Bias-Adjusted Maximum Likelihood Estimation

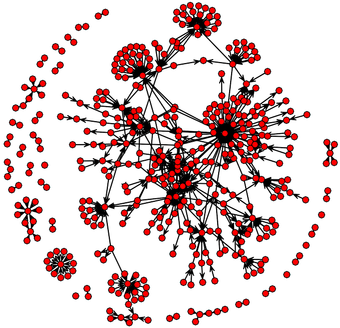
Improving Estimation for Exponential-Family Random Graph Models (ERGMs)

Ruth M Hummel
David R Hunter

Department of Statistics, Penn State University

MURI meeting, May 25, 2010

Motivation: Why model networks?



A statistical model for observed network data y^{obs} allows us to:

- Summarize: Give a parsimonious quantitative summary of the data and, ideally, how precisely we know this summary
- Predict: Describe or simulate other networks that could have arisen from the same process

Motivation: The likelihood function and MLE

The ERG model class:

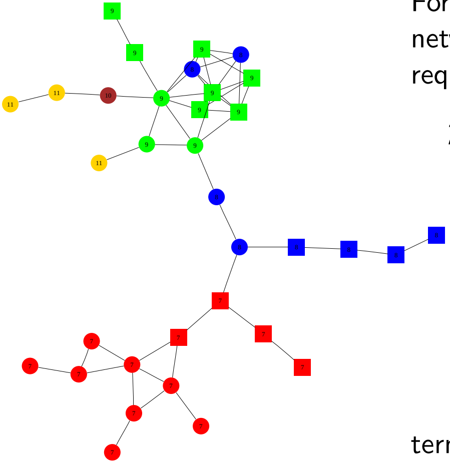
$$P_\theta(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}, \text{ where } \kappa(\theta) = \sum_{\substack{\text{all possible} \\ \text{graphs } z}} \exp\{\theta^t g(z)\}$$

- θ is a parameter vector to be estimated.
- $g(y)$ is a user-defined vector of graph statistics.
- The loglikelihood function is

$$\ell(\theta) = \theta^t g(y^{\text{obs}}) - \log \kappa(\theta).$$

- The MLE is the maximizer $\hat{\theta}$ of the likelihood.

The likelihood is sometimes intractable



For this undirected, 34-node network, computing $\ell(\theta)$ directly requires summation of

*7,547,924,849,643,082,704,483,
109,161,976,537,781,833,842,
440,832,880,856,752,412,600,
491,248,324,784,297,704,172,
253,450,355,317,535,082,936,
750,061,527,689,799,541,169,
259,849,585,265,122,868,502,
865,392,087,298,790,653,952*

terms.

The pseudolikelihood: A tractable alternative

- Some algebra based on the ERGM gives, for all $i \neq j$,

$$\log \frac{P(Y_{ij} = 1 \mid Y_{ij}^c)}{P(Y_{ij} = 0 \mid Y_{ij}^c)} = \theta^t \left[g(Y_{ij}^+) - g(Y_{ij}^-) \right].$$

- The pseudolikelihood ignores the conditioning, assuming instead

$$\log \frac{P(Y_{ij} = 1)}{P(Y_{ij} = 0)} = \theta^t \left[g(Y_{ij}^+) - g(Y_{ij}^-) \right] \equiv \theta^t \delta(Y)_{ij}$$

independently for all $i \neq j$.

- Thus, the pseudolikelihood equals

$$\prod_{i \neq j} \frac{\exp\left\{\theta^t \delta(y^{\text{obs}})_{ij}\right\}^{y_{ij}^{\text{obs}}}}{1 + \exp\left\{\theta^t \delta(y^{\text{obs}})_{ij}\right\}}$$

Evidence of bias in MPLE compared to MLE

Van Duijn, Gile, and Handcock (2009, *Social Networks*) compare MLE to MPLE.

- They cite a small but compelling set of explorations of the MPLE, suggesting that there may be large differences between the MPLE and the approximate MLE, sometimes even in cases where the dependence is not thought to be a concern.
- They explore the bias in the MLE and MPLE compared to the "truth"
- They introduce a bias-corrected version of the MPLE (the "MBLE").
- A similar bias-correction is possible for the MLE, though it is a bit less straightforward.

bias-correction via Firth

The bias-correction we employ (which might be better described as a preemptive bias-*mitigation*, rather than correction) follows from Firth (1993). The idea is to maximize a penalized likelihood which induces a bias in the score function in order to reverse the some of the anticipated bias in the maximizer. The penalized likelihood is:

$$\ell_{bc}(\theta) = \ell(\theta) + 1/2 \log |I(\theta)|$$

The resulting maximizer is also the Bayesian maximum posterior estimator based on assigning a Jeffreys prior to the parameter.

The intuition behind this modification for an exponential family model is the following: Since the score function, $U(\eta)$, can be written

$$U(\eta) = \ell'(\eta) = g(Y) - \kappa'(\eta),$$

it is clear that the *shape* of $U(\eta)$ is not affected by the sufficient statistic, $g(Y)$. For this reason, any anticipated bias in the MLE can be offset by shifting the score function by the amount $bias * \nabla U$. (Here $\nabla U = -i(\eta)$.) This adjustment is illustrated in the following figure, taken from Firth (1993):

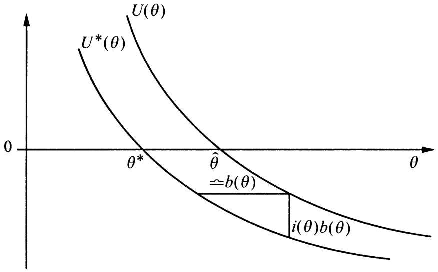
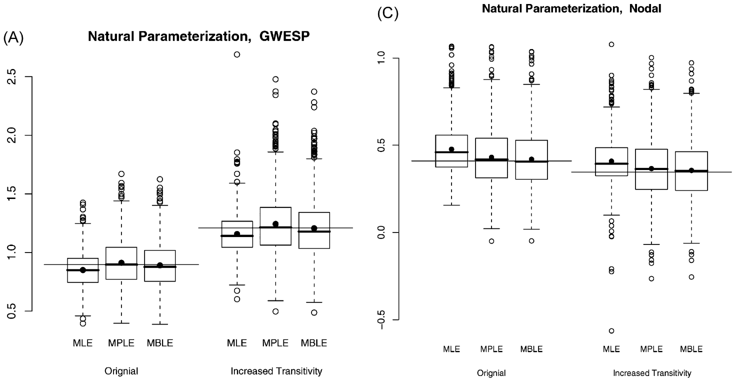


Figure: Modification of the unbiased score function

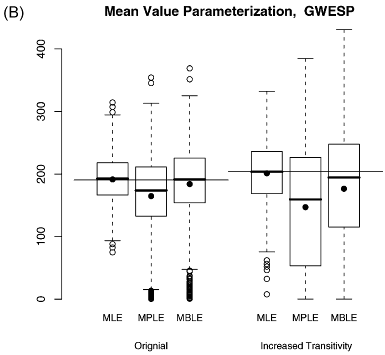
Evidence of bias in MLE (and MPLE) compared to “truth”

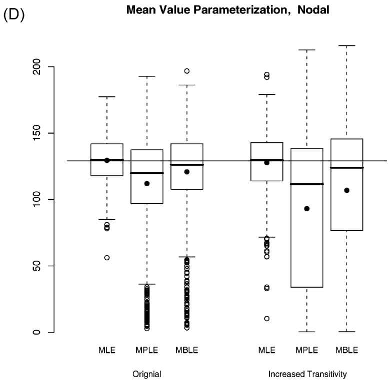
Taken from van Duijn, et al. (2009), these boxplots show the bias of the MLE for selected parameters in two networks (“original” and “transitivity”) for the canonical parameter space. (The true parameter is shown as a horizontal line.) Note that the bias is greatest in the MLE.



Evidence of bias in MLE (and MPLE) compared to "truth"

Here we see that there is no bias of the MLE for selected parameters in two networks ("original" and "transitivity") for the mean value parameter space. (This is by definition, since the mean-value MLE **is** the observed statistic.)





Comparison on Lazega collaboration network

In order to compare our present extended results to the results found for just the MBLE and the ordinary MPLE and MLE in the van Duijn, et al. paper, we duplicate their results on the corporate lawyer partnerships data and include the analysis for the bias-corrected MLE (pMLE).

Lazega collaboration network

The Lazega collaboration data are collaborations in the late 1980's between 36 New England lawyers determined by their responses to the question *"With which members of your firm have you spent time together on at least one case, have you been assigned to the same case, have they read or used your work product or have you have read or used their work product?"*

Additional member attributes collected include the attorneys' **gender**, **age**, **status** (36 are partners; 35 are associates), **seniority**, **years with the firm**, **practice** (litigation or corporate), **office location** (Boston, Hartford, or Providence), and **law school attended** (Yale or Harvard, University of Connecticut, or any other).

Following van Duijn, et al., we simulate networks based on a "truth" for the following model:

Model terms	"True" parameter value
edges	-6.506
GWESP	0.897
seniority (nodal covariate)	0.853
practice (nodal covariate)	0.410
practice (homophily effect)	0.759
gender (homophily effect)	0.702
office (homophily effect)	1.145

Preliminary results:

Results based on very few simulations show no improvement in the MLE yet...

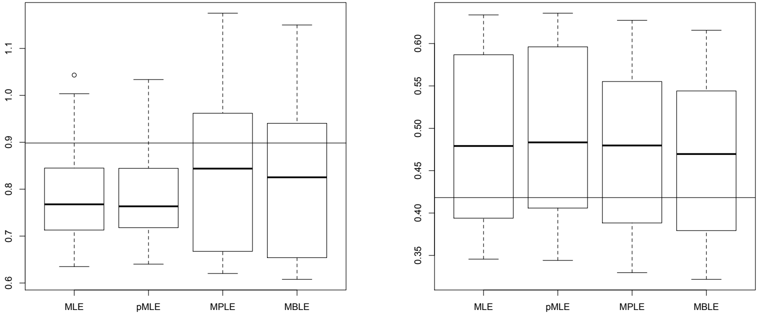


Figure: Distribution of the GWESP and Nodal Practice canonical parameter; true parameter shown as horizontal line.

Preliminary results:

Here you can see that the number of sub-simulations for calculating the mean value parameter is clearly not sufficient, as the mean for the uncorrected MLE should be unbiased...

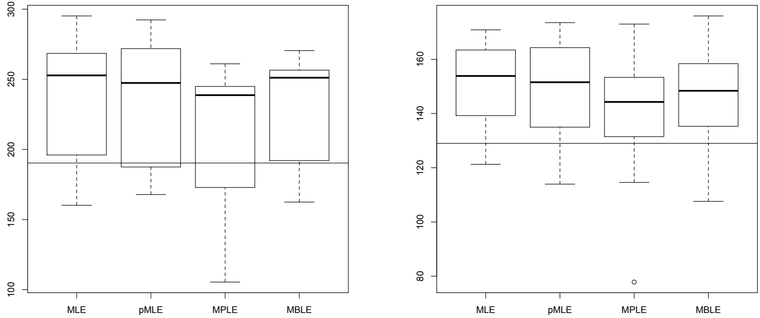


Figure: Distribution of the GWESP and Nodal Practice mean value parameter; true parameter shown as horizontal line.

Current extensions:

- increasing the simulations for the current network
- applying the same to the "increased transitivity" version of the collaboration network as used in van Duijn, et al.
- applying the same to a larger biological network
- applying the same to a friendship network

A few words about Contrastive Divergence (CD)

Consider the idea of MCMC MLE:

- Suppose we fix η_0. A bit of algebra shows that

$$-\log E_{\eta_0}\left[\exp\left\{(\eta-\eta_0)^t g(Y)\right\}\right] = \ell(\eta) - \ell(\eta_0). \qquad (1)$$

- The Law of Large Numbers suggests obtaining a sample of Y from the model using θ_0 as the parameter, then approximating the expectation by a sample mean.
- Q: How do we sample from $g(Y)$ using θ_0 as the parameter?
  A: Run MCMC infinitely long.

A few words about Contrastive Divergence (CD)

Consider the idea of MCMC MLE:

- Suppose we fix η_0. A bit of algebra shows that

$$-\log E_{\eta_0}\left[\exp\left\{(\eta-\eta_0)^t g(Y)\right\}\right] = \ell(\eta) - \ell(\eta_0). \qquad (1)$$

- The Law of Large Numbers suggests obtaining a sample of Y from the model using θ_0 as the parameter, then approximating the expectation by a sample mean.
- Q: How do we sample from $g(Y)$ using θ_0 as the parameter?
  A: Run MCMC infinitely long.
- But what if we only run MCMC for a single step (starting at y^{obs}), for a randomly chosen Y_{ij}?
- For this Y_{ij}, we're sampling from the conditional distribution given $(y^{\text{obs}})^c_{ij}$.

A few words about Contrastive Divergence (CD)

To summarize:

- Running an infinitely long Markov chain leads to the loglikelihood.
- Running a 1-step Markov chain leads to the pseudolikelihood.

Thus, if we alternately sample and then optimize the resulting "likelihood-like" function, we can view MLE and MPLE as two ends of a spectrum, the "contrastive divergence" spectrum. (MLE is CD-∞ and MPLE is CD-1.)

A few words about Contrastive Divergence (CD)

Considering CD-1. . .

Q: Is it better to

1. Repeatedly pick $i \neq j$ at random, or
2. Cycle through all possible $i \neq j$ in some systematic fashion?

A: The latter. The reason boils down to the following well-known identity for any two random variables Y and Z:

$$\mathrm{Var}(Y) = \mathrm{Var}\left[E(Y \mid Z)\right] + E\left[\mathrm{Var}(Y \mid Z)\right].$$

Here, "Y" is the likelihood-like quantity based on the randomly sampled networks and "Z" is the selected pair $i \neq j$.