Statistiques résumées géométriques pour le choix de modèle ABC entre des champs de Gibbs cachés

Julien Stoehr [1] & Pierre Pudlo [1] & Lionel Cucala [1]

[1] *I3M UMR CNRS 5149, Université Montpellier 2, Place E. Bataillon 34095 Montpellier CEDEX, France*
julien.stoehr@univ-montp2.fr
pierre.pudlo@univ-montp2.fr
lionel.cucala@univ-montp2.fr

Résumé. Choisir entre différentes structures de dépendance d'un champ de Markov caché est difficile, en raison de la constante de normalisation de la vraisemblance, incalculable explicitement, et de la somme sur tous les champs latents possibles. Les méthodes bayésiennes approchées, aussi connues sous l'acronyme de méthodes ABC, fournissent une procédure de choix de modèle dans le paradigme bayésien. Ces méthodes sont basées sur la comparaison de données observées avec de nombreuses simulations numériques, au travers de statistiques résumées. Lorsque le champ de Gibbs est directement observé, Grelaud *et al.* (2009) exhibent des statistiques résumées exhaustives qui garantissent, de façon immédiate, la consistance de l'algorithme. En revanche, lorsque le champ aléatoire est caché, ces statistiques ne sont plus exhaustives. Nous fournirons de nouvelles statistiques résumées basées sur la géométrie de l'image, et plus précisément sur des clusters de pixels. Afin d'évaluer leur efficacité, nous introduirons un taux d'erreur conditionnel mesurant la puissance locale des algorithmes ABC.

Mots-clés. Méthodes ABC, choix de modèle, champ de Gibbs caché, statistique résumée, taux d'erreur de classification

Abstract. Selecting between different dependence structures of a hidden Markov random field can be very challenging, due to the intractable normalizing constants in the likelihoods and the sum over all possible latent random fields. Approximate Bayesian Computation (ABC) algorithms provide a model choice method in the Bayesian paradigm. The scheme compares the observed data and many numerical simulations through summary statistics. When the Gibbs random field is directly observed, Grelaud *et al.* (2009) exhibit sufficient summary statistics that immediately guarantee the consistency of the ABC algorithm. But, when the random field is hidden, those statistics are not sufficient anymore. We provide new summary statistics based on the geometry of the image, more precisely a clustering analysis of pixels. To assess their efficiency, we also derive a conditional misclassification rate evaluating the local power of ABC algorithms.

Keywords. Approximate Bayesian Computation, model choice, hidden Gibbs random fields, summary statistics, misclassification rate

1 Introduction

Malgré un large éventail d'applications, les champs de Gibbs (Besag, 1974) présentent des difficultés majeures du point de vue de l'inférence (Grelaud *et al.*, 2009 ; Everitt, 2012 ; Cucala et Marin, 2013). Le choix entre deux structures de dépendance différentes peut s'avérer très complexe, en raison des constantes de normalisation des vraisemblances incalculables explicitement, pour tout réseau non trivialement petit. Lorsque le champ de Gibbs est observé, ce problème peut être qualifié de doublement complexe, car la vraisemblance, mais aussi, la probabilité *a posteriori* sont impossibles à expliciter. Dans notre cas, comme dans le cas des modèles de Markov cachés, le champ aléatoire n'est pas directement observé. Le problème présente alors un troisième niveau de complexité.

Il existe actuellement très peu de travaux sur ce problème de choix de modèle (Forbes et Peyrard, 2003 ; Cucala et Marin, 2013). Nous nous intéresserons dans cette présentation aux méthodes ABC qui fournissent une procédure de choix de modèle dans le paradigme bayésien.

L'algorithme ABC compare les données observées y^{obs} avec un grand nombre de simulations numériques y au travers de statistiques résumées $S(y)$. Sa consistance est immédiate lorsque $S(y)$ est exhaustive en ce qui concerne le problème de choix du modèle. Dans le cas contraire, les statistiques doivent être choisies avec soin (Robert *et al.*, 2011 ; Marin et al, 2014). Lorsque le champ de Gibbs est directement observé, Grelaud *et al.* (2009) exhibent des statistiques exhaustives pour le problème de choix de modèle directement basées sur le potentiel des distributions de Gibbs. Toutefois, lorsque le champ aléatoire est caché, cette propriété n'est pas conservée en raison de la perte d'exhausitvité de ces statistiques résumées. Robert *et al.* (2011) ont montré que l'utilisation des méthodes ABC avec des statistiques non exhaustives, pour le problème du choix de modèle bayésien, peut conduire à des résultats erronés Après ce premier avertissement , les auteurs ont développé des résultats théoriques sur la question. En effet Marin *et al.* (2014) fournissent des conditions génériques, beaucoup plus faibles que l'exhaustivité, qui impliquent la consistance de la procédure ABC.

Par ailleurs, Blum (2010) et Fearnhead et Prangle (2012) ont montré que la qualité de l'approximation fournie par ABC diminue avec la dimension de $S(y)$. Ainsi, un compromis doit être trouvé entre statistiques informatives et de petite dimension. D'autres travaux ont été réalisés pour construire ou sélectionner le vecteur $S(y)$ parmi un large ensemble de statistiques résumées (*e.g.*, Blum *et al*, 2013 ; Prangle *et al.*, 2013). En ce qui nous concerne, nous nous intéresserons à la sélection de statistiques résumées les plus informatives possible parmi des ensembles de petite dimension dans le cadre particulier des champs de Potts.

Cette présentation introduira de nouvelles statistiques basées sur la géométrie de l'image, et plus précisément sur les composantes connexes de l'image. Nous évaluerons ensuite leur efficacité du point de vue du choix de modèle ABC entre champs de Gibbs cachés. Les conditions proposées par Marin *et al.* (2014) pour assurer la convergence

étant difficiles à vérifier dans la pratique, nous proposerons une nouvelle méthode basée sur un taux d'erreur de classification conditionnelle pour valider la procédure.

2 Champs de Potts cachés

La vraisemblance du modèle de Potts caché de paramètre β sur le graphe $\mathcal{G}$ et de bruit distribué suivant P_α, noté HPM$(\mathcal{G}, \alpha, \beta)$, est donnée par

$$f(y|\alpha,\beta,\mathcal{G}) = \sum_{x\in\mathscr{X}} \pi\left(x|\mathcal{G},\beta\right)\pi_\alpha(y|x), \text{ où } \pi(x|\mathcal{G},\beta) = \frac{1}{Z(\mathcal{G},\beta)}\exp\left(\beta\sum_{i\overset{\mathcal{G}}{\sim}j}\mathbb{1}\{x_i = x_j\}\right),$$

est la densité du modèle de Potts. La constante de normalisation $Z(\mathcal{G},\beta)$ est une somme sur tous les champs latent x possibles, donc incalculable en pratique. La vraisemblance du modèle caché est aussi fonction de $\pi_\alpha(y|x) = \prod_i P_\alpha(y_i|x_i)$ qui désigne la distribution conditionnelle des observations. Dans cette présentation, nous nous intéresserons au cas particulier d'un modèle de bruit discret donné par

$$P_\alpha(y_i|x_i) = \frac{\exp\left\{\alpha(2\mathbb{1}\{x_i = y_i\} - 1)\right\}}{\exp(\alpha) + (K-1)\exp(-\alpha)}.$$

L'objectif de notre étude est de sélectionner entre deux modèles $\mathcal{M}_4$ et $\mathcal{M}_8$ le champ de Gibbs caché qui correspond le mieux à une image donnée y^{obs}, où $\mathcal{M}_4$ est un HPM$(\mathcal{G}_4, \alpha, \beta)$ avec une loi *a priori* π_4 sur les paramètres (α, β) et $\mathcal{M}_8$ est un HPM$(\mathcal{G}_8, \alpha, \beta)$ avec une loi *a priori* π_8 sur les paramètres (α, β). Les graphes $\mathcal{G}_4$ et $\mathcal{G}_8$ correspondent respectivement aux structures de dépendance des quatre et huit voisins les plus proches.

Le calcul de la probabilité *a posteriori* présente un triple niveaux de complexité : l'intégrale sur l'espace des paramètres Θ_m, et les sommes sur l'ensemble des champs latents possibles intervenant dans les fonctions $Z(\mathcal{G},\beta)$ et $f(y|\alpha,\beta,\mathcal{G})$.

3 Méthodes ABC pour le choix de modèle

Une des réponses à ce problème vient des méthodes bayésienne approchées (Grelaud *et al.*, 2009). ABC simule des données y pour de nombreuses valeurs du paramètre θ_m sous chacun des modèles m (Algorithme 1).

Algorithm 1: Simulation d'un table de référence ABC

Result: Une table de référence de taille n_{REF}
for $j \leftarrow 1$ **to** n_{REF} **do**
  simuler m suivant π;
  simuler $\theta = (\alpha, \beta)$ suivant π_m;
  simuler y suivant la vraisemblance $f_m(\cdot|\theta)$;
  calculer $S(y)$;
  poser $(m^j, \theta^j, S(y^j)) \leftarrow (m, \theta, S(y))$;
end

Parmi les particules simulées de la table de référence, ABC conserve celles issues des modèles sous lesquels $S(y)$ est proche de $S(y^{\text{obs}})$ au sens d'une distance ρ, où S est un vecteur de statistiques résumées. Pour ce faire, on trie les particules m^j de la table de référence suivant $\rho(S(y^j), S(y^{\text{obs}}))$ et on conserve les n_{POST} premières. Notons ϵ le quantile de la distance, *i.e.* la distance de la $n^{\text{ème}}_{\text{POST}}$ particule la plus proche. Les particules acceptées m^j sont distribuées suivant

$$\pi\left(m \;\middle|\; \rho(S(y^j), S(y^{\text{obs}})) < \epsilon\right),$$

et

$$\widehat{\pi}_{ABC}\left(m \;\middle|\; S\left(y^{\text{obs}}\right)\right) = \frac{\sum_{j=1}^{n_{\text{REF}}} \mathbb{1}\{m^j = m, \rho(S(y^j), S(y^{\text{obs}})) < \epsilon\}}{\sum_{j=1}^{n_{\text{REF}}} \mathbb{1}\{\rho(S(y^j), S(y^{\text{obs}})) < \epsilon\}}$$

est un estimateur Monte Carlo de la probabilité *a posteriori* du modèle m.

4 Composantes connexes et taux d'erreur conditionnel

Considérons une image y et un graphe $\mathcal{G}$. Nous définissons le graphe $\Gamma(\mathcal{G}, y)$ induit par $\mathcal{G}$ sur y par: il existe une arête entre les pixels i et j dans $\Gamma(\mathcal{G}, y)$ si, et seulement si, il existe une arête entre i et j dans $\mathcal{G}$ et si les deux pixels partagent la même couleur, *i.e.* $y_i = y_j$.

Nous définissons alors deux statistiques résumées à valeurs dans $\mathbb{N}$: $T(\mathcal{G}, y)$ est le nombre de composantes connexes de $\Gamma(\mathcal{G}, y)$ et $U(\mathcal{G}, y)$ la taille de la plus grosse composante connexe de $\Gamma(\mathcal{G}, y)$.

Étant donné une table de référence et un ensemble de statistiques résumées, pour toute nouvelle observation, ABC peut prédire une indice de modèle $\hat{m}_\ell(y)$ suivant la règle

du maximum *a posteriori* ABC. Confronté à des données observées y^{obs}, l'utilisateur d'un algorithme ABC est intéressé par l'erreur qu'il commet en croyant $\hat{m}_\ell(y^{\text{obs}})$ calculé avec la table de référence dont il dispose, et un ensemble de statistiques résumées donné $S_\ell(y)$. Lors de cet exposé, nous préconiserons l'usage d'un taux d'erreur de classification conditionnel à $S_\ell(y^{\text{obs}})$, à savoir

$$\tau(S_\ell(y^{\text{obs}})) = \sum_m \pi(m|S_\ell(y^{\text{obs}}))\tau(m, S_\ell(y^{\text{obs}})),$$

où $\tau(m, S_\ell(y^{\text{obs}})) = \mathbb{P}\Big(\hat{m}_\ell(Y) \neq m \Big| \rho\left(S_\ell(Y), S_\ell(y^{\text{obs}})\right) \leq \eta\Big)$.

La précision globale de la procédure de choix de modèle ABC basée sur le $\ell^{\text{ième}}$ ensemble de statistiques résumées peut être évaluée avec l'espérance par rapport au modèle bayésien de $\tau(m, S_\ell(y^{\text{obs}}))$, alors que $\tau(S_\ell(y^{\text{obs}}))$ décrit le comportement de la procédure localement autour de y^{obs}. La comparaison de ces taux d'erreur fournit une procédure pour sélectionner quelques statistiques pertinentes, efficaces au voisinage de y^{obs}.

On utilisera deux expériences numériques, l'une pour deux couleurs, l'autre pour 16 couleurs, afin d'illustrer l'efficacité de nos statistiques géométriques d'un point de vue global et local grâce aux taux d'erreur de classification précédents. Les points importants que l'on notera sont une diminution significative de l'erreur intégrée lorsque l'on ajoute des statistiques géométriques et une dépendence locale de l'efficacité de ces méthodes ; la conclusion générale étant l'amélioration notable et quasi systématique des procédures existantes.

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] Blum, M. G. B. (2010), *Approximate Bayesian Computation: A Nonparametric Perspective*, Journal of the American Statistical Association, 105(491), 1178–1187.

[3] Blum, M. G. B., Nunes, M. A., Prangle, D. et Sisson, S. A. (2013), *A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation*, Statistical Science, 28(2), 189–208.

[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] Everitt, R. G. (2012), *Bayesian Parameter Estimation for Latent Markov Random Fields and Social Networks*, Journal of Computational and Graphical Statistics, 21(4), 940–960.

[6] Fearnhead, P. et Prangle, D. (2012), *Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(3), 419–474.

[7] 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.
[8] Grelaud, A., Robert, C. P., Marin, J. M., Rodolphe, F., et Taly, J. F. (2009), *ABC likelihood-free methods for model choice in Gibbs random fields*, Bayesian Analysis, 4(2), 317–336.
[9] Marin, J. M., Pillai, N. S., Robert, C. P. et Rousseau, J. (2014), *Relevant statistics for Bayesian model choice*, To appear in the Journal of the Royal Statistical Society, Series B.
[10] Prangle, D., Fearnhead, P., Cox, M. P., Biggs, P. J. et French N. P. (2013), *Semi-automatic selection of summary statistics for ABC model choice*, Statistical Applications in Genetics and Molecular Biology, 1–16.
[11] Robert, C. P., Cornuet, J. M., Marin, J. M. et Pillai, N. S. (2011), *Lack of confidence in approximate Bayesian computation model choice*, Proceedings of the National Academy of Sciences, 108(37), 15112–15117.