Efficient MCMC for Gibbs Random Fields using Pre-Computation

Aidan Boland, Nial Friel

School of Mathematics and Statistics,
University College Dublin, Ireland

Introduction

Gibbs random fields (GRFs) are very difficult to work with due to the intractability of the normalising constant. Let $y = y_1, ..., y_N$ be realised data defined on a set of nodes $1, ..., N$ of a graph. The likelihood of a Gibbs random field given a vector of parameters $\theta = (\theta_1, ..., \theta_m)$ is

$$f(y|\theta) = \frac{\exp(\theta^T s(y))}{Z(\theta)} = \frac{q(y|\theta)}{Z(\theta)}, \tag{1}$$

where $s(y) = (s_1(y), ...s_m(y))$ is a vector of sufficient statistics. The normalising constant for the likelihood $z(\theta) = \sum_{y \in Y} \exp(\theta^T s(y))$ depends on θ, and is intractable for all but trivially small situations.

Bayesian Inference

The aim is to make inferences on the parameters θ through the posterior distribution

$$\pi(\theta|y) \propto \frac{q_\theta(y)}{Z(\theta)}\pi(\theta),$$

where $\pi(\theta)$ denotes the prior distribution of θ. Inference on the posterior distribution can be made by obtaining representative samples from this distribution of interest. One method to obtain samples is the MCMC method, the Metropolis-Hastings algorithm. The M-H algorithm works by first proposing a new value for θ and then accepting this new value with the probability:

$$\alpha(\theta, \theta') = \min\left(1, \frac{q_{\theta'}(y)\pi(\theta')h(\theta|\theta')}{q_\theta(y)\pi(\theta)h(\theta'|\theta)} \times \frac{Z(\theta)}{Z(\theta')}\right).$$

However for Gibbs random fields this is unworkable due to the presence of the normalising constants. One method to overcome this intractability is the exchange algorithm of Murray et al. [4], this algorithm works by introducing an auxiliary variable and sampling from an augmented distribution whose marginal distribution for θ is the posterior of interest.

1. Gibbs update
   - i Draw $\theta' \sim h(\theta,)$
   - ii Simulate $x \sim q(\cdot|\theta')$
2. Propose exchange from θ to θ' with probability:

$$\min\left(1, \frac{q(x|\theta)\pi(\theta')h(\theta|\theta')q(y|\theta')}{q(y|\theta)\pi(\theta)h(\theta'|\theta)q(x|\theta')} \times \frac{Z(\theta)Z(\theta')}{Z(\theta)Z(\theta')}\right)$$

In step 2, all of the intractable normalising constants cancel above and below in the fraction. This allows us to make inferences on Gibbs random fields, however the algorithm requires draws from the likelihood at each step. Simulating from the likelihood can be time consuming especially for large data.

Pre-Computation

Moores et al. [3] addressed the computational expense of repeated simulations for Gibbs random fields by using a pre-processing step to learn about the distribution of the summary statistics $(s(x))$ of simulated data, this method was applied using ABC. Moores et al. [2] extended this model to MCMC methods using path sampling to estimate the ratio of intractable normalising constants. However the path sampling method is only suitable for single parameter models.
The method presented here uses importance sampling (2) which enables the extension to multi-parameter models. First the grid values at which to pre-sample must be chosen, for GRF's this can be done using an estimation of the gradient to explore the posterior and focus the grid in areas of high probability.

$$\nabla \log \pi(\theta|y) \approx s(y) - \frac{1}{N}\sum_{i=1}^{N} s(x_i) + \nabla\pi(\theta).$$

For each grid value $\dot\theta_m \in (\dot\theta_1, ..., \dot\theta_M)$, N realisations, $(x_m^1, ..., x_m^N)$ are simulated from the likelihood function, $f(\cdot|\dot\theta_m)$.
It is straightforward to show by importance sampling that an unbiased estimator of $\frac{Z(\dot\theta_k)}{Z(\dot\theta_m)}$ is given by

$$\frac{Z(\dot\theta_k)}{Z(\dot\theta_m)} \approx \widehat{\frac{Z(\dot\theta_k)}{Z(\dot\theta_m)}} = \frac{1}{N}\sum_{n=1}^{N}\frac{q_{\dot\theta_k}(x_m^n)}{q_{\dot\theta_m}(x_m^n)} \tag{2}$$

Using the pre-computed data, any normalising ratio $\frac{Z(\theta)}{Z(\theta')}$ can be estimated. A path of pre-computed values is decided, $(\dot\theta_{t(1)}, ..., \dot\theta_{t(C)})$, where $t(1) = \arg\min_s ||\theta - \dot\theta_s||$ and $t(C) = \arg\min_s ||\theta' - \dot\theta_s||$.

Pre-computation (contd.)

The normalising ratio is then approximated using the Equation (3), where the ratios at both ends are calculated using the pre-computed data and importance sampling.

$$\frac{Z(\theta)}{Z(\theta')} \approx \widehat{\frac{Z(\theta)}{Z(\dot\theta_{t(1)})}}\,\widehat{\frac{Z(\dot\theta_{t(1)})}{Z(\dot\theta_{t(2)})}} \cdots \widehat{\frac{Z(\dot\theta_{t(C-1)})}{Z(\dot\theta_{t(C)})}}\left(\widehat{\frac{Z(\theta')}{Z(\dot\theta_{t(C)})}}\right)^{-1} \tag{3}$$

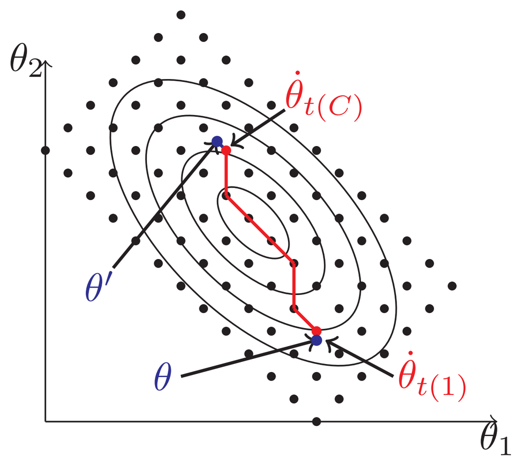

Figure 1: Shows an example of a path between two parameter vectors in two-dimensional space. The solid lines represent the target distribution and the black dots represent the pre-computed parameter values. The red line shows the path. note that there are multiple possible paths between the two vectors.

Noisy Approximate Exchange

1. Estimate normalising ratio using pre-computed data.
   - (a) Draw $\theta' \sim h(\theta,)$
   - (b) Find the closest pre-computed parameter for both θ and θ'
     - i. $k = \arg\min_s ||\theta - \dot\theta_s||$
     - ii. $m = \arg\min_s ||\theta' - \dot\theta_s||$
   - (c) Calculate $\widehat{\frac{Z(\theta)}{Z(\theta')}}$ using pre-computed data (3).
2. Propose exchange from θ to θ' with probability:

$$\hat\alpha(\theta, \theta', \mathbf{x}) = \min\left(1, \frac{q_{\theta'}(y)\pi(\theta')h(\theta|\theta')}{q_\theta(y)\pi(\theta)h(\theta'|\theta)} \times \widehat{\frac{Z(\theta)}{Z(\theta')}}\right)$$

Theoretical Convergence

Theoretical convergence is available for the noisy algorithm, some assumptions are first made.

- There is a constant c_π such that $1/c_\pi \le \pi(\theta) \le c_\pi$.
- There is a constant c_h such that $1/c_h \le h(\theta'|\theta) \le c_h$.

When the above two assumptions are satisfied we have that Θ is a bounded set, in this case we let $T = \sup_{\theta\in\Theta}||\theta||$ and let $\psi = \sup||s(\cdot)||$. This means that $0 \le \exp(-T\psi) \le q_\theta(x) \le \exp(T\psi)$ for any θ and $s(x)$. It follows that for any θ, θ' and x,

$$\frac{q_\theta(x)}{q_{\theta'}(x)} = \exp\{(\theta - \theta')s(x))\} \le \exp\{(2T)\psi\} = K_1.$$

For two neighbouring pre-computed points, $\dot\theta_k$ and $\dot\theta_m$, $|\dot\theta_k - \dot\theta_m| = \epsilon$. It follows that for any two neighbouring grid points $\dot\theta_k$, $\dot\theta_m$ and any dataset x,

$$\frac{q_{\dot\theta_k}(x)}{q_{\dot\theta_m}(x)} = \exp\left\{(\dot\theta_k - \dot\theta_m)s(x))\right\} \le \exp\{\epsilon\psi\} = K_2.$$

It also holds that $|\theta - \dot\theta_m| < \epsilon$, where $\dot\theta_m = \arg\min_s ||\theta - s||$ and so for any θ, $\dot\theta_m$ and x,

$$\frac{q_\theta(x)}{q_{\dot\theta_m}(x)} \le \exp\{(\theta - \dot\theta_m)s(x)\} \le \exp\{\epsilon\psi\} = K_2.$$

Using the above inequalities a bound can be placed between the true acceptance probability and the noisy acceptance probability,

$$\mathbb{E}_{\mathbf{x}}|\hat\alpha(\theta, \theta', \mathbf{x}) - \alpha(\theta, \theta')| \le \frac{c_\pi^2 c_h^2 K_1^4 K_2^2}{\sqrt{N}}\left(\frac{1}{\sqrt{N}} + K_2^4\right).$$

This allows the use of the theory of Alquier et al. [1] to place a bound between the true transition kernel P and approximate transition kernel $\hat P$.

$$\sup_{\theta_0\in\Theta}||\delta_{\theta_0}P^n - \delta_0\hat P^n|| \le \frac{D(c_\pi, c_h, K_1, K_2)}{\sqrt{N}}\left(\frac{1}{\sqrt{N}} + K_2^4\right),$$

where $D(c_\pi, c_h, K) = c_\pi^2 c_h^2 K_1^4 K_2^2\left(\lambda - \frac{B\rho^\lambda}{1-\rho}\right)$, is explicitly known, with $\lambda = \left(\frac{\log(1/2)}{\log(\rho)}\right)$.

Autologistic Example

The autologistic model is a GRF model for spatial binary data. The likelihood of the autologistic model is given by,

$$\begin{aligned} f(y|\theta) &\propto \exp(\theta^T s(y)) \\ &= \exp(\theta_1 s_1(y) + \theta_2 s_2(y)), \end{aligned}$$

where $s_1(y) = \sum_{i=1}^N y_i$ and $s_2(y) = \sum_{i\sim j} y_i y_j$ with $i \sim j$ denoting node i and node j are neighbours. θ_1 controls the relative abundance of -1 and $+1$ values while θ_2 controls the level of spatial aggregation.

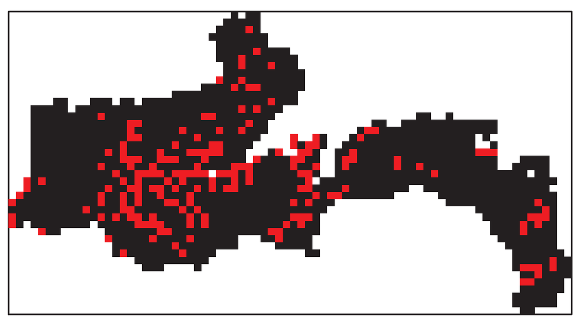

Figure 2: Presence (red) and absence (black) of red deer in the Grampian region of Scotland.

We implement the autologistic model using red deer census data, presence or absence of deer by 1km square in the Grampian region of Scotland. Figure 2 shows the observed data, a red square indicates the presence of deer, while a black square indicates the absence of deer.

Results

A long long run of the exchange algorithm (4 hours) was used as a 'ground truth'. The pre-computing grid points (Figure 3 (left)) were chosen using the method described in pre-computing Section. A total of 124 parameter values were chosen as the values to pre-sample from. It took just over 45 seconds to choose the grid and calculate ratios for all pairs of parameter values. Table 1 shows the parameter estimates, and Figure 3 (right) shows the estimated total variation over time between the two algorithms and the long run of the exchange algorithm. The noisy exchange algorithm outperforms the exchange algorithm as it converges much quicker. The table shows that the noisy exchange algorithm result in more accurate mean and variance parameter estimates when compared to the exchange algorithm when run for the same amount of time.

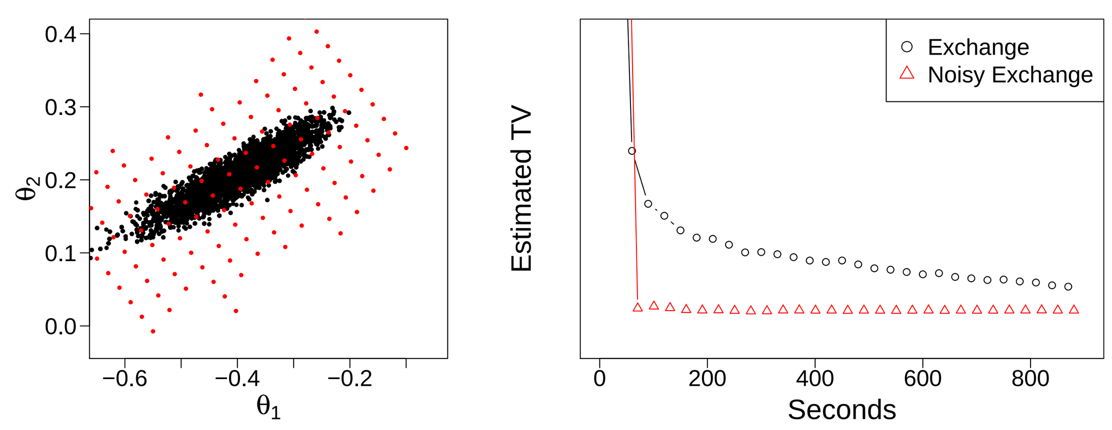

Figure 3: Grid for pre-sampling (left) and estimated total variation over time (right). The plot on the right shows that when the estimate total variation distance between the algorithms and the long exchange is compared, the noisy exchange algorithm outperforms the exchange algorithm.

Table 1: Posterior means and variances for the deer data. The table shows that the mean and variance estimates of the noisy exchange are closer to the 'ground truth' long exchange run.

	θ_1		θ_2	
	Mean	Var	Mean	Var
Exchange (long)	-0.39343	0.00467	0.20843	0.00095
Exchange	-0.38527	0.00441	0.21166	0.00090
Noisy Exchange	-0.39067	0.00449	0.20866	0.00089

Funding

SimSci is Funded under the programme for Research in Third-level Institutions and co-funded under the European Regional Development fund

References

[1] P. Alquier, N. Friel, R. Everitt, and A. Boland. Noisy Monte Carlo: convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 21(6):1–19, 2014.

[2] Matthew T Moores, Anthony N Pettitt, and Kerrie Mengersen. Scalable bayesian inference for the inverse temperature of a hidden potts model. *arXiv preprint arXiv:1503.08066*, 2015.

[3] MatthewT. Moores, ChristopherC. Drovandi, Kerrie Mengersen, and ChristianP. Robert. Pre-processing for approximate bayesian computation in image analysis. *Statistics and Computing*, 25(1):23–33, 2015. ISSN 0960-3174. doi: 10.1007/s11222-014-9525-6.

[4] I. Murray, Z. Ghahramani, and D. MacKay. Mcmc for doubly-intractable distributions. *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia, AUAI Press, 2006.

Investing In Your Future