Critères de choix de modèle pour champs de Gibbs cachés

Julien Stoehr [1] & Jean-Michel Marin [1] & Pierre Pudlo [1]

[1] *I3M UMR CNRS 5149, Université de Montpellier, Place E. Bataillon 34095 Montpellier CEDEX, France*
julien.stoehr@univ-montp2.fr
jean-michel.marin@univ-montp2.fr
pierre.pudlo@univ-montp2.fr

Résumé. La question du choix de modèle pour un champ de Gibbs caché est difficile. La structure de dépendance markovienne ne permet pas le calcul explicite de la constante de normalisation de la vraisemblance et de la somme sur tous les champs latents possibles. Les critères de choix de modèle de type BIC (Schwartz, 1978) ne sont donc pas estimables directement. Des approximations de BIC basées sur le principe des champs moyens ont été proposées pour rendre le calcul possible (Stanford et Raftery, 2002 ; Forbes et Peyrard, 2003). L'approximation consiste à remplacer la loi du modèle par une loi produit sur un ensemble de variables aléatoires réelles indépendantes. Dans le cas de la segmentation d'image, cela revient à factoriser la loi sur l'image en un produit de loi sur les pixels. Nous proposerons une extension de ces approximations lorsque la vraisemblance est remplacée par une loi produit sur des sous ensembles de variables aléatoires, *i.e.*, des blocs de l'image.

Mots-clés. Champ de Gibbs caché, choix de modèle, critère BIC, approximation de type champ moyen.

Abstract. Selecting between different models for a hidden Markov random field can be very challenging. Due to the Markovian dependance structure, the normalizing constants in the likelihoods and the sum over all possible latent random fields are intractable. Selection criterion like BIC (Schwartz, 1978) cannot be computed exactly. Approximations of BIC based on mean field theory have been proposed to make the computation tractable (Stanford et Raftery, 2002 ; Forbes et Peyrard, 2003). The approximation consists in replacing the true likelihood by a product distribution on a system of independent variables. For example, in image segmentation the distribution on the image is factorized as the product of conditional likelihoods of pixels. We will show an extension of the previous approximations when the likelihood is replaced by a product distribution on subsets of variables, *i.e.*, blocks of the image.

Keywords. Hidden Gibbs random field, model choice, Bayesian Information Criterion, mean field approximation.

1 Introduction

En dépit d'un large éventail d'applications, les champs de Gibbs (Besag, 1974) présentent des difficultés majeures tant du point de vue de l'inférence sur les paramètres que de la sélection de modèle. Il existe actuellement peu de travaux sur le problème de choix de modèle (*e.g.*, Stanford et Raftery, 2002 ; Forbes et Peyrard, 2003 ; Cucala et Marin, 2013 ; Stoehr *et al.*, 2015). Choisir parmi une collection de modèles peut s'avérer très compliqué en raison de la structure de dépendance markovienne qui ne permet pas de calculer explicitement la constante de normalisation du modèle, rendant impossible l'évaluation de la vraisemblance intégrée. Afin de palier cette difficulté, différentes approximations ont été proposées. Compte tenu de sa simplicité, celle couramment utilisée est le critère BIC (Bayesian Information Criterion) proposé par Schwartz (1978) et basé sur une méthode de Laplace pour approcher la vraisemblance intégrée du modèle.

Dans le cadre de champs de Gibbs cachés, la difficulté vient de la maximisation de la log-vraisemblance intégrée sur le latent faisant intervenir des distributions de Gibbs, marginale ou conditionnelle, que l'on ne peut expliciter. La solution standard consiste alors à approcher BIC en remplaçant ces distributions par des lois produits plus simples. Stanford et Raftery (2002) proposent de substituer la loi à maximiser par la pseudo-vraisemblance introduite par Qian et Titterington (1991) pour définir le critère PLIC (Penalized Pseudolikelihood Criterion). Cette approche est généralisée par Forbes et Peyrard (2003) qui construisent une famille d'approximations du critère BIC basées sur une généralisation du principe de champ moyen issu de la physique statistique.

Ces deux approches consistent à remplacer le champ de Gibbs par un ensemble de variables aléatoires réelles indépendantes. Dans le cadre de la segmentation d'image, cela revient à remplacer la distribution de l'image par le produit de lois conditionnelles sur les pixels. On se propose d'étendre ces approches en considérant des distributions qui se factorisent sur des blocs de l'image. On illustrera en particulier les limites des approximations proposées par Stanford et Raftery (2002) et Forbes et Peyrard (2003) et on étudiera l'influence de la taille des blocs sur différents exemples.

2 Champ de Gibbs caché

Un champ de Gibbs (Besag, 1974) est un modèle probabiliste markovien défini sur un graphe de dépendance $\mathcal{G}$. La vraisemblance du modèle est donnée par

$$\pi(x \mid \beta, \mathcal{G}) = \frac{1}{Z(\beta, \mathcal{G})} \exp\left(\beta^T S(x \mid \mathcal{G})\right),$$

où $\beta \in \mathbb{R}^p$ est un vecteur de paramètres et $S(\cdot)$ est une fonction, appelée potentiel, définie sur les cliques du graphe $\mathcal{G}$ et à valeurs dans $\mathbb{R}^p$. La constante de normalisation $Z(\beta, \mathcal{G})$, appelée fonction de partition, est une somme sur tous les champs latents x possibles, ce qui rend son calcul impossible en pratique du fait de la complexité combinatoire.

Dans les cas d'intérêt, le champ x n'est pas observé directement. On dispose d'une copie bruité y supposée conditionnellement indépendante de x, dont la densité est de la forme

$$f(y \mid x, \theta) = \prod_{i \in \mathcal{S}} f_i(y_i \mid x_i, \theta),$$

où $\mathcal{S}$ désigne l'ensemble des sommets du graphe $\mathcal{G}$, aussi appelés pixels.

3 Critère BIC

Pour choisir parmi une collection de M modèles, l'approche bayésienne standard consiste à sélectionner le modèle ayant la plus grande probabilité *a posteriori*. D'après le théorème de Bayes, cela équivaut, sous l'hypothèse d'un prior uniforme sur l'espace des modèles, à sélectionner le modèle m ayant la vraisemblance intégrée la plus grande

$$\pi(y \mid m) = \int \pi(y \mid \psi, m)\pi(\psi \mid m)\mathrm{d}\psi,$$

où ψ est le paramètre du modèle m et $\pi(\psi \mid m)$ est la loi *a priori* sur l'espace des paramètres du modèle m.

Lorsque y est un champ de Gibbs, cette intégrale ne peut être calculée directement du fait de la constante de normalisation du modèle. Le critère BIC fournit une approximation de cette vraisemblance intégrée basée sur la méthode de Laplace (Schwartz, 1978)

$$2 \log \pi(y \mid m) \simeq \mathrm{BIC}(m) = 2 \log \pi(y \mid \hat{\psi}^{\mathrm{ml}}, m) - d_m \log(|\mathcal{S}|),$$

où $\hat{\psi}^{\mathrm{ml}}$ est le maximum de vraisemblance de $\pi(y \mid \psi, m)$ et d_m est la dimension de l'espace des paramètres du modèle et $|\mathcal{S}|$ le nombre de pixels.

La difficulté vient alors du fait que ni le maximum de vraisemblance $\hat{\psi}^{\mathrm{ml}}$ ni la loi de l'observation $\pi(y \mid \psi, m)$ ne sont accessibles directement car cela nécessite d'évaluer une intégrale sur le latent

$$\pi(y \mid \psi, m) = \int f(y \mid x, \theta)\pi(x \mid \beta, \mathcal{G})\mathrm{d}x. \tag{1}$$

La difficulté du calcul de BIC se résume donc au calcul de (1). Différentes options, pouvant être classées en trois familles, ont été présentées dans la littérature pour approcher cette intégrale sur le latent.

La solution la plus simple à mettre en oeuvre est d'approcher (1) par une méthode de Monte-Carlo. Il suffit pour cela de moyenner le long d'une chaîne de Markov de loi stationnaire $\pi(\cdot \mid \beta, \mathcal{G})$, c'est à dire

$$\pi(\cdot \mid \beta, \mathcal{G}) \simeq \frac{1}{N} \sum_{j=1}^{N} \delta_{x_j}, \text{ où } x_j \sim \pi(\cdot \mid \beta, \mathcal{G}).$$

Cette méthode est très peu avantageuse du fait de son coût de calcul. En effet la taille de l'échantillon MCMC N doit être en pratique très grand afin d'explorer correctement l'espace des champs latents, en particulier ceux qui contribuent le plus à la somme (1).

La seconde famille d'approximations consiste à voir la réalisation sur le graphe $\mathcal{G}$ comme la réalisation d'un modèle de mélange indépendant. On oublie donc toute structure de dépendance induite par $\mathcal{G}$ et la loi du latent est approchée par le produit des lois marginales des sommets du graphe

$$\pi(x \mid \beta, \mathcal{G}) \simeq \prod_{i \in \mathcal{S}} \pi(x_i). \tag{2}$$

Forbes et Peyrard (2003) illustrent les performances d'un tel critère pour le choix du nombre de composantes possibles du latent. En revanche, une telle approximation ne peut être utilisée pour choisir entre plusieurs structures de dépendance.

La troisième famille d'approximations a pour but de se ramener à un système de variables indépendantes pour construire une loi produit approchant au mieux $\pi(x \mid \beta, \mathcal{G})$ au sens de Kullback-Leibler. La solution optimale est obtenue avec le principe de champ moyen issu de la physique statistique. Celeux *et al.* (2003) étendent ce principe en fixant les voisins d'un sommet du graphe $\mathcal{G}$ à des constantes $\tilde{x}$. En toute généralité, cela conduit à l'approximation de la distribution marginale du latent

$$\pi(x \mid \beta, \mathcal{G}) \simeq \prod_{i \in \mathcal{S}} \pi(x_i \mid \beta, \mathcal{G}, \tilde{x}). \tag{3}$$

Il s'agit de l'approche mise en oeuvre par Stanford et Raftery (2002) puis par Forbes et Peyrard (2003). Le critère PLIC (Penalized Pseudolikelihood Information Criterion) proposé par Stanford et Raftery (2002) s'obtient en prenant pour $\tilde{x}$ dans (3) un estimateur ICM (Besag, 1986) du latent. Ce critère est à rapprocher du critère $\mathrm{BIC}^{\tilde{x}}$ de Forbes et Peyrard (2003) obtenu en choisisant pour $\tilde{x}$ une réalisation suivant la loi conditionnelle $\pi(x \mid y, \beta, \mathcal{G})$. Un tel critère ne pénalise néanmoins pas suffisamment les modèles les plus complexes et conduit à des résultats insatisfaisants en pratique. En effet le choix de $\tilde{x} \sim \pi(x \mid y, \beta, \mathcal{G})$ conduit à approcher $\pi(x \mid y, \beta, \mathcal{G})$ et non $\pi(x \mid \beta, \mathcal{G})$. Nous proposons d'approcher la loi marginale du latent par une moyenne de ces approximations en tirant des $\tilde{x}^{(j)}$ suivant $\pi(\cdot \mid \psi^{\mathrm{ml}})$

$$\pi(x \mid \beta, \mathcal{G}) \simeq \frac{1}{N} \sum_{j=1}^{N} \prod_{i \in \mathcal{S}} \pi(x_i \mid \beta, \mathcal{G}, \tilde{x}^{(j)}), \tag{4}$$

Contrairement aux méthodes de Monte-Carlo naïves, N n'a pas besoin d'être grand ici. Au cours de cette présentation, nous discuterons du choix de N et de la loi de proposition des $\tilde{x}$.

4 Extension vers des produits par blocs

L'idée présentée à la section précédente est de remplacer la distribution de Gibbs par la distribution d'un système de variables aléatoires réelles indépendantes afin d'éviter le problème du calcul de la constante de normalisation ou de moyenner de telles approximations. Si le calcul de la constante de normalisation reste impossible en toute généralité, Friel et Rue (2007) proposent un algorithme récursif qui permet de calculer cette constante pour des graphes trivialement petits (*e.g.*, un réseau régulier de taille inférieure 20×20). Partant de ce constat, l'idée générale est de remplacer la loi produit sur les éléments de $\mathcal{S}$ par une loi produit sur une partition de $\mathcal{S}$, *i.e.*, des blocs de petites tailles.

Notons $\mathcal{S} = \sqcup_i A_i$ cette partition. Nous commencerons par étendre le cas du produit des lois marginales sur les sommets (2) à un produit de lois marginales sur des blocs

$$\pi(x \mid \beta, \mathcal{G}) \simeq \prod_i \pi(x_{A_i}).$$

On s'intéressera ensuite à l'approximation obtenue en conditionnant les éléments de la partition au bord

$$\pi(x \mid \beta, \mathcal{G}) \simeq \prod_i \pi(x_{A_i} \mid \beta, \mathcal{G}, \tilde{x}),$$

ce qui généralise (3). Nous présenteront l'influence de telles approximations sur les critères de choix de modèle existants. Deux exemples de choix de modèle seront considérés : la sélection du nombre de composantes de l'état latent et la sélection de la structure de dépendance. On sintéressera en particulier au choix de $\tilde{x}$ et à l'impact de la taille des blocs de la partition sur des exemples d'images simulées et d'images réelles avec bruit simulé.

Bibliographie

[1] Besag, J. (1974), *Spatial interaction and the statistical analysis of lattice systems (with Discussion)*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 36(2), 192–236.

[2] Besag, J. (1986), *On the statistical analysis of dirty pictures*, Journal of the Royal Statistical Society. Series B (Methodological), 259–302.

[3] Celeux, G., Forbes, F., et Peyrard, N. (2003), *EM procedures using mean field-like approximations for Markov model-based image segmentation*, Pattern recognition, 36(1), 131–144.

[4] Cucala, L. et Marin, J. M. (2013), *Bayesian Inference on a Mixture Model With Spatial Dependence*, Journal of Computational and Graphical Statistics, 22(3), 584–597.

[5] Forbes, F. et Peyrard, N. (2003), *Hidden Markov random field model selection criteria based on mean field-like approximations*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 25(9), 1089–1101.

[6] Friel, N. et Rue, H. (2007), *Recursive computing and simulation-free inference for general factorizable models*, Biometrika, 94(3), 661–672.
[7] Qian, W. et Titterington, D. M. (1991), *Estimation of Parameters in Hidden Markov Models*, Philosophical Trans. Royal Soc. London A, 337, 407–428.
[8] Schwartz, G. (1978), *Estimating the Dimension of a Model*, The Annals of Statistics, 6, 461–464.
[9] Stanford, D. C. and Raftery A. E. (2002), *Approximate Bayes factors for image segmentation: The pseudolikelihood information criterion (PLIC)*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(11), 1517–1520.
[10] Stoehr, J., Pudlo, P. et Cucala, L. (2015), *Adaptive ABC model choice and geometric summary statistics for hidden Gibbs random fields*, Statistics and Computing, 25(1), 129–141.