Hidden Gibbs random fields model selection using Block Likelihood Information Criterion

Jean-Michel Marin

Université de Montpellier
Institut Montpelliérain Alexander Grothendieck (IMAG)
Institut de Biologie Computationnelle (IBC)
Labex Numev

Part of the thesis of Julien Stoehr
joint work with Pierre Pudlo

Introduction

Discrete Gibbs or Markov random fields have appeared as convenient statistical model to analyse different types of spatially correlated data.

Hidden random fields: we observe only a noisy version $\boldsymbol{y}$ of an unobserved discrete latent process $\boldsymbol{x}$

Discrete Gibbs or Markov random fields suffer from major computational difficulties

Intractable normalizing constant

For parameter estimation:

Richard Everitt (2012) Bayesian Parameter Estimation for Latent Markov Random Fields and Social Networks, Journal of Computational and Graphical Statistics

Model choice questions: selecting the number of latent states and the dependency structure of hidden Potts model

Use the Bayesian Information Criterion

Plan

- Discrete hidden Gibbs or Markov random fields
- Block Likelihood Information Criterion
  - Background on Bayesian Information Criterion
  - Gibbs distribution approximations
  - Related model choice criteria
- Comparison of BIC approximations
  - Hidden Potts models
  - First experiment: selection of the number of colors
  - Second experiment: selection of the dependency structure
  - Third experiment: BLIC versus ABC

Discrete hidden Gibbs or Markov random fields

A discrete Markov random field $\mathbf{X}$ with respect to $\mathscr{G}$:

- a collection of random variables X_i taking values in $\mathscr{X} = \{0, \ldots, K-1\}$ indexed by a finite set of sites $\mathscr{S} = \{1, \ldots, n\}$
- the dependency between the sites is given by an undirected graph $\mathscr{G}$ which induces a topology on $\mathscr{S}$:

$$\mathbf{P}\left(X_i = x_i \mid \mathbf{X}_{-i} = \mathbf{x}_{-i}\right) = \mathbf{P}\left(X_i = x_i \;\middle|\; \mathbf{X}_{\mathscr{N}(i)} = \mathbf{x}_{\mathscr{N}(i)}\right),$$

where $\mathscr{N}(i)$ denotes the set of all the neighbor sites to i in $\mathscr{G}$: i and j are neighbor if and only if i and j are linked by an edge in $\mathscr{G}$.

Markov random fields $\Longleftrightarrow$ Undirected graphical models

A discrete Gibbs random fields $\mathbf{X}$ with respect to $\mathscr{G}$

- a collection of random variables X_i taking values in $\mathscr{X} = \{0, \ldots, K-1\}$ indexed by a finite set of sites $\mathscr{S} = \{1, \ldots, n\}$
- the pdf of $\mathbf{X}$ factorizes with respects to the cliques of $\mathscr{G}$:

$$\mathbf{P}(\mathbf{X} = \mathbf{x} \mid \mathscr{G}) = \pi(\mathbf{x} \mid \psi, \mathscr{G}) = \frac{1}{Z(\psi, \mathscr{G})} \exp\left\{ -\sum_{c \in \mathscr{C}_{\mathscr{G}}} H_c(\mathbf{x}_c \mid \psi) \right\}$$

  - $\mathscr{C}_{\mathscr{G}}$ is the set of maximal cliques of $\mathscr{G}$,
  - ψ is a vector of parameters,
  - the H_c functions denote the energy functions.

If $\mathbf{P}(\mathbf{X} = \mathbf{x} \mid \mathscr{G}) > 0$ for all $\mathbf{x}$, the Hammersley-Clifford theorem proves that Markov and Gibbs random fields are equivalent with regards to the same graph.

Intractable normalizing constant (the partition function)

$$Z(\psi, \mathscr{G}) = \sum_{\mathbf{x} \in \mathscr{X}^n} \exp\left\{ - \sum_{c \in \mathscr{C}_{\mathscr{G}}} H_c\left(\mathbf{x}_c \mid \psi\right) \right\}$$

Summation over the numerous possible realizations of the random field **X** cannot be computed directly

Hidden Markov random fields

$\boldsymbol{x}$ is latent, we observe $\boldsymbol{y}$ and assume that

$$\pi(\boldsymbol{y} \mid \boldsymbol{x}, \phi) = \prod_{i \in \mathscr{S}} \pi(y_i \mid x_i, \phi)$$

Emission distribution $\pi(y_i \mid x_i, \phi)$: discrete, Gaussian, Poisson...

Likelihood

$$\pi(\mathbf{y} \mid \phi, \psi) = \sum_{\mathbf{x} \in \mathscr{X}^n} \pi(\mathbf{y} \mid \mathbf{x}, \phi) \frac{1}{Z(\psi, \mathscr{G})} \exp\left\{ - \sum_{c \in \mathscr{C}_{\mathscr{G}}} H_c(\mathbf{x}_c \mid \psi) \right\}.$$

Double intractable issue!

Core of bayesian model choice: the integrated likelihood

$$\int \sum_{\mathbf{x} \in \mathscr{X}^n} \pi(\mathbf{y} \mid \mathbf{x}, \phi) \frac{1}{Z(\psi, \mathscr{G})} \exp\left\{ - \sum_{c \in \mathscr{C}_{\mathscr{G}}} H_c(\mathbf{x}_c \mid \psi) \right\} \pi(\phi, \psi) d\phi d\psi$$

Triple intractable problem!

Block Likelihood Information Criterion

Background on Bayesian Information Criterion

$\mathbf{y} = \{y_1, \ldots, y_n\}$ an iid sample
Finite set of models $\{m : 1, \ldots, M\}$

$$\pi(m \mid \mathbf{y}) = \frac{\pi(m) e(\mathbf{y} \mid m)}{\sum_{m'} \pi(m') e(\mathbf{y} \mid m')}$$

$$e(\mathbf{y} \mid m) = \int \pi_m(\mathbf{y} \mid \theta_m) \pi_m(\theta_m) \, d\theta_m$$

Laplace approximation

$$\log e\left(\mathbf{y} \mid m\right) = \log \pi_m\left(\mathbf{y} \,\middle|\, \hat{\theta}_m\right) - \frac{d_m}{2}\log(n) + R_m\left(\hat{\theta}_m\right) + \mathcal{O}\left(n^{-\frac{1}{2}}\right)$$

$\hat{\theta}_m$ is the maximum likelihood estimator of θ_m

d_m is the number of free parameters for model m

R_m is bounded as the sample size grows to infinity

BIC

$$-2\log e\left(\mathbf{y} \mid m\right) \simeq \mathbf{BIC}(m) = -2\log \pi_m\left(\mathbf{y} \,\middle|\, \hat{\theta}_m\right) + d_m \log(n)$$

Penalty term: $d_m \log(n)$ increases with the complexity of the model

Consistency of BIC: iid processes from the exponential families, mixture models, Markov chains...

For selecting the neighborhood system of an observed Gibbs random fields: Csiszar and Talata (2006) proposed to replace the likelihood by the pseudo-likelihood and modify the penalty term.

Gibbs distribution approximations

Replace the Gibbs distribution by tractable surrogates

Pseudo-likelihood (Besag, 1975), composite likelihood (Lindsay, 1988): replace the original Markov distribution by a product of easily normalized distribution

Conditional composite likelihoods are not a genuine probability distribution for Gibbs random field

$\Longrightarrow$ **the focus hereafter is solely on valid probability function**

Idea: minimize the Kullback-Leibler divergence over a restricted class of tractable probability distribution

$\Longrightarrow$ Mean field approaches: minimize the Kullback-Leibler divergence over the set of probability functions that factorize on sites

$\Longrightarrow$ Celeux, Forbes and Peyrard (2003)

$$\mathbf{P}^{\text{MF-like}}_{\tilde{\mathbf{x}}}(\mathbf{x} \mid \psi, \mathscr{G}) = \prod_{i \in \mathscr{S}} \pi\left(x_i ; \tilde{\mathbf{x}}_{\mathscr{N}(i)}, \psi, \mathscr{G}\right)$$

$\pi\left(x_i ; \tilde{\mathbf{x}}_{\mathscr{N}(i)}, \psi, \mathscr{G}\right) = \mathbf{P}\left(X_i = x_i \;\middle|\; \mathbf{X}_{\mathscr{N}(i)} = \tilde{\mathbf{x}}_{\mathscr{N}(i)}\right)$
$\tilde{\mathbf{x}}$ is a fixed point of an iterative algorithm

Use tractable approximations that factorize over larger sets of nodes

$A(1), \ldots, A(C)$ a partition

$$P_{\tilde{x}}(x \mid \psi, \mathscr{G}) = \prod_{\ell=1}^{C} \pi\left(x_{A(\ell)}; \tilde{x}_{B(\ell)}, \psi, \mathscr{G}\right)$$

$\tilde{x}$ is a constant field

$B(\ell)$ is either the set of neighbor of $A(\ell)$ or the empty set

For parameter estimation

Nial Friel (2012) Bayesian inference for Gibbs random fields using composite likelihoods. Proceedings of the Winter Simulation Conference 2012

If $\mathrm{B}(\ell) = \emptyset$, we are cancelling the edges in $\mathscr{G}$ that link elements of $\mathcal{A}(\ell)$ to elements of any other subset of $\mathscr{S}$.

The Gibbs distribution is then simply replaced by the product of the likelihood restricted to $\mathcal{A}(\ell)$.

$$\mathbf{P}_{\tilde{x}}\left(\mathbf{y} \mid \psi, \phi, \mathscr{G}\right) = \sum_{x \in \mathscr{X}^n} \pi\left(\mathbf{y} \mid x, \phi\right) \mathbf{P}_{\tilde{x}}\left(x \mid \psi, \mathscr{G}\right)$$

$$= \prod_{\ell=1}^{C} \sum_{x_{A(\ell)}} \left\{ \prod_{i \in A(\ell)} \pi\left(y_i \mid x_i, \phi\right) \right\} \pi\left(x_{A(\ell)}; \tilde{x}_{B(\ell)}, \psi, \mathscr{G}\right)$$

$$= \prod_{\ell=1}^{C} \sum_{x_{A(\ell)}} \pi\left(\mathbf{y}_{A(\ell)} \mid x_{A(\ell)}, \phi\right) \pi\left(x_{A(\ell)}; \tilde{x}_{B(\ell)}, \psi, \mathscr{G}\right).$$

Block Likelihood Information Criterion (BLIC)

$$\text{BIC} \approx -2 \log \mathbf{P}_{\tilde{x}}\left(\mathbf{y} \mid \theta^*, \mathscr{G}\right) + d \log(|\mathscr{S}|) = \text{BLIC}^{\tilde{x}}\left(\theta^*\right)$$

$\theta^* = (\phi^*, \psi^*)$ is a parameter value to specify

d the number of parameters

Nial Friel and Havard Rue (2007) Recursive computing and simulation-free inference for general factorizable models, Biometrika

Each term of the product can be computed as long as the blocks are small enough!

$$\pi\left(\mathbf{x}_{A(\ell)};\tilde{\mathbf{x}}_{B(\ell)},\psi,\mathscr{G}\right)=\frac{1}{Z\left(\psi,\mathscr{G},\tilde{\mathbf{x}}_{B(\ell)}\right)}\exp\left\{\psi^{T}\mathbf{S}\left(\mathbf{x}_{A(\ell)};\tilde{\mathbf{x}}_{B(\ell)}\right)\right\}$$

$\mathbf{S}\left(\mathbf{x}_{A(\ell)};\tilde{x}_{B(\ell)}\right)$ is the restriction of $\mathbf{S}$ to the subgraph defined on the set $A(\ell)$ and conditioned on the fixed border $\tilde{\mathbf{x}}_{B(\ell)}$

$$\sum_{x_{A(\ell)}} \pi\left(y_{A(\ell)} \mid x_{A(\ell)}, \phi\right) \pi\left(x_{A(\ell)}; \tilde{x}_{B(\ell)}, \psi, \mathscr{G}\right)$$

$$= \frac{1}{Z\left(\psi, \mathscr{G}, \tilde{x}_{B(\ell)}\right)} \underbrace{\sum_{x_{A(\ell)}} \exp\left\{\log \pi\left(y_{A(\ell)} \mid x_{A(\ell)}, \phi\right) + \psi^T S\left(x_{A(\ell)}; \tilde{x}_{B(\ell)}\right)\right\}}_{=Z\left(\theta, \mathscr{G}, y_{A(\ell)}, \tilde{x}_{B(\ell)}\right)}.$$

$Z\left(\theta, \mathscr{G}, y_{A(\ell)}, \tilde{x}_{B(\ell)}\right)$ corresponds to the normalizing constant of the conditional random field $X_{A(\ell)}$ knowing $Y_{A(\ell)} = y_{A(\ell)}$

Initial model with an extra potential on singletons

$$\mathrm{BLIC}^{\tilde{x}}(\theta^*) =$$

$$-2\sum_{\ell=1}^{C}\left\{\log Z\left(\theta^*, \mathscr{G}, \mathbf{y}_{A(\ell)}, \tilde{\mathbf{x}}_{B(\ell)}\right) - \log Z\left(\psi^*, \mathscr{G}, \tilde{\mathbf{x}}_{B(\ell)}\right)\right\} + d\log(|\mathscr{S}|)$$

Related model choice criteria

Our approach encompasses the Pseudo-Likelihood Information Criterion (PLIC) of Stanford and Raftery (2002) as well as the mean field-like approximations $\mathrm{BIC}^{\text{MF-like}}$ proposed by Forbes and Peyrard (2003).

They consider the finest partition of $\mathscr{S}$ and propose ingenious solutions for choosing $\tilde{\mathbf{x}}$ and estimating θ_*.

Stanford and Raftery (2002) suggest to set $(\tilde{\mathbf{x}}, \theta_*)$ to the final estimates of the Iterated Conditional Modes algorithm of Besag (1986).

Forbes and Peyrard (2003) put forward the use of the output $(\hat{\theta}^{\text{MF-like}}, \tilde{\mathbf{x}}^{\text{MF-like}})$ of the mean-field EM algorithm of Celeux, Forbes and Peyrard(2003).

$$\mathrm{PLIC} = \mathrm{BLIC}^{\tilde{\mathbf{x}}^{\mathrm{ICM}}}\left(\widehat{\theta}^{\mathrm{ICM}}\right)$$

$$\mathrm{BIC}^{\text{MF-like}} = \mathrm{BLIC}^{\tilde{\mathbf{x}}^{\text{MF-like}}}\left(\widehat{\theta}^{\text{MF-like}}\right)$$

Comparison of BIC approximations

Hidden Potts models

$$\pi(\mathbf{x} \mid \psi, \mathscr{G}) = \frac{1}{Z(\psi, \mathscr{G})} \exp\left\{ -\psi \sum_{i \overset{\mathscr{G}}{\sim} j} \mathbb{1}\{x_i = x_j\} \right\}$$

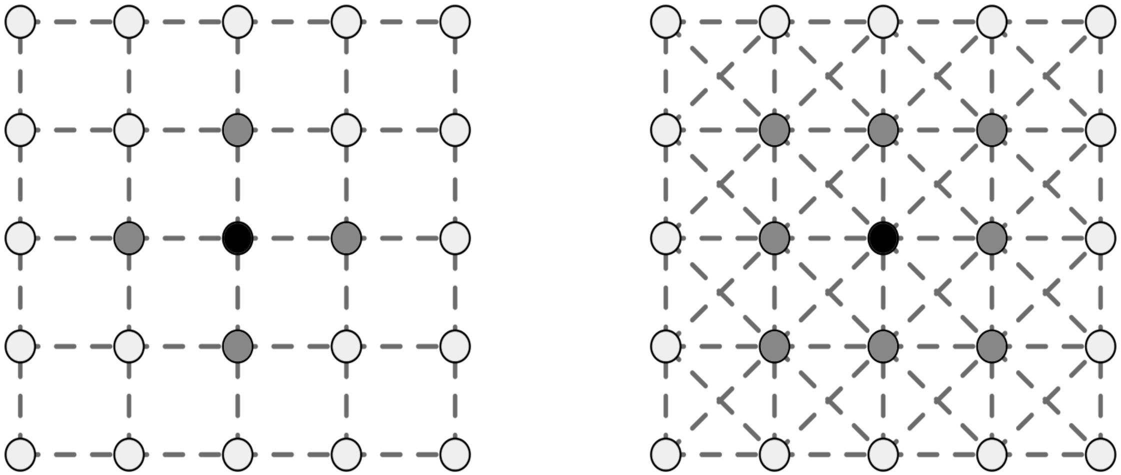
where the sum $i \overset{\mathscr{G}}{\sim} j$ is over the set of edges of the graph $\mathscr{G}$.

In the statistical physic literature, ψ is interpreted as the inverse of a temperature, and when the temperature drops below a fixed threshold, values x_i of a typical realization of the field are almost all equal.

Neighborhood graphs $\mathscr{G}$ of hidden Potts model

The four closest neighbour graph $\mathscr{G}_4$

The eight closest neighbour graph $\mathscr{G}_8$



$\mathbf{y}^{\text{obs}}$, $n = 100 \times 100$ pixels image, such that

$$y_i \mid x_i = k \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right) \quad k \in \{0, \dots, K-1\},$$

$$\mathcal{M} = \{\text{HPM}\,(\mathscr{G}, \theta, K) : \; K = K_{\min}, \dots, K_{\max} \; ; \; \mathscr{G} \in \{\mathscr{G}_4, \mathscr{G}_8\}\},$$

θ^* and the field $\tilde{\mathbf{x}}$: mean-field EM

EM-like algorithm has been initialized with a simple K-means procedure

$\mathsf{A}(\ell)$: square block of dimension $\mathsf{b} \times \mathsf{b}$.

Block Likelihood Criterion is indexed by the dimension of the blocks: $\text{BLIC}^{\text{MF-like}}_{\mathsf{b}\times\mathsf{b}}$.

$\text{BIC}^{\text{MF-like}} = \text{BLIC}^{\text{MF-like}}_{\mathsf{1}\times\mathsf{1}}$

$\text{B}(\ell) = \emptyset$, we note our criterion $\text{BLIC}_{\mathsf{b}\times\mathsf{b}}$
$\text{BLIC}_{\mathsf{1}\times\mathsf{1}}$ is the BIC approximations corresponding to a finite independent mixture model

Simulated images obtained using the Swendsen-Wang algorithm

First experiment: selection of the number of colors

Dependency structure is known
Select the number K if hidden states

$K = 4$, $\mu_k = k$ and $\sigma_k = 0.5$
for $\mathscr{G}_4 \quad \to \quad \psi = 1$
for $\mathscr{G}_8 \quad \to \quad \psi = 0.4$

The images present homogeneous regions and then the observations exhibit some spatial structure

$\mathrm{HPM}(\mathscr{G}_4, \theta, 4)$

K	**2**	**3**	**4**	**5**	**6**	**7**
$\mathrm{BIC}^{\text{MF-like}}$	0	0	39	23	16	22
$\mathrm{BLIC}^{\text{MF-like}}_{2\times2}$	0	0	58	18	8	16
$\mathrm{BLIC}_{1\times1}$	0	0	97	1	2	0
$\mathrm{BLIC}_{2\times2}$	0	0	100	0	0	0

$\mathrm{HPM}(\mathscr{G}_8, \theta, 4)$

K	**2**	**3**	**4**	**5**	**6**	**7**
$\mathrm{BIC}^{\text{MF-like}}$	0	0	43	18	19	20
$\mathrm{BLIC}_{2\times 2}^{\text{MF-like}}$	0	0	52	14	17	17
$\mathrm{BLIC}_{1\times 1}$	0	3	90	1	4	2
$\mathrm{BLIC}_{2\times 2}$	0	1	99	0	0	0
$\mathrm{BLIC}_{4\times 4}$	0	0	100	0	0	0

Second experiment: selection of the dependency structure

K is known

Discriminate between the two dependency structures

$$\text{HPM}(\mathscr{G}_4, \theta, 4)$$

	$\mathscr{G}_4$	$\mathscr{G}_8$
$\text{BLIC}_{1\times 1}$	46	54
$\text{BIC}^{\text{MF-like}}$	100	0
$\text{BLIC}_{2\times 2}^{\text{MF-like}}$	100	0
$\text{BLIC}_{2\times 2}$	100	0

$$\mathrm{HPM}(\mathscr{G}_8, \theta, 4)$$

	$\mathscr{G}_4$	$\mathscr{G}_8$
$\mathrm{BIC}^{\text{MF-like}}$	0	100
$\mathrm{BLIC}_{2\times 2}^{\text{MF-like}}$	0	100
$\mathrm{BLIC}_{2\times 2}$	59	41
$\mathrm{BLIC}_{4\times 4}$	0	100

Third experiment: BLIC *versus* ABC

K is known
Discriminate between the two dependency structures

$K = 2$, $\mu_k = k$ and $\sigma_k = 0.39$

for $\mathscr{G}_4 \quad \to \quad \pi(\psi) = \mathcal{U}[0, 1]$
for $\mathscr{G}_8 \quad \to \quad \pi(\psi) = \mathcal{U}[0, 0.35]$

1000 realizations from HPM$(\mathscr{G}_4, \theta, 2)$ and HPM$(\mathscr{G}_8, \theta, 2)$

ABC approximations

Train size	$5,000$	$100,000$
2D statistics	14.2%	13.8%
4D statistics	10.8%	9.8%
6D statistics	8.6%	6.9%

Clever geometric summary statistics: number of connected components, size of the biggest connected components.

BLIC approximation

$$\text{BLIC}_{4\times 4} \longrightarrow 7.7\%$$