Using latent variables to account for heterogeneity in exponential family random graph models

Johan Koskinen

Abstract

We consider relaxing the homogeneity assumption in exponential family random graph models (ERGMs) using binary latent class indicators. This may be interpreted as combining a posteriori blockmodelling with ERGMs, relaxing the independence assumptions of the former and the homogeneity assumptions of the latter. We propose a Markov chain Monte Carlo algorithm for drawing from the joint posterior of the model parameters and latent class indicators.

1. Introduction

Researchers in social science have long recognised the potential of using graph theory for studying the social interaction among social units [13]. A social network may be conceptualised as consisting of a set of vertices \( V = \{1, \ldots, n\} \), representing the social units, e.g. people, that are pairwise relationally connected by ties, represented by an edge set \( E \subseteq \mathcal{N} \), where \( \mathcal{N} = \binom{V}{2} \), for undirected relations and \( \mathcal{N} = V^{(2)} \) for directed relations. We define the adjacency matrix \( x = \{x_v : e \in \mathcal{N} \} \) as the collection of edge indicators \( x_v = 1_E(e), x \in \mathcal{X} = \{0, 1\}^\mathcal{N} \). White et al. [19] proposed summarising the structural information by reducing the graph to a blockmodel, a schematic representation of positions \( \rho : V \rightarrow \mathcal{P} = \{0, \ldots, P - 1\} \) and a blockmodel that is a graph \( G(\mathcal{P}, \mathcal{B}) \). Strictly speaking \( G(\mathcal{P}, \mathcal{B}) \) is such that \( x_{ij} = 1_B(\{\rho(i), \rho(j)\}), \) for \( i \neq j \). A priori blockmodels and a posteriori blockmodels are now well established practice in social network analysis [1]. Stochastic a priori blockmodels [2], [7] have since been extended to stochastic a posterior blockmodels, where the target of inference is to infer the positions of the vertices from their observed relations [21], [16], [23]. For stochastic blockmodels the (pairs of) tie-indicators are independent but [6] sought to relax this usually unrealistic assumption.

Exponential random graph models (ERGMs) take dependencies between the elements of \( x \) into account. Dependencies are specified according to a dependence graph \( \mathcal{D} = G(\mathcal{N}, \mathcal{E}) \), for example \( \{i, j\}, \{k, \ell\} \in \mathcal{E} \), iff \( \{i, j\} \cap \{k, \ell\} \neq \emptyset \) in the case of Markov graphs [3]. This specifies a log-linear model \( p(x) \) with parameters

\(^1\)University of Melbourne, E-mail: johank@unimelb.edu.au
\[ p(x|\theta) = \exp\{\theta^T z(x) - \kappa(\theta)\}, \]

where where \(z\) is a \(p \times 1\) vector valued function of \(x\) and the normalising constant
\[ \kappa(\theta) = \log(c(\theta)), \quad c(\theta) = \sum_{y \in X} \exp\{\theta^T z(y)\}, \]

is a function of the parameter vector \(\theta \in \Theta \subseteq \mathbb{R}^p\). For Markov graphs \(z\) consists of counts of the number of edges, stars and triangles (for non-Markov specifications see [22]).

The homogeneity assumption may be relaxed if attributes of the nodes are observed [18] but unexplained heterogeneity has detrimental effects on estimation.

2. Bayesian inference for ERGM

Given a fixed \(x\) we may perform inference for \(\theta\) by exploring a sample from the posterior distribution \(\pi(\theta|x) \propto p(x|\theta)\pi(\theta)\), using the linked importance sampler auxiliary variable (LISA) Metropolis-Hastings proposed in [10]. LISA combines the auxiliary variable MCMC [15] with the linked importance sampler [17] (a similar algorithm is proposed in [14]). In brief, LISA relies on the introduction of an auxiliary variable \(\omega\) defined on the space \( \prod_{i=1}^{n} X^K \times \{K\} \times \{K\} \), constructed such that there are two pmfs, a forward distribution \(P_{\theta,\psi}^F(\omega)\) and a backward distribution \(P_{\theta,\psi}^B(\omega)\), each indexed by a pair \((\theta, \psi)\) of parameter vectors. Furthermore, the normalising constants of these distributions are \(c(\theta)\) and \(c(\psi)\) respectively. In LISA, instead of producing draws from \(\pi(\theta|x)\) directly, we produce draws from the joint distribution
\[ \pi(\theta, \omega|x) \propto p(x|\theta)P_{\theta,\psi}^B(\omega) = \exp\{\theta^T z(x)\}/c(\theta)Q_{\psi,\theta}^B(\omega)/c(\psi) \]

This may be accomplished for a Metropolis-Hastings sampler that in each iteration, with present state \((\theta^{(t)}, \omega^{(t)})\), proposes a move to \((\theta^*, \omega^*)\) drawn from the proposal distribution \(f(\theta^*, \omega^*)|\theta^{(t)}, \omega^{(t)})\), and this move is accepted with probability \(\min\{1, H\}\), where
\[ H = \frac{\pi(\theta^*, \omega^*|x) f(\theta^{(t)}, \omega^{(t)}|\theta^*, \omega^*)}{\pi(\theta^{(t)}, \omega^{(t)}|x) f(\theta^*, \omega^*|\theta^{(t)}, \omega^{(t)})}. \]

Evaluating this expression involves a ratio \(c(\theta^*)/c(\theta^{(t)})\) of undefined normalising constants but if \(f\) is chosen to be \(f(\theta^*, \omega^*|\theta^{(t)}, \omega^{(t)}) = f(\omega^*|\theta^*)f(\theta^*|\theta^{(t)})\), where \(\theta^*|\theta^{(t)} \sim N(\theta^{(t)}, \Sigma)\) and \(\omega^*|\theta^* \sim Q_{\psi,\theta}^B(\omega^*)/c(\theta^*)\), the ratio \(H\) simplifies to \(e^{\theta^T z(x) - \theta^{(t)} z(x)}\Pi_{\psi}(\omega^*, \theta^*)/\Pi_{\psi}(\omega^{(t)}, \theta^{(t)})\), where \(\Pi_{\psi}(\omega, \theta) = Q_{\psi,\theta}^B(\omega)/Q_{\psi,\theta}^B(\omega)\). Note that \(H\) has the interpretation that \(E[\Pi_{\psi}(\omega^*, \theta^*)] = c(\psi)/c(\theta^*)\), where the expectancy is with respect to \(P_{\theta,\psi}^E(\omega^*)\). This is also why we may characterise the algorithm as a Metropolis-Hastings that accepts a proposal with a Hastings ratio where \(\Pi_{\psi}(\omega, \theta) = \lambda_{\text{LIS}}(\theta, \psi; \omega)\), and where \(\lambda_{\text{LIS}}\) is the linked importance
sampler (LIS) estimator [17] of the ratio of normalising constants. Given a sample \( \omega = (y, \mu, \nu) \), the LIS estimate of \( \lambda(\theta, \psi) \) is given by

\[
\lambda_{\text{LIS}}(\theta, \psi; \omega) = \prod_{j=1}^{m-1} \frac{\sum_{i=1}^{K} w(y_j^{(i)}; \theta(j), \theta(j + 1))}{\sum_{i=1}^{K} w(y_{j+1}^{(i)}; \theta(j + 1), \theta(j))},
\]

where the weights \( w \) are derived from the sampling process that we now proceed to describe.

The LIS estimator is based on \( K \) sample points from \( m \) Markov chains \( y_j = (y_j^{(1)})_{i=1}^{K} \) with different target distributions, drawn using Metropolis-Hasting transition probabilities \( T_{\theta(t)} \) and \( T_{\psi(t)} \), for a smooth mapping connecting \( \theta \) and \( \psi \) as in path sampling [4].

The \( m \) samples are connected in points \( \mu_1, \ldots, \mu_m \) and \( \nu_1, \ldots, \nu_m \), such that given \( \mu_j \) and \( (y_j^{(1)})_{i=1}^{K} \), we set \( y_{j+1}^{(i+1)} := y_j^{(\mu_i)} \). Given \( \nu_j \) and \( y_j^{(\nu_j)} \) we create the chain \( (y_j^{(1)})_{i=1}^{K} \) by simulating forward from \( y_j^{(\nu_j)} \) using \( T_{\theta(j)}(y_j^{(\nu_j)}, y_j^{(\nu_j + 1)}), T_{\psi(j)}(y_j^{(\nu_j + 1)}, y_j^{(\nu_j + 2)}), \) etc., until we have produced \( y_j^{(K)} \). We also simulate backwards from \( y_j^{(\nu_j)} \) using the reversed transition kernels \( T_{\theta(j)}(y_j^{(i)}, y_j^{(i - 1)}) \), until we have produced \( y_j^{(1)} \). The implied pmf of a chain \( y_j = (y_j^{(1)})_{i=1}^{K} \) conditional on the insertion point and the linking state is

\[
P(y_j | \nu_j, y_j^{(\nu_j)}) = \prod_{i=1}^{K} T_{\theta(j)}(y_j^{(i + 1)}, y_j^{(i)}) \prod_{i=\nu_j}^{K} T_{\psi(j)}(y_j^{(i)}, y_j^{(i + 1)}).
\]

To choose which of the \( K \) sample points that should provide the link to the next chain, we choose \( \mu_j \) with probabilities

\[
\eta(\mu_j | y_j) = w(y_j^{(\mu_j)}; \theta(j), \theta(j + 1))/\sum_{i=1}^{K} w(y_j^{(i)}; \theta(j), \theta(j + 1)),
\]

where \( w(y; \theta, \theta^*) = q(y; \theta)^{-1/2}q(y; \theta^*)^{1/2} \) and insertion points \( \nu_j \) are chosen uniformly on \( \{1, \ldots, K\} \). The initial state \( y_1^{(\nu_1)} \) of the first chain is chosen according to \( p(y_1^{(\nu_1)} | \theta) \).

3. Latent variable

For \( P = 2 \), assuming \( a = (a_i)_{i \in V} \) is a collection of indicators \( a_i = \rho(i) \), including

the statistics \( z_L(x) = \text{deg}(x), z_M(x; a) = \sum_{i < j} x_{ij}(a_i + a_j), \) and \( z_H(x; a) = \sum_{i < j} x_{ij} a_i a_j \), parameters \( (\theta_L, \theta_M, \theta_H)^T \in R^p \), defines a Bernoulli blockmodel (BBM) where the probability of a tie \( \{i, j\} \in E \) is \( p_{a_i a_j} \), with \( \logit(p_{a_i a_j}) = \theta_L + \theta_M(a_i + a_j) + \theta_H a_i a_j \). If \( a \) is an unobserved estimation may be done as in [16]

but when we introduce parameters that correct for lack of independence this is no longer possible (nor is the ML-approach of [20] as \( \times \) becomes a non-trivial
function of both $\theta$ and the binary $a$). Without loss of generality we may assume that the dependence is described by the alternating $k$-triangle statistic [22] $z_T(x; a) = (1 + e^{-a})^{-1}\{\deg(x) - \sum_{i<j} x_{ij} e^{-\alpha S_{ij}}\}$, for a smoothing constant $\alpha > 0$, and a parameter $\theta_T$, where $S_{ij} = \frac{1}{2}[k : \{i,k\}, \{j,k\} \in E \setminus \{i,j\}]$. For the purpose of estimation we assume that $a$ is another parameter to be estimated, defining $\eta = (\hat{\theta}^T, a^T)^T$, where $\hat{\theta} = (\hat{\theta}_L, \theta_M, \hat{\theta}_H, \theta_T)^T$, as our target of inference for the model $p(x | \eta) = \exp\{g(x; \eta) - \beta(\eta)\}$. This now defines a curved ERGM [8] with natural parametrisation $g(x; \eta) = \beta(\eta)^T h(x)$, for $\beta(\eta) = (\theta_L, \theta_T, \theta_{1,2}, \ldots, \theta_{n,n(n-1)})^T$ and $h(x) = (z(x), z_T(x; a), x_{1,2}, \ldots, x_{n,n(n-1)})^T$, where $\theta_{ij} = \theta_M(a_i + a_j) + \theta_H a_i a_j$. LISA only requires that we may evaluate $p(x | \eta) c(\eta)$ for fixed values of $\eta$, in which case $p(x | \eta)$ reduces to a regular ERGM with parameter $\theta$ and statistics $z(x; a)$.

The model is not fully identified as $\theta^*$ may be set so that $p(x | \eta) = p(x | \eta^*)$, $a^* = 1 - a$. This may lead to label switching [12] which may be solved in different ways for BBMs [16] but here there is the additional issue that the model may be “separated” (c.p. [5]) for some realisation of $a$ so that the posterior may be improper for improper priors [11]. To counter these two issues we assume the following partially informative priors: $\pi(\theta_L, \theta_T) \propto c$; $\theta_M | \theta_H, \theta_L, \theta_T \sim N(0, \lambda)$; and $\theta_H | \theta_M, \theta_L, \theta_T \sim N(0, \lambda)$ truncated to the left in $\theta_L$, for shrinkage factor $\lambda$.

To illustrate the application of the algorithm we fit the three models in Table 1 to the well-known Kapferer’s tailor data that has $n = 39$ actors [9] ($\lambda = 10$ in all models). Model II only includes $z_L$ and $z_T(x; a)$ and is fitted according to [10]; to set $\psi = (\hat{\theta}^T, \hat{a}^T)^T$ for Model III, we have used the predicted $\hat{a} (= \hat{a})$ from Model I, and $\hat{\theta}$ is obtained as $\theta_{M, L, H}(\hat{a})$ assuming $\hat{a}$ to be true. A proposal distribution that is consistent with the form of the prior is to draw, in iteration $j$, $\theta_M, \theta_L, \theta_H, \theta_T \sim N(\theta_{M}^{(j)}, \theta_{L}^{(j)}), \theta^{(j)}_{T}, \Sigma_{124})$, and $\theta_H^* \sim N(\theta_H^{(j)}, \sigma_3)$ truncated to the left in $\theta_M^{*}$. To set $\Sigma_{124}$ and $\sigma_3$ we have used the rescaled information matrix $I(\theta_{M, L, H}(\hat{a})^{-1})$. A nearest neighbour proposal is used for $a^* | a^{(j)}$, where $a^*_i := 1 - a_i^{(j)}$ for a number (usually one) of $i \in V$. In the LIS part we have used a linear map $\eta(t) = \eta t + (1-t)\psi$ (as described in [10]). In other words, for the purposes of implementing LIS, $a_i$ is allowed to be continuous on $[0, 1]$. To improve mixing, $\theta$ and $a$ are updated in separate blocks which give satisfactory mixing, with the caveats in [10] regarding drawing $p(y_i | \psi, \theta)$ (in lieu of perfect sampling a burn-in of 50,000 is used for model III and for Model II the pseudo perfect sampling scheme of [10] was used; details of the performance may be obtained from the author).

The allocations for the vertices are stable and the measure $H = \frac{8}{n(n-1)} \sum_{i<j} \hat{\pi}_{ij}(1 - \hat{\pi}_{ij})$ [16], where $\hat{\pi}_{ij}$ is the MCMC estimator of $\Pr(a_i = a_j | x)$, is 0.21 and 0.05 for Model I and III respectively. The estimates in Table 1 differ mostly in magnitude between models, most notably for $\theta_L$, but not substantively. The correlation of $\Pr(a_i = 1 | x)$ between Model I and III is 0.976.

4. Summary

We have proposed an algorithm for performing Bayesian inference for ERGMs with latent variables meant to capture unexplained heterogeneity. We have illustrated the application of the algorithm to a well-known data set assuming two classes.
Table 1: Summaries of posteriors for three models fitted to Kapferer’s tailors [9]

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std</td>
<td>Mean</td>
</tr>
<tr>
<td>$\theta_L$</td>
<td>0.83</td>
<td>0.28</td>
<td>-4.24</td>
</tr>
<tr>
<td>$\theta_M$</td>
<td>-2.13</td>
<td>0.25</td>
<td>-2.38</td>
</tr>
<tr>
<td>$\theta_H$</td>
<td>2.04</td>
<td>0.39</td>
<td>2.65</td>
</tr>
<tr>
<td>$\theta_T$</td>
<td>1.37</td>
<td>0.17</td>
<td>1.10</td>
</tr>
</tbody>
</table>

Further work needs to be done in order to investigate the performance of the algorithm but the principled solution also opens up for further research into how to assess the choice of $P$; how to make full use of the posterior predictive distributions; and if latent classes may improve the fit of the ERGM.

References


