36720 Notes: Exponential-family Random Graph Models (ERGMs) Part 1

Nicolás Kim

3 October 2016

1 Overview

- Definitions
  - Graph modeling
  - Examples: Erdös-Renyi, p_1, 2-star, triangle
- Properties
  - Edge prediction
  - Moments
- Estimation
  - MLE equation
  - Stochastic approximation
  - MCMCMLE

2 Definitions

Definition For any graph $G = (V, E)$ on n nodes, we refer to its *adjacency matrix* as A_{ij} (if viewed as a random variable) or a_{ij} (as a realization of A_{ij}).[1] We say $A_{ij} = 1$ if $(i, j) \in E$, and $A_{ij} = 0$ if $(i, j) \notin E$.

Definition A *graph model* is a probability distribution over the space of graphs.

[1] Uppercase letters will be used throughout these notes to refer to random variables, and lowercase letters will be used to refer to constants or realizations of random variables.

2.1 How do we model a random graph?

We can start by selecting a subset of features to model. In doing so, we are hoping that only some of the possible features are actually important, as far as the probability distribution is concerned.

In other words, we will postulate that a set of statistics $\{T_1, T_2, \ldots, T_d\}$ are sufficient.

Definition According to Neyman's factorization theorem, T is *sufficient for a parameter* θ if

$$p_\theta(x) = h(x)g(\theta; T(x)).$$

In other words, the density of the random variable X can be factored such that $g(\cdot)$ depends on x only through $T(x)$.

Remark If T is sufficient for θ, then the likelihood ratio can be simplified as:

$$\frac{p_\theta(x)}{p_{\theta_0}(x)} = \frac{h(x)g(\theta; T(x))}{h(x)g(\theta_0; T(x))} \tag{1}$$

$$= \frac{g(\theta; T(x))}{g(\theta_0; T(x))}. \tag{2}$$

So the likelihood ratio only depends on x through $T(x)$. Furthermore, if we wish to maximize the likelihood, there is no information in x that is necessary apart from what $T(x)$ provides.

One way to create a model where our desired statistics $T(\cdot)$ are sufficient is by using them in an exponential family distribution.

Definition A model $p_\theta(x)$ is an *exponential family* distribution if it can be written in the form

$$p_\theta(x) \propto \exp\left\{\sum_{i=1}^{d} T_i(x)\theta_i\right\} \tag{3}$$

$$= e^{T(x)\cdot\theta}. \tag{4}$$

This implies that if $p_\theta(x)$ is an exponential family model,

$$p_\theta(x) = \frac{e^{T(x)\cdot\theta}}{\sum_{x'} e^{T(x')\cdot\theta}} \tag{5}$$

$$= \frac{e^{T(x)\cdot\theta}}{Z(\theta)} \tag{6}$$

$$= e^{T(x)\cdot\theta - \psi(\theta)}. \tag{7}$$

Remark If a model has this form, then T is sufficient (by construction).

Remark If T is to be sufficient and its support does not change with θ, then $p_\theta(x)$ must be an exponential family distribution. The more precise statement follows.

Theorem 2.1 (Fisher-Pitman-Koopman-Darmois)**.** *Let $T = (T_1, T_2, \ldots, T_d)$ be a finite set of sufficient statistics for a model $p_\theta(x)$ with support that does not depend on θ. Then, $p_\theta(x)$ must either be an exponential family distribution, or a uniform distribution.*

Definition *Exponential-family Random Graph Models* (ERGMs) are exponential families over graphs. In other words, the sufficient statistics are functions of the graph/adjacency matrix.

A recipe for creating an ERGM is therefore:

1. Pick d (distinct) functions of the graph; they might be chosen through appeals to theory, experience, guesswork, tradition, referee pressure, trial and error, etc.

2. Then, calculate these statistics, and forget the original graph for all within-model work: simulating, testing, estimation, etc.

2.2 Some ERGMs

Example (Random graph/Erdös-Renyi). One parameter θ = probability of an edge. Assuming a directed graph with self-edges, the sufficient statistic is

$$T(\mathbf{a}) = \sum_{i,j} a_{ij},$$

so the model is

$$p_\theta(\mathbf{a}) = \prod_{i=1}^{n}\prod_{j=1}^{n} \theta^{a_{ij}}(1-\theta)^{1-a_{ij}} \tag{8}$$

$$= \exp\left\{ \sum_{i,j} a_{ij}\log\theta + (1-a_{ij})\log(1-\theta) \right\} \tag{9}$$

$$= \exp\left\{ n^2\log(1-\theta) + \sum_{i,j} a_{ij}\log\frac{\theta}{1-\theta} \right\}. \tag{10}$$

Example (Block models). Parameters are the entries of the affinity matrix $\mathbf{b} = [b_{rs}]_{rs}$ and the sufficient statistics are the edge counts between the blocks, $[e_{rs}]_{rs}$.

Example (p_1 model). The sufficient statistics are the out-degree of each node (denoted a_i), the in-degree of each node (denoted b_i), and the total number of reciprocated edges (denoted r). An edge is reciprocated if both $(i,j) \in E$ and $(j,i) \in E$.

If we do not include r, this is instead called a configuration model.

Example (Graph motif counts). It is common to use both the edge and 2-star counts, or the edge and triangle counts as the sufficient statistics for an ERGM. For instance, in the edge-triangle model, we have

$$p_\theta(\mathbf{a}) \propto \exp\{\theta_1 e(\mathbf{a}) + \theta_2 t(\mathbf{a})\},$$

where $e(\mathbf{a})$ is the number of edges and $t(\mathbf{a})$ is the number of triangles.

- These small motifs are used for practical reasons only.
- Motif count models can also include attributes on the nodes of the graph, e.g. the number of edges between nodes of the same type (this is called homophily).
- Can combine these models with e.g. the node degree, to get something closer to the p_1 model.
- Can include global characteristics (like the graph diameter), but this is not often seen.

Example (Not SBMs). The stochastic block model (with latent block assignments) is a mixture of exponential families, so it is not itself an exponential family.

3 Properties

3.1 Edge/link prediction

Assuming we have seen most of the graph, how do we guess whether there is an edge between a particular pair of nodes? I.e. we want to predict if $(i,j) \in E$ or $(i,j) \notin E$.

Denote by $\mathbf{a}_{+ij}$ the adjacency matrix $\mathbf{a}$, but with the edge (i,j) set so that $a_{ij} = 1$. Likewise, let $\mathbf{a}_{-ij}$ be $\mathbf{a}$ but with $a_{ij} = 0$. Then,

$$p_\theta(\mathbf{a}_{+ij}) = e^{T(\mathbf{a}_{+ij})\cdot\theta}/Z(\theta)$$

and

$$p_\theta(\mathbf{a}_{-ij}) = e^{T(\mathbf{a}_{-ij})\cdot\theta}/Z(\theta)$$

so that

$$\frac{p_\theta(\mathbf{a}_{+ij})}{p_\theta(\mathbf{a}_{-ij})} = e^{(T(\mathbf{a}_{+ij}) - T(\mathbf{a}_{-ij}))\cdot\theta} \tag{11}$$

$$= e^{\Delta_{ij}\cdot\theta} \tag{12}$$

$$\implies \log\frac{p_\theta(\mathbf{a}_{+ij})}{p_\theta(\mathbf{a}_{-ij})} = \Delta_{ij}\cdot\theta, \tag{13}$$

which is a logistic regression. This lends an easy interpretation to the parameters: "For any given configuration of the graph, if I toggle an edge and it leads to an increase in the statistics, it is more likely to see that configuration."

However, we cannot apply any causal interpretation, since there is no reason to think that any given edge (i, j) was generated after all of the other edges in the graph.

3.2 Moments of the sufficient statistics

As we saw before, the normalization factor of an exponential family is

$$Z(\theta) = \sum_x \exp\left\{\sum_i T_i(x)\theta_i\right\}.$$

Taking its derivative, we get

$$\begin{aligned}
\frac{\partial Z(\theta)}{\partial \theta_i} &= \sum_x \frac{\partial}{\partial \theta_i} \exp\left\{\sum_i T_i(x)\theta_i\right\} && (14)\\
&= \sum_x \exp\left\{\sum_{j\neq i} T_j(x)\theta_j\right\} \frac{\partial}{\partial \theta_i} \exp\left\{T_i(x)\theta_i\right\} && (15)\\
&= \sum_x \exp\left\{\sum_{j\neq i} T_j(x)\theta_j\right\} T_i(x) \exp\left\{T_i(x)\theta_i\right\} && (16)\\
&= \sum_x T_i(x) \exp\left\{\sum_i T_i(x)\theta_i\right\} && (17)\\
&= \sum_x T_i(x) \left(\frac{\exp\left\{\sum_i T_i(x)\theta_i\right\}}{Z(\theta)}\right) Z(\theta) && (18)\\
&= \sum_x T_i(x) Z(\theta) p_\theta(x) && (19)\\
&= Z(\theta) \sum_x T_i(x) p_\theta(x) && (20)\\
&= Z(\theta)\mathbb{E}_\theta[T_i]. && (21)
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}_\theta[T_i] &= \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta_i} Z(\theta) && (22)\\
&= \frac{\partial}{\partial \theta_i} \log Z(\theta). && (23)
\end{aligned}$$

Remark We can get the higher moments by taking higher-order derivatives.

4 Estimation

4.1 MLE equation

The likelihood of an exponential family is

$$L(\theta) = p_\theta(x) = e^{T(x)\cdot\theta}/Z(\theta).$$

We find the MLE by maximizing $L(\theta)$:

$$\left.\frac{\partial}{\partial\theta_i}\frac{e^{T(x)\cdot\theta}}{Z(\theta)}\right|_{\theta=\hat{\theta}} = 0 \tag{24}$$

$$\iff \left.\frac{Z(\theta)\frac{\partial}{\partial\theta_i}[e^{T(x)\cdot\theta}] - \frac{\partial}{\partial\theta_i}[Z(\theta)]e^{T(x)\cdot\theta}}{(Z(\theta))^2}\right|_{\theta=\hat{\theta}} = 0 \tag{25}$$

$$\iff Z(\hat{\theta})T_i(x)e^{T(x)\cdot\hat{\theta}} - e^{T(x)\cdot\hat{\theta}}\mathbb{E}_{\hat{\theta}}[T_i]Z(\hat{\theta}) = 0 \tag{26}$$

$$\iff T_i(x) = \mathbb{E}_{\hat{\theta}}[T_i]. \tag{27}$$

Note that all of this applies to general exponential families, not just ERGMs.

So to find the MLE of θ, we "just" need to solve for $\hat{\theta}$ in the equation $T_i(x) = \mathbb{E}_{\hat{\theta}}[T_i]$, where

$$\mathbb{E}_\theta[T_i] = \frac{\partial}{\partial\theta_i}\log Z(\theta) \tag{28}$$

$$= \frac{\partial}{\partial\theta_i}\log\sum_x e^{\theta\cdot T}. \tag{29}$$

How big is this sum? For graphs, "x" is a full graph, so there are $2^{\binom{n}{2}}$ terms in this sum, which is the number of undirected simple graphs on n nodes. This sum is too large to simply brute-force it.

What can we do to get around this? In some special cases, including the edge-triangle ERGM, the block model, and the Erdös-Renyi random graph model, we can get a closed-form solution for the MLE.

4.2 Stochastic approximation

When this fails, we can try simulating:

- Start with an initial graph configuration $\mathbf{a}^{(0)}$.
- Pick an edge (i, j) at random.
- Flip that edge with probability

  $$\frac{p_\theta(\mathbf{a}^{(0)}_{+ij})}{p_\theta(\mathbf{a}^{(0)}_{-ij})},$$

  which does not involve $Z(\theta)$.

This is a Gibbs sampling procedure for finding the correct equilibrium distribution. In particular, since $Z(\theta)$ is not involved, we circumvent the issue of computing its value.

Now, to solve $T(x) = \mathbb{E}_{\hat{\theta}}[T]$ for $\hat{\theta}$,

(i) Start with a guestimate $\hat{\theta}^{(0)}$.

(ii) Use simulation to get many graphs from $\hat{\theta}^{(0)}$.

(iii) Approximate $\mathbb{E}_{\hat{\theta}}[T]$ by sample averages.

(iv) Adjust $\hat{\theta}^{(0)}$ to $\hat{\theta}^{(1)}$ to bring $\mathbb{E}_{\hat{\theta}}[T]$ closer to $T(x)$.

This procedure is stochastic approximation; it can be implemented via the Robbins-Monro algorithm.

4.3 MCMCMLE

By the definition of the MLE,

$$\hat{\theta} = \operatorname*{argmax}_{\theta} L(\theta) = \operatorname*{argmax}_{\theta} \frac{L(\theta)}{L(\theta_0)}$$

where θ_0 is some fixed initial guess for θ. We can rewrite the likelihood ratio as

$$\frac{L(\theta)}{L(\theta_0)} = \frac{e^{T(x)\cdot\theta}/Z(\theta)}{e^{T(x)\cdot\theta_0}/Z(\theta_0)} \tag{30}$$

$$= \exp\left\{T(x)\cdot(\theta-\theta_0)\right\} / \frac{Z(\theta)}{Z(\theta_0)}, \tag{31}$$

and we can rewrite the denominator as

$$\frac{Z(\theta)}{Z(\theta_0)} = \sum_{x'} \frac{e^{T(x')\cdot\theta}}{Z(\theta_0)} \tag{32}$$

$$= \sum_{x'} \frac{e^{T(x')\cdot(\theta-\theta_0+\theta_0)}}{Z(\theta_0)} \tag{33}$$

$$= \sum_{x'} e^{T(x')\cdot(\theta-\theta_0)} \frac{e^{T(x')\cdot\theta_0}}{Z(\theta_0)} \tag{34}$$

$$= \mathbb{E}_{\theta_0}\left[e^{T(x')\cdot(\theta-\theta_0)}\right]. \tag{35}$$

We can estimate this by simulating with θ_0 *only*. That is, we never need to use any of the updates of θ.

However, if $\theta - \theta_0$ is large, this estimate will oscillate. We would either need a ton of samples, or to update θ_0 to a better value at some point. But in principle, we do not need anything but our initial guess θ_0 to perform this step.[2]

[2]For an implementation of this see the statnet project by Morris, Handcock, et al.