Pattern detection and characterization for astronomical data through probabilistic modelling and statistical inference

Radu Stoica, Shuyan Liu, Lauri Juhan Liivamägi, Enn Saar, Elmo Tempel, Florent Deleflie, Marc Fouchard, Daniel Hestroffer, Irina Kovalenko, Alain Vienne

Pattern detection and characterization for astronomical data through probabilistic modelling and statistical inference

Radu S. Stoica[1,3], S. Liu[2], L. J. Liivamägi[4,6], E. Saar[4,7] E. Tempel[4], F. Deleflie[3], M. Fouchard[1,3], D. Hestroffer[3], I. Kovalenko[3], A. Vienne[1,3]

[1]Université de Lille, France

[2]Université Paris 1 Panthéon Sorbonne, France

[3]Institut de Mécanique Céleste et Calcul d'Ephémérides, Observatoire de Paris, France

[4]Tartu Observatory, Estonia

[6]Institute of Physics, University of Tartu, Estonia

[7]Estonian Academy of Sciences, Estonia

Abstract

The paper presents several problems coming from astronomy that may be tackled using probability theory and statistics. Due to the nature of data the common question of these problems is what is the pattern hidden in the data. The probabilistic framework allows a statistical and morphological description of these patterns.

Keywords: spatial Markov models, MCMC simulation, statistical Bayesian inference.

1. Introduction. Spatial data are sets of observations made of elements having two components : location and characteristic. The location component gives the coordinates where the observation took place. The characteristic component is represented usually by a multi-dimensional real vector containing measures associated to an astronomical object at the corresponding location. Astronomical data is a very good example of spatial data. Depending, on the studied object, examples of such characteristics are : mass, radius, luminosity, morphological type, brightness, time occurrence, age, light curve, temperature, etc.

In many situations, the spatial character of the data induces a strong morphological component to the possible answers that might be given to questions issued from the data analysis. Therefore, the question almost always arising is what is the pattern hidden in the data ?

This paper shows several examples of how this questions arises in different cases of astronomical data analysis. One common point of these questions is that their respective answers are based on the probability and statistics theory.

The probabilistic framework allows the construction of powerful methodologies based on the following steps : data observation and exploratory analysis, hypothesis and model formulation, building simulation algorithms, statistical inference and hypothesis verification. The main advantage of using the probabilistic approach is that the local exploratory analysis of the data can be naturally transformed in rich information. This transformation is achieved by means of the integration mechanism implicitly given by the probability density

describing the model.

Let $\boldsymbol{d}$ be the observed data set and let us assume that the pattern we are looking for it is the realisation of a stochastic model described by the probability density

$$p(\boldsymbol{x}, \theta|\boldsymbol{d}) = p(\boldsymbol{x}|\boldsymbol{d}, \theta)p(\theta). \tag{1}$$

with $p(\boldsymbol{x}|\boldsymbol{d}, \theta)$ and $p(\theta)$, the conditional law of the pattern and the prior density for the model parameters, respectively.

The conditional law in (1) can be written as

$$p(\boldsymbol{x}|\boldsymbol{d}, \theta) = \exp\left[-\frac{U_{\boldsymbol{d}}(\boldsymbol{x}|\theta) + U_i(\boldsymbol{x}|\theta)}{Z_{\boldsymbol{d}}(\theta)}\right] \tag{2}$$

where the term $U_{\boldsymbol{d}}(\boldsymbol{x}|\theta)$ is called the data energy, the term $U_i(\boldsymbol{x}|\theta)$ is the interaction energy. The interaction energy builds the pattern, while the data energy position the pattern in the data field.

The key hypothesis in defining the pattern model (2) is that the pattern $\boldsymbol{x}$ is a complex entity made of rather simple objects $\{x_1, x_2, \ldots, x_n\}$ that interact. Typical example of interactions are attraction or repulsion of two objects if the objects are enough close. There is a lot of freedom in defining such interactions, provided the model (2) is well defined. Depending on the type of interaction the objects exhibit, the pattern is built from local interactions, the model becomes Markovian. This may lead to the writing of the energy functions in a simplified form, known under the name of Hammersley-Clifford factorization [5, 17].

The main challenge to be overcome whenever adopting this modelling framework is that the normalizing constant in (2) given by $Z_{\boldsymbol{d}}(\theta)$ cannot be always computed in an analytical close form. The solution to this drawback is to build MCMC simulation algorithms in order to do statistical inference. If an appropriate optimization technique is available, the hidden pattern in the data is given by the joint estimator :

$$(\widehat{\boldsymbol{x}}, \widehat{\theta}) = \arg\min_{\Omega\times\Psi}\left\{\frac{U_{\boldsymbol{d}}(\boldsymbol{x}|\theta) + U_i(\boldsymbol{x}|\theta)}{Z_{\boldsymbol{d}}(\theta)} - \log p(\theta)\right\}$$

with Ω and Ψ the corresponding state spaces for the pattern and the parameters, respectively.

2. Marked point processes for studying galaxy structure distribution. Marked point process are probabilistic models dealing with configurations of random points having random characteristics or marks [1, 17, 5]. By their definition, these processes are particularly adapted for analysing spatial point patterns [3].

It is a fact, that at very large scale, the galaxy distribution is not uniform [4]. The galaxy positions are spread in our Universe forming intricate structures such as clusters, walls and filaments. The galactic filamentary pattern can be seen as the realisation of a marked point process whose realisations are configuration of segments that align and connect.

The Bisous model is a marked point process constructed for modelling such patterns [8]. It was successfully applied for detecting them and for characterizing the filaments from a statistical and morphological perspective [10, 11, 16]. The filamentary pattern is estimated by the segments configuration maximising the probability density describing the model. For this purpose a simulated annealing procedure based on a tailored Metropolis-Hastings procedure was built. Non-informative priors were chosen for the model parameters. The morphological characterisation of the filamentary pattern was done using the sufficient statistics of the model : the total number of segments, the number of segments connected at one extremity and the number of segments connected at both extremities. Several other cosmological questions were addressed using the Bisous model. In [15, 13] it was shown that the spin axes of spiral galaxies are preferentially aligned with galactic filaments, whereas the minor semi-axes of elliptical galaxies are preferentially perpendicular

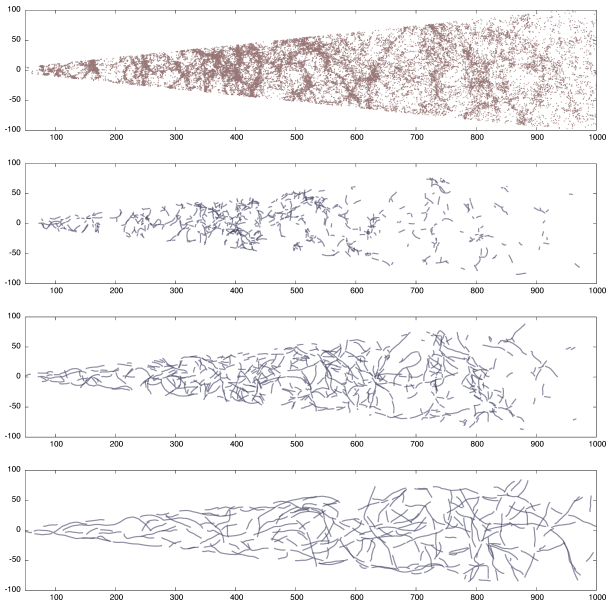

Figure 1: Example of multi-scale filaments in GAMA data : the model proposed in [16] in order to state the existence of cosmic filaments at different resolution levels.

to filaments. These results indicate that spiral galaxies form through peaceful accretion of matter that is falling to filaments. The formation of elliptical galaxies is dominated by galaxy mergers that happen along the filaments. The paper [14] investigated the connection between Bisous filaments and underlying velocity field using N-body simulations. It was showed that Bisous filaments detected purely from dark matter halo distribution is very well aligned with underlying velocity field. It confirms that the filaments detected in galaxy/halo distribution are physical systems, not only visual structures. The distribution of galaxies along the filaments was analysed in [12]. For that, the two point correlation function was used. It was showed that the galaxy distribution along filaments is not uniform : there is regularity in this galaxy distribution, where the distance between galaxies is roughly 10 Mpc. The Figure 1 shows a detection result on a sample of the GAMA (Galaxy And Mass Assembly) survey.

3. Spatial statistics for studying spatial debris. Spatial debris are small celestial bodies produced by the artificial satellites around the Earth. These debris are pieces that can be detached from a satellite by an obsolescence effect or that originate from an accidental collision of two satellites. Our undergoing project is to study the debris distribution using statistics for spatio-temporal point processes [2]. The main question these tools may answer is to check whether these debris behaviour is completely random, repulsive

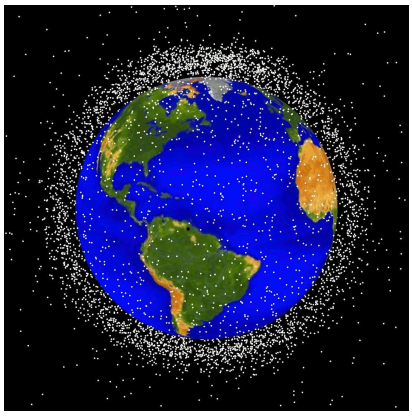

Figure 2: Cloud of spatial debris around the Earth.

or clustering. The most challenging problem is to use spatial statistics and modelling to detect the regions with a high probability for the collision of the existing debris with a new satellite. The Figure 2 illustrates the debris distributions around the Earth.

4. Heavy tail distributions for studying the Öpik-Oort cloud comets planetary perturbations. The comets, in the Solar System, have very elongated orbits. Due to this property, a comet may come close to a giant planet (Jupiter, Saturn, Uranus or Neptune). Such a close encounter with a planet is able to modify significantly the comet trajectory. This mechanism is the main factor responsible of the transport of the comet in the Solar System. Since, this is a highly chaotic mechanism, any individual motion cannot be truthfully modelled over few encounters with any planet. An alternative solution to this problem is to consider probabilistic modelling and statistical inference.

In [9], such a study has been initiated. A sample of Öpik-Oort cloud comets affected by the four giant planets was analysed. The considered phase space was given by the perihelion distance and the ecliptical inclination of the comets trajectories. The obtained results indicated that heavy tail distributions should be used to model the planetary perturbations around the trajectories of the giant planets. The perturbations in this region of the phase space were forming a spatial pattern that was naturally explained by the Öpik theory [6]. Actually, another study on a simplified model is carried out. This study aims to understand the separate effect of each planet on the trajectory of a comet. A partial result of this study is showed in Figure 3. For the perturbations with a perihelion distance of 5.1 A.U., perihelion argument, w, and inclination angle, i a parameter estimation à la [9] was done. The obtained results exhibit a pattern made of two "arrows" oriented from left to right. The symmetry axes of these two arrows are rather close to a perihelion argument of $w = 20$ and $w = 160$ degrees, respectively. These values correspond to a trajectory crossing the Jupiter's orbit. Hence, the perturbations distributions in this region tend to exhibit a heavy tail character. The final conclusions of these two studies should be used to obtain a general probabilistic model for the planetary perturbations.

5. MCMC Bayesian inference for studying orbit detection. The study of binary asteroids is of particular interest. The process of their formation and evolution have began since the formation of the Solar System. Therefore, the detection of large number of them and their study may have important implications for understanding and verifications of theoretical models of dynamical evolution of the Solar System.

Furthermore, the study of binary asteroids is important to prevent possible collisions with the Earth. Orbit determination is one of the classical problems of celestial mechanics. This is an inverse problem : determine the orbital parameters from observations. Moreover, calculate the orbit of a single asteroid with great accu-

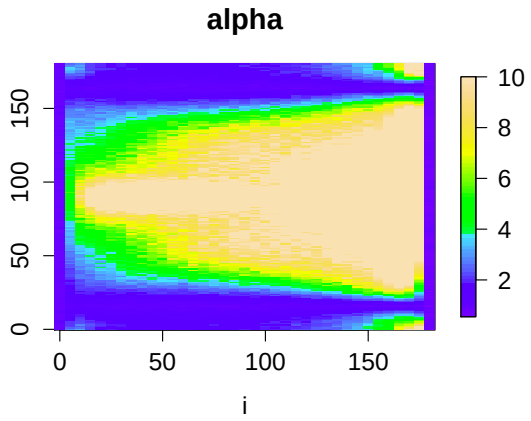

Figure 3: Tail exponent for planetary distributions for perturbation with a perihelion distance of 5.1 A.U.. The x and y axes represent, the inclination angle and the perihelion argument, respectively. The values lower than 2 indicate a heavy tail character of the distribution.

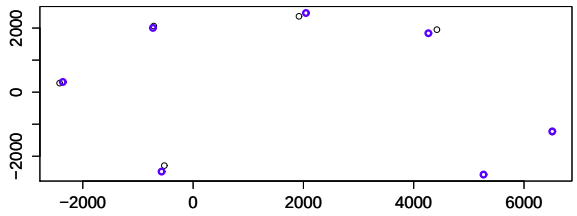

Figure 4: Orbit detection for the 66652 Borasisi binary system : observed positions (black) and computed positions (blue) with the parameters of the detected orbit.

racy is much easier than the orbit of binaries. This problem is more complicated due to the relative motion of its components. In addition, the determined orbit allows to derive essential physical characteristics of the considered objects, such as mass and density.

The statistical ranging method for heliocentric orbit of simple asteroid has been investigated by [18]. For binary asteroids relative orbit determination the method was applied by [7]. For their purpose, the authors used Bayesian modelling and MCMC inference. Our current approach is work in progress. The Figure 4 shows a partial result of our method. We aim to continue on the directions opened by previously cited authors. Nevertheless, the main distinction of our method consists in the modelling choice : we sample directly the orbital parameters conditioned by the observation, while in the previous methods the orbits were derived through the sampling of observational coordinates.

6. Conclusions. The present paper shows by means of several examples, the key role that probability and statistical methodology plays in analysing astronomical data. The choice of the mathematical models depends on the problems to be tackled : marked point processes, regular variation tails distributions, spatio-temporal processes and many others. Almost all these models need special algorithms to be simulated. In most of

the cases, the models are sampled using a MCMC algorithm. The mathematical problems to be solved are typical for the probabilistic framework : existence and properties of the models, convergence properties of the simulation algorithms and precision of the inference procedures built for these models. The common point of all these problems is that the solution we are looking for is represented by a pattern in a chaotic dynamics.

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] A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp (eds.). *Handbook of Spatial Statistics.* Chapman and Hall/CRC, Boca Raton, 2010.

[3] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns.* John Wiley and Sons, 2008.

[4] V. J. Martinez and E. Saar. *Statistics of the galaxy distribution.* Chapman and Hall, 2002.

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

[6] E. J. Öpik. *Interplanetary encounters : close-range gravitational interactions.* Elsevier Science, 1976.

[7] D. Oszkiewicz, K. Muinonen, J. Virtanen, and M. Granvik. Asteroid orbital ranging using markov-chain monte carlo. *Meteoritics and Planetary Science*, 44:1897–1904, 2009.

[8] R. S. Stoica, P. Gregori, and J. Mateu. Simulated annealing and object point processes : tools for analysis of spatial patterns. *Stochastic Processes and their Applications*, 115:1860–1882, 2005.

[9] R. S. Stoica, S. Liu, Yu. Davydov, M. Fouchard, A. Vienne, and G. B. Valsecchi. Order statistics and heavy-tailed distributions for planetary perturbations on Oort cloud comets. *Astronomy and Astrophysics*, 513(A14):1–9, 2010.

[10] R. S. Stoica, V. J. Martinez, and E. Saar. A three dimensional object point process for detection of cosmic filaments. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 55:189–205, 2007.

[11] R. S. Stoica, V. J. Martinez, and E. Saar. Filaments in observed and mock galaxy catalogues. *Astronomy and Astrophysics*, 510(A38):1–12, 2010.

[12] E. Tempel, R. Kipper, E. Saar, M. Bussov, A. Hektor, and J. Pelt. Galaxy filaments as pearl necklaces. *Astronomy and Astrophysics*, 572:A8, 2014.

[13] E. Tempel and N. I. Libeskind. Galaxy spin alignment in filaments and sheets: observational evidence. *The Astrophysical Journal Letters*, 775:L42, 2013.

[14] E. Tempel, N. I. Libeskind, Y. Hoffman, L. J. Liivamägi, and A. Tamm. Orientation of cosmic web filaments with respect to the underlying velocity field. *Monthly Notices of the Royal Astronomical Society*, 437:L11–L15, 2014.

[15] E. Tempel, R. S. Stoica, and E. Saar. Evidence for spin alignment of spiral and elliptical galaxies in filaments. *Monthly Notices of the Royal Astronomical Society*, 428:1827–1836, 2013.

[16] E. Tempel, R. S. Stoica, E. Saar, V. J. Martinez, L. J. Liivamägi, and G. Castellan. Detecting filamentary pattern in the cosmic web: a catalogue of filaments for the SDSS. *Monthly Notices of the Royal Astronomical Society*, 438(4):3465–3482, 2014.

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

[18] J. Virtanen, K. Muinonen, and E. Bowell. Statistical ranging of asteroid orbits. *Icarus*, 154:412–431, 2001.