Simultaneous model and pattern learning in spatial data

Joint PhD subject proposed by Inria Nancy and Université de Lorraine

1 PhD Subject

Typical examples of spatial data are digital images, epidemiological data or catalogues of celestial bodies in astronomy. One of the most frequent questions related to these types of data sets is the detection and the characterization of the "hidden" pattern in the data. Such patterns may be the collection of cells in some biological images, the set of clusters exhibited by a surveyed disease or the filamentary network outlined by the galaxy positions within the observed Universe.

Within a probabilistic context, this problem is tackled by assuming the pattern to be the outcome $\boldsymbol{y}$ of a model $\mathbb{Y}$ such as a random field, a random graph or a marked point process. In many situations, the Gibbsian framework allows to write a probability density describing it:

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

with $U : \Omega \to \mathbb{R}^+$ the energy function, θ the model parameters and $c(\theta)$ the normalising constant. The energy function can be written as the sum

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

The first term in the sum is called data (or likelihood) term and is related to the positioning of the objects forming the pattern $\boldsymbol{y}$ in the spatial data field $\boldsymbol{d}$. The second term is called interaction term and is related to the general structure of the pattern. This term is also interpreted as a prior controlling the objects interactions generating the hidden pattern we are looking for.

Being in the posession of a model (1), the hidden pattern estimator is given by

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

The dual formulation of the pattern detection question is the estimation of the parameters of the assumed model. Let us now consider that an object

pattern $\boldsymbol{y}$ is observed. The observed pattern is supposed to be the realisation of a probabilistic model given by the probability density $p(\boldsymbol{y}|\theta)$. Let $p(\theta|\boldsymbol{y})$ be the conditional distribution of the model parameters or the posterior law

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

with $p(\theta)$ the prior density for the model parameters and $Z(\boldsymbol{y})$ the normalisation constant.

The maximisation of (2) is not a straightforward procedure since it requires the evaluation of the ratio $c(\theta)/c(\psi)$.

The aim of this thesis is to derive an algorithm able to solve simultaneously pattern and parameter learning. In the case of Markov random fields, this problem was tackled by [4, 10, 11, 9]. The principle of these methods is also known under the name "stochastic gradient". A link between stochastic gradient based methods and EM algorithms was established by [4, 2]. The stochastic gradient was applied for estimation parameters of point processes by [6, 3].

Our aim is to study the approach of [4, 10] in order to investigate its impact for proposing new methodological tools for pattern detection in spatial data. Depending on the application domain, this may demand the adaptation of this approach to marked point processes. For further reading on marked point processes we recommend [1, 8, 5], while for applications in image analysis, cosmology and environmental sciences we suggest [7] and the references included.

2 Directors

Madalina Deaconu
Researcher Inria Nancy
madalina.deaconu@inria.fr
http://www.iecl.univ-lorraine.fr/ Madalina.Deaconu/

Radu S. Stoica
Professor Université de Lorraine
radu-stefan.stoica@univ-lorraine.fr
https://sites.google.com/site/radustefanstoica/

3 PhD candidate

The ideal candidate possesses excellent skills in applied mathematics, especially in probability and statistics, while being highly motivated by practical applications. The knowledge of an object oriented programming language such C++ or being familiar with mathematical software such as Matlab, Scilab or R, are skills that may be also interesting.

References

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

[2] B. Delyon, M. Lavielle, and 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, and M.N.M. van Lieshout, editors, *Stochastic Geometry, Likelihood and Computation.* CRC Press/Chapman and Hall, Boca Raton, 1999.

[4] S. Lakshmanan and 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 and R. P. Waagepetersen. *Statistical inference and simulation for spatial point processes.* Chapman and Hall/CRC, Boca Raton, 2004.

[6] R. A. Moyeed and 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.