36-720 Statistical Network Models - October 5 Lecture Notes

Abulhair Saparov
asaparov@andrew.cmu.edu

Background from previous lecture

A distribution $p(x)$ parameterized by θ is said to be in the exponential family with sufficient statistics $T \triangleq \{T_1, \ldots, T_d\}$ when

$$p_\theta(x) = \frac{\exp\{\theta^\top T(x)\}}{\sum_{x'} \exp\{\theta^\top T(x)\}} = \frac{1}{Z(\theta)} \exp\left\{\sum_{i=0}^{d} \theta_i T_i(x)\right\} = \exp\left\{\theta^\top T(x) - t(\theta)\right\}$$

The **moment generating function** of T is $M_\theta(\phi) = \mathbb{E}_\theta[\exp\{\theta^\top T\}]$, so in the exponential family, $M_\theta(\phi) = Z(\theta + \phi)/Z(\theta)$. The value of θ that maximizes the likelihood $\hat{\theta} \triangleq \arg\max_\theta p(x|\theta)$ satisfies

$$\mathbb{E}_{\hat{\theta}}[T] = T(x).$$

That is, the most likely θ makes the expected values of the sufficient statistics equal to the observed values.

Exponential family distributions maximize entropy

Def 1. *The **entropy** of a random variable X is defined as*

$$H[X] \triangleq -\mathbb{E}[\log X],$$

which for discrete X is $\sum_x p(x) \log p(x)$ (where the quantity $p(x) \log p(x) = 0$ when $p(x) = 0$).

Hereafter, we only consider discrete X. The entropy has a number of useful properties:

1. $H[X] \geq 0$, with $H[X] = 0$ if and only if $p(x) = 1$ for some x (X is degenerate/deterministic),
2. $H[X] \leq \log |\mathcal{X}|$ where $\mathcal{X}$ is the set of possible values of X,
3. $H[X] = \log |\mathcal{X}|$ if and only if $p(x) = \frac{1}{|\mathcal{X}|}$ for all x (X is the uniform distribution).

These properties can be demonstrated by observing $H[X] = \log |\mathcal{X}| - D(p(x)||u(x))$ where $u(x)$ is the uniform distribution on $\mathcal{X}$, and $D(p||q)$ is the Kullback-Leibler divergence from distributions q to p.

The *maximum entropy problem* is to find the distribution for X which maximizes $H[X]$ under the constraints $\mathbb{E}[T_1] = t_1$, $\mathbb{E}[T_2] = t_2$, …, $\mathbb{E}[T_d] = t_d$. The constraints can be written in vectorized form: $\mathbb{E}[T] = t$.

$$\max_{p(x)} H[X] \qquad \text{such that } \mathbb{E}[T] = t.$$

We can use Lagrange multipliers to rewrite this as an unconstrained optimization

$$\max_{P \in \mathbb{R}^{|\mathcal{X}|}, \lambda, \theta} \mathcal{L}(P, \lambda, \theta) \text{ where}$$

$$\mathcal{L}(P, \lambda, \theta) = -\sum_{x \in \mathcal{X}} P_x \log P_x + \lambda \left(\sum_{x \in \mathcal{X}} P_x - 1\right) + \theta^\top \left(\sum_{x \in \mathcal{X}} T(x) P_x - t\right).$$

$$\frac{\partial \mathcal{L}}{\partial P_x} = -\left(\log P_x + \frac{P_x}{P_x}\right) + \lambda + \theta^\top T(x).$$

At the optimal P_x^*, λ^*, and θ^*, we have

$$\begin{aligned} 0 &= -\left(\log P_x^* + 1\right) + \lambda^* + (\theta^*)^\top T(x), \\ P_x^* &= \exp\left\{(\theta^*)^\top T(x) + \lambda^* - 1\right\}, \\ \text{and so } \exp\{\lambda^* - 1\} &= \frac{1}{\sum_{x \in \mathcal{X}} \exp\{(\theta^*)^\top T(x)\}}. \end{aligned}$$

Thus, the distribution that maximizes the entropy and satisfies the constraints is

$$p(x) = \frac{1}{Z(\theta^*)} \exp\{(\theta^*)^\top T(x)\}.$$

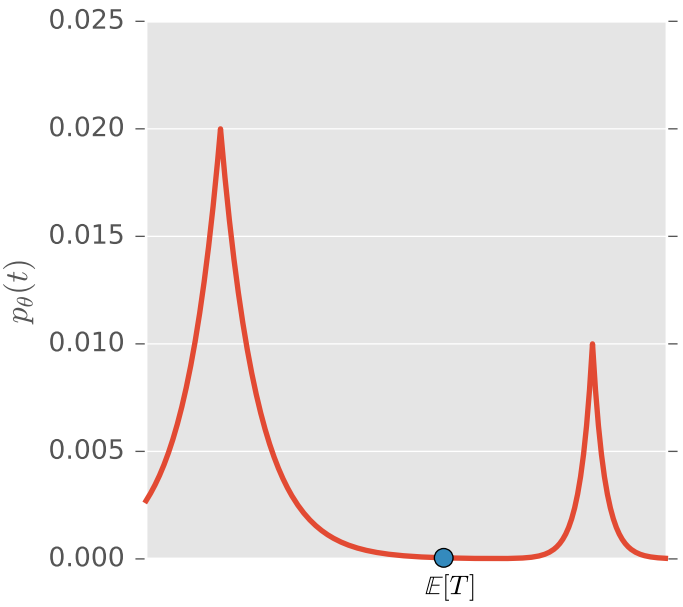
If t are the observed values, then θ^* is the maximum likelihood estimate $\hat{\theta}$ due to the constraint $\mathbb{E}[T] = t$. Therefore, the most random (highest entropy) distribution under the expectation-value constraint $\mathbb{E}[T] = t$ is always an exponential family

$$p(x) = \frac{1}{Z(\theta^*)} \exp\{(\theta^*)^\top T(x)\} \qquad \text{with } \mathbb{E}_{\theta^*}[T] = t.$$

The maximum likelihood estimate for the exponential family is the maximum entropy distribution where the expected value of the sufficient statistics match the observed values. The maximum entropy distribution can be appealing since, in some sense, it assumes as little as possible about the structure of the underlying process.

Problems with exponential-family random graph models

Degeneracy: In many typical applications of exponential-family random graph models (ERGMs), Markov chain Monte Carlo methods for performing inference tended to take a very long time to mix and reach the stationary distribution. Handcock et al. [1] found that at the MLE, where $\mathbb{E}_{\hat{\theta}}[T] = T(x)$, for real-world data x, the distribution $p_\theta(t)$ is frequently multimodal, and the probability mass at the expected value $p_\theta(\mathbb{E}[T])$ can be very low.



Thus, the Monte Carlo algorithm will spend much of its time near the modes of the distribution, but very little time near the expected value. The probability of the algorithm jumping from one mode to another can

be very small. The distribution over graphs in these cases has very low entropy. Mark Handcock called this distribution *near-degenerate* for this reason.

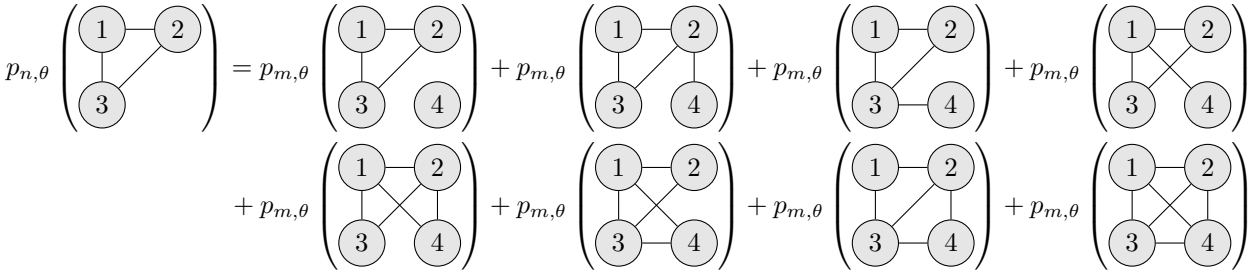
If $p_\theta[T = t_{observed}]$ is very low, then maximizing the likelihood does not help much, as the model effectively implies that the observed data is incredibly unlikely anyway. This is true even though $\mathbb{E}_\theta[T] = t_{observed}$. This behavior suggests that there is no distribution in that model for which your data is a typical output, and the model does not explain the data well.

Degeneracy occurs when the observed data essentially cannot be fit by the selected model, and the model attempts to position modes of the distribution around the observed data, which is the best it can do. Often, but not always, one mode occurs around nearly-full graphs, and another mode occurs around nearly-empty graphs. [1] has examples, where the sufficient statistics are the number of edges and number of triangles. They suggest to change the sufficient statistics, and to use "geometrically-weighted" statistics [5, 2]. For example, a statistic that counts the number of k-stars in a graph tends to greatly overestimate the probability of graphs with high-degree vertices, under the exponential family random graph model. To alleviate this, the impact of these features can be reduced with geometrically decreasing weights, which have been found to work well empirically. Similar geometrically decreasing weights have been proposed for other motif statistics, such as the number of triangles.

Projectivity: Consider a dataset of two sample sizes, x and (x, y), where the sample size of x is n and the sample size of (x, y) is $m > n$. A model is called **projective** when for all θ and x,

$$p_{n,\theta}(x) = \sum_y p_{m,\theta}(x, y).[4]$$

Figure 1: An example of a model $p_{n,\theta}$ that is projective for a triangle graph x. Note that if the model is called projective, this property must hold for all x.



Projectivity provides useful properties that allow the study of the asymptotic and non-asymptotic properties of the model. For example, it is a necessary condition for the Kolmogorov extension theorem [4, 3]. Many models that we study are projective, such as all models in which the observations are independent and identically distributed. Projectivity also holds in conditionally-specified models where the conditional distribution of new observations given the old observations is specified. However, ERGMs do not fall into either of these categories.

Under what conditions does an ERGM satisfy projectivity? Intuitively, an ERGM with sufficient statistics T is projective when the sufficient statistics for a supergraph G can be written as a sum of the sufficient statistics of any partition of G into two subgraphs, along with a term for the edges between the partitions. To make this more precise, we need a few definitions.

Def 2. *The **volume factor** for a fixed value of a sufficient statistic t is the number of graphs in a set of graphs $\mathcal{X}$ whose sufficient statistic matches the fixed value.*

$$v_n(t) \triangleq |\{x \in \mathcal{X} : t_n(x) = t\}|.$$

Def 3. *The **joint volume factor** is defined for some $\delta \in \mathbb{R}$,*

$$v_{n,m}(t, \delta) \triangleq |\{(x, y) \in \mathcal{X} : t_n(x) = t, t_m(x, y) - t_n(x) = \delta\}|,$$

where y can be imagined as subgraphs of (x, y).

Def 4. *The **conditional volume factor** is*

$$v_{n,m}(x, \delta) = |\{y : t_m(x, y) - t_n(x) = \delta, (x, y) \in \mathcal{X}\}|.$$

Def 5. *A sufficient statistic t is said to have **separable increments** if $t_m(x, y) - t_n(x)$ has the same range for all $x \in \mathcal{X}$, and the conditional volume factor is constant in x:*

$$v_{n,m}(x, \delta) = v_{n,m}(x', \delta),$$

for all $x, x' \in \mathcal{X}$, and $\delta \in \mathbb{R}$.

As an example, this clearly would not hold for a statistic that counts the number of triangles. If we imagine a k-star graph, the addition of a new vertex can dramatically increase the number of triangles from 0 to k.

Thm 1. *An exponential family model is projective if and only if all sufficient statistics T have separable increments. [4]*

Thm 2. *No sufficient statistic that counts a motif with at least 3 vertices has separable increments.*

Any ERGM where all sufficient statistics are sums over dyads (pairs of nodes) is projective. So for instance, block models are projective.

References

[1] Mark S Handcock. Assessing degeneracy in statistical models of social networks. *Center for Statistics and the Social Sciences*, January 2003.

[2] David R. Hunter, Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. Ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3):1–29, 12 2007.

[3] Olav Kallenberg. *Foundations of modern probability*, pages 91–93. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

[4] Cosma Rohilla Shalizi and Alessandro Rinaldo. Consistency under sampling of exponential random graph models. *Ann. Statist.*, 41(2):508–535, 04 2013.

[5] Tom A. B Snijders, Philippa E. Pattison, Garry L Robins, and Mark S Handcock. New specifications for exponential random graph models. April 2004.