

Estimation Monte Carlo dans les processus ponctuels marqués en biologie tissulaire

Mathieu Emily, Radu Stoica

► **To cite this version:**

Mathieu Emily, Radu Stoica. Estimation Monte Carlo dans les processus ponctuels marqués en biologie tissulaire. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France, France. 2009. <inria-00386607>

HAL Id: inria-00386607

<https://hal.inria.fr/inria-00386607>

Submitted on 22 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ESTIMATION MONTE CARLO DANS LES PROCESSUS PONCTUELS MARQUÉS EN BIOLOGIE TISSULAIRE

Mathieu Emily* & Radu Stefan Stoica†

* *Laboratoire de Statistique, Université Haute Bretagne, Place du Recteur H. Le Moal, CS 24307, 35043 Rennes Cedex*

† *Laboratoire Paul Painlevé, Université Lille 1, 59655 Villeneuve d'Ascq Cedex*

Résumé

Nous proposons une approche de modélisation par processus ponctuels marqués pour l'analyse de clichés de tissus biologiques. Une telle modélisation présente de nombreux avantages. Premièrement, l'ensemble des contraintes physiologiques et mécaniques, relatives aux tissus, peuvent être intégrées naturellement dans une densité de probabilité de type gibbsien. Ensuite, ce formalisme est parfaitement adapté à l'inférence paramétrique de type Monte Carlo. Il est possible ainsi, de caractériser statistiquement la morphologie des structures géométriques complexes. Notre but est d'utiliser ce modèle sur des données réelles, représentant le développement d'un tissu pré-tumoral.

Abstract

We propose a marked point process approach to model a stained biological tissue. These models allow to take into account the tissue physiological and mechanical constraints, by means of a Gibbsian probability density. Then, parameter estimation can be achieved using Monte Carlo based inference. Thus, this framework allows statistical description of the morphology of complex structures. Our goal is to apply the model on real data, representing the behaviour of a pre-tumoral tissue.

Mot-clés: Processus ponctuels marqués, processus de Gibbs, estimation de paramètres, simulation et inférence Monte Carlo, tissu biologique

1 Problématique

L'étude d'une coupe histologique est un acte médical réalisé par un oeil d'expert. Cependant, comme le montre l'exemple de la Figure 1, certains tissus présentent une organisation spatiale spécifique. Sur cette figure nous constatons que le tissu est composé de cellules de deux types différents: des cellules claires et des cellules foncées. L'observation de ce tissu fait apparaître clairement un agrégat de cellules foncées au milieu. Cet amas est entouré de cellules blanches. Nous constatons également la présence de petits agrégats de cellules blanches à l'intérieur de l'amas de cellules foncées. Par cet exemple, nous pouvons donc extraire trois caractéristiques tissulaires principales: (i) agrégats de cellules voisines de même type, (ii) taille d'agrégats variables, (iii) cellules de forme relativement homogène.

Afin d'aider l'acte médical par une approche complémentaire de diagnostic, nous utilisons le formalisme des processus ponctuels marqués pour modéliser le développement d'un tissu biologique. Ces processus sont particulièrement adaptés à l'étude de particules en interaction (telles des cellules biologiques dans notre cas). Depuis une vingtaine d'années, ce type de processus a trouvé de nombreuses applications notamment en écologie (Kuuluvainen et Pukkala (1989), Kühlmann-Berenzon et al. (2005)), dans l'étude de la distribution d'arbres dans une forêt (Penttinen et al. (1992), Stoyan et Stoyan (1998), Picard et al. (2008)), ainsi qu'en reconnaissance de formes (Stoica et al. (2004) (2007a) (2007b)). Récemment, une telle modélisation a également été développée pour estimer les interactions entre les cellules d'un tissu biologique (Emily et François (2007)).

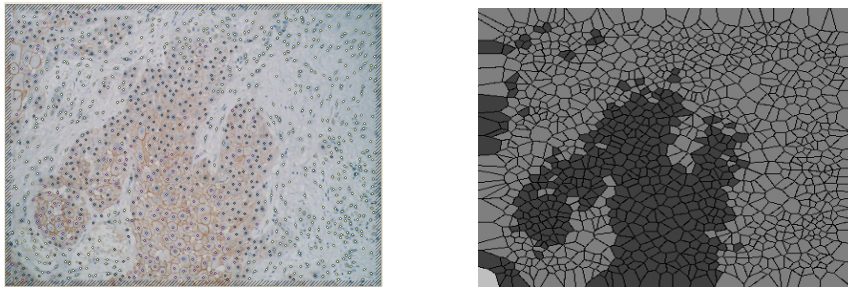


Figure 1: *Exemple de coupe histologique d'un tissu. Figure de gauche: image réelle. Figure de droite: tessellation de Voronoï à partir des noyaux et des types des cellules.*

Dans cet article, nous proposons une extension de ce dernier modèle pour prendre en compte les caractéristiques observées sur la Figure 1. Un algorithme de type Metropolis-Hastings est construit pour la simulation du modèle. L'estimation des paramètres est faite à partir de l'approximation Monte Carlo de la fonction de vraisemblance. Cette procédure sera testée sur des données simulées ainsi que sur un ensemble de données des tissus cancéreux.

2 Modèle de tissu biologique

Soit $K \subset \mathbb{R}^2$ un domaine d'observation de volume $0 < \nu(K) < \infty$. Un processus ponctuel simple sur K est une configuration finie de points $\{k_i \in K, i = 1, \dots, n\}$ telle que $k_i \neq k_j$ pour $i \neq j$. Pour former des objets plus complexes, nous pouvons attacher des caractéristiques ou des marques à ces points. Soit (M, \mathcal{M}, ν_M) l'espace de probabilité qui décrit les marques. Une configuration aléatoire finie de points marqués est un processus ponctuel marqué si le processus des positions des objets seulement, est un processus ponctuel simple.

Dans notre cas, les points correspondent aux positions des noyaux cellulaires dans une région tissulaire alors que les marques sont représentées par le codage de deux types de cellules, pour lequel on choisit naturellement $M = \{0, 1\}$. Ainsi une telle configuration de cellules est donnée par $\underline{\varphi} = \{(k_1, m_1), \dots, (k_n, m_n)\}$, avec $k_i \in K$ et $m_i \in M$, la position du noyau et le type de cellule associé, respectivement, pour tout $1 \leq i \leq n$.

Le plus simple processus ponctuel marqué dans K est le processus ponctuel marqué de Poisson d'intensité unité. Ainsi, le nombre d'objets est choisi par rapport à une loi de Poisson de paramètre $\nu(K)$, ensuite les points représentant les positions des objets sont uniformément distribués dans K . Les marques sont choisies indépendamment pour chaque point, selon ν_M . Des modèles plus réalistes peuvent être construits en précisant une densité de probabilité. Pour une présentation exhaustive des processus ponctuels marqués, le lecteur est renvoyé à la lecture de (Stoyan *et al.* (1995), Van Lieshout (2000) et Daley et Vere-Jones (2003)).

Le processus ponctuel marqué de Poisson n'est pas adapté à notre problème car l'organisation structurée des cellules sous-entend une interaction entre celles-ci. Par conséquent, nous proposons un modèle de processus ponctuel marqué de type plus proche voisin défini comme une perturbation du processus de Poisson. Cette perturbation se traduit par une densité de probabilité prenant en compte ces interactions cellulaires, et qui a l'expression suivante :

$$\forall \underline{\varphi} \in K \times M \quad f(\underline{\varphi}, \Theta) = \frac{h(\underline{\varphi}, \Theta)}{Z(\Theta)} = \frac{\exp(-H(\underline{\varphi}, \Theta))}{Z(\Theta)},$$

avec

$$H(\underline{\varphi}, \Theta) = \theta_1 \sum_{i=1}^n p(k_i, m_i) + \theta_2 \sum_{i \sim j} q(m_i, m_j) + \theta_3 \sum_{i=1}^n r(k_i, m_i). \quad (1)$$

Le vecteur $\Theta = \{\theta_1, \theta_2, \theta_3\}$ correspond à l'ensemble des paramètres, $Z(\Theta)$ est la constante de normalisation de la densité ainsi la fonction h représente la densité non-normalisée. La fonction H , nommée fonction d'énergie, permet de modéliser explicitement les trois caractéristiques principales décrivant l'organisation spatiale d'un tissu biologique tel que celui observé dans la Figure 1. Plus précisément, p correspond à la fonction de contrainte de forme pour chaque cellule i du tissu. Le symbole $i \sim j$ signifiant que les cellules

i et j sont voisines au sens de Voronoï, la fonction q modélise les interactions qui existent entre cellules voisines. Enfin la fonction r permet de considérer la non-homogénéité des clusters formés par les cellules.

3 Simulation et estimation

La principale difficulté dans l'étude des processus ponctuels marqués réside dans le fait que la constante de normalisation, $Z(\Theta)$ est incalculable. Cependant, il est tout à fait possible d'utiliser des techniques MCMC pour la simulation et l'estimation de paramètres pour contourner cette difficulté.

Simulation. Pour la simulation de notre processus, l'utilisation d'une topologie de voisinage de Voronoï et de fonctions d'énergie à portée finie garantissent la convergence d'une dynamique de Metropolis-Hastings. Cet algorithme a été implémenté et certaines propriétés géométriques du diagramme de Voronoï ont été utilisées pour accélérer la convergence (Emily et François (2007)).

Estimation. Afin d'estimer les paramètres du modèle $\Theta = \{\theta_1, \theta_2, \theta_3\}$, nous utilisons le maximum de vraisemblance par méthode Monte Carlo. Ceci revient à approximer la fonction de vraisemblance originelle par son équivalent Monte Carlo et maximiser cette nouvelle fonction à la place (Geyer (1999)). En effet en considérant un jeu de paramètres connu Θ_0 , on peut montrer que la log-vraisemblance, $l(\Theta, \underline{\varphi})$ associée à un jeu de paramètres Θ , pour une configuration $\underline{\varphi}$ s'écrit:

$$l(\Theta, \underline{\varphi}) = \log \left(\frac{h(\underline{\varphi}, \Theta)}{h(\underline{\varphi}, \Theta_0)} \right) - \log \left(\mathbb{E} \left(\frac{h(\underline{\psi}, \Theta)}{h(\underline{\psi}, \Theta_0)} \right) \right).$$

En utilisant l'approximation Monte Carlo

$$\mathbb{E} \left(\frac{h(\underline{\psi}, \Theta)}{h(\underline{\psi}, \Theta_0)} \right) \sim \frac{1}{n} \sum_{i=1}^n \frac{h(\underline{\psi}_i, \Theta)}{h(\underline{\psi}_i, \Theta_0)},$$

avec $\underline{\psi}_i$ des configurations simulées sous Θ_0 , nous obtenons l'équivalent Monte Carlo de la fonction de vraisemblance. Il est ainsi possible de proposer comme estimateur pour les paramètres du modèle, le maximum de la vraisemblance approximée.

4 Conclusion et perspectives

Un échantillon simulé du modèle est présentée à la Figure 2. Cette simulation est obtenue suite à 20.000 itérations de l'algorithme de Metropolis-Hastings, avec une configuration

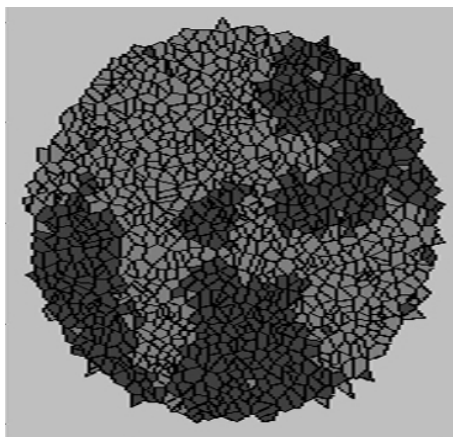


Figure 2: Exemple de simulation du modèle proposé

initiale poissonnienne dans le domaine K , ici un disque centré dans l'origine et de rayon 10. Le jeu de paramètres utilisés a été $(\theta_1, \theta_2, \theta_3) = (10, 0.3, 0.01)$.

La procédure statistique décrite à la section précédente nécessite de fixer un jeu de paramètres Θ_0 . Bien que le choix de ce jeu de paramètre soit critique, nous avons pu constater, à partir de simulation, qu'un choix de Θ_0 proche du vrai Θ est satisfaisant.

Nous souhaitons par la suite, utiliser le modèle comme une aide au diagnostic. En effet l'estimation des paramètres peut être utilisée pour détecter des comportements tissulaires "anormaux". Nous savons qu'un clone tumoral se développe de façon anarchique, modifiant ainsi la structure type existante. Par notre méthodologie, nous souhaitons classifier des tissus en utilisant les paramètres du modèle. Ainsi, le modèle pourrait mettre en évidence différents caractères du développement cellulaire comme par exemple, tissu sain ou tissu pré-tumoral.

Bibliographie

- [1] Kuuluvainen, T. et Pukkala T. (1989), Simulation of within-tree and between-tree shading of direct radiation in a forest canopy: effect of crown shape and sun elevation, *Ecological modelling*, 49, 89–100.
- [2] Kühlmann-Berenzon, S., Heikkinen, J., Sarkka, A. (2005) An Additive Edge Correction for the Influence Potential of Trees, *Biometrical Journal*, 47, 517–526.
- [3] Penttinen, A. K., Stoyan, D. et Henttonen, H. M. (1992). Marked point processes in forest statistics. *Forest Sci.*, 38, 806–824.
- [4] Stoyan, D. et Stoyan, H. (1998). Non-Homogeneous Gibbs Process Models for Forestry - A Case Study, *Biometrical Journal*, 40, 521–531.
- [5] Picard, N., Bar-Hen, A., Mortier, F., Chadoeuf J. (2008). The Multi-scale Marked Area-interaction Point Process: A Model for the Spatial Pattern of Trees. Published online in *Scandinavian Journal of Statistics*.

- [6] Stoica, R. S., Descombes, X. et Zerubia, J. (2004). A Gibbs point process for road extraction in remotely sensed images. *International Journal of Computer Vision*, 57(2), 121–136.
- [7] Stoica R. S., Martinez V. et E. Saar. A (2007a) Three dimensional object point process for detection of cosmic filaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(4), 459–477.
- [8] Stoica, R. S. , Gay E. et Kretzschmar A. (2007b) Cluster detection in spatial data based on Monte Carlo inference. *Biometrical Journal*, 49(4), 505–519.
- [9] Emily, M. et François O. (2007). A statistical approach to estimating the strength of cell-cell interactions under the differential adhesion hypothesis. *Theoretical Biology and Medical Modelling*, 4:37.
- [10] Daley, D. J. et Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes* , vol. I: Elementary Theory and Methods. 2nd edn. Springer, New York.
- [11] Van Lieshout M. N. M. (2000). *Markov point processes and their applications*, Imperial College Press/World Scientific Publishing, London/Singapore.
- [12] Stoyan D., Kendall W. S. et Mecke J. (1995) *Stochastic geometry and its applications*, John Wiley and Sons.
- [13] Geyer, C. J. (1999) *Likelihood inference for spatial point processes*, in *Stochastic geometry, likelihood and computation*, eds O. Barndorff-Nielsen and W. S. Kendall and M. N. M. van Lieshout, CRC Press/Chapman and Hall, Boca Raton.