PhD thesis, Cornell University, September.
- [28] Schweinberger, M., Snijders, T. A. B. (2003). Settings in Social Networks: A Measurement Model. *Sociological Methodology*, 33(1), 307–341.
- [29] Snijders, T. A. B., Nowicki, K. (1997). Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification*, 14(1), 75–100.
- [30] Tallberg, C. (2005). A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology*, 29(1), 1–23.
- [31] van Duijn, M. A. J. (1995). Estimation of a Random Effects Model for Directed Graphs. in *Toeval zit overal: programmatuur voor random-coëfficiënt modellen*, eds. Snijders, T. A. B., Engel, B., Van Houwelingen, J. C., Keen, A., Stemerding, G. J., Verbeek, M., Groningen: IEC ProGAMMA, 113–131.
- [32] van Duijn, M. A. J., Snijders, T. A. B., Zijlstra, B. H. (2004). p_2 : a random effects model with covariates for directed graphs. *Statistica Neerlandica*, 58(2), 234–254.

- [33] Wasserman, S., Faust, K. (1994). *Social Network Analysis: Methods and Applications(Structural Analysis in the Social Sciences) First Edition*, Cambridge : Cambridge University Press.
- [34] Wyatt, D., Choudhury, T., Bilmes, J. (2008). Learning Hidden Curved Exponential Random Graph Models to Infer Face-to-Face Interaction Networks from Situated Speech Data. *Proceedings of AAAI*, 2, 732–738.
- [35] Zijlstra, B. J. H., van Duijn, M. A. J., Snijders, T. A. B. (2009). MCMC estimation for the p_2 network regression model with crossed random effects. *British Journal of Mathematical and Statistical Psychology*, 62(1), 143–166.