Some Bayesian Approaches for ERGM

Ranran Wang, UW

MURI-UCI August 25, 2009

Outline

- Introduction to ERGM
- Current methods of parameter estimation:
  - MCMCMLE: Markov chain Monte-Carlo estimation
  - MPLE: Maximum pseudo-likelihood estimation
- Bayesian Approaches:
  - Exponential families and variational inference
  - Approximation of intractable families
  - Application on ERGM
  - Simulation study

Introduction to ERGM

Network Notation

- m actors; $n = \frac{m(m-1)}{2}$ dyads
- Sociomatrix (adjacency matrix) Y: $\{y_{i,j}\}_{i,j=1,\cdots,n}$
- Edge set $\{(i,j) : y_{i,j} = 1\}$.
- Undirected network: $\{y_{i,j} = y_{j,i} = 1\}$

ERGM

Exponential Family Random Graph Model (Frank and Strauss, 1986; Wasserman and Pattison, 1996; Handcock, Hunter, Butts, Goodreau and Morris, 2008):

$$\log[P(Y = y_{obs};\eta)] = \eta^T \phi(y_{obs}) - \kappa(\eta, \mathcal{Y}), \quad y \in \mathcal{Y}$$

where

- **Y** is the random matrix
- $\eta \in \Omega \subset \mathbb{R}^q$ is the vector of model parameters
- $\phi(y)$ is a q-vector of statistics
- $\kappa(\eta, \mathcal{Y}) = \log \sum_{z\in\mathcal{Y}} \exp\{\eta^T \phi(z)\}$ is the normalizing factor, which is difficult to calculate.
- R package: **statnet**

Current estimation approaches for ERGM

MCMC-MLE (Geyer and Thompson 1992, Snijders, 2002; Hunter, Handcock, Butts, Goodreau and Morris, 2008):

1. Set an initial value η_0, for parameter η.
2. Generate MCMC samples of size m from P_{η_0} by Metropolis algorithm.
3. Iterate to obtain a maximizer $\tilde{\eta}$ of the approximate log-likelihood ratio:

$$(\eta - \eta_0)^T \phi(y_{obs}) - \log \left[\frac{1}{m} \sum_{i=1}^{m} \exp \left\{ (\eta - \eta_0)^T \phi(Y_i) \right\} \right]$$

4. If the estimated variance of the approximate log-likelihood ratio is too large in comparison to the estimated log-likelihood for $\tilde{\eta}$, return to step 2 with $\eta_0 = \tilde{\eta}$.
5. Return $\tilde{\eta}$ as MCMCMLE.

MPLE (Besag, 1975; Strauss and Ikeda, 1990):

Conditional formulation:

$$\text{logit}[P(Y_{ij} = 1 | Y_{ij}^C = y_{ij}^C)] = \eta^T \delta(y_{ij}^C).$$

where $\delta(y_{ij}^C) = \phi(y_{ij}^+) - \phi(y_{ij}^-)$, the change in $\phi(y)$ when y_{ij} changes from 0 to 1 while the rest of network remains y_{ij}^C.

Comparison

Simulation study: van Duijn, Gile and Handcock (2008)

MCMC-MLE	MPLE
• Slow-mixing • Highly depends on initial values • Be able to model various network characteristics together.	• Deterministic model; computation is fast • Unstable • Dyadic-independent model; could not capture higher-order network characteristics.

Idea:
Use prior specifications to deemphasize degenerate parameter values

Let pr(η) be an arbitrary prior distribution for η..
Choice of prior distributions for η?

- pr(η) based on social theory or knowledge
- Many conjugate prior families
  $\Rightarrow$ Gutiérrez-Peña and Smith (1997), Yanagimoto and Ohnishi (2005)

Standard conjugate prior (Diaconis and Ylvisaker 1979):
Let $h(\nu, \gamma)$ be the $(q+1)$ parameter exponential family with distribution:

$$\text{pr}(\eta; \nu, \gamma) = \frac{\exp\{\nu^T \eta + \gamma\psi(\eta)\}}{c(\gamma, \nu)} \qquad \eta \in \Lambda, \gamma > 0$$

where $\psi(\cdot)$ is a prespecified function (e.g., $-\log(c(\eta))$.

Reexpressing conjugate priors

$$\mathrm{pr}(\eta; \eta_0, \gamma) = \frac{\exp\{-\gamma D(\eta_0, \eta)\}}{d(\gamma, \eta_0)} \qquad \eta \in \Lambda, \gamma > 0$$

where $D(\eta_0, \eta)$ is the Kullback-Leibler divergence from the model $P_\eta(Y = y)$ to the model $P_{\eta_0}(Y = y)$.

This can be translated into a prior on the mean-values:

$$\mathrm{pr}(\mu; \mu_0, \gamma) = \frac{\exp\{-\gamma D(\mu, \mu_0)\}}{d(\gamma, \mu_0)} \qquad \mu \in \mathrm{int}(\mathrm{C}), \gamma > 0$$

Posterior distributions

$$\mathrm{pr}(\mu | Y = y; \mu_0, \gamma) = \frac{\exp\{-D(g(y), \mu) - \gamma D(\mu, \mu_0)\}}{d(\gamma + 1, \mu_0)} \qquad \mu \in \mathrm{int}(\mathrm{C}), \gamma > 0$$

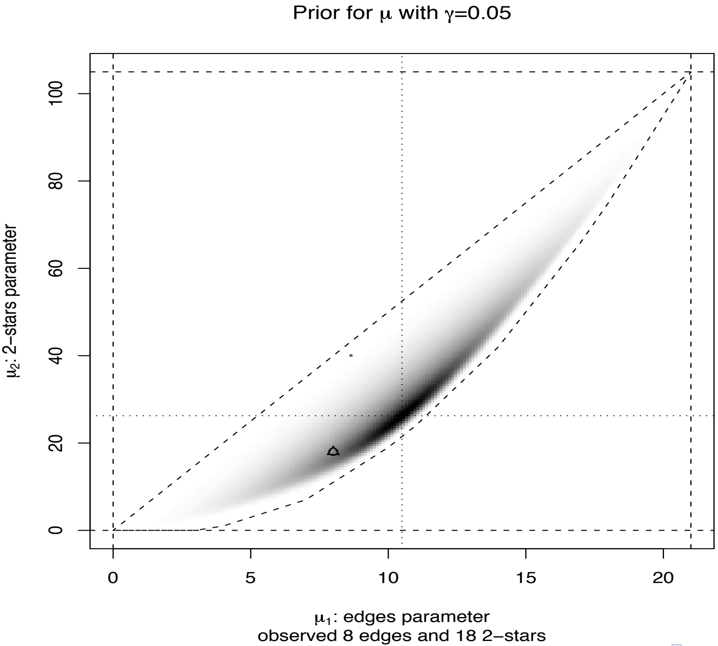
$$\mathbf{E}(\nu; \nu_0, \gamma) = \nu_0$$

$$\mathbf{E}(\mu; \mu_0, \gamma) = \mu_0$$

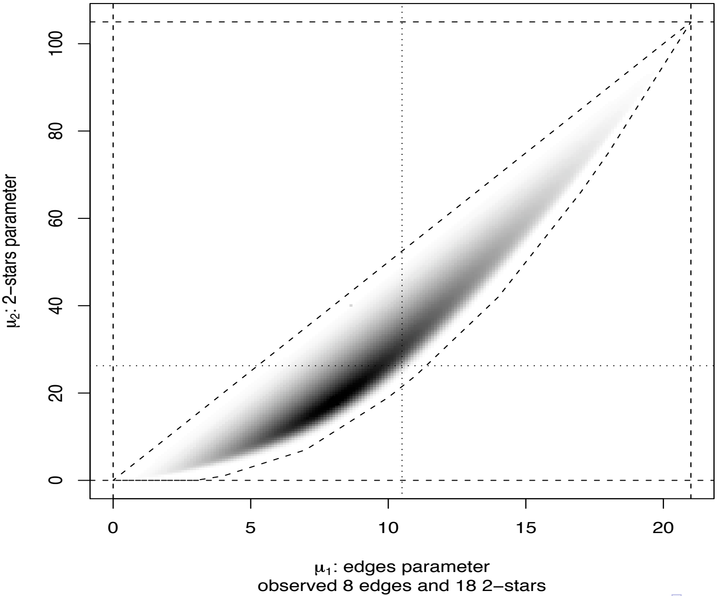
$$\mathbf{E}(\mu | Y = y; \mu_0, \gamma) = \frac{g(y) + \gamma\mu_0}{1 + \gamma}$$

Estimation

Under (component-wise) squared error loss in μ, the posterior mean is optimal.


Prior for μ with γ=0.05
μ2: 2−stars parameter
0
20
40
60
80
100
0
5
10
15
20
μ1: edges parameter
observed 8 edges and 18 2−stars

Posterior for μ with γ=0.05



Non-degeneracy prior

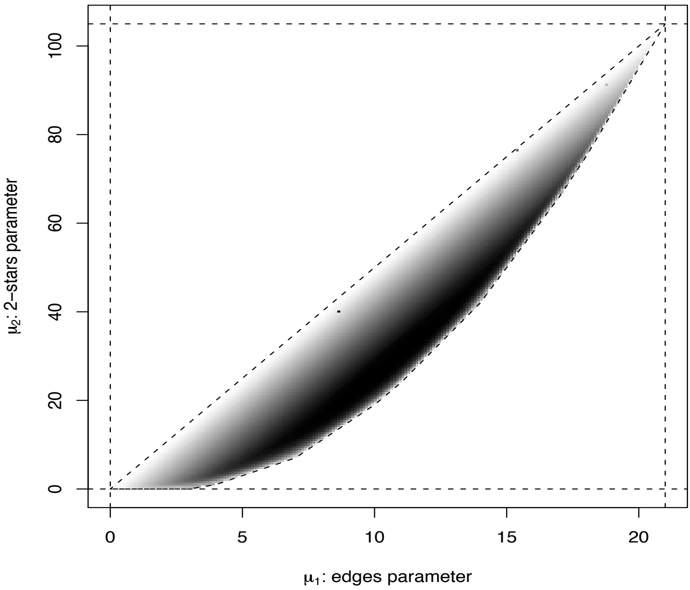
Define the *non-degeneracy prior*

$$Pr(\eta) \propto P_\eta(Y \in \mathrm{int(C)})$$

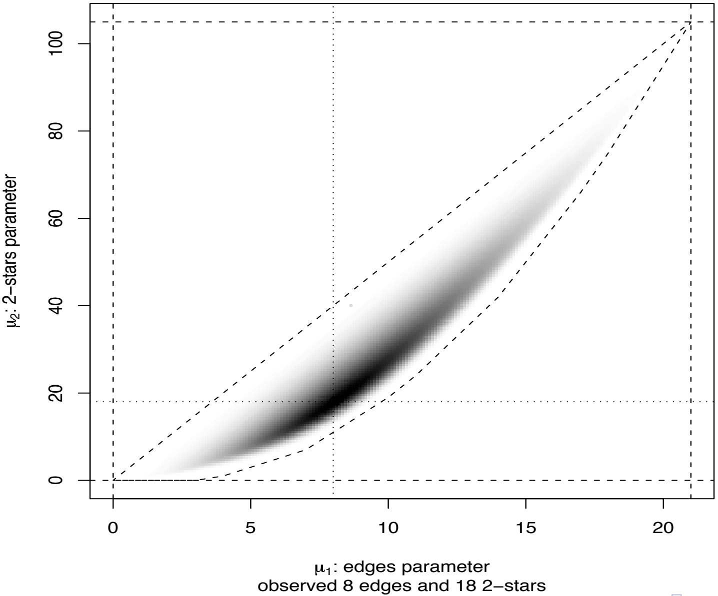
$\eta \in \Lambda$

– a natural “reference prior" for random network models

Figure 7: Non−degeneracy Prior for μ



Non–degeneracy Posterior for μ



- Consider extending the exponential family to include the standard exponential families that form the faces of C.

  – The MLE is admissible as an estimator of μ under squared-error loss. $\Rightarrow$ Meeden, Geyer, et. al. (1998)

  – The MLE is the Bayes estimator of μ under the “non-degeneracy" prior distribution.

Implementation of Bayesian Posterior models

The Bayesian posterior of η has density

$$\pi(\eta|y) \propto \exp[\eta \cdot (\delta\mu_0 + g(y)) - (1+\delta)\kappa(\eta)].$$

To generate samples by a Metropolis-Hasting algorithm, we need to calculate a Metropolis-Hastings ratio:

$$H(\eta'|\eta) = \frac{\exp[\eta' \cdot (\delta\mu_0 + g(y))]/\exp((1+\delta)\kappa(\eta'))}{\exp[\eta \cdot (\delta\mu_0 + g(y))]/\exp((1+\delta)\kappa(\eta))} \frac{q(\eta|\eta')}{q(\eta'|\eta)}, \tag{1}$$

where $q(\eta'|\eta)$ is the proposal density. However, (1) contains intractable normalizing constant $\kappa(\eta)$, which needs to be approximated. A straightforward approach is to approximate $\kappa(\eta') - \kappa(\eta)$ by MCMC (Geyer and Thompson, 1992), but the computation will be extremely expensive.

Auxiliary variable approach

Moller et al. (2006) proposed an efficient MCMC algorithm based on auxiliary variables. The goal is to sample from a posterior density

$$\pi(\eta|y) \propto \pi(\eta)\exp(\eta g(y) - \kappa(\eta)).$$

- Suppose x is an auxiliary variable defined on the same state space as that of y. It has conditional density $f(x|\eta, y)$ and posterior density

$$p(\eta, x|y) \propto p(\eta, x, y) = f(x|\eta, y)\pi(\eta, y) = f(x|\eta, y)\pi(\eta)p(y|\eta).$$

- If (η, x) is the current state of the algorithm, propose first η' with density $p(\eta'|\eta, x)$ and next x' with density $p(x'|\eta', \eta, x)$. Here, we take the proposal density for auxiliary variable x' to be the same as likelihood, i.e.

$$p(x'|\eta', \eta, x) = p(x'|\eta') = \exp(\eta' g(x'))/\exp(\kappa(\eta')).$$

- The Metropolis-Hasting ratio becomes

$$
\begin{aligned}
H(\eta', x'|\eta, x) &= \frac{p(\eta', x'|y)}{p(\eta, x|y)} \frac{q(\eta, x|\eta', x')}{q(\eta', x'|\eta, x)} \\
&= \frac{f(x'|\eta', y)p(\eta', y)}{f(x|\eta, y)p(\eta, y)} \frac{p(x|\eta)p(\eta|\eta', x')}{p(x'|\eta')p(\eta'|\eta, x)} \\
&= \frac{f(x'|\eta', y)\pi(\eta') \exp(\eta' g(y))/ \exp(\kappa(\eta'))}{f(x|\eta, y)\pi(\eta) \exp(\eta g(y))/ \exp(\kappa(\eta))} \\
&\cdot \frac{\exp(\eta g(x))/ \exp(\kappa(\eta)) \cdot p(\eta|\eta', x')}{\exp(\eta' g(x'))/ \exp(\kappa(\eta')) \cdot p(\eta'|\eta, x)}
\end{aligned}
$$

- Finally, we have the M-H ratio as

$$
H(\eta', x'|\eta, x) = \frac{f(x'|\eta', y)\pi(\eta') \exp(\eta' g(y)) \exp(\eta g(x))p(\eta|\eta', x')}{f(x|\eta, y)\pi(\eta) \exp(\eta g(y)) \exp(\eta' g(x'))p(\eta'|\eta, x)} \tag{2}
$$

does not depend on normalizing constants.

Note that:

For simplicity, we can assume that

$$p(\eta'|\eta, x) = p(\eta'|\eta)$$

does not depend on x.

Appropriate auxiliary density $f(x|\eta, y)$ and proposal density $p(\eta'|\eta)$ must be chosen so that the algorithm has good mixing and convergence properties.

Application to ERGM with uniform prior

2-star ERGM

Likelihood: $p(y|\eta) = \exp(\eta g(y) - \kappa(\eta))$

Uniform prior: $\eta \in \Theta =[-1, 1]^2$.

Suppose η is the current state of the parameter, and η' is the proposal. The algorithm to sample from posterior is as follows:

1. Approximate conditional density by

$$f(x|\eta, y) = \exp[\widetilde{\eta} g(x) - \kappa(\widetilde{\eta})],$$

   where $\widetilde{\eta}$ is MPLE.
2. Sample proposals η' from Normal distribution with mean η, so that $p(\eta|\eta')/p(\eta'|\eta) = 1$. The standard deviations is adjustable.
3. Sample x' from $p(x'|\eta') = \exp(\eta' g(x') - \kappa(\eta')$ by M-H sampling.
4. The M-H ratio then reduces to

$$H(\eta', x'|\eta, x) = I[\eta' \in \Theta] \frac{\exp(\widetilde{\eta} g(x') + \eta' g(y) + \eta g(x))}{\exp(\widetilde{\eta} g(x) + \eta g(y) + \eta' g(x'))}.$$

5. Accept η' with probability $\min\{1, H(\eta', x'|\eta, x)\}$.

Laplace Approximations with Conjugate Priors

Basics

Let $a(\eta)$ be a known function of η e.g. $a(\eta) = \eta$, or $a(\eta) = \eta^2$.

We wish to compute the posterior mean of $a(\eta)$:

$$\mathbb{E}_\eta[a(\eta)|y] = \int_{\eta\in\Theta} a(\eta)p(\eta|y)d\eta.$$

The posterior distribution $\pi(\eta|y)$ is given by

$$\pi(\eta|y) = \frac{\exp\{\mu\eta - \kappa(\eta)\}}{\int_{\eta\in\Theta}\exp\{\mu\eta - \kappa(\eta)\}d\eta}.$$

where the posterior mean and effective degrees-of-freedom are.

$$\mu = \frac{\delta\mu_0 + g(y)}{1+\delta} \qquad \phi = 1 + \delta$$

Let $f(\eta|y) = \exp\{\mu\eta - \kappa(\eta)\}$. The posterior expectation can be written as

$$\mathbb{E}_\eta[a(\eta)|y] = \frac{\int a(\eta)f(\eta|y)d\eta}{\int f(\eta|y)d\eta}.$$

Define

$$\begin{aligned} h^*(\eta) &= \log a(\eta) + \log f(\eta|y), \\ h(\eta) &= \log f(\eta|y). \end{aligned}$$

The Laplace approximation to $\mathbb{E}[a(\eta)|y]$ has the form

$$\hat{\mathbb{E}}_\eta[a(\eta)|y] = \frac{a(\eta^*) f(\eta^*|y)| -\nabla^2 h^*(\eta^*)|^{-1/2}}{f(\hat{\eta}|y)| -\nabla^2 h(\hat{\eta})|^{-1/2}}, \tag{3}$$

where

$$\eta^* := \mathrm{argsup}_\eta h^*(\eta);$$
$$\hat{\eta} := \mathrm{argsup}_\eta h(\eta).$$

And $\nabla^2 h^*(\eta^*)$ and $\nabla^2 h(\hat{\eta})$ are the Hessian matrices of h^* evaluated at η^* and h evaluated at $\hat{\eta}$, respectively.

Likelihood approximation

We know that

$$h(\eta) = \log f(\eta|y) = \mu\eta - \kappa(\eta).$$

We can approximate the difference $r(\eta, \eta^0) = h(\eta) - h(\eta^0)$ by

$$\hat{r}_m(\eta, \eta^0) = \mu(\eta - \eta^0) - \log \big[\frac{1}{m}\sum_{i=1}^{m} \exp\{(\eta - \eta^0)g(Y_i)\big], \quad (4)$$

where $Y_1, \cdots, Y_m$ are $\mathrm{i.i.d}$ $P(Y = y; \eta^0)$.

Iterative algorithms to estimate modes

Due to the conjugacy properties, we can estimate the mode of $h(\eta)$ by stochastic approximation analogous to the Markov chain Monte Carlo maximum likelihood estimation (Hunter and Handcock) procedure we apply in likelihood distributions. The Hessian matrix can be approximated by the same procedure.

We derive that

$$\nabla h(\eta) = \phi[\mu - \frac{\partial \kappa(\eta)}{\partial \eta}].$$

So,

$$\begin{aligned} \nabla h(\eta) &= \phi[\mu - \mathbb{E}_\eta g(y)] \\ \nabla^2 h(\eta) &= -\phi \mathrm{Var}_\eta g(y) = -\phi I(\eta) \end{aligned}$$

where $I(\eta)$ is the Fisher information matrix.

Newton-Raphson algorithm yields

$$\begin{aligned} \eta^{(k+1)} &= \eta^{(k)} - \left[\nabla^2 h(\eta^{(k)})\right]^{-1} \nabla h(\eta^{(k)}) \\ &= \eta^{(k)} + \left[I(\eta^{(k)}\right]^{-1} \Big[\mu - \mathbb{E}_\eta g(y)\Big] \end{aligned}$$

Analogous to sampling distributions, the update step in iterative algorithm for posterior distribution becomes: Sample $y_1, \cdots, y_m$ i.i.d $P(Y = y; \eta^{(k)})$. Update as follows:

$$\eta^{(k+1)} = \eta^{(k)} + \left\{ \hat{I}(\eta^{(k)}) \right\}^{-1} \left[\mu - \sum_i \omega_i^{(k)} g_i \right], \tag{5}$$

where g_{obs} and g_i denote $g(y_{obs})$ and $g(y_i)$ respectively.

$$\omega_i^{(k)} = \frac{\exp\{[\eta^{(k)} - \eta^0] g_i\}}{\sum_{j=1}^m \exp\{[\eta^{(k)} - \eta^0] g_j\}}$$

(weight by inverse probability) and

$$\hat{I}(\eta^{(k)}) = \Big\{ \sum_{i=1}^m \omega_i^{(k)} g_i g_i^t - \big(\sum_{i=1}^m \omega_i^{(k)} g_i \big) \big(\sum_{i=1}^m \omega_i^{(k)} g_i \big)^t \Big\}$$

(approximated Fisher information matrix).

Assume $a(\eta) = \eta$.

$$\begin{aligned}\nabla h^*(\eta) &= \frac{1}{\eta} + \phi[\mu - \mathbb{E}_\eta g(y)] \\ \nabla^2 h^*(\eta) &= -\frac{1}{\eta^2} - \phi I(\eta)\end{aligned}$$

Then, the approximate Fisher scoring method is implemented as

$$\eta^{(k+1)} = \eta^{(k)} + \Big[\frac{1/\phi}{{\eta^{(k)}}^2} + \hat{I}(\eta^{(k)})\Big]^{-1}\Big[\frac{1/\phi}{\eta^{(k)}} + \mu - \sum_i \omega_i^{(k)} g_i\Big], \tag{6}$$

Final algorithm

- We estimate the mode of $h(\eta)$, $\hat{\eta}$ by equation (5).
- By the same MCMC samples, we estimate the mode of $h^*(\eta)$, η^* by equation (6).
- By equation (4), the Laplace approximation to $\mathbb{E}[a(\eta)|y]$ can be calculated as

$$\hat{\mathbb{E}}_\eta[a(\eta)|y] = a(\eta^*) \frac{\exp\left\{\phi\mu(\eta^* - \hat{\eta})\right\}}{\left[\frac{1}{m}\sum_{i=1}^m \exp\{(\eta^* - \hat{\eta})g(Y_i)\}\right]^\phi} \frac{|-\nabla^2 h^*(\eta^*)|^{-1/2}}{|-\nabla^2 h(\hat{\eta})|^{-1/2}}$$

In particular we can easily compute:

$$\hat{\mathbb{E}}_\eta[\eta|y] \quad \text{posterior mean}$$

$$\hat{\mathbb{SD}}_\eta[\eta|y] \quad \text{posterior standard deviation}$$

based on a single MCMC run (e.g., like that for the MC-MLE).