

Delayed rejection algorithm to estimate Bayesian social networks

Algoritmo del rifiuto ritardato per la stima di reti sociali bayesiane

Alberto Caimo and Antonietta Mira

Abstract The Delayed rejection (DR) strategy is a modification of the Metropolis-Hastings (MH) algorithms that reduces the variance of the resulting Markov chain Monte Carlo estimators and allows partial adaptation of the proposal distribution. This strategy is exploited to estimate dyadic independence network models leading to an average 40% variance reduction relative to the competing MH algorithm, confirming that DR dominates, in terms of Peskun ordering, the MH algorithm.

Abstract *La strategia del rifiuto ritardato è una modifica degli algoritmi di Metropolis-Hastings (MH) che riduce la varianza degli stimatori Markov chain Monte Carlo e permette un parziale adattamento della “proposal distribution”. Questa strategia è adottata per la stima di modelli di rete con indipendenza diadica portando ad una riduzione media della varianza del 40% rispetto al concorrente algoritmo di MH e confermando quindi che il rifiuto ritardato migliora, in termini di ordinamento di Peskun, l'algoritmo MH.*

Key words: Delayed rejection, Peskun ordering, dyadic independence networks.

1 Dyadic independence network models

Statistical network models (see Salter-Townshend et al. (2012) for a recent review) are widely used in many scientific areas as they give the possibility to investigate how an observed network may be related to local relational structures. From a statistical viewpoint, networks are relational data represented as a graph consists of a set of n nodes and a set of m edges which define some sort of relations between pairs of

Alberto Caimo

Faculty of Economics, University of Lugano, Switzerland, e-mail: alberto.caimo@usi.ch

Antonietta Mira

Faculty of Economics, University of Lugano, Switzerland, e-mail: antonietta.mira@usi.ch

nodes called dyads. The network structure is described by an $n \times n$ adjacency matrix y with elements:

$$y_{ij} = \begin{cases} 1, & \text{if node } i \text{ and node } j \text{ are connected} \\ 0, & \text{otherwise.} \end{cases}$$

If $y_{ij} = y_{ji}$ then the adjacency matrix is symmetric and the graph is undirected, otherwise the graph is directed and it is often called digraph.

The interest in network modeling began with a series of papers by Gilbert (1959); Erdős and Rényi (1961) who set up the definition of *random graph* under the basic assumption that the presence of an edge between two nodes occurs with a fixed probability θ : The resulting likelihood is a binomial distribution and implies a uniform distribution over the sample space of possible graphs with exactly m edges.

The p_1 model proposed by Holland and Leinhardt (1981) is supposed to satisfy the dyadic independence property according to which dyads are assumed to be statistically independent. This model expresses the presence of three tendencies: the propensity with which a node i will be connected to j ; the propensity with which a node i will attract others; and the degree of reciprocated edges. The p_2 model (van Duijn et al. (2004)) is a random effects version of the p_1 which recognizes dependence between dyads with same nodes.

2 MCMC methods for Bayesian social networks

Bayesian methods are becoming increasingly popular as techniques for modeling social networks (Caimo and Friel (2011, 2013)). Following the Bayesian paradigm, prior distribution is assigned to θ . The posterior distribution of θ given the data y is:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}.$$

Direct evaluation of $p(\theta|y)$ requires the calculation of both the likelihood $p(y|\theta)$, which is computationally demanding if not intractable, and the marginal likelihood $p(y)$ which is typically intractable.

2.1 Delayed rejection strategy

Markov chain Monte Carlo (MCMC) algorithms are general simulation methods for sampling from posterior distributions and computing posterior quantities of interest.

Delayed rejection (DR) is a modification of the Metropolis-Hastings algorithm (MH) Tierney (1994); Green and Mira (2001); Mira (2001a) aimed at improving efficiency of the resulting MCMC estimators relative to Peskun (1973); Tierney (1998); Mira (2001b) asymptotic variance ordering. The basic idea is that, upon re-

jection in a MH, instead of advancing time and retaining the same position, a second stage move is proposed. The acceptance probability of the second stage candidate preserves reversibility of the Markov chain with respect to the target distribution of interest (the posterior, in a Bayesian setting). This delaying rejection mechanism can be iterated for a fixed or random number of stages. The higher stage proposal distributions are allowed to depend on the candidates so far proposed and rejected. Thus DR allows partial local adaptation of the proposal within each time step of the Markov chain still retaining reversibility and the Markovian property.

The advantage of DR over alternative ways of combining different MH proposals or kernels such as mixing and cycling Tierney (1994), is that a hierarchy between kernels can be exploited so that kernels that are easier to compute (in terms of CPU time) are tried first, thus saving in terms of simulation time. Or moves that are more “bold” (bigger variance of the proposal, for example) are tried at earlier stages thus allowing the sampler to explore the state space more efficiently following a sort of ‘first bold’ versus ‘second timid’ tennis-service strategy.

To simplify the notation indicate the posterior distribution of interest, $p(\theta|y)$, simply as $p(\theta)$, dropping the conditioning on y . Suppose the current position of the Markov chain is $X_t = \theta$. As in a regular MH, a candidate move θ_1 is generated from a proposal $q_1(\theta, \cdot)$ and accepted with probability

$$\alpha_1(\theta, \theta_1) = 1 \wedge \frac{p(\theta_1)q_1(\theta_1, \theta)}{p(\theta)q_1(\theta, \theta_1)} = 1 \wedge \frac{N_1}{D_1}.$$

Upon rejection, instead of retaining the same position, $X_{t+1} = \theta$, as we would do in a standard MH, a second stage move θ_2 is generated from a proposal distribution that is allowed to depend, not only on the current position of the chain, but also on what we have just proposed and rejected: $q_2(\theta, \theta_1, \cdot)$. The second stage acceptance probability is:

$$\alpha_2(\theta, \theta_1, \theta_2) = 1 \wedge \frac{p(\theta_2)q_1(\theta_2, \theta_1)q_2(\theta_2, \theta_1, \theta)[1 - \alpha_1(\theta_2, \theta_1)]}{p(\theta)q_1(\theta, \theta_1)q_2(\theta, \theta_1, \theta_2)[1 - \alpha_1(\theta, \theta_1)]} = 1 \wedge \frac{N_2}{D_2}.$$

This process of delaying rejection can be iterated and the i -th stage acceptance probability is, following Mira (2001a):

$$\alpha_i(\theta, \theta_1 \dots \theta_i) = 1 \wedge \frac{N_i}{D_i} = 1 \wedge \left\{ \frac{p(\theta_i)q_1(\theta_i, \theta_{i-1})q_2(\theta_i, \theta_{i-1}, \theta_{i-2}) \dots q_i(\theta_i, \theta_{i-1} \dots \theta)}{p(\theta)q_1(\theta, \theta_1)q_2(\theta, \theta_1, \theta_2) \dots q_i(\theta, \theta_1 \dots \theta_i)} \right. \\ \left. \frac{[1 - \alpha_1(\theta_i, \theta_{i-1})][1 - \alpha_2(\theta_i, \theta_{i-1}, \theta_{i-2})] \dots [1 - \alpha_{i-1}(\theta_i, \dots, \theta_1)]}{[1 - \alpha_1(\theta, \theta_1)][1 - \alpha_2(\theta, \theta_1, \theta_2)] \dots [1 - \alpha_{i-1}(\theta, \theta_1, \dots, \theta_{i-1})]} \right\}$$

If i -th stage is reached, it means that $N_j < D_j$ for $j = 1, \dots, i-1$, therefore $\alpha_j(\theta, \theta_1 \dots \theta_j)$ is simply N_j/D_j , $j = 1, \dots, i-1$ and a recursive formula can be obtained: $D_i = q_i(\theta \dots \theta_i)(D_{i-1} - N_{i-1})$ which leads to:

$$D_i = q_i(\theta \dots \theta_i)[q_{i-1}(\theta \dots \theta_{i-1})[q_{i-2}(\theta, \dots, \theta_{i-2}) \dots \\ [q_2(\theta, \theta_1, \theta_2)[q_1(\theta, \theta_1)p(\theta) - N_1] - N_2] - N_3] \dots - N_{i-1}].$$

Since reversibility with respect to p is preserved separately at each stage, the process of delaying rejection can be interrupted at any stage. The user can either decide, in advance, to try at most, a fixed number of times to move away from the current position or, alternatively, upon each rejection, toss a p -coin (i.e. a coin with head probability equal to p), and if the outcome is head move to a higher stage proposal, otherwise stay put.

Tierney and Mira (1999) prove that the DR strategy provides MCMC estimators with smaller asymptotic variance than standard MH. This better performance holds no matter what is the function f whose expectation relative to p we want to estimate (provided f is squared integrable with respect to p).

3 Example

One popular statistical network model is based exclusively on the node degrees. If y is the observed network graph with degree sequence $d_1, \dots, d_n$ the likelihood of this model can be written as:

$$p(y|\theta) = \frac{\exp\{\sum_i^n \theta^i d_i\}}{\prod_{i < j} 1 + \exp\{\theta^i + \theta^j\}}$$

The Zachary's karate club network (Figure 1, Zachary (1977)) will be used to compare MH and DR in terms of efficiency. Each iteration of the MH algorithm consists in generating θ_1 from some proposal distribution $q_1(\cdot)$. A block-update sampler with normal proposal is used to simultaneously update all the parameter of the posterior distribution. The DR algorithm that we propose here consists of two stages. The first stage proposal distribution is the same as the proposal used in the competing MH sampler and the first stage acceptance probability is:

$$\alpha_1(\theta, \theta_1) = 1 \wedge \frac{p(y|\theta_1)p(\theta_1)q_1(\theta, \theta_1)}{p(y|\theta)p(\theta)q_1(\theta_1, \theta)}$$

The second stage of the DR algorithm, consists in generating θ_2 from a different proposal distribution $q_2(\cdot)$. The second stage acceptance probability is defined as:

$$\alpha_2(\theta, \theta_1, \theta_2) = 1 \wedge \frac{p(y|\theta_2)p(\theta_2)q_1(\theta_1, \theta_2)q_2(\theta, \theta_1)[1 - \alpha_1(\theta_2, \theta_1)]}{p(y|\theta)p(\theta)q_1(\theta_1, \theta)q_2(\theta_2, \theta)[1 - \alpha_1(\theta, \theta_1)]}$$

The posterior estimates displayed in Table 1 were obtained using very flat normal prior for each parameter: $p(\theta^i) \sim \mathcal{N}(0, \sigma^2)$ where $\sigma^2 = 100$, a normal proposal distribution $q_1(\theta_1|\theta) \sim \mathcal{N}(\theta, \sigma_{h1}^2 I_n)$ where $\sigma_{h1}^2 = 0.055$ and I_n is the n -dimensional identity matrix, and 30,000 MCMC iterations. The second stage proposal is as a deterministic move in the opposite direction of the move at first stage so that: $\theta_2 = 2\theta - \theta_1$. We consider, as function of interest, the identity function, i.e. $f(\theta^i) = \theta^i, \forall i$

in other words, we focus on estimating the posterior mean of each parameter of interest. The efficiency ratios displayed in Table 1 are calculated as:

$$\frac{ESS_{DR}/ESS_{MH}}{TIME_{DR}/TIME_{MH}}$$

where ESS_{DR} and ESS_{MH} are the effective sample sizes (Kass et al. (1998)) obtained by the DR and MH respectively, and $TIME_{DR}$ and $TIME_{MH}$ are the CPU simulation times obtained by the DR and MH respectively. The results of Table 1 are consistent with previous results demonstrating the social prominence of the actors associated to node 1 and 34.

Table 1 Posterior mean (PM), posterior standard deviation (PSD) and efficiency ratio (ER) estimates for each parameter.

	θ^1	θ^2	θ^3	θ^4	θ^5	θ^6	θ^7	θ^8	θ^9	θ^{10}	θ^{11}	θ^{12}	θ^{13}	θ^{14}	θ^{15}	θ^{16}	θ^{17}
PM	1.42	0.13	0.37	-0.63	-1.71	-1.28	-1.28	-1.31	-0.89	-2.27	-1.78	-3.29	-2.44	-0.93	-2.35	-2.33	-2.37
PSD	0.42	0.47	0.45	0.53	0.67	0.60	0.65	0.62	0.59	0.81	0.74	1.11	0.88	0.57	0.81	0.82	0.87
ER	1.32	1.35	1.27	1.45	1.48	1.48	1.30	1.22	1.26	1.66	1.28	1.09	1.20	1.33	1.22	1.46	1.46
	θ^{18}	θ^{19}	θ^{20}	θ^{21}	θ^{22}	θ^{23}	θ^{24}	θ^{25}	θ^{26}	θ^{27}	θ^{28}	θ^{29}	θ^{30}	θ^{31}	θ^{32}	θ^{33}	θ^{34}
PM	-2.31	-2.31	-1.75	-2.28	-2.33	-2.39	-0.91	-1.81	-1.72	-2.31	-1.31	-1.77	-1.30	-1.31	-0.61	0.74	1.57
PSD	0.82	0.80	0.70	0.81	0.83	0.89	0.57	0.69	0.72	0.81	0.64	0.69	0.62	0.62	0.52	0.43	0.41
ER	1.30	1.69	1.29	1.68	1.50	1.09	1.34	1.45	1.54	1.61	1.50	1.32	1.60	1.28	1.46	1.34	1.36

4 Conclusions

The better performance of the delayed rejection strategy relative to the standard MH sampler is demonstrated in the setting of Bayesian social network models. A very simple deterministic second stage proposal is considered for the delayed rejection but more creative strategies could be elaborated. Performance comparison takes CPU simulation time into account and considers the effective sample size of the two

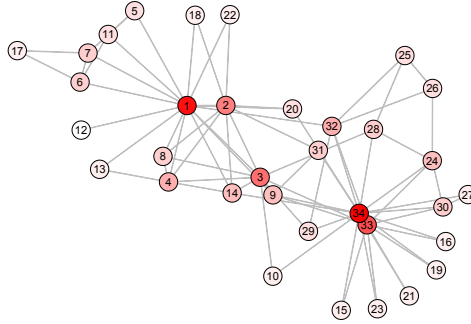


Fig. 1 Zachary karate club network graph: social network of friendships between 34 members of a karate club at a US university in the 1970. The more red the color of a node, the greater its degree.

competing samplers. For all the 34 model parameters the DR algorithm leads to a decrease of autocorrelation along the path of the simulated Markov chain which, in turn, translates into a smaller asymptotic variance and a higher effective sample size. Averaging over the different parameters, the efficiency ratio of the DR is higher than the one of the corresponding MH by approximately 40%.

References

- Caimo, A. and Friel, N. (2011), “Bayesian Inference for Exponential Random Graph Models,” *Social Networks*, 33, 41 – 55.
- (2013), “Bayesian model selection for exponential random graph models,” *Social Networks*, 35, 11 – 24.
- Erdős, P. and Rényi, A. (1961), “On the strength of connectedness of a random graph,” *Acta Mathematica Scientia Hungary*, 12, 261–267.
- Gilbert, E. N. (1959), “Random graphs,” *Annals of Mathematical Statistics*, 30, 1141–1144.
- Green, P. J. and Mira, A. (2001), “Delayed rejection in reversible jump Metropolis–Hastings,” *Biometrika*, 88, 1035–1053.
- Holland, P. W. and Leinhardt, S. (1981), “An exponential family of probability distributions for directed graphs (with discussion),” *Journal of the American Statistical Association*, 76, 33–65.
- Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998), “Markov chain monte carlo in practice: A roundtable discussion,” *The American Statistician*, 52, 93–100.
- Mira, A. (2001a), “On Metropolis-Hastings algorithms with delayed rejection,” *Metron*, 59, 231–241.
- (2001b), “Ordering and improving the performance of Monte Carlo Markov chains,” *Statistical Science*, 340–350.
- Peskun, P. (1973), “Optimum monte-carlo sampling using markov chains,” *Biometrika*, 60, 607–612.
- Salter-Townshend, M., White, A., Gollini, I., and Murphy, T. B. (2012), “Review of statistical network analysis: models, algorithms, and software,” *Statistical Analysis and Data Mining*, 5, 243–264.
- Tierney, L. (1994), “Markov chains for exploring posterior distributions,” *the Annals of Statistics*, 1701–1728.
- (1998), “A note on Metropolis-Hastings kernels for general state spaces,” *Annals of Applied Probability*, 1–9.
- Tierney, L. and Mira, A. (1999), “Some adaptive Monte Carlo methods for Bayesian inference,” *Statistics in medicine*, 18, 2507–2515.
- van Duijn, M. A., Snijders, T. A. B., and Zijlstra, B. H. (2004), “p2: a random effects model with covariates for directed graphs,” *Statistica Neerlandica*, 58, 234–254.
- Zachary, W. (1977), “An Information Flow Model for Conflict and Fission in Small Groups,” *Journal of Anthropological Research*, 33, 452–473.