Parameter estimation in Bayesian Exponential-Family Models of Random Graphs

Min-Yi Chen
Class : Social Physics
National Chenchi University
May 22, 2023

1 Introduction

Statisticians often utilize sufficient statistics to analyze the structure of a network. These statistics include measures such as the degree of the network, k-star, and more. In this discussion, I will introduce the framework of the Exponential Random Graph Model (ERGM) and its Bayesian approach. Additionally, I will explain the simulation algorithm in detail, which includes the exchange algorithm, Gibbs sampling, and the underlying theorem supporting these methods.

1.1 ERGM notation

As a beginner, we start from the simple case with the following notations[1].

- We consider a finite population of nodes $\mathcal{N} = \{1, \ldots, N\}$ for $N \geq 2$.
- Denote the collection of attributes of population members by $x_{\mathcal{N}} \in \mathcal{X}_{\mathcal{N}} \subseteq \mathbb{R}^q$.
- The network is simple which it is undirected and exclude self-loop(self-edges).
- Edges between nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$ are random variable denoted by $\mathcal{Y}_{i,j}$.
- Denote weight $\mathcal{Y} = \{0, 1\}$ where o indicated the absence of an edge and 1 indicated the presence of an edge.
- Sample space $\Omega = \mathcal{Y}_{\mathcal{N}} = \{0, 1\}^{\frac{N}{2}}$.

We consider exponential families of densities with respect to a σ-finite reference measure v with support $\mathcal{Y}_{\mathcal{N}}$ by sufficient statistics $s : \mathcal{X}_{\mathcal{N}} \times \mathcal{Y}_{\mathcal{N}} \mapsto \mathbb{R}^q$ and a map $\eta : \Theta \times \mathcal{N} \mapsto \mathbb{R}^q$ with $\Theta \subseteq \{\theta \in \mathbb{R}^q; \psi(\theta, \mathcal{N}) < \infty\}$:

$$\frac{d\mathbb{P}_{\mathcal{N},\eta(\theta,\mathcal{N})}}{dv} = \exp(\langle \eta(\theta, \mathcal{N}), s(x_{\mathcal{N}}, y_{\mathcal{N}}) \rangle - \psi(\theta, \mathcal{N})) \tag{1}$$

where $\langle \eta(\theta, \mathcal{N}), s(x_{\mathcal{N}}, y_{\mathcal{N}}) \rangle$ denotes the inner product of natural parameter $\eta(\theta, \mathcal{N})$ and sufficient statistics $s(x_{\mathcal{N}}, y_{\mathcal{N}})$ and

$$\psi(\theta, \mathcal{N}) = \log \int_{\mathcal{Y}_{\mathcal{N}}} \exp(\langle \eta(\theta, \mathcal{N}), s(x_{\mathcal{N}}, y'_{\mathcal{N}}) \rangle) dv(y'_{\mathcal{N}}) \tag{2}$$

$\psi(\cdot)$ is normalization. We denote $\frac{d\mathbb{P}}{dv}$, since $\mathbb{P}$ mapps from subset of $\mathcal{F}$ to $[0, 1]$, not subset of Ω to $[0, 1]$. We have the probability mass function :

$$\mathbb{P}_{\mathcal{N}, \eta(\theta, \mathcal{N})}(Y_{\mathcal{N}} = y_{\mathcal{N}}) = \exp(\langle \eta(\theta, \mathcal{N}), s(x_{\mathcal{N}}, y_{\mathcal{N}}) \rangle - \psi(\theta, \mathcal{N})) v(y_{\mathcal{N}}) \tag{3}$$

where

$$\psi(\theta, \mathcal{N}) = \log \sum_{y'_{\mathcal{N}} \in \mathcal{Y}_{\mathcal{N}}} \exp(\langle \eta(\theta, \mathcal{N}), s(x_{\mathcal{N}}, y'_{\mathcal{N}}) \rangle) v(y'_{\mathcal{N}}) \tag{4}$$

which is a Lebesgue integral.

1.2 The advantages of ERGMs

Those are the advantages of ERGMs which are mentioned below.

1. Language
   Exponential families provide a convenient language for formulating ideas about network data and dependencies theorem.

2. Unifying statistical framwork
   Exponential family framwork is a unifying statistical framework that includes a wide range of random graph models.

3. Computational advantages
   Exponential families have useful convexity properties

   - The natural parameter space of exponential families is a convex set.
   - The negative loglikelihood function is a strictly convex function on the interior of the natural parameter space.
     These convexity properties imply that maximum likelihood estimation of natural parameters is a convex minimization program with a unique solution.
   - Exponential families admit data reduction by sufficiency.
   - Exponential-family likelihood functions depend on the data only through minimal sufficient statistics.
     Likelihood-based estimation algorithms are agnostic to structure of the sample space of network data, which can be a large and discrete set, as long as observed and expected minimal sufficient statistics can be computed exactly or approximately.

4. Theoretical advantages
   Network data are complex, so we wanted to keep statistical theory as simple as possible. The exponential family framework enables theoreticians to do so.

5. Practical advantages
   ERGMs are widely use in practice.

2 Bayesian parameter estimation

2.1 Bayesian ERGM and Computational Intractable

Generally we can rewrite the equation 3 in the Bayesian approach as:

$$\pi(y|x,\theta) = \frac{\exp\{\eta(\theta)^T s(y,x)\}}{Z(\theta)} \tag{5}$$

where $Z(\theta) = \sum_{y' \in \mathcal{Y}} \exp\{\eta(\theta)^T s(y',x)\}$ is the normalising constant ensuring the probability mass function sums to one.

Given an observation y of the network, inference is performed by analysing the posterior distribution $\pi(\theta|y)$:

$$\begin{aligned} \pi(\theta|y) &= \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)} \\ &= \frac{\exp\{\eta(\theta)^T s(y,x)\}\pi(\theta)}{Z(\theta)\pi(y)} \end{aligned} \tag{6}$$

where $\pi(y) = \int_\Theta \pi(y|\theta)\pi(\theta)d\theta$ is the model evidence. However, the posterior distribution is generally not available in a closed-form expression. This is because the posterior distributions has two sources of intractability which are refered to as $doubly - intractable$.

1. $\pi(y)$ is intractable since y is random.
2. The likelihood is intractable via the normalising constant $Z(\theta)$ is intractable. Since $\mathcal{Y}$ is an infinite-dimensional distribution, and the normalizing constant involves an integral over all possible partitions of the data.

2.2 Exchange Algorithm

Excange algorithm[2][3] is an MCMC scheme designed to circumvent computational intractable. Exchange algorithm introduce an auxiliary variable $y' \sim \pi(\cdot|\theta')$, and where a simulated network is drawn from the ERGM with parameter θ'.

The algorithm targets an augmented posterior

$$\pi(\theta, \theta', y, y') \propto \pi(\theta|y)h(\theta'|\theta)\pi(y'|\theta') \tag{7}$$

where

- $\pi(\theta|y)$ is the original (target) posterior.
- $h(\theta'|\theta)$ is an arbitrary, normalisable proposal function.
- $\pi(y'|\theta')$ is the likelihood of the auxiliary variable.

And we assume $h(\theta'|\theta)$ is symmetric , so that we can use detailed balance in MCMC to estimate the parameters.

The algorithm proceeds as follows :

At each iteration :

1. Perform a Gibbs' update of (θ', y') by :
   drawing $\theta' \sim h(\cdot|\theta)$
   drawing $y' \sim \pi(\cdot|\theta')$
2. Exchange θ and θ' with probability $\min(1, AR(\theta', \theta, y', y))$

We denote $AR(\theta', \theta, y', y)$:

$$\begin{aligned} AR(\theta', \theta, y', y) &= \frac{\pi(\theta'|y)}{\pi(\theta|y)} \cdot \frac{\pi(y'|\theta)}{\pi(y'|\theta')} \\ &= \frac{\exp\{\theta'^T s(y)\}\pi(\theta')Z(\theta)}{\exp\{\theta^T s(y)\}\pi(\theta)Z(\theta')} \cdot \frac{\exp\{\theta^T s(y')\}Z(\theta')}{\exp\{\theta'^T s(y)\}Z(\theta)} \\ &= \exp\{[\theta' - \theta]^T [s(y) - s(y')]\}\frac{\pi(\theta')}{\pi(\theta)} \end{aligned} \tag{8}$$

and $AR(\theta', \theta, y', y)$ cancel out the intractable normalizing constant. Compare with the origin Metropolis acceptance ratio

$$A(x', x) = \min(1, \frac{P(x')g(x|x')}{P(x)g(x'|x)}) \tag{9}$$

we can conclude that the acceptance ratio $AR(\theta', \theta, y', y)$ in the exchange algorithm is to compare:

$$\begin{aligned} \pi_1(\theta, \theta', y, y') &\propto \pi(\theta|y)h(\theta'|\theta)\pi(y'|\theta') \\ \pi_2(\theta, \theta', y, y') &\propto \pi(\theta'|y)h(\theta|\theta')\pi(y'|\theta) \end{aligned} \tag{10}$$

, and since $h(\theta'|\theta)$ is symmetric

$$h(\theta'|\theta) = h(\theta|\theta') \tag{11}$$

$\pi_1(\cdot)$, $\pi_2(\cdot)$ can be shown as:

$$\begin{aligned} \pi_1(\theta, \theta', y, y') &\propto \pi(\theta|y)\pi(y'|\theta') \\ \pi_2(\theta, \theta', y, y') &\propto \pi(\theta'|y)\pi(y'|\theta) \end{aligned} \tag{12}$$

So that

$$AR(\theta', \theta, y', y) = \frac{\pi_2(\theta, \theta', y, y')}{\pi_1(\theta, \theta', y, y')} \tag{13}$$

Algorithm 1: Exchange Algorithm for a Bayesian ERGM

Input: number of MCMC iteration T, initial value θ_0

for $t = 1, \ldots, T$ **do**
  draw $\theta' \sim h(\cdot|\theta)$;
  draw $y' \sim \pi(\cdot|\theta')$;
  set $\theta_t = \theta'$ with probability $\min(1, AR(\theta', \theta, y', y))$;
  else, set $\theta_t = \theta_{t-1}$
end

3 Gibbs Sampling

In this section, I will provide a description of Gibbs sampling, a Markov Chain Monte Carlo (MCMC) algorithm, including its formula and the underlying theory.

3.1 Invariant, Equilibrium and Stationary

Before introducing Gibbs sampling, we need to have knowledge of Markov chains. When discussing Markov chains, the terms **invariant**, **equilibrium** and **stationary** frequently arise. Some people may mistakenly assume these terms have the same meaning; however, they are slightly different.

Theorem 3.1. *Invariant [4]*
We say λ is invariant if

$$\lambda P = \lambda$$

Theorem 3.2. *Stationary [4]*
Let $(X_n)_{n\geq 0}$ be $Markov(\lambda, P)$ and suppose that λ is invariant for P. Then $(X_{m+n})_{n\geq 0}$ is also $Markov(\lambda, P)$

Invariant and Stationary are similar. It just a terms used by different people. Koller discusses the distinction between invariant and stationary in her book "Probabilistic Graphical Models: Principles and Techniques." [5]

Definition 3.3. *A distribution $\pi(X)$ is a stationary distribution for Markov Chain $\mathcal{T}$ if it satisfies:*

$$\pi(X = j) = \sum_{i\in I} \pi(X = i)\mathcal{T}(i \to j), \text{ where } \mathcal{T} \text{ is the transition probability[?]}$$

A stationary distribution is also called an invariant distribution

However, in most cases, when we use the term "stationary," we imply that after reaching state n, the system remains unchanged even after transitioning to state $n+m$. (Theorem3.2)

Theorem 3.4. *Equilibrium [4]*
Let I be finite. Suppose for some $i \in I$ that

$$\mathcal{T}^{(n)}(i \to j) \to \pi_j \text{ as } n \to \infty \ \forall j \in I$$

Then $\pi = (\pi_j : j \in I)$ is an invariant distribution.

Equilibrium is used to describe the state where X_t is approximately equal to X_{t+1}, indicating convergence. The equation below demonstrates this equilibrium:

$$\mathcal{T}(i \to j) = \lim_{n\to\infty} \mathcal{T}^{(n)}(i \to j) = \pi_j$$

3.2 Gibbs sampler in Bayesian inference

Recall Naive Bayes again, we have

$$P(\theta|\mathcal{X}) = \frac{P(\mathcal{X}|\theta)P(\theta)}{P(\mathcal{X})}$$

The objective of Gibbs sampling is to iteratively approach an approximation of $P(\theta|\mathcal{X})$ as $\pi(\theta)$. Let $\theta = (\theta^{(1)}, \ldots, \theta^{(d)})$, $d \in \mathbb{N}$. Gibbs sampling update θ one dimension per time during iteration. We have

- Before update: $\theta = (\theta^{(1)}, \ldots, \theta^{(i)}, \ldots, \theta^{(d)})$
- After update: $\theta' = (\theta^{(1)}, \ldots, \theta'^{(i)}, \ldots, \theta^{(d)})$

We define the process as $\mathcal{T}_i(\theta \to \theta')$, where θ' represents the new state, and the probability of transitioning to θ' is denoted as

$$\mathcal{T}_i(\theta \to \theta') = \frac{\pi(\theta^{(1)},\ldots,\theta^{(i-1)},\theta'^{(i)},\ldots,\theta^{(d)})}{\sum_{\mathcal{Z}} \pi(\theta^{(1)},\ldots,\theta^{(i-1)},\mathcal{Z},\ldots,\theta^{(d)})}$$

A practical example is exhibited below.

Example 3.5. *Let* $\theta = (\theta^{(1)}, \theta^{(2)}, \theta^{(3)})$, $\theta^{(i)} \in \{1, 2, 3, 4, 5\}$,$\theta = (2, 3, 5)$

$$\mathcal{T}_2(\theta \to \theta') = \frac{\pi(2,\theta'^{(2)},5)}{\pi(2,1,5)+\ldots+\pi(2,5,5)}$$

Now that we understand how Gibbs sampling works, two questions arise.

1. Since the kernels $\mathcal{T}i$ and $\mathcal{T}i+1$ are different, we need to consider whether this multiple-kernel chain is stationary.
2. We want to determine whether $P(\theta|\mathcal{X})$ is a stationary distribution, given that $\mathcal{T}_i(\theta^{(i)} \to \theta'^{(i)})$ does not depend on $x_i \in X_i$.

To demonstrate the stationary of the posterior distribution with Gibbs sampling, we first discuss the transition kernel.

Let $(\theta^{(-i)}, \theta^{(i)})$ be i state and $(\theta^{(-i)}, \theta'^{(i)})$ be i' state after Gibbs sampling.

$$\mathcal{T}_i((\theta^{(-i)}, \theta^{(i)}) \to (\theta^{(-i)}, \theta'^{(i)})) = P(\theta'^{(i)}|\theta^{(-i)}) = \pi_x(\theta)$$

We can describe the transition kernel as a function

$$\mathcal{T}_i : (a, b) \to (x, d), \text{ if } a \neq c,\ P(\mathcal{T}_i(a, b) = (c, d)) = 0$$

Thus, set $x = (x_{-i}, x_i)$ and $y = (y_{-i}, y_i)$ if $x_i \neq y_i$,$P(\mathcal{T}_i(x_{-i}, x_i) = (y_{-i}, y_i)) = 0$

With the statement above, we write

$$\pi(x) = \textstyle\sum_y \mathcal{T}_i(y \to x)\pi(y)$$

where $\mathcal{T}_i(y \to x) = \mathcal{T}_i((x_{-i}, y_i) \to (x_{-i}, x_i)) = \frac{\pi((x_{-i},x_i))}{\sum_Z \pi((x_{-i},Z))}$

To write it clearly, we have

$\pi(x) = \sum_{(y_{-i}=x_{-i},y_i)} \mathcal{T}_i(y \to x)\pi(y) = \sum_{(y_{-i}=x_{-i},y_i)} \frac{\pi(x_{-i},x_i)}{\sum_Z \pi(x_{-i},Z)}\pi(x_{-i}, y_i)$

$= \frac{\pi(x_{-i},x_i)}{\sum_Z \pi(x_{-i},Z)} \sum_{(y_{-i}=x_{-i},y_i)} \pi(x_{-i}, y_i)$

Since $\sum_{(y_{-i}=x_{-i},y_i)} \pi(x_{-i}, y_i) = \sum_Z \pi(x_{-i}, Z)$,we have

$\pi(x) = \pi(x_{-i}, x_i)$,thus, we show that the posterior distribution is stationary.

3.3 Unique stationary distribution

While we have shown that Gibbs sampling is a stationary process, it is important to verify whether the posterior distribution is indeed the correct distribution. To support this claim, we need to prove that the Markov chain Monte Carlo (MCMC) method, based on Markov chains, possesses a unique stationary distribution.

I have found some proofs[5][4] regarding the uniqueness of the stationary distribution in the context of Gibbs sampling, and I organized these proofs by myself.

In "Probabilistic Graphical Models: Principles and Techniques", Koller explains how Markov Chain has a unique stationary distribution with a theorem related to regular Markov Chain.

Definition 3.6. *Regular Markov Chain[5]*
A Markov Chain is said to be regular if there exist some number k such that, for every $i, j \in I$, the probability of getting from i to j in exactly k steps is > 0.

And the theorem is described below:

Theorem 3.7. *If a finite state Markov Chain is regular, then it has a unique stationary distribution.[5]*

There are some theorems in Norries work that can be used to prove theorem 3.7. Theorem 3.7 holds because $\mathcal{T}^{(k)}(i \to j) > 0$ in regular Markov Chain. Then, by theorem3.8.

Theorem 3.8. *Let $\mathcal{T}$ be irreducible. Then the following are equivalent:*

1. *every state is positive recurrent;*
2. *some state i is positive recurrent;*
3. *$\mathcal{T}$ has an invariant distribution, π say.*

We can deduce that the regular Markov chain is irreducible. Additionally, the ergodic theorem can be employed to demonstrate that the stationary distribution is the unique stationary distribution.

Theorem 3.9. *Ergodic theorem [4]*
Let P be irreducible and let λ be any distribution. If $(X_n)_{n\geq 0}$ is Markov(λ, P) then

$$\mathbb{P}(\tfrac{V_i(n)}{n} \to \tfrac{1}{m_i} \text{ as } n \to \infty) = 1$$

where $m_i = \mathbb{E}_i(T_i)$ is the expected return time to state i and $V_i(n)$ is the number of visits to state i before n. Moreover, in the positive recurrent case, for any bounded function $f : I \to \mathbb{R}$ we have

$$\mathbb{P}(\tfrac{1}{n}\textstyle\sum_{k=0}^{n-1} f(X_k) \to \bar{f} \text{ as } n \to \infty) = 1$$

where

$$\bar{f} = \textstyle\sum_{i\in I} \pi_i f_i$$

and where $(\pi_i : i \in I)$ is the unique invariant distribution.

3.4 Detailed Balalnce

In MCMC methods, we frequently employ the concept of detailed balance to verify that the Markov chain possesses the desired stationary distribution. Detailed balance is the underlying principle that necessitates the symmetry of the function $h(\cdot)$ in equation11.

Theorem 3.10. *Reversible Markov Chain [5]*
A finite-state Markov chain $\mathcal{T}$ is reversible if there exists a unique distribution π such that, for all $i, j \in I$:

$$\pi(i)\mathcal{T}(i \to j) = \pi(j)\mathcal{T}(j \to i)$$

This equation is called the detailed balance. It is useful in application to denote the acceptance rate by the relation

$$\frac{\pi(i)}{\pi(j)} = \frac{\mathcal{T}(j \to i)}{\mathcal{T}(i \to j)} \tag{14}$$

e.g. the acceptance rate $AR(\theta', \theta, y', y)$ in equation 13 above.

According to Theorem 3.10, reversibility implies the existence of a unique stationary distribution π. However, it is important to note that reversibility alone does not guarantee convergence of the chain. This can be illustrated by considering Example 3.11.

Example 3.11. $\mathcal{T} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, *Then $\mathcal{T}^{(2)} = I$, so $\mathcal{T}^{(2n)} = I$ and $\mathcal{T}^{(2n+1)} = \mathcal{T}$ $\forall n$. Thus, there is no convergence to equilibrium.*

Additionally, the condition of detailed balance is **stronger** than the concept of invariant.

1. invariant : $\pi(x) = \sum_y \pi(y)\mathcal{T}(y \to x)$
2. detailed balance : $\pi(x)\mathcal{T}(x \to y) = \pi(y)\mathcal{T}(y, x)$

If detailed balance holds, $\pi(x) = \sum_y \pi(y)\mathcal{T}(y \to x) = \pi(x)\sum_y \mathcal{T}(x \to y)$. We have $\sum_y \mathcal{T}(x \to y) = 1$ and $\pi(x) = \pi(x)$.

4 Conclusion

To conclude, my final project holds great significance as a cherished keepsake, representing the culmination of my entire university experience. It serves to unify the various strands of knowledge and skills I have acquired during these transformative years. Upon completing my PhD, I plan to revisit this article and engage in a candid critique of my own self-righteousness as a senior. I will promptly revise it, recognizing the urgency before I attain the status of a renowned scholar. Ha ha!

References

[1] Michael Schweinberger, Pavel N. Krivitsky, Carter T. Butts, and Jonathan R. Stewart. Exponential-Family Models of Random Graphs: Inference in Finite, Super and Infinite Population Scenarios. *Statistical Science*, 35(4):627 – 662, 2020.

[2] Alberto Caimo and Nial Friel. Bergm: Bayesian exponential random graphs in r. *Journal of Statistical Software*, 61(2):1–25, 2014.

[3] Alberto Caimo and Nial Friel. Bayesian inference for exponential random graph models. *Social Networks*, 33(1):41–55, 2011.

[4] J. R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, UK, 1997.

[5] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press, 2009.