Estimation simultanée des paramètres et détection de structures dans les données spatiales

Sujet de thèse commun proposé par Inria Nancy et l'Université de Lorraine

1 Sujet de thèse

Des exemples typiques de données spatiales sont des images numériques, des données épidémiologiques ou des catalogues de corps célestes en astronomie. L'une des plus fréquente question liée à ces types d'ensembles de données, est la détection et la caractérisation de la structure "cachée" dans les données. Ces modèles peuvent être : la collection de cellules dans une certaine image biologique, l'ensemble d'agrégats manifesté par une maladie étudiée ou le réseau filamenteux délimité par les positions des galaxies observées dans l'univers.
Dans le contexte probabiliste, ce problème est abordé en supposant que le modèle est la réalisation $\boldsymbol{y}$ d'un processus stochastique $\mathbb{Y}$ qui peut être un champ aléatoire ou un processus ponctuel marqué. Dans de nombreuses situations, la modélisation de type Gibbs permet d'écrire une telle densité de probabilité sous la forme

$$p(\boldsymbol{y}|\theta) = \frac{\exp[-U(\boldsymbol{y}|\theta)]}{c(\theta)} \tag{1}$$

où $U : \Omega \to \mathbb{R}^+$ est la fonction d'énergie, θ les paramètres du modèle et $c(\theta)$ la constante de normalisation. La fonction d'énergie peut s'exprimer sous la forme d'une somme

$$U(\boldsymbol{y}|\theta) = U_{\boldsymbol{d}}(\boldsymbol{y}|\theta) + U_i(\boldsymbol{y}|\theta).$$

Le premier terme de la somme est appelé terme d'attache aux données (ou vraisemblance) et est lié au positionnement des objets formant le motif $\boldsymbol{y}$ dans le champ de données spatiales $\boldsymbol{d}$. Le second terme est appelé terme d'interaction et est lié à la structure générale. Ce terme est également interprété comme un contrôle a priori des interactions des objets, générant le motif caché que nous recherchons.

Pour le modèle (1), l'estimateur de motif caché est donné par

$$\widehat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}\in\Omega}\{p(\boldsymbol{y}|\theta)\} = \arg\min_{\boldsymbol{y}\in\Omega}\{U(\boldsymbol{y}|\theta)\}.$$

La formulation duale de la question de détection du modèle est le problème d'estimation de paramètres. Considérons maintenant qu'une structure $\boldsymbol{y}$ est observée. On suppose que l'observation est la réalisation d'un processus ponctuel marqué donné par la densité de probabilité $p(\boldsymbol{y}|\theta)$. Soit $p(\theta|\boldsymbol{y})$ la loi conditionnelle des paramètres du modèle ou de la loi a posteriori

$$p(\theta|\boldsymbol{y}) = \frac{\exp[-U(\boldsymbol{y}|\theta)]p(\theta)}{Z(\boldsymbol{y})c(\theta)} \tag{2}$$

avec $p(\theta)$ la densité a priori pour le paramètres du modèle et $Z(\boldsymbol{y})$ la constante de normalisation.

La maximisation de (2) n'est pas une procédure aisée car elle nécessite l'évaluation du ratio $c(\theta)/c(\psi)$.

L'objectif de cette thèse est de construire un algorithme capable de résoudre simultanément la détection de motifs et l'estimation de paramètres. Dans le cas de champ de Markov aléatoires, ce problème a été abordé par [4, 10, 11, 9]. Le principe de ces méthodes est également connu sous le nom "gradient stochastique". Un lien entre les méthodes basées sur le gradient stochastique et les algorithmes EM a été établi par [4, 2]. Le gradient stochastique a été appliqué pour l'estimation des paramètres des processus ponctuels par [6, 3].

Notre objectif est d'étudier l'approche de [4, 10] afin de caractériser son application à notre méthodologie de détection de structure. Cela demande une adaptation de cette approche aux processus ponctuels.

Pour une lecture plus approfondie sur les processus ponctuels marqués, nous recommandons [1, 8, 5], alors que pour les applications en analyse d'images, cosmologie et sciences de l'environnement, nous sugérons [7] et les références incluses.

2 Encadrants

Madalina Deaconu
Chargé de recherche

Inria Nancy - Grand Est
madalina.deaconu@inria.fr

Radu S. Stoica
Professeur
Université de Lorraine
radu-stefan.stoica@univ-lorraine.fr

3 Prérequis

La durée de la thèse est de trois ans. L'étudiant doit posséder de solides compétences en mathématiques appliquées, en particulier en statistiques et probabilités, tout en étant fortement motivé par des applications pratiques. La connaissance rigoureuse d'un langage de programmation orienté objet tel que C ++ est également nécessaire. Être familier avec le logiciel mathématique tel que Matlab, Scilab ou R, serait un plus.

Références

[1] S. N. Chiu, D. Stoyan, W. S. Kendall, et J. Mecke. *Stochastic Geometry and its Applications. Third Edition.* John Wiley and Sons, 2013.

[2] B. Delyon, M. Lavielle, et E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1) :94–128, 1999.

[3] C. J. Geyer. Likelihood inference for spatial point processes. In O. Barndorff-Nielsen, W.S. Kendall, et M.N.M. van Lieshout, editors, *Stochastic Geometry, Likelihood and Computation.* CRC Press/Chapman and Hall, Boca Raton, 1999.

[4] S. Lakshmanan et H. Derrin. Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8) :799–813, 1989.

[5] J. Møller et R. P. Waagepetersen. *Statistical inference and simulation for spatial point processes.* Chapman and Hall/CRC, Boca Raton, 2004.

[6] R. A. Moyeed et A. J. Baddeley. Stochastic approximation of the MLE for a spatial point pattern. *Scandinavian Journal of Statistics*, 18 :39–50, 1991.

[7] R. S. Stoica. *Modélisation probabiliste et inférence statistique pour l'analyse des données spatialisées.* Habilitation à diriger des recherches - Université de Lille, 2014.

[8] M. N. M. van Lieshout. *Markov Point Processes and their Applications.* Imperial College Press, London, 2000.

[9] G. Winkler. *Image analysis, random fields and Markov chain Monte Carlo methods (second edition).* Springer, 2003.

[10] L. Younes. Estimation and annealing for gibbsian fields. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 24 :269–294, 1988.

[11] L. Younes. Parametric inference for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields*, 82 :625–645, 1989.