On model choice for hidden Markov random fields: approximate Bayesian computation *versus* BIC approximations

Julien Stoehr[1]

This is a joint work with Pierre Pudlo[1] and Jean-Michel Marin[1].

[1]I3M, Université de Montpellier

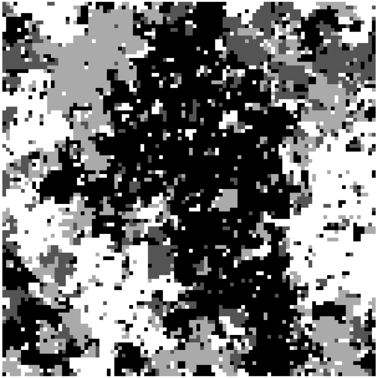
Journée du réseau AIGM, 30th June 2015

PLAN

Plan

Gibbs random fields

- **Gibbs random fields:** models useful to analyse different types of spatially correlated data.
- **Potts model (1952)** describes the spatial dependency of discrete random variable on the vertices of an undirected graph.



Hidden Potts model and model choice

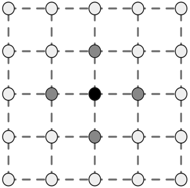
HPM$(\mathcal{G}, \alpha, \beta)\sim$ hidden Potts model where

- $\mathcal{G}$ graph of the depency structure,
- α noise parameter between the observed and the latent random field,
- β interaction parameter on the edges of $\mathcal{G}$.

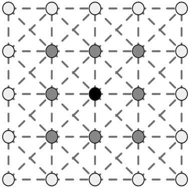
Aim: given an observation y select

the number of latent states K and/or **the latent dependency structure** $\mathcal{G}$.

$\mathcal{M}_4 = \text{HPM}(\mathcal{G}_4, \alpha, \beta)$ where $\mathcal{G}_4$ is



$\mathcal{M}_8 = \text{HPM}(\mathcal{G}_8, \alpha, \beta)$ where $\mathcal{G}_8$ is



Intractable likelihood

Bayesian distribution set

- Prior on the model space, $\pi(1), \ldots, \pi(M)$,
- Prior on the parameter space of each model, $\pi_m(\theta_m)$,
- Likelihood of the data y within each model, $\pi_m(y \mid \theta_m)$

Bayesian analysis

The posterior probability of a model is given by

$$\pi\left(m \mid y\right) \propto \pi(m) \int \pi_m(y \mid \theta_m)\pi_m(\theta_m)\mathrm{d}\theta_m.$$

Triple intractable problem !

Intractable likelihood

Intractable Gibbs distribution: $\pi(x \mid \beta_m, \mathcal{G})$

$$Z(\beta_m, \mathcal{G}) = \sum_{x \in \mathcal{X}} \exp\left(\beta_m \sum_{i \overset{\mathcal{G}}{\sim} j} \mathbb{1}\{x_i = x_j\}\right)$$

Intractable evidence:

$$\pi_m(y \mid \theta_m) = \sum_{x \in \mathcal{X}} f(y \mid x, \alpha_m)\pi(x \mid \beta_m, \mathcal{G})$$

Intractable posterior distribution:

$$\pi\left(m \mid y\right) \propto \pi(m) \int \pi_m(y \mid \theta_m)\pi_m(\theta_m)\mathrm{d}\theta_m$$

Plan

ABC = approximate Bayesian computation

Aim

A simulation based approach that can addresses the model choice issue in the Bayesian paradigm,

$$\pi(m \mid y) \propto \int \underbrace{\pi(m)\pi_m(y|\theta_m)\pi_m(\theta_m)}_{(*)} \, \mathrm{d}\theta_m.$$

Selecting the model that best fits the observed data y^{obs}

$$\widehat{m}_{\text{MAP}}(y^{\text{obs}}) = \arg\max_m \pi(m|y^{\text{obs}}).$$

ABC = approximate Bayesian computation

Aim

A simulation based approach that can addresses the model choice issue in the Bayesian paradigm,

$$\pi(m \mid y) \propto \int \underbrace{\pi(m)\pi_m(y|\theta_m)\pi_m(\theta_m)}_{(*)} \, \mathrm{d}\theta_m.$$

Selecting the model that best fits the observed data y^{obs}

$$\widehat{m}_{\text{MAP}}(y^{\text{obs}}) = \arg\max_m \pi(m|y^{\text{obs}}).$$

A first naive algorithm

- Draw a large set of particles (m, θ_m, y) from $(*)$.
- Keep the ones such that $y = y^{\text{obs}}$.

ABC = approximate Bayesian computation

Aim

A simulation based approach that can addresses the model choice issue in the Bayesian paradigm,

$$\pi(m \mid y) \propto \int \underbrace{\pi(m)\pi_m(y|\theta_m)\pi_m(\theta_m)}_{(*)} \,\mathrm{d}\theta_m.$$

Selecting the model that best fits the observed data y^{obs}

$$\widehat{m}_{\text{MAP}}(y^{\text{obs}}) = \arg\max_m \pi(m|y^{\text{obs}}).$$

A first naive algorithm

- Draw a large set of particles (m, θ_m, y) from $(*)$.
- Keep the ones such that $\rho(S(y), S(y^{\text{obs}})) < \epsilon$

ABC in practice

Algorithm 1: Simulation of the ABC reference table

Output: A reference table of size n_{REF}

for $j \leftarrow 1$ **to** n_{REF} **do**
  draw m from the prior π;
  draw θ from the prior π_m;
  draw y from the likelihood $\pi_m(\cdot|\theta)$;
  compute $S(y)$;
  save $(m_j, \theta_j, S(y_j)) \leftarrow (m, \theta, S(y))$;
end
return the table of $(m_j, \theta_j, S(y_j))$, $j = 1, \dots, n_{\text{REF}}$

- The reference table serves as a **training database**
- Computer memory: one saves **only the simulated vectors of summary statistics**.

ABC in practice

Algorithm 2: Uncalibrated ABC model choice

Output: A sample of size k distributed according to the ABC approximation of the posterior

simulate the reference table $\mathcal{T}$ according to Algorithm 1;
sort the replicates of $\mathcal{T}$ according to $\rho(S(y_j), S(y^{\text{obs}}))$;
keep the k first replicates;
return the relative frequencies of each model among the k first replicates and the most frequent model;

- ABC algorithm = a k**-nearest neighbor method** (Biau *et al.*, 2013).

ABC in practice

- The relative frequency of model m returned by Algorithm 2 converges to

  $$\pi(m \mid S(y^{\text{obs}}))$$

- When the summary statistics are **not sufficient**, it can **greatly differ** from $\pi(m \mid y^{\text{obs}})$ (Didelot *et al.*, 2011 ; Robert *et al.*, 2011).

- Marin *et al.* (2013) provide conditions on $S(\cdot)$ for the consistency of the MAP based on $\pi(m \mid S(y^{\text{obs}}))$.

ABC in practice

- The relative frequency of model m returned by Algorithm 2 converges to

$$\pi(m \mid S(y^{\text{obs}})) \neq \pi(m \mid y^{\text{obs}})$$

- When the summary statistics are **not sufficient**, it can **greatly differ** from $\pi(m \mid y^{\text{obs}})$ (Didelot *et al.*, 2011 ; Robert *et al.*, 2011).

- Marin *et al.* (2013) provide conditions on $S(\cdot)$ for the consistency of the MAP based on $\pi(m \mid S(y^{\text{obs}}))$.

ABC in practice

- The frequencies returned by Algorithm 2 should be used to **construct a knn classifier** $\widehat{m}$ that predicts the model number.

- Calibration of k should be done by **minimizing the misclassification error rate** of the classifier

- Evaluation of the misclassification rate on a **validation reference table**, independent of the reference table.

Trade off to find when no sufficient statistics

$$\pi\left(m \mid S(y^{\text{obs}})\right) \neq \pi\left(m \mid y^{\text{obs}}\right)$$

A trade off has to be found between the **loss of information** and the **dimension of** $S(\cdot)$.

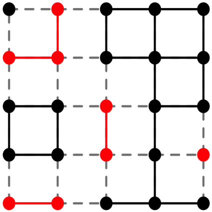
- $S(\cdot)$ of ***low*** dimension $\Rightarrow \pi\left(m \mid S(y^{\text{obs}})\right)$ is a bad approximation.
- $S(\cdot)$ of ***high*** dimension $\Rightarrow \pi\left(m \mid S(y^{\text{obs}})\right)$ is a good approximation approximation but it's difficult to draw y such that $S(y) \approx S(y^{\text{obs}})$.

Four geometric summary statistics

$$\Gamma(\mathcal{G}, y) : \quad i \sim j \iff i \overset{\mathcal{G}}{\sim} j \text{ et } y_i = y_j$$

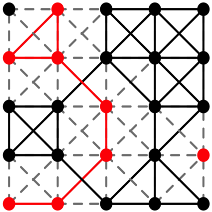
- **number of connected components:** $T(\mathcal{G}, y)$
- **size of the biggest** connected component: $U(\mathcal{G}, y)$

$\Gamma(\mathcal{G}_4, y)$



$T(\mathcal{G}_4, y) = 7$
$U(\mathcal{G}_4, y) = 12$

$\Gamma(\mathcal{G}_8, y)$



$T(\mathcal{G}_8, y) = 4$
$U(\mathcal{G}_8, y) = 16$

Sets of summary statistics to compare

Aim of ABC

Selecting the hidden Gibbs model that **better fits** a given observation.

Our aim

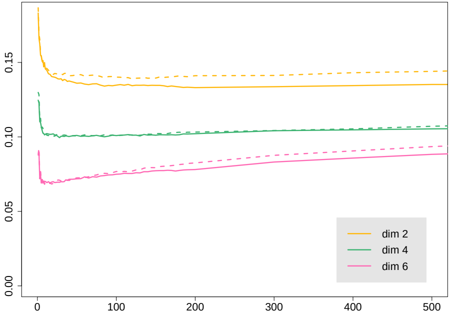
Selecting the **most informative set** of summary statistics.

Summary statistics	Grelaud, *et al.*	Number of conn. comp.	Size of the biggest conn. comp.
$S_{2D}(y)$ (dim $= 2$)	✓		
$S_{4D}(y)$ (dim $= 4$)	✓	✓	
$S_{6D}(y)$ (dim $= 6$)	✓	✓	✓

ABC experiment

Settings

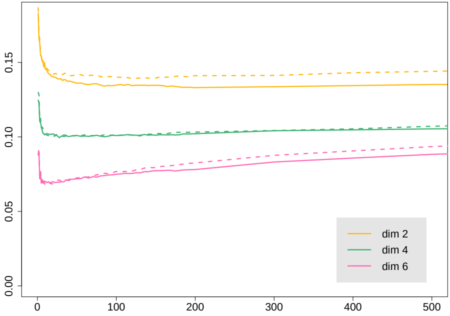
- 2 colors,
- $y_i \mid x_i = c \sim \mathcal{N}(c, \sigma^2)$, $c \in \{0; 1\}$
- Training reference table: 50 000 or 100 000,
- Validation reference table: 20 000.

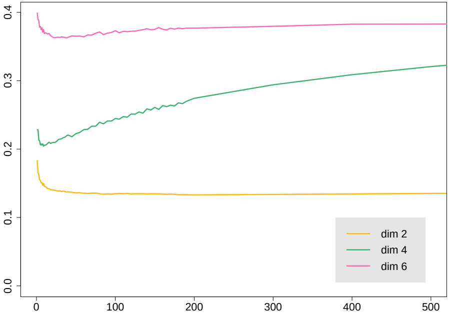


ABC experiment

Settings

- 2 colors,
- $y_i \mid x_i = c \sim \mathcal{N}(c, \sigma^2)$, $c \in \{0;1\}$
- Training reference table: 50 000 or 100 000,
- Validation reference table: 20 000.





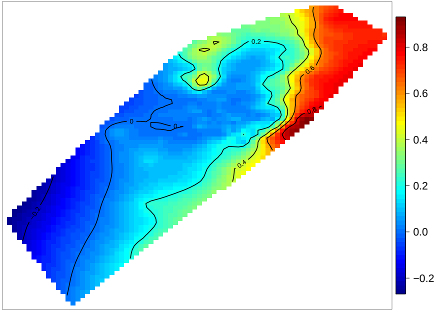
ABC experiment

Settings

- 2 colors,
- $y_i \mid x_i = c \sim \mathcal{N}(c, \sigma^2)$, $c \in \{0; 1\}$

Prior error rates

Train size	5,000	100,000
2D statistics	14.2%	13.8%
4D statistics	10.8%	9.8%
6D statistics	8.6%	6.9%
Adaptive ABC	8.2%	6.7%



Reference

- Stoehr, J., Pudlo, P., and Cucala, L. (2014). *Adaptive ABC model choice and geometric summary statistics for hidden Gibbs random fields*. Statistics and Computing, 25(1), 129-141.

Plan

Bayesian Information Criterion

Principle: approximate the integrated likelihood using Laplace's method (Schwarz, 1978)

$$\mathrm{BIC}(m) = 2\log \pi_m(y \mid \hat{\theta}_m^{mle}) - d_m \log(|\mathcal{S}|),$$

where

$$\pi_m(y \mid \hat{\theta}_m^{mle}) = \int_{\mathcal{X}} f(y \mid x, \hat{\alpha}_m^{mle}) \pi(x \mid \hat{\beta}_m^{mle}, \mathcal{G}) \mathrm{d}x.$$

Bayesian Information Criterion

Principle: approximate the integrated likelihood using Laplace's method (Schwarz, 1978)

$$\text{BIC}(m) = 2 \log \pi_m(y \mid \hat{\theta}_m^{mle}) - d_m \log(|\mathcal{S}|),$$

where

$$\pi_m(y \mid \hat{\theta}_m^{mle}) = \int_{\mathcal{X}} f(y \mid x, \hat{\alpha}_m^{mle}) \pi(x \mid \hat{\beta}_m^{mle}, \mathcal{G}) \mathrm{d}x.$$

Solutions:

- Monte Carlo draws from $\pi(x \mid \hat{\beta}^{mle}, \mathcal{G})$
- Likelihood approximations (*e.g.*, Stanford and Raftery 2002, Celeux *et al.*, 2003, Forbes and Peyrard, 2003, Varin and Vidoni, 2005)

Pseudolikelihood *versus* Mean-field approximation

Pseudolikelihood (Besag, 1975)

$$f_{\mathrm{CL}}(x \mid \beta, \mathcal{G}) = \prod_{i=1}^{C} \pi(x_{A(i)} \mid x_{B(i)}, \beta, \mathcal{G}).$$

- Not a genuine probability distribution.

Mean field approximation: minimizes the Kullback-Leibler divergence between a given distribution P and the Gibbs distribution $\pi(\cdot \mid \beta, \mathcal{G})$ over the set of probability distributions that factorize

$$P(x) = \prod_{i \in \mathcal{S}} P_i(x_i), \text{ where } P_i \in \mathcal{M}_1^+(\mathcal{X}_i) \text{ and } P \in \mathcal{M}_1^+(\mathcal{X}).$$

BIC based on Mean field-like approximations

Mean field-like approximation:

$$P^{\text{MFL}}(x \mid \beta, \mathcal{G}) = \prod_{i \in \mathcal{S}} \pi(x_i \mid X_{\mathcal{N}(i)} = \tilde{x}_{\mathcal{N}(i)}, \beta, \mathcal{G}).$$

Notable solutions

- *Approximate Bayes factors for image segmentation: The pseudolikelihood information criterion (PLIC)*, Stanford and Raftery (IEEE PAMI, 2002)
- *Hidden Markov random field model selection criteria based on mean field-like approximations*, Forbes and Peyrard (IEEE PAMI, 2003)

Block Likelihood Information Criterion (BLIC)

Thrust: working with distributions that factorize on blocks

$$P = \prod_{i=1}^{C} P_{A(i)}, \text{ where } P_{A(i)} \in \mathcal{M}_1^+(\mathcal{X}_{A(i)}) \text{ and } P \in \mathcal{M}_1^+(\mathcal{X}).$$

Approximation:

$$\pi(y \mid \hat{\theta}^{mle}, m) \approx \prod_{i=1}^{C} \frac{\sum_{x_{A(i)}} f(y_{A(i)} \mid x_{A(i)}, \hat{\alpha}^{mle}) \exp\left(\hat{\beta}^{mle} \sum_{i \overset{\mathcal{G}}{\sim} j} \mathbb{1}\{x_i = x_j\}\right)}{Z\left(\mathcal{G}_{\text{block}}, \hat{\beta}^{mle}\right)}$$

Idea: opportunity to compute normalizing constants if blocks are small enough (*e.g.*, Friel and Rue, Biometrika, 2007).

PLAN

Selection of K

Noise distribution: $y_i \mid x_i = c \sim \mathcal{N}(c, 0.25)$.

- **Data set:** 100 draws from $\pi(x \mid \beta = 1, \mathcal{G}_4)$ when $K = 4$

K	2	3	4	5	6	7	8
PLIC	0	9	91	0	0	0	0
BICp	0	0	39	23	16	22	0
BLIC(2×2)	0	0	100	0	0	0	0

- **Data set:** 100 draws from $\pi(x \mid \beta = 0.4, \mathcal{G}_8)$ when $K = 4$

K	2	3	4	5	6	7	8
PLIC	0	7	93	0	0	0	0
BICp	0	0	43	18	19	20	0
BLIC(2×2)	0	1	99	0	0	0	0
BLIC(4×4)	0	0	100	0	0	0	0

Selection of $\mathcal{G}$

Noise distribution: $y_i \mid x_i = c \sim \mathcal{N}(k, 0.25)$.

- **Data set:** 100 draws from $\pi(x \mid \beta = 1, \mathcal{G}_4)$ when $K = 4$

	$\mathcal{G}_4$	$\mathcal{G}_8$
PLIC	53	47
BICp	100	0
BLIC(2×2)	100	0

- **Data set:** 100 draws from $\pi(x \mid \beta = 0.4, \mathcal{G}_8)$ when $K = 4$

	$\mathcal{G}_4$	$\mathcal{G}_8$
PLIC	0	100
BICp	0	100
BLIC(2×2)	59	41
BLIC(4×4)	0	100

Comparison ABC *versus* BIC approximations

- 2 colors,
- $y_i \mid x_i = c \sim \mathcal{N}(c, 0.15)$, $c \in \{0; 1\}$,
- $\pi(m) \sim \mathcal{U}(\{\mathcal{G}_4, \mathcal{G}_8\})$, $\pi_{\mathcal{G}_4}(\beta) \sim \mathcal{U}([0; 1])$ and $\pi_{\mathcal{G}_8}(\beta) \sim \mathcal{U}([0; 0.4])$,
- 2000 draws from the corresponding Gibbs distribution using Swendsen Wang algorithm (5000 iterations).

Train size	5,000	100,000	Criterion	Error rate
2D statistics	14.2%	13.8%	PLIC	19.8%
4D statistics	10.8%	9.8%	BICp	7.6%
6D statistics	8.6%	6.9%	BICw	7.1%
Adaptive ABC	8.2%	6.7%	BLIC(4x4)	7.7%

PLAN

Take home message

ABC

- ABC model choice = classification problem
- A local error which assesses the accuracy of the classifier at y^{obs}
- Calibrating the number of neighbors in ABC provides better results than a fixed quantile of the distances $\Rightarrow$ reduce significantly the number of simulations.

Latent Markov random fields

- New class of summary statistics based on cluster analysis

BIC approximations

- BIC approximations provide good results while being computationaly efficient.
- Block approximations seem preferable to single sites approximations to select the number of hiddent states.