# ABC methods for model choice in Gibbs random fields

Jean-Michel Marin

Institut de Mathématiques et Modélisation
Université Montpellier 2

Joint work with Aude Grelaud, Christian Robert, François Rodolphe,
Jean-François Taly

We consider a finite set of sites $\mathcal{S} = \{1, \cdots, n\}$.

At each site $i \in \mathcal{S}$, we observe $x_i \in \mathcal{X}_i$ where $\mathcal{X}_i$ is a finite set of states.

We also consider an undirected graph $\mathcal{G}$: the sites $i$ and $i^{'}$ are said neighbours, if there is a vertex between $i$ and $i^{'}$.

A clique $c$ is a subset of $\mathcal{S}$ where all elements are mutual neighbours (Daroch, 1980).

We denote by $\mathcal{C}$ the set of all cliques of the undirected graph $\mathcal{G}$.

Gibbs Random Fields (GRFs) are probabilistic models associated with densities

$$f(\mathbf{x}) = \frac{1}{Z}\exp\{-U(\mathbf{x})\} = \frac{1}{Z}\exp\left\{-\sum_{c\in\mathcal{C}} U_c(\mathbf{x})\right\},$$

where $U(\mathbf{x}) = \sum_{c\in\mathcal{C}} U_c(\mathbf{x})$ is the potential and $Z$ is the corresponding normalising constant

$$Z = \sum_{\mathbf{x}\in\mathcal{X}} \exp\left\{-\sum_{c\in\mathcal{C}} U_c(\mathbf{x})\right\}.$$

If the density $f$ of a Markov Random Field (MRF) is everywhere positive, then the Hammersley-Clifford theorem establishes that there exists a GRF representation of this MRF (Besag, 1974).

We consider here GRF with potential $U(\mathbf{x}) = -\boldsymbol{\theta}^{\mathrm{T}} S(\mathbf{x})$ where $\boldsymbol{\theta} \in \mathbb{R}^p$ is a scale parameter, $S(\cdot)$ is a function taking values in $\mathbb{R}^p$.

$S(\mathbf{x})$ is defined on the cliques of the neighbourhood system in that $S(\mathbf{x}) = \sum_{c \in \mathcal{C}} S_c(\mathbf{x})$.

In that case, we have

$$f(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z_{\boldsymbol{\theta}}} \exp\{\boldsymbol{\theta}^{\mathrm{T}} S(\mathbf{x})\}\,,$$

the normalising constant $Z_{\boldsymbol{\theta}}$ now depends on the scale parameter $\boldsymbol{\theta}$.

GRF are used to model the dependency within spatially correlated data, with applications in epidemiology and image analysis, among others (Rue and Held, 2005).

They often use a Potts model defined by a sufficient statistic $S$ taking values in $\mathbb{R}$ in that

$$S(\mathbf{x}) = \sum_{i' \sim i} \mathbb{I}_{\{x_i = x_{i'}\}} \, ,$$

where $\sum_{i' \sim i}$ indicates that the summation is taken over all the neighbour pairs.

$\mathcal{X}_i = \{1, \cdots, K\}$, $K = 2$ corresponding to the Ising model, and $\theta$ is a scalar.

$S(\cdot)$ monitors the number of identical neighbours over $\mathcal{X}$.

In most realistic settings, the summation

$$Z_{\boldsymbol{\theta}} = \sum_{\mathbf{x}\in\mathcal{X}} \exp\{\boldsymbol{\theta}^{\mathrm{T}} S(\mathbf{x})\}$$

involves too many terms to be manageable.

Selecting a model with sufficient statistic $S_0$ versus a model with sufficient statistics $S_1$ relies on the Bayes factor

$$BF_{m_0/m_1}(\mathbf{x}) = \int \exp\{\boldsymbol{\theta}_0^{\mathrm{T}} S_0(\mathbf{x})\}/Z_{\boldsymbol{\theta}_0,0}\pi_0(\mathrm{d}\boldsymbol{\theta}_0) \Big/ \int \exp\{\boldsymbol{\theta}_1^{\mathrm{T}} S_1(\mathbf{x})\}/Z_{\boldsymbol{\theta}_1,1}\pi_1(\mathrm{d}\boldsymbol{\theta}_1)$$

This quantity is not easily computable.

For a fixed neighbourhood or model, the unavailability of $Z_{\boldsymbol{\theta}}$ complicates inference on the scale parameter $\boldsymbol{\theta}$.

The difficulty is increased manifold when several neighbourhood structures are under comparison.

We propose a procedure based on an ABC algorithm aimed at selecting a model.

We consider the toy example of an iid sequence [with trivial neighbourhood structure] tested against a Markov chain model [with nearest neighbour structure].

In a model choice perspective, we face $M$ Gibbs random fields in competition.

Each model $m$ is associated with sufficient statistic $S_m$ $(0 \leq m \leq M-1)$, i.e. with corresponding likelihood

$$f_m(\mathbf{x}|\theta_m) = \exp\left\{\theta_m^{\mathrm{T}} S_m(\mathbf{x})\right\} / Z_{\theta_m,m}\,,$$

where $\theta_m \in \Theta_m$ and $Z_{\theta_m,m}$ is the unknown normalising constant.

The choice between those models is driven by the posterior probabilities of the models.

We consider an extended parameter space $\Theta = \cup_{m=0}^{M-1} \{m\} \times \Theta_m$ that includes the model index $\mathcal{M}$,

We define a prior distribution on the model index $\pi(\mathcal{M} = m)$ as well as a prior distribution on the parameter conditional on the value $m$ of the model index, $\pi_m(\theta_m)$, defined on the parameter space $\Theta_m$.

The computational target is thus the model posterior probability

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) \propto \int_{\Theta_m} f_m(\mathbf{x}|\theta_m)\pi_m(\theta_m)\,\mathrm{d}\theta_m\,\pi(\mathcal{M} = m)\,,$$

the marginal of the posterior distribution on $(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})$ given $\mathbf{x}$.

If $S(\mathbf{x})$ is a sufficient statistic for the joint parameters $(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})$,

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) = \mathbb{P}(\mathcal{M} = m|S(\mathbf{x}))\,.$$

Each model has its own sufficient statistic $S_m(\cdot)$.

Then, for each model, the vector of statistics $S(\cdot) = (S_0(\cdot), \ldots, S_{M-1}(\cdot))$ is obviously sufficient.

We have shown that the statistic $S(\mathbf{x})$ is also sufficient for the joint parameters $(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})$.

That the concatenation of the sufficient statistics of each model is also a sufficient statistic for the joint parameters is a property that is specific to Gibbs random field models.

When we consider $M$ models from generic exponential families, this property of the concatenated sufficient statistic rarely holds.

## ABC algorithm for model choice (ABC-MC)

1. Generate $m^*$ from the prior $\pi(\mathcal{M} = m)$.
2. Generate $\theta^*_{m^*}$ from the prior $\pi_{m^*}(\cdot)$.
3. Generate $\mathbf{x}^*$ from the model $f_{m^*}(\cdot|\theta^*_{m^*})$.
4. Compute the distance $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*))$.
5. Accept $(\theta^*_{m^*}, m^*)$ if $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) \leq \epsilon$, otherwise, start again in 1.

Simulating a data set $\mathbf{x}^*$ from $f_{m^*}(\cdot|\theta^*_{m^*})$ at step 3 is non-trivial for GRFs (Møller and Waagepetersen, 2003).

It is often possible to use a Gibbs sampler updating one clique at a time conditional on the others.

This algorithm results in an approximate generation from the joint posterior distribution

$$\pi\left\{(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})|\rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) \leq \epsilon\right\} .$$

When it is possible to achieve $\epsilon = 0$, the algorithm is exact since $S$ is a sufficient statistic.

Once a sample of $N$ values of $(\theta^{i*}_{m^{i*}}, m^{i*})$ $(1 \leq i \leq N)$ is generated from this algorithm, a standard Monte Carlo approximation of the posterior probabilities is provided by the empirical frequencies of visits to the model, namely

$$\widehat{\mathbb{P}}(\mathcal{M} = m|\mathbf{x}^0) = \sharp\{m^{i*} = m\}/N\,,$$

where $\sharp\{m^{i*} = m\}$ denotes the number of simulated $m^{i*}$'s equal to $m$.

$$BF_{m_0/m_1}(\mathbf{x}^0) = \frac{\mathbb{P}(\mathcal{M}=m_0|\mathbf{x}^0)}{\mathbb{P}(\mathcal{M}=m_1|\mathbf{x}^0)} \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}$$

$$\overline{BF}_{m_0/m_1}(\mathbf{x}^0) = \frac{\sharp\{m^{i*}=m_0\}}{\sharp\{m^{i*}=m_1\}} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}\,,$$

This estimate is only defined when $\sharp\{m^{i*}=m_1\} \neq 0$.

To bypass this difficulty, the substitute

$$\widehat{BF}_{m_0/m_1}(\mathbf{x}^0) = \frac{1+\sharp\{m^{i*}=m_0\}}{1+\sharp\{m^{i*}=m_1\}} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}$$

is particularly interesting because we can evaluate its bias.

We set $N_0 = \sharp\{m^{i*} = m_0\}$ and $N_1 = \sharp\{m^{i*} = m_1\}$.

If $\pi(\mathcal{M} = m_1) = \pi(\mathcal{M} = m_0)$, then $N_1$ is a binomial $\mathcal{B}(N, \rho)$ random variable with probability $\rho = (1 + BF_{m_0/m_1}(\mathbf{x}^0))^{-1}$ and

$$\mathbb{E}\left[\frac{N_0 + 1}{N_1 + 1}\right] = BF_{m_0/m_1}(\mathbf{x}^0) + \frac{1}{\rho(N+1)} - \frac{N+2}{\rho(N+1)}(1-\rho)^{N+1}\,.$$

The bias of $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ is $\{1 - (N+2)(1-\rho)^{N+1}\}/(N+1)\rho$, which goes to zero as $N$ goes to infinity.

$\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ can be seen as the ratio of the posterior means on the model probabilities under a $\mathcal{D}ir(1, \cdots, 1)$ prior.

$\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ suffers from a large variance when $BF_{m_0/m_1}(\mathbf{x}^0)$ is very large since.

When $\mathbb{P}(\mathcal{M} = m_1|\mathbf{x}^0)$ is very small, $\sharp\{m^{i*} = m_1\}$ is most often equal to zero.

We can used a reweighting scheme.

If the choice of $m^*$ in the ABC algorithm is driven by the probability distribution $\mathbb{P}(\mathcal{M} = m_1) = \varrho = 1 - \mathbb{P}(\mathcal{M} = m_0)$ rather than by $\pi(\mathcal{M} = m_1) = 1 - \pi(\mathcal{M} = m_0)$, the value of $\sharp\{m^{i*} = m_1\}$ can be increased and later corrected by considering instead

$$\widetilde{BF}_{m_0/m_1}(\mathbf{x}^0) = \frac{1 + \sharp\{m^{i*} = m_0\}}{1 + \sharp\{m^{i*} = m_1\}} \times \frac{\varrho}{1 - \varrho}\,.$$

Two step ABC:

If a first run of the ABC algorithm exhibits a very large value of $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$, the estimate $\widetilde{BF}_{m_0/m_1}(\mathbf{x}^0)$ produced by a second run with

$$\varrho \propto 1 \Big/ \hat{\mathbb{P}}(\mathcal{M} = m_1|\mathbf{x}^0)$$

will be more stable than the original $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$.

Results on a toy example:

Our first example compares an iid Bernoulli model with a two-state first-order Markov chain.

Both models are special cases of GRF, the first one with a trivial neighbourhood structure and the other one with a nearest neighbourhood structure.

Furthermore, the normalising constant $Z_{\theta_m,m}$ can be computed in closed form, as well as the posterior probabilities of both models.

We consider a sequence $\mathbf{x} = (x_1, .., x_n)$ of binary variables. Under model $\mathcal{M} = 0$, the GRF representation of the Bernoulli distribution $\mathcal{B}(\exp(\theta_0)/\{1 + \exp(\theta_0)\})$ is

$$f_0(\mathbf{x}|\theta_0) = \exp\left(\theta_0 \sum_{i=1}^{n} \mathbb{I}_{\{x_i=1\}}\right) \Big/ \{1 + \exp(\theta_0)\}^n \, .$$

For $\theta_0 \sim \mathcal{U}(-5, 5)$, the posterior probability of this model is available since the marginal when $S_0(\mathbf{x}) = s_0$ ($s_0 \neq 0$) is given by

$$\frac{1}{10} \sum_{k=0}^{s_0-1} \binom{s_0 - 1}{k} \frac{(-1)^{s_0-1-k}}{n-1-k} \left[(1 + e^5)^{k-n+1} - (1 + e^{-5})^{k-n+1}\right] \, .$$

Model $\mathcal{M} = 1$ is chosen as a Markov chain.

We assume a uniform distribution on $x_1$ and

$$f_1(\mathbf{x}|\theta_1) = \frac{1}{2}\exp\left(\theta_1 \sum_{i=2}^{n} \mathbb{I}_{\{x_i = x_{i-1}\}}\right) \Big/ \{1+\exp(\theta_1)\}^{n-1}\,.$$

For $\theta_1 \sim \mathcal{U}(0,6)$, the posterior probability of this model is once again available, the likelihood being of the same form as when $\mathcal{M} = 0$.

We simulated $2,000$ datasets $\mathbf{x}^0 = (x_1, \cdots, x_n)$ with $n = 100$ under each model, using parameters simulated from the priors.

For each of those $2,000$ datasets $\mathbf{x}^0$, the ABC-MC algorithm was run for $4 \times 10^6$ loops, meaning that $4 \times 10^6$ sets $(m^*, \theta^*_{m^*}, \mathbf{x}^*)$ were exactly simulated from the joint distribution.

A random number of those were accepted when $S(\mathbf{x}^*) = S(\mathbf{x}^0)$. (In the worst case scenario, the number of acceptances was 12!)

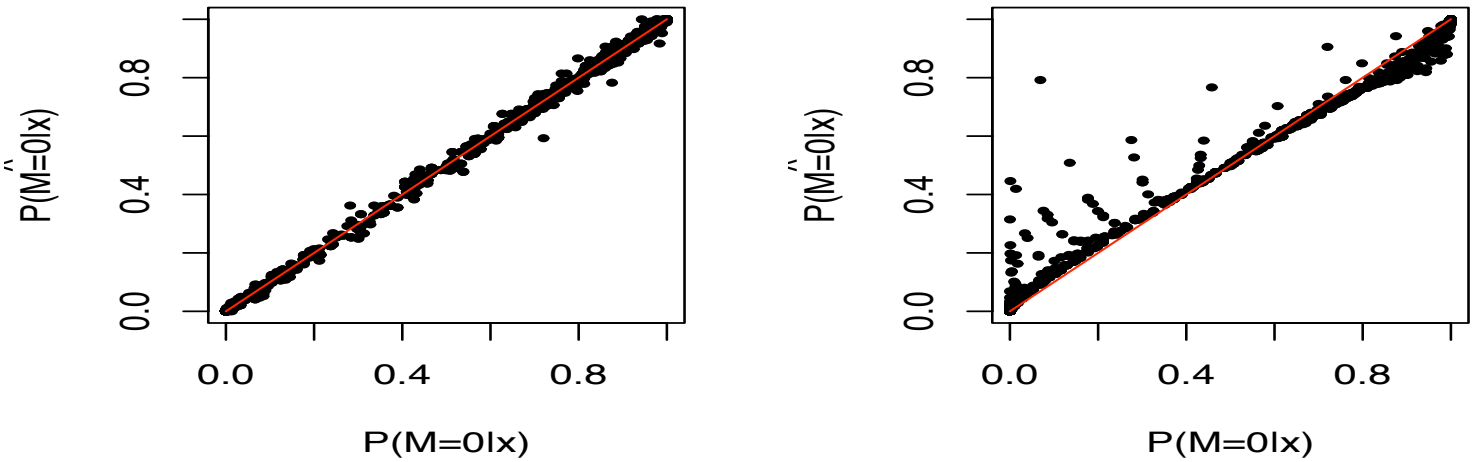


Figure 1: *(left)* Comparison of the true $\mathbb{P}(\mathcal{M} = 0|\mathbf{x}^0)$ with $\widehat{\mathbb{P}}(\mathcal{M} = 0|\mathbf{x}^0)$ over $2,000$ simulated sequences and $4 \times 10^6$ proposals from the prior. The red line is the diagonal. *(right)* Same comparison when using a tolerance $\epsilon$ corresponding to the 1% quantile on the distances.

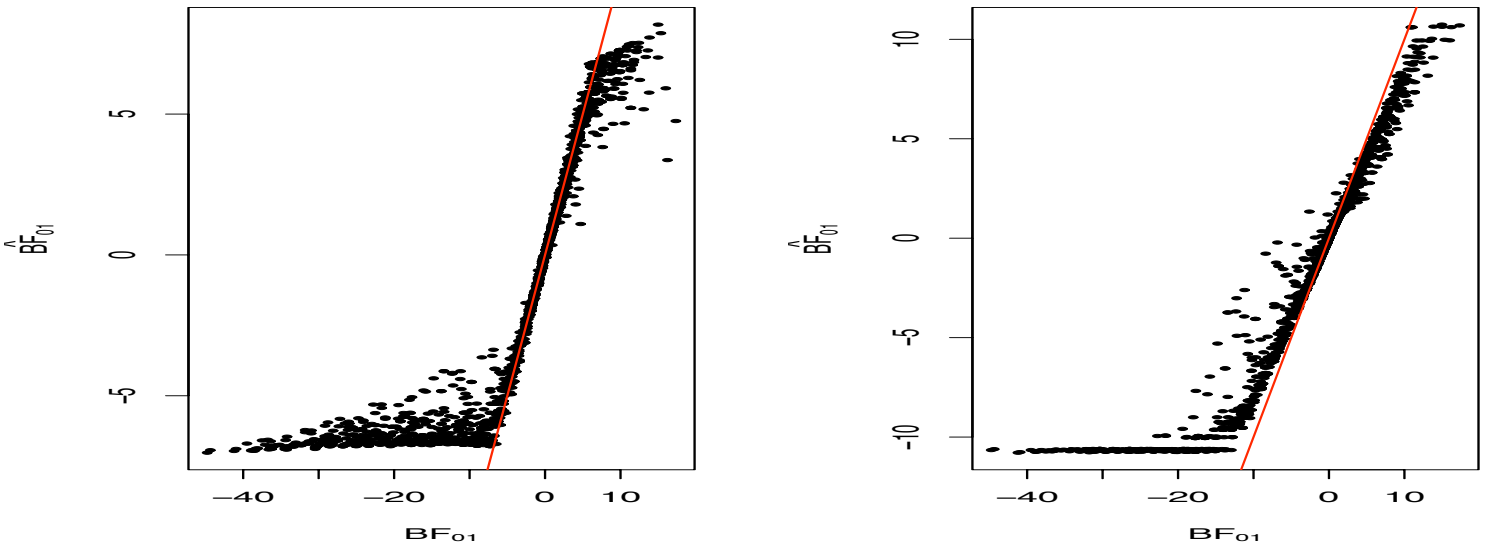


Figure 2: *(left)* Comparison of the true $BF_{m_0/m_1}(\mathbf{x}^0)$ with $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ (in logarithmic scales) over $2,000$ simulated sequences and $4 \times 10^6$ proposals from the prior. The red line is the diagonal. *(right)* Same comparison when using a tolerance corresponding to the 1% quantile on the distances.