Accelerating Bayesian estimation for network Poisson models using frequentist variational estimates

Joint work with S. Robin

Sophie Donnet. INRAE

Journée en l'honneur de Eric Parent

Outline

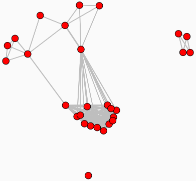
- Aim: Bayesian inference in a latent variable model

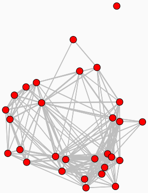
$$p(\theta, Z|Y)$$

- For the model of interest : SBM-Poisson avec covariables
  - easy to find variational frequentist estimators
  - Variational Bayes not easy to find
- Idea
  - Build a proxy posterior distribution from the variational frequentist estimation
  - Use this proxy to sample more efficiently from the true posterior distribution

Equid social networks [RSF+15]

Interactions between all pairs of individuals recorded during several days (44 for the 28 zebras and 82 for the 29 onagers).





Equid social networks

Data at hand.

- $Y = (Y_{ij})_{1 \leq i,j, \leq n} = n \times n$ matrix. Y_{ij}: interaction strength between individual i and j
- x_{ij} = vector of covariates for the pair (i, j)

Equid social networks

Data at hand.

- $Y = (Y_{ij})_{1\leq i,j,\leq n} = n \times n$ matrix. Y_{ij}: interaction strength between individual i and j
- $x_{ij} =$ vector of covariates for the pair (i, j)

Three binary variables indicating whether the two individuals share

- the same sex (x^1),
- in the same age category (x^2)
- the same status (x^3)
  - Onagers: T: territorial male, N: non-lactating, L:lactating

Equid social networks

Data at hand.

- $Y = (Y_{ij})_{1 \leq i,j, \leq n} = n \times n$ matrix. Y_{ij}: interaction strength between individual i and j
- x_{ij} = vector of covariates for the pair (i, j)

Three binary variables indicating whether the two individuals share

- the same sex (x^1),
- in the same age category (x^2)
- the same status (x^3)
    - Onagers: T: territorial male, N: non-lactating, L:lactating

Question of interest

- Does the status of the individuals contributes to shape the interaction network?
- Are the two networks structured by the same attributes?

Stochastic block model

SBM. Very popular tool for network analysis [HL79, NS01]

Principle. Model-based node clustering:

- Generative probabilistic model
- Introduce heterogeneity in the “social” behavior

Stochastic Block model

Data at hand.

- $Y = (Y_{ij})_{1\leq i,j,\leq n} = n \times n$ matrix. Y_{ij}: interaction strength between individual i and j
- x_{ij} = vector of covariates for the pair (i, j)

SBM with K groups

- $\forall i$, $Z_i = k$ if node i belongs to cluster k. $(Z_i)_i$ i.i.d.

$$\mathbb{P}(Z_i = k) = \nu_k$$

- $(Y_{ij})_{1\leq i,j\leq n}$ = conditionally independent :

$$Y_{ij} \mid Z_i = k, Z_j = \ell \sim \mathcal{P}(\exp\{\alpha_{kl} + x_{ij}^\intercal \beta\}).$$

$\rightarrow$ Node clusters are independent from covariates' effect

Bayesian inference

Prior distribution

$$\gamma = (\alpha, \beta) \sim \mathcal{N}(\gamma_0, V_0)$$
$$\nu \sim \mathcal{D}ir(e_{01}, \cdots e_{0K})$$

$\rightarrow$ Get $Z, \theta | Y$?

A mixed strategy

1. Deriving an approximation of the posterior distribution $\widetilde{p}(\theta, Z)$ from a frequentist variational maximum likelihood estimate

   $\rightarrow$ VEM-based proxy

2. Designing an efficient MC algorithm to sample from $p(\theta, Z \mid Y)$ taking advantage of $\widetilde{p}(\theta, Z)$

A VEM-based proxy for $p(\theta, Z \mid Y)$

Construction of a proxy

$$\widetilde{p}_Y(Z, \theta) := \widetilde{q}(Z)\widetilde{p}_Y(\theta) = \widetilde{q}(Z)\widetilde{p}_Y(\nu)\widetilde{p}_Y(\gamma).$$

- use VEM to get $\widetilde{q}(Z)$
- use a Laplace approximation of the "variational bound" to get $\widetilde{p}_Y(\theta)$

$\widetilde{q}(Z)$ from a variational frequentist estimation

Principle

Maximisation of the likelihood $\log p_\theta(Y)$ replaced by maximisation of the lower bound

$$\begin{aligned} J(Y;\theta,\widetilde{q}) &= \log p_\theta(Y) - \mathrm{KL}(\widetilde{q}(Z) \,||\, p_\theta(Z \mid Y)) \\ &= \mathbb{E}_{\widetilde{q}}[\log p_\theta(Y,Z)] + \mathcal{H}\left(\widetilde{q}(Z)\right) \end{aligned}$$

where $\widetilde{q}$ factorizable: $\widetilde{q}(Z) = \prod_i \widetilde{q}_i(Z_i)$

Output

- $(\widetilde{\theta}, \widetilde{q}) := \arg\max_{\theta,q} J(Y;\theta,q)$
- $\widetilde{q}(Z) \approx p(Z|Y,\widetilde{\theta})$

A VEM- Laplace based proxy for $p(\theta \mid Y)$

- Laplace approximation: popular approximation of the posterior distribution
- Taylor expansion of the log-likelihood $\log p_\theta(Y)$
- Unavailable in our model $\rightarrow$ replace it with the lower bound

$$
\begin{aligned}
p(\theta \mid Y) &\propto \exp\left(\log \pi(\theta) + \log p_\theta(Y)\right) \\
&\simeq \exp\left(\log \pi(\theta) + J(Y;\theta,\widetilde{q})\right) \\
&\propto \exp\left(\log \pi(\theta) + \frac{1}{2}(\theta - \widetilde{\theta})^\intercal \left(\partial^2_{\theta^2} J(Y;\widetilde{\theta},\widetilde{q})\right)(\theta - \widetilde{\theta})\right),
\end{aligned}
$$

Expression of the proxy for $p(\theta \mid Y)$

- For $\gamma = (\alpha, \beta)$
  - Gaussian prior distribution on γ
  - Laplace approximation
  - $$\widetilde{p}(\gamma) := \mathcal{N}\left(\left(V_0^{-1} + \widetilde{V}_Y^{-1}\right)^{-1}\left(V_0^{-1}\gamma_0 + \widetilde{V}_Y^{-1}\widetilde{\gamma}\right), \left(V_0^{-1} + \widetilde{V}_Y^{-1}\right)^{-1}\right).$$
- For ν
  - VEM algorithm provides an estimate of the number of nodes belonging to each class k: $\widetilde{N}_k := \sum_i \widetilde{\tau}_{ik}$
  - Conjugacy properties of the Dirichlet distribution
  - $$\widetilde{p}_Y(\nu) := \mathcal{D}(e_0 + \widetilde{e}), \qquad \text{where} \quad \widetilde{e} = (\widetilde{N}_k)_{1 \le k \le K}.$$

A few remarks

- $\widetilde{p}_Y$ combines $\pi(\theta)$ and Y.
- Posterior dependence between the components of γ represented.
- $\widetilde{p}_Y$ neglects the probabilistic dependence involving Z.
- Computational cost of the computation $\widetilde{p}_Y$ reduces to VEM
- $\widetilde{p}_Y$ easily to simulate $+$ density function explicit expression.

$\rightarrow$ $\widetilde{p}_Y$ not a satisfactory approximation of $p(\theta, Z \mid Y)$ but will be used to drastically accelerate the posterior sampling of the true posterior distribution p_Y.

Sequential Monte Carlo sampling

Aim : Generate a sample $(\theta^m, Z^m)_{m=1,\ldots,M}$ from $p(\theta, Z \mid Y)$.

Naive Importance sampling.

- Use $\widetilde{p}(\theta, Z)$ to sample directly from the posterior

$\rightarrow$ Poor effective sample size (ESS): few particles with non-zero weight

Sequential Monte Carlo [DDJ06].

$$\begin{tikzcd} q_0 \arrow[r, bend left] & q_{h-1} \arrow[r, bend left] & q_h \arrow[r, bend left] & p(\theta, Z \mid Y) \end{tikzcd}$$

- Define a sequence of distributions $(q_h)_{0 \leq h \leq H}$
- Sequentially sample: $S_h = (\theta^{h,m}, Z^{h,m})_{1 \leq m \leq M}$ from q_h using S_{h-1}

Proposed path sampling scheme

Set $0 = \rho_0 < \rho_1 < \cdots < \rho_{H-1} < \rho_H = 1$,

Standard distribution path

$$\begin{tikzcd}
\pi(\theta, Z) \arrow[r, bend left] & \pi(\theta, Z)p(Y|\theta, Z)^{\rho_h} \arrow[r, bend left] & p(\theta, Z \mid Y)
\end{tikzcd}$$

- Starts from the prior distribution
- Sequentially includes data (through the likelihood)

Proposed path sampling scheme

Set $0 = \rho_0 < \rho_1 < \cdots < \rho_{H-1} < \rho_H = 1$,

Our distribution path:

$$\begin{tikzcd} \widetilde{p}(\theta, Z) \arrow[r, bend left] & \widetilde{p}(\theta, Z)^{1-\rho_h} p(\theta, Z|Y)^{\rho_h} \arrow[r, bend left] & p(\theta, Z \mid Y) \end{tikzcd}$$

- Starts from the proxy
- Sequentially transforms the proxy into the true posterior

SMC [DDJ06]

Aim: At each step h, provides $\mathcal{E}_h = \{(U_h^m, w_h^m)\}_m$, weighted sample of q_h using $\mathcal{E}_{h-1}$.

At iteration h : 3 steps

- Moving the particles using a transition kernel,
- Re-weighting the particles : to correct the discrepancy between the sampling distribution and q_h (weights W_m)
- Selecting the particles: reduce the variability of the importance sampling weights and avoid degeneracy.

Theoretical justification: [DDJ06]. At each step h, construct a distribution for the whole particle path with marginal p_h.

Advantages

- Adaptative choice of the sequenceρ_h

Simulation study

- Illustrating the fact that our strategy drastically decreases the computational time with respect to a classical annealing-scheme (starting from the prior distribution)
- Equivalently, that $\widetilde{p}$ can be "corrected" into the true posterior distribution at a low computational cost.
- Remark : robustness of the sampling strategy with respect to the mis-specification of $\widetilde{p}_Y$ tested in a previous working paper.

Simulation design.

Simulate $S = 100$ networks similar to the datasets.

Analysis

1. Sample with a standard annealing scheme starting from prior [SMC from prior]
2. - Derive the proxy of the posterior distribution with R-package blockmodels [Leg16] + our approach
   - Sample with the presented strategy [SMC from approx]

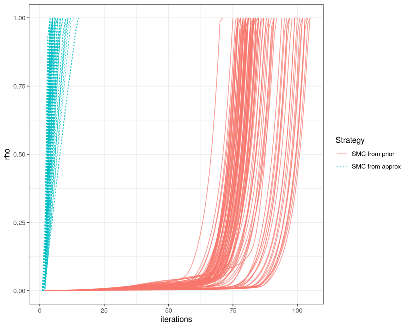
Implementation

- $M = 2000$ particles
- Codes written in R.

Computational time

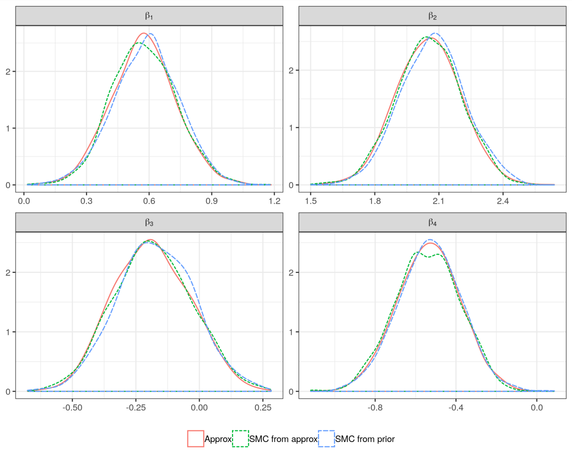
Compare the number of iterations in [SMC from approx] or [SMC from prior]

- Number of iterations = rough indicator of quality of the proxy



- In average [SMC from approx]: 15 times faster than [SMC from prior]
- Proxy: Less than 1 minute (including model selection)
- [SMC from approx]: 32 seconds

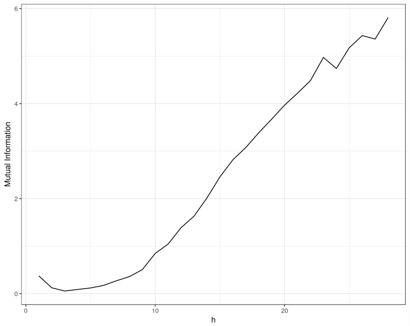
Marginal posterior distributions of the β's



- [SMC from prior] and [SMC from approx] are similar.
- $\widetilde{p}$ already a good approximation of the marginal true posterior

Mutual information

- SMC used to learn dependencies.
- $$\mathrm{MI}_h(Z) = KL\left[p_h(Z); \prod_{i=1}^{n} p_h(Z_i)\right].$$



Equid social networks: answers

Network analysis with covariates raises two typical questions

- The actual effect of each of these covariates on the structure of the network? $\rightarrow$ inference on the β

Answer

- The sex (x^1) is the only significant effect for the zebra network (model posterior probability $= 98.2\%$),
- Combination of the sex and the age (x^1, x^2) contributes to structure the onager network (model posterior probability $\simeq 100\%$)

- The existence of some residual structure in the network, once accounted for the effect of the covariates. $\rightarrow$ residual representation

  [LRO18]

Answer

- A remaining individual effect, not related to the sex or the age

Conclusion

Variational approximations.

- Efficient algorithms, reasonably easy to implement
- Good empirical behavior but few theoretical guaranties

Pragmatic point-of-view:

- Use V(B)EM as a first step for regular statistical inference
- Can be applied to any other way to approach the posterior (for instance max. of lik.)

Extensions

- Other (latent variable) models

References i

P. Del Moral, A. Doucet, and A. Jasra.

Sequential Monte Carlo samplers.

Journal of the Royal Statistical Society. Series B: Statistical Methodology, 68(3):411–436, 2006.

P.W. Holland and S. Leinhardt.

Structural sociometry.

Perspectives on Social Network Research, pages 63–83, 1979.

Jean-Benoist Leger.

Blockmodels: A R-package for estimating in latent block model and stochastic block model, with various probability functions, with or without covariates.

Technical report, arXiv:1602.07587, 2016.

P. Latouche, S. Robin, and S. Ouadah.

Goodness of fit of logistic regression models for random graphs.

Journal of Computational and Graphical Statistics, 27(1):98–109, 2018.

K. Nowicki and T.A.B. Snijders.

Estimation and prediction for stochastic block-structures.

Journal of the American Statistical Association, 96:1077–87, 2001.

D. I Rubenstein, S. R Sundaresan, I. R Fischhoff, C. Tantipathananandh, and T. Y Berger-Wolf.

Similar but different: dynamic social network analysis highlights fundamental differences between the fission-fusion societies of two equid species, the onager and Grevy's zebra.

PloS one, 10(10):e0138645, 2015.