Network structure from rich but noisy data

OxWaSP Module 1: Computational Statistics and Statistical Computing

Maud Lemercier and Natalia Garcia Martin

19 October 2018

Overview

Motivation

- Study of social networks
- Noisy data : measured interactions $\neq$ actual interactions
- $P(\theta|\text{data}) \propto \sum_A P(\text{data}|A, \theta_Y)P(A|\theta_A)P(\theta)$
- Introduce α = TP rate and β = FP rate

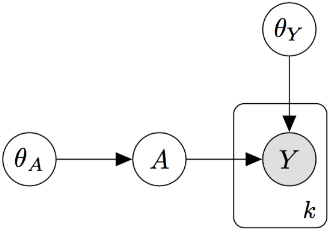

FIGURE – Graphical model of noisy data.

Network models

Bernoulli (Erdös Rényi) random network

- $\rho =$ probability of edge formation
- $\forall i, j \in [n], i < j, a_{ij} \overset{iid}{\sim} \mathrm{Bern}(\rho)$
- $\ell(A|p) = \prod_{i<j} \rho^{a_{ij}} (1-\rho)^{1-a_{ij}}$
- $d_v \sim \mathrm{Binom}(n-1, \rho)$

Exponential random graph models (ERGM)

- $P(A|\theta) = \frac{1}{Z(\theta)} \exp\left(\sum_{i=1}^{m} \theta_i T_i(a)\right) = \frac{1}{Z(\theta)} \exp\left[\theta^t T(a)\right]$
- Bernoulli graph : 1-dimensional case with $\theta = \frac{\rho}{1-\rho}$ and $T(a) = \sum_{i<j} a_{ij}$

FIGURE – (Left) A graph generated with the Erdös Rény model (Right) A business network among 16 Florentine families.

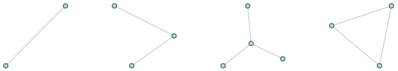

FIGURE – Example of edge, 2-star, 3-star and triangle.

Challenges

- Noisy data :
  - Inference with latent variables
  - Complicates MAP and ML computations
- Approaches :
  - The EM algorithm
  - Aggregate data from repeated observations
- ERGM models :
  - Intractable normalising constant $Z(\theta)$ due to large number of possible networks
- Approaches :
  - Introducing a well-chosen auxiliary variable in Metropolis-Hastings algorithms

Algorithms

Algorithm 1 Expectation Maximization algorithm

Given (α, β, ρ) from the previous iteration

1. Expectation step

$$Q_{i,j} \leftarrow \frac{\rho\alpha^{E_{i,j}}(1-\alpha)^{k-E_{i,j}}}{\rho\alpha^{E_{i,j}}(1-\alpha)^{k-E_{i,j}}+(1-\rho)\beta^{E_{i,j}}(1-\beta)^{k-E_{i,j}}}$$

2. Maximization step

$$\alpha \leftarrow \frac{\sum_{i<j} E_{i,j}Q_{i,j}}{k\sum_{i<j} Q_{i,j}}$$

$$\beta \leftarrow \frac{\sum_{i<j} E_{i,j}(1-Q_{i,j})}{k\sum_{i<j}(1-Q_{i,j})}$$

$$\rho \leftarrow \frac{1}{\binom{n}{2}} \sum_{i<j} Q_{i,j}$$

Algorithms

Algorithm 2 Exchange algorithm

Given $\theta_n \in \Theta$ at the n^{th} iteration
1. Propose $\theta' \sim q(\cdot|\theta_n)$
2. Generate the auxiliary variable $u \sim \frac{h(\cdot|\theta')}{Z(\theta')}$
3. Accept θ_{n+1} with probability
$\alpha = min\left\{1, \frac{p(\theta')h(x|\theta')h(u|\theta_n)q(\theta_n|\theta')}{p(\theta_n)h(x|\theta_n)h(u|\theta')q(\theta'|\theta_n)}\right\}$
Reject otherwise

Algorithm 3 Double Metropolis-Hastings algorithm

Given $\theta_n \in \Theta$ at the n^{th} iteration
1. Propose $\theta' \sim q(\cdot|\theta_n)$
2. Generate the auxiliary variable using m MH-updates $u \sim T^m_{\theta'}(.|x)$
3. Accept θ_{n+1} with probability
$\alpha = min\left\{1, \frac{p(\theta')T^m_{\theta'}(x|u)h(u|\theta_n)q(\theta_n|\theta')}{p(\theta_n)h(x|\theta_n)T^m_{\theta'}(u|x)q(\theta'|\theta_n)}\right\}$
Reject otherwise

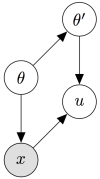

FIGURE – Augmented model.

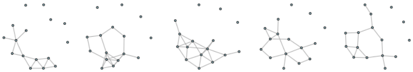

FIGURE – Example of 4 Gibbs updates.

Experiments and results

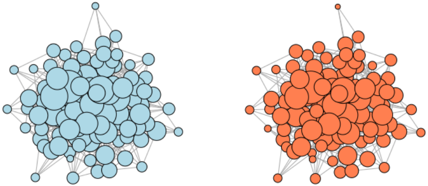

FIGURE – (Left) Ground truth underlying network (Right) Inferred network using the EM algorithm·

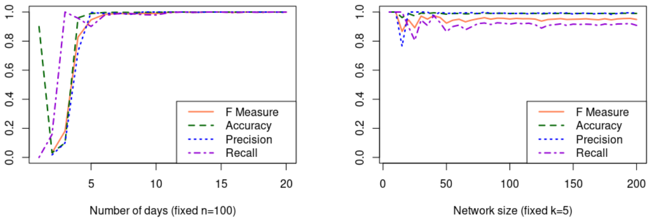

FIGURE – (Left) Performance metrics versus the number of repeated observations (Right) Performance versus network size.

Experiments and results

Dataset

- A Florentine Business Network (ergm R package)
- Graph with 16 nodes
- Sufficient statistics $S(x) = \{S_1(x), S_2(x), S_3(x), S_4(x)\}$ representing edges, 2-stars, 3-stars and triangles

Double Metropolis-Hastings algorithm

- 30000 iterations
- 10 cycles of Gibbs updates
- Random initialisation
- Uniform priors $\mathcal{U}(-5, 5)$
- Random walk scale $\sigma = 0.05$

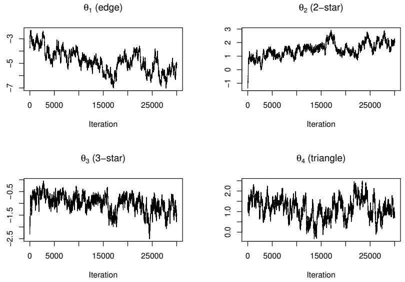

FIGURE – Parameter trace plots for the coefficients of the ERGM model.

Experiments and results

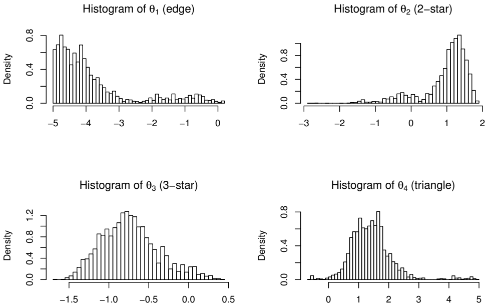

FIGURE – Histograms for the coefficients of the ERGM model.

Results

Parameter	Mean	SD
θ_1	-4.80	0.07
θ_2	1.48	0.04
θ_3	-0.98	0.02
θ_4	1.16	0.03

TABLE – Post. means and standard deviations

Gold standard

Parameter	Mean	SD
θ_1	-4.39	0.007
θ_2	1.25	0.004
θ_3	-0.84	0.002
θ_4	1.22	0.004

TABLE – Post. means and standard deviations

Future work and conclusion

- Quantify the scalability of the algorithms with the number of nodes
- Generalise bayesian inference methods to ERGM models with noisy data