A start of Variational Methods for ERGM

Ranran Wang, UW

MURI-UCI April 24, 2009

Outline

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

Introduction to ERGM

Network Notations

- 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 high-order network characteristics.

Variational method

Exponential families and variational representations
Basics of exponential family:

$$\log[p(x;\theta)] = \langle \theta, \phi(x) \rangle - \kappa(\theta).$$

- Sufficient statistics: $\phi(x)$.
- Log-partition function: $\kappa(\theta) = \log \sum_{x \in \mathcal{X}} \exp\langle \theta, \phi(x) \rangle$.
- Mean value parametrization: $\mu \in \mathbb{R}^q := \mathbb{E}(\phi(x))$
- Mean value space (convex hull):

$$\mathcal{M} = \big\{ \mu \in \mathbb{R}^q | \ \exists p(\cdot) \ s.t. \ \sum_{\mathcal{X}} \phi(x) p(x) = \mu \big\}.$$

The log-partition function is smooth and convex in terms of θ.

Suppose $\theta = (\theta_\alpha, \theta_\beta, \cdots)$ and $\phi(x) = (\phi_\alpha(x), \phi_\beta(x), \cdots)$:

$$\frac{\partial \kappa}{\partial \theta_\alpha}(\theta) = \mathbb{E}[\phi_\alpha(x)] := \sum_{x \in \mathcal{X}} \phi_\alpha(x) p(x; \theta). \tag{1}$$

$$\frac{\partial \kappa}{\partial \theta_\alpha \partial \theta_\beta}(\theta) = \mathbb{E}[\phi_\alpha(x)\phi_\beta(x)] - \mathbb{E}[\phi_\alpha(x)]\mathbb{E}[\phi_\beta(x)]. \tag{2}$$

So, $\mu(\theta)$ can be reexpressed as

$$\mu(\theta) = \frac{\partial \kappa}{\partial \theta}(\theta)$$

and it has gradient

$$\frac{\partial^2 \kappa}{\partial \theta^T \partial \theta}(\theta).$$

(Barndorff-Nielson, 1978; Handcock, 2003; Wainwright and Jordan, 2003)

Exp: Ising model on graph $\mathcal{G}(V, E)$

$$\log p(x, \theta) = \{\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - \kappa(\theta)\}, \tag{3}$$

where:

- x_s, associated with $s \in V$ is a Bernoulli random variable;
- components x_s and x_t are allowed to interact directly only if s and t are joined by an edge in the graph.

The relevant mean parameters in this representation are as follows:

$$\mu_s = \mathbb{E}_\theta[x_s] = p(x_s = 1; \theta), \quad \mu_{st} = \mathbb{E}_\theta[x_s x_t] = p(x_s = 1, x_t = 1; \theta).$$

For each edge (s, t), the triplet $\{\mu_s, \mu_t, \mu_{st}\}$ uniquely determines a joint marginal $p(x_s, x_t; \mu)$ as follows:

$$p(x_s, x_t; \mu) = \begin{bmatrix} (1 + \mu_{st} - \mu_s - \mu_t) & (\mu_t - \mu_{st}) \\ (\mu_s - \mu_{st}) & \mu_{st} \end{bmatrix}.$$

To ensure the joint marginal, we impose non-negativity constraints on all four entries, as follows:

$$\begin{aligned} 1 + \mu_{st} - \mu_s - \mu_t &\geq 0 \\ \mu_{st} &\geq 0 \\ \mu_{s(/t)} - \mu_{st} &\geq 0 \end{aligned}$$

The inequalities above define $\mathcal{M}$.

Variational inference and mean value estimation

For any $\mu \in ri\mathcal{M}$ (ri: relative interior), we have following lower bound:

$$\kappa(\theta) = \sup_{\mu \in \mathcal{M}} \langle \theta, \mu \rangle - \kappa^*(\mu) \tag{4}$$

$$\begin{aligned}
\kappa(\theta) &= \log \sum_{x \in \mathcal{X}} \frac{\exp\{\langle \theta, \phi(x) \rangle\}}{p(x;\theta)} p(x;\theta) \\
&\geq \sum_{x \in \mathcal{X}} \log \left(\frac{\exp\{\langle \theta, \phi(x) \rangle\}}{p(x;\theta)}\right) p(x;\theta) \\
&= \sum_{x \in \mathcal{X}} \langle \theta, \phi(x) \rangle p(x;\theta) - \sum_{x \in \mathcal{X}} \log(p(x;\theta)) p(x;\theta) \\
&= \mathbb{E}\langle \theta, \phi(x) \rangle - \mathbb{E}[\log(p(x;\theta))] = \langle \theta, \mu \rangle - \kappa^*(\mu).
\end{aligned}$$

The inequality follows from Jensen's inequality, and the last equality follows from $\mathbb{E}(\phi(x)) = \mu$ and $\kappa^*(\mu) = \mathbb{E}[\log(p(x;\theta(\mu)))]$, the negative entropy of distribution $p(x;\theta)$.

Why variational method?

- Variational representation turns the problem of calculating intractable summation/integrals to optimization problem (finding lower bound of κ over $\mathcal{M}$).
- The problem of computing mean parameters can be solved simultaneously.

Two main difficulties:

- The constraint set $\mathcal{M}$ of realizable mean parameters is difficult to characterize in an explicit manner.
- $\kappa^*(\mu)$ is lack of explicit form and needs proper approximation.

Mean value estimation

- μ is obtained by solving the optimization problem in (4).
- However, the dual function κ^* lacks an explicit form in many cases.
- We restrict the choice of μ to a tractable subset $\mathcal{M}_t(H)$ of $\mathcal{M}(G)$, where H is the tractable subgraph of G. The lower bound in (4) will then be computable.
- The solution of the optimization problem

$$\sup_{\mu \in \mathcal{M}_t(H)} \{\langle \mu, \theta \rangle - \kappa_H^*(\mu)\}$$

  specifies optimal approximation $\tilde{\mu}_t$ of μ.
- The optimal $\tilde{\mu}_t$, in fact, minimizes the Kullback-Leibler divergence between the tractable $\mathcal{M}_t$ and the target constraint $\mathcal{M}$, and KL divergence between their natural parameter spaces as well.

Ising model on Graph: Approximation of κ^*

Exp: Ising model on Graph: Approximation of κ^*

Assume the tractable graph H_0 is fully disconnected, then the mean value parameter set is

$$\mathcal{M}_0(H_0) = \{(\mu_s, \mu_{st}) | 0 \leq \mu_s \leq 1, \mu_{st} = \mu_s \mu_t\}$$

Here, $\mu_s = p(x_s = 1)$ and $\mu_{st} = p(x_s = 1, x_t = 1) = \mu_s \mu_t$. So, the distribution on H_0 is fully factorizable.

Deriving from Bernoulli distribution,

$$\kappa^*_{H_0}(\mu) = \sum_{s \in V} [\mu_s \log \mu_s + (1 - \mu_s) \log(1 - \mu_s)].$$

By (4),

$$\kappa(\theta) = \max_{\{\mu_s\} \in [0,1]^n} \Big\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t - \sum_{s \in V} [\mu_s \log \mu_s + (1 - \mu_s) \log(1 - \mu_s)] \Big\}. \tag{5}$$

After taking gradient and setting it to zero, we have following updates for μ:

$$\text{logit}(\mu_s) \leftarrow \theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st} \mu_t. \tag{6}$$

Apply (6) iteratively (coordinate ascent) to each node until convergence is reached.

Applications to ERGM

Dependence Graph

- G_Y is a graph with m actors and $n = \frac{m(m-1)}{2}$ dyads
- Construct a dependence graph D_Y to describe the dependence structure of G_Y: $D_Y = \mathcal{G}(V(D), E(D))$.
  - Each dyad $(i, j), i < j$ on G is an actor on D.
  - Each actor $(ij) \in V(D)$ has a binary variable y_{ij}.
  - Each edge on D exists if (ij) and (kl) as actors on D_Y share a common value, i.e (ij) and (kl) as dyads on G_Y share a node.
- Frank and Strauss, 1986.

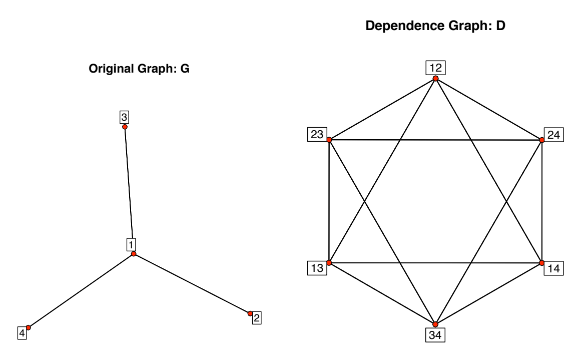

Figure 1: Dependence Graph D

Exp: Erdos-Renyi Model: For an undirected random graph $Y = \{Y_{ij}\}$, all dyads are mutually independent, so the dependency graph D is fully disconnected. Each $y_{ij}, (ij) \in D(V)$ is a Bernoulli random variable. The model can be written as

$$\log[P_\theta(Y = y)] = \sum_{i<j} \theta_{ij} y_{ij} - \kappa(\theta, \mathcal{Y}), \; y \in \mathcal{Y}.$$

Calculating entropy of Bernoulli distribution, we have

$$\kappa^*(\mu) = \sum_{i<j} [\mu_{ij} \log(\mu_{ij}) + (1 - \mu_{ij}) \log(1 - \mu_{ij})], \tag{7}$$

where $\mu_{ij} = P(Y_{ij} = 1)$. Then,

$$\kappa(\theta) \;=\; \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - \kappa^*(\mu)\} = \sum_{i<j} \log(1 + \exp(\theta_{ij})),$$

when $\theta_{ij} = \log(\frac{\mu_{ij}}{1-\mu_{ij}})$.

2-star ERGM model

Analogous to Ising model, on dependence graph $D = \mathcal{G}(V(D), E(D))$,

$$\log P(Y, \theta) = \sum_{s \in V(D)} \theta_s y_s + \sum_{(s,t) \in E(D)} \theta_{st} y_s y_t - \kappa(\theta),\ s : (ij) \in V(G).$$

If $\theta_s = \eta_1, s \in V$ and $\theta_{st} = \eta_2, (s,t) \in E$,

$$\log P(Y, \eta) = \{\eta_1 \sum_{i<j} y_{ij} + \eta_2 \sum_i \sum_{j,k>i} y_{ij} y_{ik} - \kappa(\eta)\},$$

which corresponds to the canonical 2-star model.

Given a graph G_Y with 6 actors and its dependency graph D_Y with 15 nodes.

For Ising model

$$\log p(x, \theta) = \{\sum_{s \in V_D} \theta_s y_s + \sum_{(s,t) \in E_D} \theta_{st} y_s y_t - \kappa(\theta)\},$$

Compare μ^{var} obtained from naive mean field algorithm to μ^{mcmc} obtained from MCMC samples for fixed θ's.

θ_{st} = 0.2, $\forall$ s,t

(ij):s	θ_s	μ_s^{mcmc}	μ_s^{var}
12	0.5	0.811	0.848
13	-0.5	0.666	0.671
14	0.5	0.852	0.848
15	-0.5	0.665	0.684
16	0.5	0.834	0.846
23	-0.5	0.671	0.671
24	0.5	0.831	0.848
25	-0.5	0.672	0.683
26	0.5	0.854	0.846
34	-0.5	0.672	0.671
35	0.5	0.855	0.837
36	-0.5	0.683	0.668
45	0.5	0.849	0.846
46	-0.5	0.672	0.683
56	0.0	0.737	0.772

For 2-star model, let $\theta_s = \eta_1 \in [-2, 2]$ and $\theta_{st} = \eta_2 \in [-2, 2]$. $\mu = P(x_s = 1), \forall s$.

Compare $\mu^{var}(\eta_1, \eta_2)$ with μ^{mcmc}.

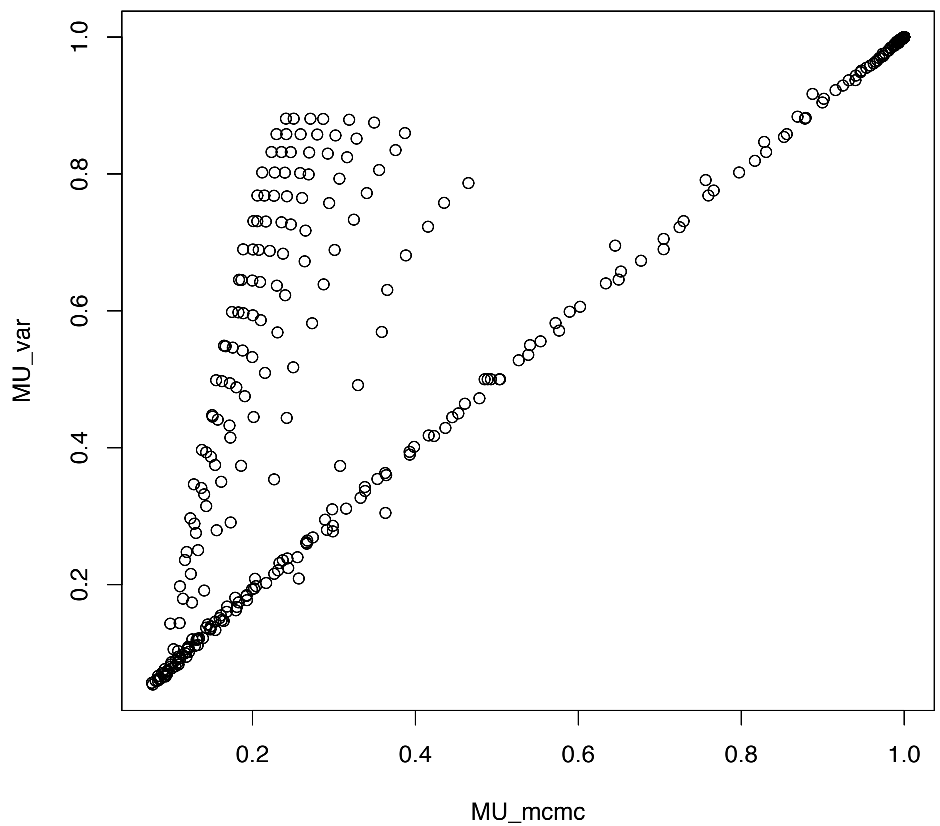

Figure 2: μ^{MCMC} vs. μ^{var}

Parameter estimation by variational inference

1. Start with $\theta^{(0)}$
2. Estimate $\widetilde{\mu}(\theta)$ from naive mean field algorithm
3. Calculate $\kappa(\theta) = \langle \theta, \widetilde{\mu} \rangle - \kappa^*(\widetilde{\mu})$ and log-likelihood $l(\theta, y)$. Also, calculate $\nabla \kappa(\theta) = \mathbb{E}_\theta(\phi(x))$ and $\nabla l(\theta, y) = \phi(x) - \mathbb{E}_\theta(\phi(x))$.
4. Update θ by gradient ascent:

$$\widetilde{\theta}^{(n+1)} = \widetilde{\theta}^{(n)} + \gamma \times \nabla l(\theta^{(n)}, y), \gamma \to 0.$$

5. Iterate until $\widetilde{\theta}$ converges.

Simulation study

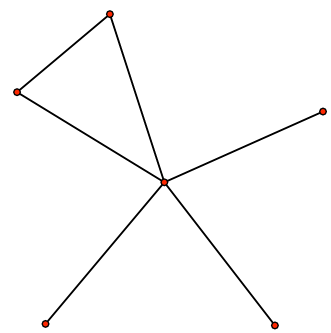

Figure 3: A sample graph with 6 edges and 12 2-stars

2-star ERGM	η_1	η_2
MLE	-1.69	0.39
MCMC-MLE	-1.74	0.40
MPLE	-7.54	2.18
Var-MLE	-1.99	0.465

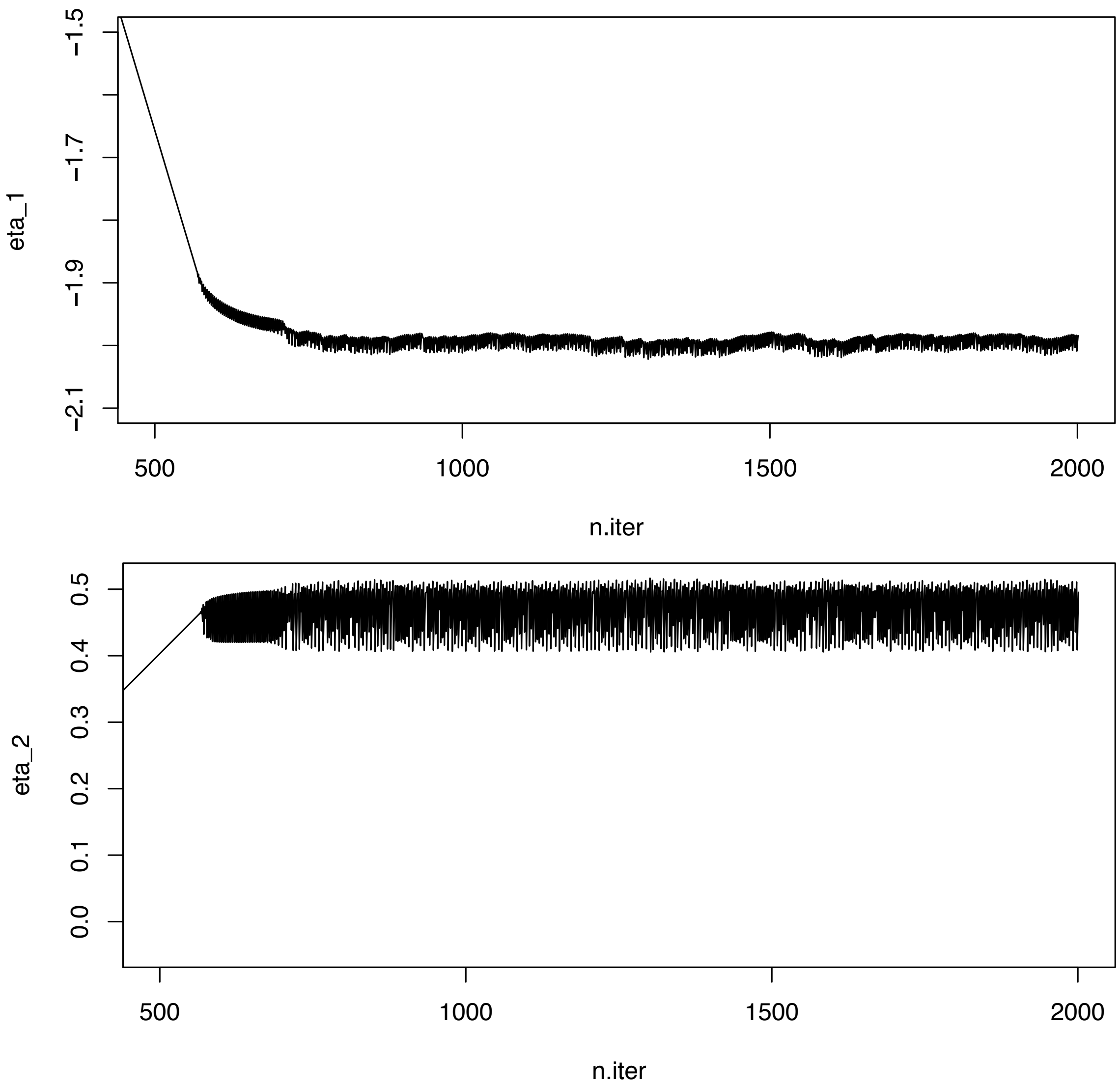

Figure 4: Convergence of Var-MLE

Discussion and Future work

Future work:

- Better approximation of A^*:
  - Structured mean field algorithm
  - Bethe entropy approximation
  - Clustered variational method
- Extension to general ERGM: clustering structure of dependence graph; constraint space
- Continuous graph: Gaussian random field
- Curved-exponential family
- Hybrid of MCMC and variational methods

Thanks for your attention!