Journal of Multivariate Analysis 153 (2017) 1–15

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

# On the classification problem for Poisson point processes

Alejandro Cholaquidis [a], Liliana Forzani [b], Pamela Llop [b,*], Leonardo Moreno [c]

[a] Centro de Matemática, Facultad de Ciencias (UdelaR), Uruguay

[b] Facultad de Ingeniería Química, UNL and researchers of CONICET, Argentina

[c] Departamento de Métodos Cuantitativos, Fcea, Iesta, Udelar, Uruguay

ARTICLE INFO

*Article history:*
Received 5 February 2016
Available online 10 September 2016

*AMS subject classifications:*
60G55
62G05
62G08

*Keywords:*
Point process
Poisson process
Nonparametric estimation
Classification

ABSTRACT

For Poisson processes taking values in a general metric space, we tackle the problem of supervised classification in two different ways: via the classical $k$-nearest neighbor rule, by introducing suitable distances between patterns of points; and via the Bayes rule, by nonparametrically estimating the intensity function of the process. In the first approach we prove that under the separability of the space, the rule turns out to be consistent. In the second case, we prove the consistency of the rule by proving the consistency of the estimated intensities. Both classifiers are shown to behave well under departures from the Poisson distribution.



## 1. Introduction

Spatial point processes are commonly used to model the spatial structure of points formed by the location of individuals in space. The growing interest in such processes is related to the wide range of areas to which they can be applied. For instance, in ecology, they can be used to model the distribution of herds of animals, the spreading of nests of birds, the speckles of trees or plants, or eroded areas in rivers or seas. In geography, the position of earthquakes or volcanoes can by modeled by these such processes. They can be also used in astronomy to model the distribution of galaxies, in telecommunications, the locations of subscribers, among others. There is a vast literature on this area: to name just a few, we refer to the recent book *Spatial Data analysis in Ecology and Agriculture Using R* [19], which contains many other possible applications and techniques as well as real data examples. In [11], the authors propose a hierarchical modeling of the interaction structure in a plant community. The current interest in this kind of process also appears in connection with the new developments in functional neuroimaging techniques (for example fMRI), where it is possible to record in real time the location of the activation zones of the brain (see for instance [12,13,24]). In this context, in order to carry out a classification of people into healthy and unhealthy ones, the differences between the neurons that fire under some stimuli can be measured by modeling them as spatial Poisson processes with different intensities. In [17], the authors present a review of several distances used to measure the differences between two spatial patterns in order to perform a clustering or classification (see also [23]). In a different application area, crime modeling and mapping using geospatial technologies (which include the use of spatial point processes) is, quoting [16], "a topic of much interest mostly to academia, but also to the private sector and the government", see also [9,1]. On this topic, we study, in Section 7, the spatial distribution of three different categories of crime which took

* Correspondence to: Facultad de Ingeniería Química, UNL, Santiago del Estero 2829, 3000 Santa Fe, Argentina.
*E-mail address:* lloppamela@gmail.com (P. Llop).

http://dx.doi.org/10.1016/j.jmva.2016.09.002

place in Chicago between 2014 and 2016 by using an open access database containing, among other variables, the spatial location of each crime.

The aim of this paper is to tackle the supervised classification problem for Poisson point processes by framing it within the functional data setting. In particular, we prove the consistency of the $k$-nearest neighbors classifier in a more general context by verifying the separability of the space and the Besicovitch condition (see [2,7] for a deeper treatment of this topic). Via some simulation studies, we show how different choices of distances lead to different results for the classification. In addition, following the ideas in [6], we also propose a nonparametric estimator of the intensity function, prove its consistency, and plug it into the Bayes rule to get a consistent classifier. This last approach is similar to the one proposed in [13] but we do not assume that the intensities vary in a parametric family. Through some simulation studies we show the good performance of the $k$-NN rule so that it can be considered as an easier to implement alternative to the classical Bayes. More precisely, the $k$-NN classifier does not require the estimation of the intensity function (which is computationally expensive) and it can be employed in more general settings. Also, both rules could be combined in a new classifier as in [3] since, although most of the classical applications of spatial point processes are for recorded locations in $\mathbb{R}^2$ or $\mathbb{R}^3$, we do not restrict our approach to this case, instead allowing the realizations of the processes to live in a general metric space (such as a functional metric space or a Riemannian manifold, among others).

The present paper is outlined as follows: in Section 2 we present the definitions and preliminary results that we will use throughout the paper. Section 3.1 is devoted to introducing an estimator of the intensity of the process in order to plug it into the Bayes rule and prove its consistency. In Section 3.2 we handle the problem of choosing a suitable distance to guarantee the separability of the space and the satisfaction of the Besicovitch condition, in order to get the consistency of the $k$-NN estimator. Section 4 is devoted to the study of the metric dimension of the space introduced in Section 3.2. In Section 5 we extend the results to a more general class of processes: the Gibbs processes. In Section 6 we perform some simulation studies in order to assess the performance of the classification rules in different scenarios, to see the effect of changing some parameters in the estimation, and to investigate its robustness when the model is not Poisson. Lastly, in Section 7, we carry out the classification in a real data scenario. All the proofs are given in the Appendix.

## 2. Definitions and preliminary results

This section introduces some definitions and tools, which we will use throughout the paper. We will start with the definition of the main object of this paper, the Poisson point process, and then we will turn to classification rules in our context. For a deeper study of Poisson processes, we refer to [8,15,18].

### 2.1. Poisson processes

Let $(S, \rho)$ be a separable and bounded metric space, endowed with a Borel measure $\nu$, let us denote by $\mathcal{B}(S)$ the Borel $\sigma$-algebra on $S$ and by $S^\infty$ the set of elements (subsets) $x$ of $S$ whose cardinality, $\#x$, is finite. That is,

$$S^\infty \equiv \{x \subset S : \#x < \infty\}.$$

Let $\lambda : S \to \mathbb{R}^+$ be an integrable function. Given a probability space $(\Omega, \mathcal{A}, \mathrm{P})$, we will say that a function $X : \Omega \to S^\infty$ is a *Poisson process* on $S$ with intensity $\lambda$ (we will denote $X \sim \mathcal{P}(S, \lambda)$) if:

- the functions $N_A : \Omega \to \{0, \dots, \infty\}$ defined by $N_A(\omega) = \#\{\omega : X(\omega) \cap A\}$ are random variables for all $A \in \mathcal{B}(S)$;
- given $n$ disjoint Borel subsets $A_1, \dots, A_n$ of $S$, the random variables $N_{A_1}, \dots, N_{A_n}$ are independent;
- $N_A$ follows a Poisson process with mean $\mu(A)$ (we will write $N_A \sim \mathcal{P}\{\mu(A)\}$), with

$$\mu(A) = \int_A \lambda(\zeta) d\nu(\zeta).$$

Let $\mathcal{S}^\infty = 2^{S^\infty}$ be the $\sigma$-algebra of part of $S^\infty$. If $X$ is a Poisson process, the distribution $\mathrm{P}_X$ of $X$ on $\mathcal{S}^\infty$ is defined by $\mathrm{P}_X(B) = \Pr(X \in B)$ for $B \in \mathcal{S}^\infty$.

A well-known result (see [18]) for point processes states that if $X_1$ and $X_2$ are Poisson processes with intensities $\lambda_1$ and $\lambda_2$, respectively, with values on a non-empty bounded metric space $(S, \rho)$ such that $\mu_i(S) < \infty$, $i = 1, 2$, the distribution of $X_1$ is absolutely continuous with respect to the distribution of $X_2$ ($\mathrm{P}_{X_1} \ll \mathrm{P}_{X_2}$) with Radon–Nikodym derivative

$$f_{X_1}(x) = \exp\{\mu_2(S) - \mu_1(S)\} \prod_{\xi \in x} \frac{\lambda_1(\xi)}{\lambda_2(\xi)},$$

with $0/0 = 0$. As a consequence, observe that if $X_2 \sim \mathcal{P}(S, 1)$, then, for all $X \sim \mathcal{P}(S, \lambda)$, $\mathrm{P}_X \ll \mathrm{P}_{X_2}$ and

$$f_X(x) = \exp\{\nu(S) - \mu(S)\} \prod_{\xi \in x} \lambda(\xi), \tag{1}$$

where $\mu(S) = \int_S \lambda d\nu$.