UDC 519.218.5+519.246.5

IDENTIFICATION OF THE POINT PROCESS INTENSITY SHAPE WITH THE PRECEDENTS MAXIMUM LIKELIHOOD DISTRIBUTIONS

V. E. Antsiperov

Kotelnikov Institute of Radioengineering and Electronics of the Russian Academy of Sciences, Mokhovaya 11-7, Moscow 125009, Russia

The paper is received on December 11, 2017

Abstract: The work is devoted to the point process intensity shape identification by a given realization of process point occurrences. This identification is supposed to be the best fitting of the registered point-set to formal description of intensity shapes of previously observed processes − precedents. As a formal description of intensity shapes, it is suggested to use the parameters of the probabilistic mixture models. The main argument in favor of such a description is the fact, that for Poisson point processes, conditional, at a given number of points, distribution of the single point occurrence coincides, with an accuracy up to normalization, with the intensity. Because the Poisson model has proven itself in many applied problems, potentially approach proposed has the large amount of applications. Moreover, since for the mixture-like approximations exist effective algorithms for mixture parameters computation, the numerical realization of the approach seems to be the most reliable in many respects. The mentioned algorithms belong to the well-known VM (VB EM) family, they implement iterative (recursive) realization of the maximum likelihood approach. We also present and discuss VM-like identification algorithms in our paper. In this connection, explicit expressions are given for the point process intensity shape iterative identification.

Key words: point process intensity shape identification, formal shape description, inhomogeneous Poison point process, finite mixture models, machine learning, effective computational schemes, EM, VB EM, VM algorithms.

Introduction

The main problem considered in the paper concerns the statistical inference about the shape of intensity of a point process, whose realization is given by a random set of discrete points [1]. It is assumed that these points occur in the state space S − some subset of R^D − the Euclidean space of dimension $D \geq 1$. Usually, if $D = 1$, then points are implied as occurring along the time axis, if $D = 2$, then points occur in some plane region, if $D = 3$ – in the space, etc. However, for further purposes, the dimension D of S is not essential and it is not specified. Consequently, for the point locations the vector notation $\vec{x} \in S$ is used, implying, if it is necessary, that $\vec{x} = (x_1, \dots, x_D)^T$ represents the list of D ordered numeric components. So, any realization of point processes comprises the random number of points $n = n(S) \geq 1$ and the random locations $\{\vec{x}_1, \dots, \vec{x}_n\}$, $\vec{x}_i \in S$ of the points. The set notation $\{\dots\}$ signifies only that the ordering of the points $\vec{x}_i$ is irrelevant, but not that the points are necessarily distinct. Such unordered lists will be further referred to as the sets.

One of the most important characteristics of any point process is its homogeneity / inhomogeneity. Usually this characteristic is defined in terms of the process intensity shape. The definition of intensity is briefly reviewed as follows: for any infinitely small neighborhood ds of $\vec{x} \in S$ the intensity $\lambda(\vec{x})$ is the average number $\langle n(ds) \rangle$ of points occurring in ds divided by its volume $d\vec{x} = dx_1 \cdot \dots \cdot dx_D$:

$$\lambda(\vec{x}) = \lim_{d\vec{x} \to 0} \frac{\langle n(ds) \rangle}{d\vec{x}} \ . \qquad (1)$$

When the intensity $\lambda(\vec{x})$ (1) is constant, the corresponding point process is said to be homogeneous, in the other case it is inhomogeneous. The inhomogeneous processes are, naturally, the most important ones for the intensity shape identification.

It is easy to show that for orderly point processes [1] the value $\lambda(\vec{x})d\vec{x}$ up to infinite smalls of a higher order gives the probability $P(n(ds) = 1)$ of occurring (exactly one) point in ds. If, in addition, the occurrences of points in nonintersecting regions of S are independent, an explicit expression can be obtained for the joint distribution density of realization $(n, \{\vec{x}_1, \dots, \vec{x}_n\})$ [2]:

$$p(n,\{\vec{x}_1,\dots,\vec{x}_n\}) = \lim_{d\vec{x}_1,\dots,d\vec{x}_n\to 0} \frac{P(n(S)=n,\ n(ds_1)=1,\dots,n(ds_n)=1)}{d\vec{x}_1\cdot\dots\cdot d\vec{x}_n} = \\ = \lambda(\vec{x}_1)\cdot\dots\cdot\lambda(\vec{x}_n)\times\exp\left(-\int_S \lambda(\vec{x})d\vec{x}\right) \quad . \qquad (2)$$

It is well-known, that any orderly point process with the independent property is, up to some technical refinements, necessarily Poisson process [2]. As for the refinements, they are important for rigorous mathematical analysis, in applications they practically always take place. So, assuming all refinements are satisfied, the Poisson process will be further considered as a model of the analyzed point process. To emphasize this fact, let's rewrite the distribution density (2) in the more familiar form:

$$p(n,\{\vec{x}_i\}|\lambda(\vec{x})) = p(\{\vec{x}_i\}|n;p(\vec{x}))\mathcal{P}(n|\Lambda)\,, \\ \mathcal{P}(n|\Lambda) = \frac{\Lambda^n}{n!}\exp(-\Lambda)\,,\ \ \Lambda = \int_S \lambda(\vec{x})d\vec{x}\ \ , \qquad (3) \\ p(\{\vec{x}_i\}|n;p(\vec{x})) = n!\prod_{i=1}^n p(\vec{x}_i),\ p(\vec{x}) = \lambda(\vec{x})/\Lambda\ ,$$

where $\mathcal{P}(n|\Lambda)$ is the standard Poisson distribution of random number of points n with the parameter Λ – the fraction of process power on S, and $p(\{\vec{x}_i\}|n;\lambda(\vec{x}))$ is a conditional distribution density of random locations of points $\{\vec{x}_1,\dots,\vec{x}_n\}$, when the process intensity is $\lambda(\vec{x})$. A couple of remarks should be made concerning (3). Firstly, if the bulk of the $\lambda(\vec{x})$ belongs to the state space S, we can spread the integration in Λ (3) to infinity and "forget" the exact form of S. Secondly, the conditional distribution $p(\{\vec{x}_i\}|n;\lambda(\vec{x}))$ has the same form as the order statistics of n independent, identically distributed random vectors on S with the common probability density $p(\vec{x})$. It follows from (3) and from these remarks that if the intensity function of a Poisson process is concentrated substantially on S, then it can be represented as the product of a factor Λ , characterizing the total power (norm) of the process, and the normalized probability distribution density $p(\vec{x})$, which characterizes the randomness in locations of independent points in R^D: $\lambda(\vec{x}) = \Lambda p(\vec{x})$. This observation will be a key point for further consideration.

Although the discussed above Poisson process is a rather special case of general point processes – it is the maximally random and, consequently, has the

simplest structure – the number of Poisson process applications is huge [3] and continues to grow rapidly. It is because the Poisson process as a convenient and flexible model for analyzing both real systems and their simulations leads often to well-interpreted results. In this paper we also use the model of the inhomogeneous Poisson point process (PPP).

History and modern trends in PPP intensity estimation

As emphasized above (3), a complete statistical description of PPP can be achieved by specifying its intensity function. Intensity function, in turn, can be parameterized by the process power $\Lambda = \bar{n}$ (intensity norm) and normalized density $p(\vec{x})$ (intensity form) (3). When application domain knowledge does not fully specify the intensity, it is necessary to estimate it from registered data.

One of the first works dedicated to the analysis and estimation of PPP intensity was Cox paper [4]. It contains the systematic application of classic (frequentist) statistical methods to the problem. A few estimates of the Poisson process intensities, grouped intensities, etc. have been proposed in this regard. But it should be noted, that all problems were solved only for the homogeneous (stationary) Poisson process and, consequently, concern only $\Lambda = \bar{n}$ estimations. The inhomogeneous case is analytically more complicated, therefore the estimation of non-constant, varying intensity (namely its shape) was not considered in the paper.

The shift in the last quarter of the 20th century in the PPP intensity estimating towards semi-numerical and numerical methods was reviewed in [5]. The author discusses the model formulations, shape estimations, numerous applications. Maximum likelihood (ML) methods are emphasized as a basis for inference whenever possible. To find the maximum likelihood estimates, it is suggested, among other things, to use the iterative Newton-Raphson methods. In some cases, the use of EM (expectation-maximization) algorithms is also considered.

The current state of intensity estimation is characterized by the trend of modern statistical methods to an algorithmic form. They increasingly correspond to the principles of the machine learning approach [6]. The good introduction in the subject is [2]. The author considers the parametric models of intensities and discusses

approaches to estimating their parameters depending on the type of problem – either it is related to the single process or to superposition of the processes. It is noted, that EM algorithm seems to be the most effective one in the superposition for Gaussian mixtures.

In this paper, we also propose a certain variant of machine learning approach to evaluation of point process intensity shape. In accordance with the principles of machine learning, in our approach we propose to use data previously obtained in the learning process. It implies getting maximum correspondence (fitting) of the registered set of data $(n, \{\vec{x}_i\})$ to a given, formally described intensity shapes of previously observed processes –precedents.

Maximum likelihood PPP intensity shape identification / estimation

In this section, we present the general approach to identification of PPP intensity shape. The approach is based on establishing the maximum likelihood fitting of the recorded data to some intensity shape from a given set. We show that such an identification necessarily includes the step of estimating some intensity parameters. It is true regardless of whether we consider the registered process or precedents, i.e. regardless of the available a priori information. However, the problem of estimating parameters for precedents is usually more complex, because in this case a priori information is often poor. In the case of a registered process, when existing precedents provide some a priori information, estimating can be somewhat simplified, but due to the limited amount of data, it should be done more carefully. The nuances and differences in the intensity identification for these two cases are discussed in the next section.

As mentioned above, the non-constant, strictly-positive intensity $\lambda(\vec{x})$ (1) completely determines PPP statistics. As is customary in the theory of statistical inferences based on models [7], we assume that each intensity belongs to some definite class of shapes, indexed by a symbol M (shape model structure), and inside the class individual intensities are determined by the parameters $\vec{\theta}$. Substantially the model M defines an a priori probability distribution of parameters $\mathcal{P}(\vec{\theta}|M)$. Some

parameters can be tightly related to the model − they have narrow marginal distributions $\mathcal{P}(\theta_j|M)$, others are weakly related − they have wide distributions, possibly not depending on M. As the parameters of the first type, we can mention some geometric characteristics − shape moments, mutual distances between shape components, etc., The second type of parameters can be associated with transformations common for all intensities − shifts, scaling, rotations, etc.

So, let us consider the intensities belonging to some parametrized set $\{\Lambda p(\vec{x}|\vec{\theta},M)\}$, where M is the class label, parameter Λ denotes the process power (norm) and $p(\vec{x}|\vec{\theta},M)$ is the positive normalized distribution density, depending on parameters $\vec{\theta}$ from parameter space T − subset of R^{Φ} (vide supra). Recall that the bulk of $p(\vec{x}|\vec{\theta},M)$ is assumed belonging to the state space S for all $\vec{\theta} \in T$. In this regard, to further use of the maximum likelihood (ML) method, it is worth replacing the likelihood function $p(n,\{\vec{x}_i\}|\lambda(\cdot))$ (3) by its logarithm (log−likelihood):

$$\begin{gathered} L(n,\{\vec{x}_i\}|\Lambda,\vec{\theta},M) = \ln\left(p\left(n,\{\vec{x}_i\}|\Lambda p(\vec{x}|\vec{\theta};M)\right)\right) = L_S(\{\vec{x}_i\}|n;\vec{\theta},M) + L_P(n|\Lambda)\,, \\ L_P(n|\Lambda) = \ln(\mathcal{P}(n|\Lambda)) = n\ln(\Lambda) - \Lambda - \ln(n!)\,, \\ L_S(\{\vec{x}_i\}|n;\vec{\theta},M) = \ln\left(p\left(\{\vec{x}_i\}|n;p(\vec{x}|\vec{\theta},M)\right)\right) = \sum_{i=1}^{n}\ln\left(p(\vec{x}_i|\vec{\theta},M)\right) + \ln(n!)\,. \end{gathered} \tag{4}$$

As it follows from (4), log−likelihood $L(n,\{\vec{x}_i\}|\Lambda,\vec{\theta},M)$ splits into the sum of two functions – $L_P(n|\Lambda)$ which depends only on Λ and $L_S(\{\vec{x}_i\}|n;\vec{\theta},M)$ depending on $(\vec{\theta},M)$. Obviously, $L_P(n|\Lambda)$ does not say anything about the shape of the intensity. Thus, the focus of the problem under consideration is the reduced log−likelihood function $L_S(\{\vec{x}_i\}|n;\vec{\theta},M)$ (4).

Due to $L_S(\{\vec{x}_i\}|n;\vec{\theta},M)$ (4) importance, it is desirable to define the function structure in more detail. In our case this means a refinement of the model for the distribution density $p(\vec{x}|\vec{\theta},M)$. Let us accept the finite mixture model [8]:

$$p(\vec{x}|\vec{\theta},M) = \sum_{k=1}^{G} w_M^k \rho_k(\vec{x}|\vec{\theta},M)\,, \tag{5}$$

where G is the number of mixture components, weights $\{w_M^k\}$ are the probabilities that the point hits a certain component, component densities $\{\rho_k(\vec{x}|\vec{\theta},M)\}$ are the conditional distributions of point location coordinates when the point belongs to the component $k = 1,\dots,G$. As is known from the theory of statistical inference [8], working with mixtures (5) it is convenient to introduce the hidden (latent) integer random variable $y \in \{1,\dots,G\}$ − component indicator. Then, considering $\{w_M^k\}$ as an indicator probability and $\rho_k(\vec{x}|\vec{\theta};M)$ as a conditional distribution $\rho(\vec{x}|y;\vec{\theta},M)$, we can tract (5) as marginal distribution of $\vec{x}$. Wherein, using the hidden variable y, we can rewrite $p(\vec{x}|\vec{\theta},M)$ (5) in a more compact form:

$$p(\vec{x}|\vec{\theta};M) = \frac{\rho(\vec{x},y|\vec{\theta},M)}{\rho(y|\vec{x};\vec{\theta},M)} = \frac{w_M^y \rho_y(\vec{x}|\vec{\theta},M)}{\rho(y|\vec{x};\vec{\theta},M)}, \quad (6)$$

although this is achieved by introducing a rather complex for the computation a posteriori distribution $\rho(y|\vec{x};\vec{\theta},M)$.

Substituting $p(\vec{x}|\vec{\theta},M)$ (6) in (4), we get the following log−likelihood function $L_S(\{\vec{x}_1,\dots,\vec{x}_n\}|n;\vec{\theta})$ structure, defined by the model (5) and "uncovered" with the entered hidden variables:

$$L_S(\{\vec{x}_i\}|n;\vec{\theta},M) = \sum_{i=1}^{n} \ln\left(\frac{\rho(\vec{x}_i,y_i|\vec{\theta},M)}{\rho(y_i|\vec{x}_i;\vec{\theta},M)}\right) + \ln(n!) = \ln\left(n!\prod_{i=1}^{n}\frac{\rho(\vec{x}_i,y_i|\vec{\theta},M)}{\rho(y_i|\vec{x}_i;\vec{\theta},M)}\right). \quad (7)$$

It is well-known, that maximum likelihood (ML) method as the method of fitting the registered data $(n,\{\vec{x}_i\})$ to any shape class M was proposed by Fisher and consists in evaluating and maximizing with respect to M of the marginal log-likelihood function $L((n,\{\vec{x}_i\})\,|M)$. We also take the ML paradigm as the basis of our approach. Hereof, one of the central problems of the approach is how to find the log-likelihood function, i.e. how to calculate it. Below we discuss the principal features of the log-likelihood function calculation, typical for machine learning and used in our approach.

The marginal log-likelihood function $L((n,\{\vec{x}_i\})\,|M)$, after a series of elementary transformations, can be written in the form:

$$L((n,\{\vec{x}_i\})\,|M) = \ln\big(p(n,\{\vec{x}_i\}|M)\big) = \ln\left(\frac{p(n,\{\vec{x}_i\};\Lambda,\vec{\theta}|M)}{p(\Lambda,\vec{\theta}|n,\{\vec{x}_i\};M)}\right) = \\ = \ln\Big(p\big(n,\{\vec{x}_i\}|\Lambda,\vec{\theta},M\big)\mathcal{P}\big(\Lambda,\vec{\theta}|M\big)\Big) - \ln\Big(p\big(\Lambda,\vec{\theta}|n,\{\vec{x}_i\};M\big)\Big), \quad (8)$$

where $\mathcal{P}\big(\Lambda,\vec{\theta}|M\big)$ is an a priori and $p\big(\Lambda,\vec{\theta}|n,\{\vec{x}_i\};M\big)$ is a posteriori probability distribution of $\big(\Lambda,\vec{\theta}\big)$ for a given class M. Since, according to assumptions made above, Λ is independent of $\big(\vec{\theta};M\big)$ (and $\{\vec{x}_i\}$) and $\vec{\theta}$ is independent of n, the following simplification take place: $\mathcal{P}\big(\Lambda,\vec{\theta}|M\big) = \mathcal{P}(\Lambda)\mathcal{P}\big(\vec{\theta}|M\big)$ and $p\big(\Lambda,\vec{\theta}|n,\{\vec{x}_i\};M\big) = p(\Lambda|n)p\big(\vec{\theta}|\{\vec{x}_i\};M\big)$. Given these circumstances and using (4), we can rewrite (8) as:

$$L((n,\{\vec{x}_i\})\,|M) = L_S\big(\{\vec{x}_i\}\big|n;\vec{\theta},M\big) + \ln\Big(\mathcal{P}\big(\vec{\theta}|M\big)\Big) - \ln\Big(p\big(\vec{\theta}|\{\vec{x}_i\};M\big)\Big) + \\ + L_P(n|\Lambda) + \ln(\mathcal{P}(\Lambda)) - \ln\big(p(\Lambda|n)\big). \quad (9)$$

Because the sum of the last three terms in (9) is independent of $\big(\vec{\theta},M\big)$ (it is $\ln\big(p(n)\big)$ and is even independent on Λ), the essential information about M is contained, as noted, in $L_S\big(\{\vec{x}_i\}\big|n;\vec{\theta},M\big)$, in a priori $\mathcal{P}\big(\vec{\theta}|M\big)$ and a posteriori $p\big(\vec{\theta}|\{\vec{x}_i\};M\big)$ distributions of $\vec{\theta}$. Since the constant terms do not affect the maximum points, the first three terms in (9) can be used in place of $L((n,\{\vec{x}_i\})\,|M)$ and the function, maximized in the framework of the method proposed, takes the form:

$$\tilde{L}((n,\{\vec{x}_i\})\,|M) = L_{S\theta}\big(\{\vec{x}_i\};\vec{\theta}\big|n;M\big) - \ln\Big(p\big(\vec{\theta}|\{\vec{x}_i\};M\big)\Big), \\ L_{S\theta}\big(\{\vec{x}_i\};\vec{\theta}\big|n;M\big) = \ln\Big(p\big(\{\vec{x}_i\};\vec{\theta}|n;M\big)\Big) = L_S\big(\{\vec{x}_i\}\big|n;\vec{\theta},M\big) + \ln\Big(\mathcal{P}\big(\vec{\theta}|M\big)\Big), \quad (10)$$

where, according (7):

$$p\big(\{\vec{x}_i\};\vec{\theta}|n;M\big) = n!\prod_{i=1}^{n}\frac{\rho(\vec{x}_i,y_i|\vec{\theta},M)}{\rho(y_i|\vec{x}_i;\vec{\theta},M)}\mathcal{P}\big(\vec{\theta}|M\big) \quad (11)$$

is the joint distribution of point coordinates $\{\vec{x}_i\}$ and parameters $\vec{\theta}$ (for given number of points n and model class M). Splitting $p\big(\{\vec{x}_i\},\vec{\theta}|n;M\big)$ (11) into a joint $p\big(\{\vec{x}_i\},\{y_i\};\vec{\theta}|n;M\big)$ and a posteriori $p\big(\{y_i\}|\{\vec{x}_i\};\vec{\theta},M\big)$ distributions of hidden

variables $\{y_i\}$, we obtain a representation, suitable for constructing variational methods (VM) of computing shortened log−likelihood function (10) [7]:

$$\begin{gathered}\tilde{L}((n,\{\vec{x}_i\})\,|M) = L_{Sh\theta}\left(\{\vec{x}_i\},\{y_i\};\vec{\theta}\middle|n;M\right) - \ln\left(p\left(\{y_i\};\vec{\theta}|\{\vec{x}_i\};M\right)\right),\\ L_{Sh\theta}\left(\{\vec{x}_i\},\{y_i\};\vec{\theta}\middle|n;M\right) = \ln\left(p\left(\{\vec{x}_i\},\{y_i\};\vec{\theta}|n;M\right)\right),\\ p\left(\{\vec{x}_i\},\{y_i\};\vec{\theta}|n;M\right) = n!\prod_{i=1}^{n}\rho\left(\vec{x}_i,y_i|\vec{\theta},M\right)\mathcal{P}\left(\vec{\theta}|M\right),\\ p\left(\{y_i\};\vec{\theta}|\{\vec{x}_i\};M\right) = \prod_{i=1}^{n}\rho\left(y_i|\vec{x}_i;\vec{\theta},M\right)p\left(\vec{\theta}|\{\vec{x}_i\};M\right).\end{gathered} \tag{12}$$

Relations (12) give us the variable part of the marginal log-likelihood function (8), expressed through all the initial model distributions and, accordingly, providing a principal way of its calculation. However, the practical realization of direct calculations can be quite complicated, especially thanks to calculations of a posteriori $p\left(\{y_i\};\vec{\theta}|\{\vec{x}_i\};M\right)$. To avoid such complications, we use the following trick. Because $\tilde{L}((n,\{\vec{x}_i\})\,|M)$ (12) does not depend on $\left(\{y_i\};\vec{\theta}\right)$, its averaging with respect to any arbitrary normalized density $q\left(\{y_i\};\vec{\theta}\right)$ will not change it. But the averaging of the terms on the right-hand side of (12) after a small correction of the probabilities under the logarithms gives the free energy of the joint distribution $F(q(\cdot),\dots)$ and the Kullback-Leibler divergence, denoted by $D_{KL}\left(q(\cdot),p(\cdot\,|\{\vec{x}_i\};M)\right)$ [9]:

$$\begin{gathered}\tilde{L}((n,\{\vec{x}_i\})\,|M) = F\left(q(\cdot),\{\vec{x}_i\},\{y_i\};\vec{\theta}|n;M\right) + D_{KL}\left(q(\cdot);p\left(\{y_i\};\vec{\theta}|\{\vec{x}_i\};M\right)\right),\\ F(q(\cdot),\dots) = \sum_{y_1=1}^{G}\dots\sum_{y_n=1}^{G}\int_T q\left(\{y_i\};\vec{\theta}\right)\ln\left(\frac{p(\{\vec{x}_i\},\{y_i\};\vec{\theta}|n;M)}{q(\{y_i\};\vec{\theta})}\right)d\vec{\theta}\,,\\ D_{KL}\left(q(\cdot),p(\cdot\,|\{\vec{x}_i\};M)\right) = \sum_{y_1=1}^{G}\dots\sum_{y_n=1}^{G}\int_T q\left(\{y_i\};\vec{\theta}\right)\ln\left(\frac{q(\{y_i\};\vec{\theta})}{p(\{y_i\};\vec{\theta}|\{\vec{x}_i\};M)}\right)d\vec{\theta}\,.\end{gathered} \tag{13}$$

Because Kullback-Leibler divergence $D_{KL}\left(q(\cdot),p(\cdot\,|\{\vec{x}_i\};M)\right)$ is nonnegative functional [9], the free energy integral $F(q(\cdot),\dots)$ always forms a rigorous lower bound on the $\tilde{L}((n,\{\vec{x}_i\})\,|M)$ and coincides with it − reaches the maximum − in the only case when $q\left(\{y_i\};\vec{\theta}\right) = \rho\left(\{y_i\};\vec{\theta}|\{\vec{x}_i\};M\right)$. Thus, the problem of calculating $\tilde{L}((n,\{\vec{x}_i\})\,|M)$ (12) can be reduced to the variational problem:

$$\tilde{L}((n,\{\vec{x}_i\})\,|M) = \max_{q(\{y_i\};\vec{\theta})} F\left(q(\cdot),\{\vec{x}_i\},\{y_i\};\vec{\theta}|n;M\right)\ . \tag{14}$$

Historically the reduction of the log-likelihood calculation problem (12) to the variational problem (14) has given rise to the development of several effective computational algorithms. Like in the well-known variational Rayleigh-Ritz method, the main idea [7] here the following. Instead of looking for approximations to the exact solution (14), which in principle is known ($\rho\left(\{y_i\};\vec{\theta}|\{\vec{x}_i\};M\right)$), but requires intractable calculations, it makes sense from the beginning of the solution (14) to constraint the set of varying functions to the tractable ones, although they will give only an approximate solution. The most popular set of such tractable functions is the family of factorized over $\{y_i\}$ and $\vec{\theta}$ distributions $q\left(\{y_i\};\vec{\theta}\right)=P_{\{y_i\}}Q(\vec{\theta})$, where $P_{\{y_i\}}$ is discrete distribution on all possible sets $\{y_i\}$, and $Q(\vec{\theta})$ is a continuous distribution on T. Taking this constraint into account, the variational problem (14) can be written (considering (12)) in the form:

$$\begin{gathered}\tilde{L}((n,\{\vec{x}_i\})\,|M)=\max_{Q(\vec{\theta})}\max_{P_{\{y_i\}}}F\left(P_{\{y_i\}},Q(\vec{\theta}),\{\vec{x}_i\},\{y_i\};\vec{\theta}|n;M\right),\\ F\left(P_{\{y_i\}},Q(\vec{\theta}),\{\vec{x}_i\},\{y_i\};\vec{\theta}|n;M\right)=\\ =\sum_{i=1}^{n}\sum_{y_i=1}^{G}P_{y_i}\int_T Q(\vec{\theta})\ln\left(\rho\left(\vec{x}_i,y_i|\vec{\theta},M\right)\right)d\vec{\theta}-\\ -\int_T Q\left(\vec{\theta}\right)\ln\left(\frac{Q(\vec{\theta})}{\mathcal{P}(\vec{\theta}|M)}\right)d\vec{\theta}-\sum_{y_1=1}^{G}\ldots\sum_{y_n=1}^{G}P_{\{y_i\}}\ln\left(P_{\{y_i\}}\right)+\ln(n!)\end{gathered}\quad,\quad(15)$$

where P_{y_i}, in contrast to $P_{\{y_i\}}$, means the marginal distribution of $P_{\{y_k\}}$ with respect to y_i. The solution in the general form of the approximate variational problem (15) can be found in two steps using, for example, the method of Lagrange multipliers (because of $P_{\{y_i\}}$ and $Q(\vec{\theta})$ unit norm constraints).

First step. Taking the usual derivatives of $F\left(P_{\{y_i\}},Q\left(\vec{\theta}\right),\ldots\right)$ (15) with respect to $P_{\{y_k\}}$ and equating them to zero:

$$\sum_{i=1}^{n}\int_T Q(\vec{\theta})\ln\left(\rho\left(\vec{x}_i,y_i|\vec{\theta},M\right)\right)d\vec{\theta}-\ln\left(P_{\{y_i\}}\right)-1-\mu_L=0\ ,\qquad(16)$$

where μ_L is Lagrange multiplier, we get:

$$P_{\{y_i\}}=\frac{1}{Z_P}\exp\left(\sum_{i=1}^{n}\int_T Q(\vec{\theta})\ln\left(\rho\left(\vec{x}_i,y_i|\vec{\theta},M\right)\right)d\vec{\theta}\right)\ ,\qquad(17)$$

$Z_P = \exp(1+\mu_L)$ is the normalization constant, partition function for discrete distribution $P_{\{y_i\}}$. It immediately follows from (17) that $P_{\{y_i\}}$ splits into a product of factors, each of which depends only on a certain $(\vec{x}_i, y_i)$ pair. Normalizing each of these factors to unity, we obtain:

$$P_{\{y_i\}} = \prod_{i=1}^{n} \pi_{iy_i}$$
$$\pi_{ik} = \frac{\exp\left(\int_T \mathcal{Q}(\vec{\theta}) \ln\left(\rho(\vec{x}_i, k|\vec{\theta}, M)\right) d\vec{\theta}\right)}{\sum_{l=1}^{G} \exp\left(\int_T \mathcal{Q}(\vec{\theta}) \ln\left(\rho(\vec{x}_i, l|\vec{\theta}, M)\right) d\vec{\theta}\right)} \quad . \qquad (18)$$

Note that π_{iy_i} (18) is the marginal distribution of $P_{\{y_k\}}$ with respect to y_i, i.e. it is exactly the same as P_{y_i}.

Second step. Substituting P_{y_i} by π_{iy_i} in (15), taking then the functional derivative of $F\left(P_{\{y_i\}}, \mathcal{Q}(\vec{\theta}), \ldots\right)$ with respect to $\mathcal{Q}(\vec{\theta})$ and equating it to zero:

$$\sum_{i=1}^{n} \sum_{k=1}^{G} \pi_{ik} \ln\left(\rho(\vec{x}_i, k|\vec{\theta}, M)\right) - \ln\left(\frac{\mathcal{Q}(\vec{\theta})}{\mathcal{P}(\vec{\theta}|M)}\right) - 1 - \nu_L = 0 \ , \qquad (19)$$

where ν_L is Lagrange multiplier, we get:

$$\mathcal{Q}(\vec{\theta}) = \frac{1}{Z_Q} n! \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{G} \pi_{ik} \ln\left(\rho(\vec{x}_i, k|\vec{\theta}, M)\right)\right) \mathcal{P}(\vec{\theta}|M) \ , \qquad (20)$$

$Z_Q = \exp(1+\nu_L)/n!$ is the normalization constant, partition function for continuous distribution $\mathcal{Q}(\vec{\theta})$. In (20) the multiplier $n!$ is introduced to represent $\mathcal{Q}(\vec{\theta})$ in the following general form (see (12)):

$$\mathcal{Q}(\vec{\theta}) = \frac{1}{Z_Q} \exp\left(\sum_{y_1=1}^{G} \ldots \sum_{y_n=1}^{G} P_{\{y_i\}} \ln\left(n! \prod_{i=1}^{n} \rho(\vec{x}_i, y_i|\vec{\theta}, M) \mathcal{P}(\vec{\theta}|M)\right)\right) =$$
$$= \frac{1}{Z_Q} \exp\left(\sum_{y_1=1}^{G} \ldots \sum_{y_n=1}^{G} P_{\{y_i\}} p\left(\{\vec{x}_i\}, \{y_i\}; \vec{\theta}|n; M\right)\right) \quad . \qquad (21)$$

Combining expressions (18) and (20), we obtain the general form of the optimal solution of the variational problem (15) (marked by an asterisk (*) to distinguish them from other distributions):

$$\pi_{ik}^{(*)} = \frac{\exp\left(\int_T \mathcal{Q}^{(*)}(\vec{\theta}) \ln\left(\rho(\vec{x}_i, k|\vec{\theta}, M)\right) d\vec{\theta}\right)}{\sum_{l=1}^{G} \exp\left(\int_T \mathcal{Q}^{(*)}(\vec{\theta}) \ln\left(\rho(\vec{x}_i, l|\vec{\theta}, M)\right) d\vec{\theta}\right)}$$
$$\mathcal{Q}^{(*)}(\vec{\theta}) = \frac{1}{Z_Q} n! \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{G} \pi_{ik}^{(*)} \ln\left(\rho(\vec{x}_i, k|\vec{\theta}, M)\right)\right) \mathcal{P}(\vec{\theta}|M) \quad , \qquad (22)$$

where it is understood that $P_{\{y_i\}}^{(*)} = \prod_{i=1}^{n} \pi_{iy_i}^{(*)}$. Substituting probabilities $\pi_{iy_i}^{(*)}$ expressed in terms of $\mathcal{Q}^{(*)}(\vec{\theta})$ (22) into free energy expression (15), we obtain its maximum value, that gives, as noted above, the approximation (from below) of the shortened log−likelihood function $\tilde{L}((n,\{\vec{x}_i\})\,|M)$ (12):

$$F^{(*)}(\{\vec{x}_i\}|n;M) = F\left(P_{\{y_i\}}^{(*)}, \mathcal{Q}^{(*)}(\vec{\theta}), \dots\right) =$$
$$= \ln\left(n!\prod_{i=1}^{n}\left[\sum_{k=1}^{G}\exp\left\{\int_T \mathcal{Q}^{(*)}(\vec{\theta})\ln\left(\rho(\vec{x}_i,k|\vec{\theta},M)\right)d\vec{\theta}\right\}\right]\right) - \quad (23)$$
$$-D_{KL}\left(\mathcal{Q}^{(*)}(\vec{\theta}), \mathcal{P}(\vec{\theta}|M)\right)$$

As it follows from (23), the approximation obtained is completely determined by a posteriori parameters distribution $\mathcal{Q}^{(*)}(\vec{\theta})$. The latter, in turn, is the randomized estimate of the parameters $\vec{\theta}$. Thus, the statement made in the beginning of the section, that intensity shape identification necessarily includes the step of estimating some intensity parameters, can be considered completely proven. Moreover, expression (23) clearly shows how parameters estimation should use a priory and a posteriori information to optimize the intensity identification procedure. Indeed, for narrow distributions $\mathcal{Q}^{(*)}(\vec{\theta})$ with a maximum in $\vec{\theta}^{(*)}$, the first term on the right-hand side of (23) is approximately a reduced log−likelihood function $L_S(\{\vec{x}_i\}|n;\vec{\theta}^{(*)},M)$ (4). The closer $\vec{\theta}^{(*)}$ to its ML estimate $\vec{\theta}_{ML}$, which depends only on the data realization $(n,\{\vec{x}_1,\dots,\vec{x}_n\})$, the larger this term. Conversely, the second term in (23) will be larger (divergence $D_{KL}\left(\mathcal{Q}^{(*)}(\vec{\theta}), \mathcal{P}(\vec{\theta}|M)\right)$ will be less), when $\mathcal{Q}^{(*)}(\vec{\theta})$ is closer to $\mathcal{P}(\vec{\theta}|M)$. Since we are looking for the maximum of sum of these two terms, we are looking for a compromise in the use of registered $(n,\{\vec{x}_1,\dots,\vec{x}_n\})$ and a priory $\mathcal{P}(\vec{\theta}|M)$ data when evaluating parameters by $\mathcal{Q}^{(*)}(\vec{\theta})$.

Computational PPP intensity shape identification / estimation

As follows from the preceding section, the problem of the ML PPP intensity shape identification by the registered data set $(n,\{\vec{x}_i\})$ consists in choosing the class $M^{(*)}$ which corresponds to the maximum free energy $F^{(*)}(\{\vec{x}_i\}|n;M)$ (23):

$$M^{(*)} = arg\max_M F^{(*)}(\{\vec{x}_i\}|n;M) \ . \quad (24)$$

Assuming that the number of classes M is not very large − in the case of a registered process it is not more than the number of available precedents, and in the case of a new precedent there are only a few model assumptions about it, we will assume that the main volume of calculations associated with (24) is in the calculation of the $F^{(*)}(\{\vec{x}_i\}|n;M)$ values, but not in the comparing them with each other (sorting the calculated set of values). Thus, concerning the computational problems within the framework of the method under discussion, one should pay attention to the calculations of $F^{(*)}(\{\vec{x}_i\}|n;M)$ (23), which, as noted above, imply an estimation of the class M parameters $\vec{\theta}$. After that the choice (24) of the most likelihood class M corresponding $F^{(*)}(\{\vec{x}_i\}|n;M)$ can be made by means of known discrete maximization methods.

So, returning to obtained in the previous section system of solutions (22), which provides calculations (23), let us note the following. Although the left-hand parts of the equations (22) are the optimal distributions in an explicit form, their right-hand parts are linked with the neighboring distributions. So, the system (22) is implicit and highly nonlinear with respect to $\pi_{ik}^{(*)}$ and $\mathcal{Q}^{(*)}(\vec{\theta})$. This is a serious problem to the analytical treatment, but it is extremely favorable for computing. Indeed, system (22) is ideal for iterative computations: starting, for example, with $\mathcal{Q}^{(0)}(\vec{\theta}) = \mathcal{P}(\vec{\theta}|M)$, we can calculate by (23) $F^{(0)}$, after that find by (22) $\pi_{ik}^{(1)}$, with its help find next approximation $\mathcal{Q}^{(1)}(\vec{\theta})$, etc.:

$$\begin{gathered}\pi_{ik}^{(j)} = \frac{\exp\left(\int_T \mathcal{Q}^{(j-1)}(\vec{\theta})\ln\left(\rho(\vec{x}_i,k|\vec{\theta},M)\right)d\vec{\theta}\right)}{\sum_{l=1}^{G}\exp\left(\int_T \mathcal{Q}^{(j-1)}(\vec{\theta})\ln\left(\rho(\vec{x}_i,l|\vec{\theta},M)\right)d\vec{\theta}\right)}, \\ \mathcal{Q}^{(j)}(\vec{\theta}) = \frac{1}{Z_Q}n!\exp\left(\sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{ik}^{(j)}\ln\left(\rho(\vec{x}_i,k|\vec{\theta},M)\right)\right)\mathcal{P}(\vec{\theta}|M), \\ F^{(j)} = \ln\left(n!\prod_{i=1}^{n}\left[\sum_{k=1}^{G}\exp\left\{\int_T \mathcal{Q}^{(j)}(\vec{\theta})\ln\left(\rho(\vec{x}_i,k|\vec{\theta},M)\right)d\vec{\theta}\right\}\right]\right) - \\ -D_{KL}\left(\mathcal{Q}^{(j)}(\vec{\theta}),\mathcal{P}(\vec{\theta}|M)\right),\end{gathered} \quad (25)$$

not forgetting at each iteration to increment counter j until $F^{(j)}$ stabilizes at some level: $\left|F^{(j)} - F^{(j-1)}\right| < \varepsilon$ for given small ε.

The only technical problem for (25) computer implementation is the continuous nature of parameters $\vec{\theta}$, which implies continuous number of $Q^{(j)}(\vec{\theta})$ values, appearing in the corresponding integrals. However, this problem can be easily overcome if the distribution $Q^{(j)}(\vec{\theta})$ depends on a finite number of parameters and so integrals of $Q^{(j)}(\vec{\theta})$ in (25) can be expressed through these parameters. Let us note that this could be done from the very beginning if the free energy $F(q(\cdot), \ldots)$ (13) would be maximized (14) over a more narrowed distributions class of factorizable parametric functions. However, it is more convenient to have the general form of the solutions (22), since it can always be restricted to a desirable subclass of variable functions. As for the question of the best $Q^{(j)}(\vec{\theta})$ parameterization, the answer depends on the problem, on the relationship between the quantities of a priori and a posteriori data, on available computing resources, etc. Below, we consider two typical cases − the analysis of precedent which leads to formal description of its intensity shape, and the case of shape identification of some registered point process using formal descriptions of available precedents intensity shapes.

The case of precedent analysis characterized by a poor a priori information, but the amount of a posteriori data is, in principal, not limited. Therefore, due to a posteriori data let $Q^{(j)}(\vec{\theta})$ be a very narrow distribution within a maximum $\vec{\theta}^{(j)}$. In other words, let us take the class of parametric $Q^{(j)}(\vec{\theta})$ as a family of Dirac δ−functions $\{\delta(\vec{\theta} - \vec{\theta}^{(j)})\}$. It follows from (21) that $\vec{\theta}^{(j)}$ can be found as maximizing exponent function in $Q^{(j)}(\vec{\theta})$:

$$\begin{gathered} \vec{\theta}^{(j)} = \max_{\vec{\theta}} \left(Q\left(\vec{\theta}, \vec{\theta}^{(j-1)}\right)\right), \\ Q\left(\vec{\theta}, \vec{\theta}^{(j-1)}\right) = \sum_{y_1=1}^{G} \ldots \sum_{y_n=1}^{G} P_{\{y_i\}}^{(j)} p\left(\{\vec{x}_i\}, \{y_i\}; \vec{\theta} | n; M\right) \end{gathered}, \tag{26}$$

where the value $Q\left(\vec{\theta}, \vec{\theta}^{(j-1)}\right)$, usually called as the Q-function, is the conditional expectation (with respect to the hidden variables $\{y_i\}$) of the complete variables

$(\{x_i\},\{y_i\})$ and $\vec{\theta}$ joint distribution logarithm. The second argument of Q-function is denoted as $\vec{\theta}^{(j-1)}$, because the distribution $P_{\{y_i\}}^{(j)}$of hidden variables as the product of $\pi_{ik}^{(j)}$–s depends exactly on $\vec{\theta}^{(j-1)}$. Indeed, substituting $Q^{(j-1)}(\vec{\theta}) = \delta(\vec{\theta} - \vec{\theta}^{(j-1)})$ into the $\pi_{ik}^{(j)}$ (25), we obtain posterior distribution of y_i precisely for $\vec{\theta}^{(j-1)}$:

$$\pi_{iy_i}^{(j)} = \frac{\rho(\vec{x}_i,y_i|\vec{\theta}^{(j-1)},M)}{\sum_{l=1}^{G}\rho(\vec{x}_i,l|\vec{\theta}^{(j-1)},M)} = \frac{\rho(\vec{x}_i,y_i|\vec{\theta}^{(j-1)},M)}{\rho(\vec{x}_i|\vec{\theta}^{(j-1)},M)} = \rho(y_i|\vec{x}_i;\vec{\theta}^{(j-1)},M) \ . \quad (27)$$

Combining expressions (26) and (27), we obtain the famous EM algorithm [10] for maximum a posterior (MAP) estimation of parameters $\vec{\theta}$:

$$\begin{matrix} E\text{:} & Q(\vec{\theta},\vec{\theta}^{(j-1)}) = \ln(n!) + \sum_{i=1}^{n}\sum_{k=1}^{G}\rho(k|\vec{x}_i;\vec{\theta}^{(j-1)},M)\ln\left(\rho(\vec{x}_i,k|\vec{\theta},M)\right) + \\ & \qquad +\ln\left(\mathcal{P}(\vec{\theta}|M)\right); \\ M\text{:} & \vec{\theta}^{(j)} = \max_{\vec{\theta}}\left(Q(\vec{\theta},\vec{\theta}^{(j-1)})\right); \end{matrix} \quad . (28)$$

If in E (28) we neglect the term $\ln\left(\mathcal{P}(\vec{\theta}|M)\right)$, whose maximum is smaller and wider in comparison with the maximum of preceding term, we obtain the corresponding (28) EM algorithm [10] for maximum likelihood (ML) estimation of parameters $\vec{\theta}$:

$$\begin{matrix} E\text{:} & Q(\vec{\theta},\vec{\theta}^{(j-1)}) = \ln(n!) + \sum_{i=1}^{n}\sum_{k=1}^{G}\rho(k|\vec{x}_i;\vec{\theta}^{(j-1)},M)\ln\left(\rho(\vec{x}_i,k|\vec{\theta},M)\right) \\ M\text{:} & \vec{\theta}^{(j)} = \max_{\vec{\theta}}\left(Q(\vec{\theta},\vec{\theta}^{(j-1)})\right) \end{matrix} \quad . (29)$$

Let us note that in calculations (28) and (29) at step E the constant $\ln(n!)$ can be ignored, since it has no effect on finding the maximum point $\vec{\theta}^{(j)}$ at step M.

Substituting $Q^{(j)}(\vec{\theta}) = \delta(\vec{\theta} - \vec{\theta}^{(j)})$ into the $F^{(j)}$ (25), we obtain free energy value at iteration j (taking into account that Dirac δ–function in numerical calculations is the Kronecker symbol $\delta_{\vec{\theta};\vec{\theta}^{(j)}}$ for which $Q^{(j)}(\vec{\theta})\ln\left(Q^{(j)}(\vec{\theta})\right) = 0$):

$$F^{(j)} = \ln\left(n!\prod_{i=1}^{n}\left[\sum_{k=1}^{G}\rho\left(\vec{x}_i, k|\vec{\theta}^{(j)}, M\right)\right]\right) + \ln\left(\mathcal{P}\left(\vec{\theta}^{(j)}|M\right)\right) =$$
$$= \ln\left(p\left(\{\vec{x}_i\}|n;\vec{\theta}^{(j)}, M\right)\right) + \ln\left(\mathcal{P}\left(\vec{\theta}^{(j)}|M\right)\right) = \ln\left(p\left(\{\vec{x}_i\};\vec{\theta}^{(j)}|n;M\right)\right). \quad (30)$$

Thus, for precedents the main calculation time is spent on computing of $F^{(j)}$ (30), which is either the reduced log−likelihood function $L_S\left(\{\vec{x}_i\}\middle|n;\vec{\theta}^{(j)}, M\right) = \ln\left(p\left(\{\vec{x}_i\}|n;\vec{\theta}^{(j)}, M\right)\right)$ (4), or its improved with $\ln\left(\mathcal{P}\left(\vec{\theta}^{(j)}|M\right)\right)$ joint log-likelihood function $L_{S\theta}\left(\{\vec{x}_i\};\vec{\theta}^{(j)}\middle|n;M\right) = \ln\left(p\left(\{\vec{x}_i\};\vec{\theta}^{(j)}|n;M\right)\right)$ (10), depending on whether we neglect or not the poor a priori information contained in the $\mathcal{P}\left(\vec{\theta}^{(j)}|M\right)$. Corresponding to these two cases, the value of $\vec{\theta}^{(j)}$ is the ML estimate in EM algorithm (29) or the MAP estimate in EM algorithm (28). Similar results were presented earlier in [11] for the one-dimensional state space S.

In the case of a registered process, which may be a variant of one of the available precedents, a priori uncertainty in the case of a correctly selected precedent, on the contrary, is small. In that case almost all parameters $\vec{\theta}$ are well defined and only some of them need to be estimated. Let us note, that due to this simplification the estimation of unknown parameters can be made with more accuracy.

Supposing, that the registered process belongs to the class M, let us construct an approximation of log-likelihood function for its realization $(n, \{\vec{x}_1, \dots, \vec{x}_n\})$ − free energy $F^{(*)}(\{\vec{x}_i\}|n;M)$ (23) in which all parameters $\vec{\theta}$, except for translation $\vec{t}$ and scale s (common for all intensity shape components) are set equal to the parameters of belonging to the same class existing precedent. In other words, consider the distribution density $p\left(\vec{x}\middle|\vec{\theta}, M\right)$ (6), characterizing the form of the intensity of the registered process, as a given function depending on two parameters:

$$p\left(\vec{x}\middle|\vec{t}, s, M\right) = p_{apr}\left(s\vec{x} + \vec{t}\middle|M\right)s^D = \frac{\rho\left(s\vec{x}+\vec{t}, y|M\right)s^D}{\rho\left(y|s\vec{x}+\vec{t};M\right)}, \quad (31)$$

where $\rho(\vec{x}, y|M)$ and $\rho(y|\vec{x};M)$ are known joint and a posteriori distribution of complete $(\vec{x}, y)$ and hidden y variables of class M precedent, exponent D is the dimension of state space S.

It follows from (31), that the main computational problem concerning algorithm (25) – problem of distribution $\mathcal{Q}^{(j)}(\vec{\theta})$ parameterization is reduced in case considered to the best parametrization of:

$$\mathcal{Q}^{(j)}(\vec{t},s) = \frac{1}{Z_Q} n! \exp\left(\sum_{i=1}^{n}\sum_{k=1}^{G} \pi_{ik}^{(j)} \ln\left(\rho(s\vec{x}_i + \vec{t}, k|M)\right)\right) s^{nD}\mathcal{P}(\vec{t},s)\,, \quad (32)$$

where we emphasize that the a priori parameters distribution $\mathcal{P}(\vec{t},s)$ does not depend on class M.

To simplify the subsequent analytic solutions, let us assume, firstly, that for $\ln(\rho(\vec{x},y|M))$ there is a good approximation quadratic in $\vec{x}$ (it holds exactly in the case of Gaussian mixtures):

$$\ln(\rho(\vec{x},y|M)) \cong \ln\left(\rho(\vec{c}_y(M),y|M)\right) - \frac{1}{2}(\vec{x}-\vec{c}_y(M))^T A_y(M)(\vec{x}-\vec{c}_y(M))\,. \quad (33)$$

Secondly, let us assume that $\ln\left(s^{nD}\mathcal{P}(\vec{t},s)\right)$ also permits a good quadratic approximation:

$$\ln\left(s^{nD}\mathcal{P}(\vec{t},s)\right) \cong \ln\left({\beta_n}^{nD}\mathcal{P}(\vec{0},\beta_n)\right) - \frac{1}{2\Delta^2}\vec{t}^{\,2} - \frac{1}{2\sigma_n^2}(s-\beta_n)^2\,, \quad (34)$$

where $\vec{t}$ and s are considered as independent. In simpler terms, we assume that the distribution $\vec{t}$ is Gaussian and the distribution s, for example, the gamma distribution, is well approximated by a Gaussian distribution.

It is easy to conclude that under assumptions made $\mathcal{Q}^{(j)}(\vec{t},s)$ (32) takes the Gaussian form:

$$\mathcal{Q}^{(j)}(\vec{t},s) = \frac{\sqrt{\det\left(\mathcal{A}^{(j)}(n;M)\right)}}{\sqrt{(2\pi)^{D+1}}}\left(-\frac{1}{2}\left(\vec{\theta}-\vec{\theta}^{(j)}\right)^T \mathcal{A}^{(j)}(n;M)\left(\vec{\theta}-\vec{\theta}^{(j)}\right)\right), \quad (35)$$

where $\vec{t}$ and s are combined into $D+1$ dimensional vector $\vec{\theta}^T = (t_1, \ldots t_D, s)$.

The maximum point $\vec{\theta}^{(j)}$ and the matrix $\mathcal{A}^{(j)}(n;M)$ in (35) can be found by twice differentiating $\ln\left(\mathcal{Q}^{(j)}(\vec{t},s)\right)$ (32) with respect to parameters $(\vec{t},s)$ and substituting into the result the approximations (33) and (34):

$$\vec{\theta}^{(j)} = \mathcal{B}^{(j)}(n;M)\vec{d}^{(j)}(n;M)\,;\ \mathcal{B}^{(j)}(n;M) = \left[\mathcal{A}^{(j)}(n;M)\right]^{-1},$$
$$\mathcal{A}^{(j)}(n;M) = \begin{pmatrix} \sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{ik}^{(j)}A_k(M) + \frac{1}{\Delta^2}E_D & \sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{ik}^{(j)}A_k(M)\vec{x}_i \\ \sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{ik}^{(j)}\vec{x}_i{}^{T}A_k(M) & \sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{ik}^{(j)}\vec{x}_i{}^{T}A_k(M)\vec{x}_i + \frac{1}{\sigma_n^2} \end{pmatrix},$$
$$\vec{d}^{(j)}(n;M) = \left(\sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{ik}^{(j)}\vec{c}_k^{T}(M)A_k(M) \quad \sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{ik}^{(j)}\vec{c}_k^{T}(M)A_k(M)\vec{x}_i + \frac{\beta_n}{\sigma_n^2}\right)^{T}, \tag{36}$$

where E_D denotes the unity matrix of dimension D.

On the base of approximations (33), (34) and the resulting Gaussian parametrization of $Q^{(j)}(\vec{t},s)$ (35), (36), one can find analytic expressions for corresponding integrals in the main algorithm (25):

$$\int_T Q^{(j)}(\vec{t},s)\ln\left(\rho(s\vec{x}+\vec{t},y|M)s^D\right)d\vec{t}ds = \ln\left(\rho\left(s^{(j)}\vec{x}+\vec{t}^{(j)},y|M\right)s^{(j)D}\right) -$$
$$-\frac{1}{2}Tr\left(\mathcal{B}^{(j)}(n;M)\mathcal{C}(\vec{x},y,n;M)\right) + \frac{1}{n}\ln\left(\mathcal{P}\left(\vec{t}^{(j)},s^{(j)}\right)\right) -$$
$$-\frac{1}{n}\int_T Q^{(j)}(\vec{t},s)\ln\left(\mathcal{P}(\vec{t},s)\right)d\vec{t}ds\,,\ \mathcal{B}^{(j)}(n;M) = \left[\mathcal{A}^{(j)}(n;M)\right]^{-1},$$
$$\mathcal{C}(\vec{x},y,n;M) = \begin{pmatrix} A_y(M) + \frac{1}{\Delta^2 n}E_D & A_y(M)\vec{x} \\ \vec{x}^{T}A_y(M) & \vec{x}^{T}A_y(M)\vec{x} + \frac{1}{\sigma_n^2 n} \end{pmatrix},$$
$$D_{KL}\left(Q^{(j)}(\vec{\theta}),\mathcal{P}(\vec{\theta}|M)\right) = \frac{1}{2}\ln\left(\det\left(\mathcal{A}^{(j)}(n;M)\right)\right) - \frac{D+1}{2}\ln(2\pi e) - \tag{37}$$
$$-\int_T Q^{(j)}(\vec{t},s)\ln\left(\mathcal{P}(\vec{t},s)\right)d\vec{t}ds$$

Substituting expressions (37) in (25), we obtain $\pi_{ik}^{(j)}$ and $F^{(j)}$ in the free from integrals form:

$$\pi_{ik}^{(j)} = \frac{\rho\left(k|s^{(j-1)}\vec{x}_i+\vec{t}^{(j-1)};M\right)\exp\left(-\frac{1}{2}Tr\left(\mathcal{B}^{(j-1)}(n;M)\mathcal{C}(\vec{x}_i,k,n;M)\right)\right)}{\sum_{l=1}^{G}\rho\left(l|s^{(j-1)}\vec{x}_i+\vec{t}^{(j-1)};M\right)\exp\left(-\frac{1}{2}Tr\left(\mathcal{B}^{(j-1)}(n;M)\mathcal{C}(\vec{x}_i,l,n;M)\right)\right)},$$
$$F^{(j)} = L_{S\theta}\left(\{\vec{x}_i\};\vec{t}^{(j)},s^{(j)}\middle|n;M\right) - \frac{1}{2}\ln\left(\det\left(\mathcal{A}^{(j)}(n;M)\right)\right) + \frac{D+1}{2}\ln(2\pi e) +$$
$$+\ln\left(\prod_{i=1}^{n}\left[\sum_{k=1}^{G}\rho\left(k|s^{(j)}\vec{x}_i+\vec{t}^{(j)};M\right)\exp\left(-\frac{1}{2}Tr\left(\mathcal{B}^{(j)}(n;M)\mathcal{C}(\vec{x}_i,k,n;M)\right)\right)\right]\right), \tag{38}$$

where notation $L_{S\theta}\left(\{\vec{x}_i\};\vec{t}^{(j)},s^{(j)}\middle|n;M\right)$ (10) for $\ln\left(p\left(\{\vec{x}_i\};\vec{t}^{(j)},s^{(j)}|n;M\right)\right)$ is used.

Collecting all the expressions (36), (37), (38) together, we obtain the realization of the iterative algorithm (25) for parameters $\left(\vec{t}^{(*)},s^{(*)}\right)$ estimation and free energy $F^{(*)}(\{\vec{x}_i\}|n;M)$ calculation in the case of a registered process with available precedents in the following final form:

Initialization: for the given class M precedent and the number of registered process points n calculate

for all precedent components $k=1,\dots,G$ center vectors $\{\vec{c}_k\}$ and matrixes $\{A_k\}$ (33),

for a priory $\left(\vec{t},s\right)$ distribution dispersion Δ^2,σ_n^2, and the expected s value β_n (34),

for all n registered point locations $\vec{x}_i$ and all precedent components matrixes $\{\mathcal{C}_{ik}\}$:

$$\mathcal{C}_{ik}=\begin{pmatrix}A_k+\frac{1}{\Delta^2 n}E_D & A_k\vec{x}_i\\ \vec{x}_i^{\,T}A_k & \vec{x}_i^{\,T}A_k\vec{x}_i+\frac{1}{\sigma_n^2 n}\end{pmatrix}, \tag{39}$$

assuming $\pi_{ik}^{(0)}=\frac{1}{G}$ for all $i=1,\dots,n$ and all $k=1,\dots,G$ initialize the following:

$$\begin{gathered}\mathcal{A}^{(0)}=\begin{pmatrix}\frac{1}{\mathrm{G}}\sum_{i=1}^{n}\sum_{k=1}^{G}A_k+\frac{1}{\Delta^2}E_D & \frac{1}{\mathrm{G}}\sum_{i=1}^{n}\sum_{k=1}^{G}A_k\vec{x}_i\\ \frac{1}{\mathrm{G}}\sum_{i=1}^{n}\sum_{k=1}^{G}\vec{x}_i^{\,T}A_k & \frac{1}{\mathrm{G}}\sum_{i=1}^{n}\sum_{k=1}^{G}\vec{x}_i^{\,T}A_k\vec{x}_i+\frac{1}{\sigma_n^2}\end{pmatrix},\\ \vec{d}^{(0)}=\left(\frac{1}{\mathrm{G}}\sum_{i=1}^{n}\sum_{k=1}^{G}\vec{c}_k^{\,T}A_k\quad\frac{1}{\mathrm{G}}\sum_{i=1}^{n}\sum_{k=1}^{G}\vec{c}_k^{\,T}A_k\vec{x}_i+\frac{\beta_n}{\sigma_n^2}\right)^T,\\ \left(\vec{t}^{(0)},s^{(0)}\right)^T=\mathcal{B}^{(0)}\vec{d}^{(0)};\ \mathcal{B}^{(0)}=\left[\mathcal{A}^{(0)}\right]^{-1},\\ F^{(0)}=L_{S\theta}\left(\{\vec{x}_i\};\vec{t}^{(0)},s^{(0)}\middle|n;M\right)-\frac{1}{2}\ln\left(\det\left(\mathcal{A}^{(0)}\right)\right)+\frac{D+1}{2}\ln(2\pi e)+\\ +\ln\left(\prod_{i=1}^{n}\left[\sum_{k=1}^{G}\rho\left(k|s^{(0)}\vec{x}_i+\vec{t}^{(0)};M\right)\exp\left(-\frac{1}{2}Tr\left(\mathcal{B}^{(0)}\mathcal{C}_{ik}\right)\right)\right]\right),\end{gathered} \tag{40}$$

set counter $j=0$ and small ε for the stop criterion;

Iteration: increment counter $j=j+1$ until $F^{(j)}$ stabilizes at some level: $\left|F^{(j)}-F^{(j-1)}\right|<\varepsilon$

$$\pi_{ik}^{(j)} = \frac{\rho\left(k|s^{(j-1)}\vec{x}_i+\vec{t}^{(j-1)};M\right)\exp\left(-\frac{1}{2}Tr\left(\mathcal{B}^{(j-1)}(n;M)\mathcal{C}_{ik}\right)\right)}{\sum_{l=1}^{G}\rho\left(l|s^{(j-1)}\vec{x}_i+\vec{t}^{(j-1)};M\right)\exp\left(-\frac{1}{2}Tr\left(\mathcal{B}^{(j-1)}(n;M)\mathcal{C}_{il}\right)\right)}$$

$$\mathcal{A}^{(j)} = \begin{pmatrix} \sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{ik}^{(j)}A_k + \frac{1}{\Delta^2}E_D & \sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{ik}^{(j)}A_k\vec{x}_i \\ \sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{ik}^{(j)}\vec{x}_i^{\,T}A_k & \sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{ik}^{(j)}\vec{x}_i^{\,T}A_k\vec{x}_i + \frac{1}{\sigma_n^2} \end{pmatrix},$$

$$\vec{d}^{(j)} = \left(\sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{ik}^{(j)}\vec{c}_k^{T}A_k \quad \sum_{i=1}^{n}\sum_{k=1}^{G}\pi_{ik}^{(j)}\vec{c}_k^{T}A_k\vec{x}_i + \frac{\beta_n}{\sigma_n^2}\right)^T \qquad (41)$$

$$\left(\vec{t}^{(j)}, s^{(j)}\right)^T = \mathcal{B}^{(j)}\vec{d}^{(0)};\ \mathcal{B}^{(j)} = \left[\mathcal{A}^{(j)}\right]^{-1},$$

$$F^{(j)} = L_{S\theta}\left(\{\vec{x}_i\};\vec{t}^{(j)}, s^{(j)}\middle|n;M\right) - \frac{1}{2}\ln\left(\det\left(\mathcal{A}^{(j)}\right)\right) + \frac{D+1}{2}\ln(2\pi e) +$$

$$+\ln\left(\prod_{i=1}^{n}\left[\sum_{k=1}^{G}\rho\left(k|s^{(j)}\vec{x}_i + \vec{t}^{(j)};M\right)\exp\left(-\frac{1}{2}Tr\left(\mathcal{B}^{(j)}\mathcal{C}_{ik}\right)\right)\right]\right),$$

The advantage of the obtained algorithm (39), (40), (41) is that as the iterations grow, the free energy necessarily increases, and being bounded from above by $\tilde{L}((n,\{\vec{x}_i\})\,|M)$ (14) always converges to a certain limit. And this means that the process of computation will always stop.

In a more simpler form, the above results for the particular case $D = 2$, in the problem of image identification were discussed in [12].

Conclusions

The proposed in the paper approach and its algorithmic implementations are clear in theoretical concepts and computationally efficient. Using the principles of the statistical, models based inferences approach, we show that accepted identification procedure can be reduced to finding the maximum likelihood (or maximum a posterior) estimates of some parameters for each precedent and the subsequent comparison of the resulting likelihood functions for all the precedents. Developed and discussed VM-like algorithms for calculating that parameters seems to be the most reliable in many respects. In this connection, explicit expressions are given for estimating parameters and likelihood functions iterative computation.

Some results of numerical modeling, reflecting the potential characteristics of the considered approach are presented in [12]. Since they are better than the results of other methods, we hope that the identification of the point process intensity shape

proposed in the paper will be useful not only theoretically, but also for developing computer algorithms for many applications.

The work was supported by a RFBR grant 15-07-04378.

References

1. Cox D.R., Isham V. Point processes. London, New York: Chapman and Hall, 1980.
2. Streit R.L. Poisson Point Processes, Imaging, Tracking, and Sensing. New York, London: Springer, 2010, DOI: 10.1007/978-1-4419-6923-1.
3. Keeler P.H. Notes on the Poisson point process. Technical report, license "CC BY-SA 3.0", 2016.
4. Cox D.R. On the Estimation of the Intensity Function of a Stationary Point Process. *Journal of the Royal Statistical Society*, *Series B*, Vol. 27, No. 2, 1965, pp. 332-337.
5. Cook R.J., Lawless J.F. The Statistical Analysis of Recurrent Events. New York: Springer, 2007, DOI: 10.1007/978-0-387-69810-6.
6. Trevor H.T., Tibshirani R., Friedman J. The Elements of Statistical Learning. 2-d ed. New York: Springer, 2009.
7. Beal M.J. Variational Algorithms for Approximate Bayesian Inference. PhD. Thesis, London: Gatsby Computational Neuroscience Unit, University College London, 2003.
8. McLachlan G. and Peel D. Finite Mixture Models. Wiley Series in Probability and Statistics, New York: John Wiley & Sons, 2000, DOI:10.1002/0471721182.
9. MacKay D.J.C. Information Theory, Inference, and Learning Algorithms (First ed.). Cambridge: Cambridge University Press, 2003.

10. Bishop C.M. Pattern Recognition and Machine Learning. New York: Springer-Verlag, 2006.

11. Antsiperov V., Mansurov G. Precedent Identification of Nonhomogeneous Point Process Intensity Dynamics. *Doklady 19 Mezhdunarodnoy konferentsii "Tsifrovaya obrabotka signalov i ee primenenie* - DSPA-2017" [Proc. 19-th International Conference “Digital Signal Processing and its Applications” (DSPA-2017)], Moscow, Russia, 2017, pp. 112-116 (in Russian).

12. Antsiperov V.E. Automatic target recognition algorithm for low-count terahertz images. *Computer optics*, Vol. 40, No 5, 2016, pp. 746-751.

For citation:
V. E. Antsiperov. Identification of the point process intensity shape with the precedents maximum likelihood distributions. *Zhurnal Radioelektroniki - Journal of Radio Electronics*. 2017. No. 12. Available at http://jre.cplire.ru/jre/dec17/11/text.pdf.