Acta Universitatis Carolinae. Mathematica et Physica

Zbyněk Pawlas

Estimation of interarrival time distribution from short time windows

Acta Universitatis Carolinae. Mathematica et Physica, Vol. 52 (2011), No. 1, 59--67

Persistent URL: http://dml.cz/dmlcz/143668

Estimation of Interarrival Time Distribution from Short Time Windows

ZBYNĚK PAWLAS

Praha

Received April 30, 2010
Revised June 9, 2010

We propose several estimators of interarrival time distribution based on observations of independent identically distributed stationary point processes in time windows with length of the same order as the mean interarrival time. This task is motivated by the situation in which a high number of neurons communicates with a target neuron. The comparison of the finite sample performance of the estimators is carried out by a simulation study for three selected models of point processes, namely Poisson point process, renewal process and mixed Poisson process.

1. Introduction

Point processes provide an important tool for modeling and analyzing data in the form of random events in time. The whole point process can be described completely by the random intervals between the events (so called interarrival times). These intervals may be dependent and not necessarily identically distributed. By means of Palm distributions, we can define the typical interarrival time. Our aim is to estimate its distribution.

Often we deal with data recorded for a long period of time and we make statistical inference about a point process from this one realization. The nonparametric estimation of the interarrival distribution from a single realization of a renewal process is examined in [5]. We are concerned with a different situation appearing when the information about a stationary point process has to be deduced from independent

MFF UK, KPMS, Sokolovská 83, 186 75 Praha 8 – Karlín

This work was supported by the research project MSM 0021620839 financed by MŠMT ČR and by grant IAA 101120604 from GA AV ČR.

Key words and phrases. Distribution function estimation, interarrival time distribution, mixed Poisson process, point process, renewal process

E-mail address: pawlas@karlin.mff.cuni.cz

and identically distributed realizations. We assume that the data are observed in relatively short time period (window). The disadvantage of this setting is the presence of length bias in the estimators. There are no observed interarrival intervals longer than the window length. Therefore, we are not able to construct the estimators for values exceeding the window length without further specific assumptions on the model. Accordingly, we concentrate on the estimation of a truncated distribution of the typical interarrival time. For finite number of observations, the estimators are compared according to their accuracy by a simulation study.

The motivation for our investigation comes from neurophysiology. Consider a homogeneous population of neurons. They communicate with one another by firing action potentials (also called spikes). Since both the shape and the amplitude of the spike are believed to carry minimal information, the main focus in neural coding is devoted to the timing of spikes. A sequence of spikes is called a spike train. Information in the nervous system is transferred by a temporal structure of spike trains generated by individual neurons (we speak about temporal coding). Thus, spike trains can be viewed as point processes where the points correspond to the times when spikes occur. A popular way to analyze a spike train is to look at the differences of times for two successive spikes (so called interspike intervals). Therefore, our aim is to estimate the characteristics of interspike intervals. Assume that a target neuron receives the information through spike trains produced by each neuron in the population. The target neuron has to respond within a short time interval. This scenario was considered in [7] where the parametric estimation of the statistical moments of interspike intervals was discussed. In the present paper we estimate the distribution of interspike intervals from independent replications of spike trains observed in a short time window.

2. Point processes

We define a point process $\Phi = \{X_i, i \in \mathbb{Z}\}$ as a random locally finite sequence of points on the real line $\mathbb{R}$. We assume $\ldots < X_{-1} < X_0 \leq 0 < X_1 < X_2 < \ldots$, i.e. no multiple points are allowed. The interarrival (or interspike) intervals are denoted by $T_i = X_{i+1} - X_i$, $i \in \mathbb{Z}$. We write shortly $\Phi(t)$ for the number of events from time 0 up to time t. More details on point processes can be found e.g. in [3], [4] or [5].

We will consider stationary point processes, i.e. the distribution is invariant under translation. Then the mean number of points in a given interval is proportional to the length of this interval, i.e. $\mathbb{E}\Phi(t) = \lambda t$. The constant λ is called the intensity of the point process Φ.

Let P_0 be the Palm distribution of a stationary point process Φ (see [4] or [5]). It can be represented as a conditional distribution of Φ under the condition that $0 \in \Phi$. We say that T is a typical interarrival time (or interval) if its distribution is same as

the distribution of X_1 under P_0. The distribution of T is called the interarrival time distribution. The corresponding distribution function will be denoted as

$$F(t) = \mathbb{P}(T \le t) = P_0(\{\varphi : \inf\{x : x \in \varphi, x > 0\} \le t\}), \quad t > 0.$$

In this section we will assume that a stationary point process Φ is observed in a single time window $[0, \Delta]$ of length $\Delta > 0$. We will consider several point process models and describe the estimation of the interarrival time distribution. This will serve as a preparation for the next section where we proceed to the estimation from multiple time windows. We suppose that $\Phi(\Delta) > 0$, for $\Phi(\Delta) = 0$ we always define the estimator to be identically 0 on the interval $[0, \Delta]$.

2.1 Poisson process

The simplest point process is the Poisson process. It is a model of the events occuring completely at random in time. For a stationary Poisson point process with finite and positive intensity λ, the interarrival intervals T_i, $i \ge 1$, are independent copies of a positive random variable T which has an exponential distribution with mean $1/\lambda$. The interarrival time distribution coincides with the distribution of T. Therefore, in order to get an estimator of $F(t) = \mathbb{P}(T \le t) = 1 - \mathrm{e}^{-\lambda t}$, it is enough to estimate parameter λ. The maximum likelihood estimator of λ is $\hat{\lambda} = \Phi(\Delta)/\Delta$ and hence the plug-in estimator of $F(t)$ becomes

$$\hat{F}^{PP}(t) = 1 - \mathrm{e}^{-\hat{\lambda} t}, \quad t \ge 0.$$

Note that this estimator is not unbiased and has other inferior properties. For example, it is not admissible under the L^2 loss function, see e.g. [6]. We consider this natural estimator only for comparison with nonparametric approaches. Our aim is not to propose estimators in this parametric model.

2.2 Renewal process

The Poisson point process can be generalized in several directions. One of them leads to renewal processes which form an important class of point processes, see [3] or [5]. The interarrival intervals T_i are independent copies of a positive random variable T with probability distribution function $F(t)$. Again, the interarrival time distribution coincides with the distribution of T. In contrast to the Poisson process, T is not necessarily exponentially distributed.

Similarly as for the Poisson process we could exploit parametric approach. However, we prefer to take $F(t)$ to be completely unknown. A naive nonparametric estimator of $F(t)$ is the empirical distribution function

$$\tilde{F}(t) = \frac{1}{\Phi(\Delta) - 1} \sum_{i=1}^{\Phi(\Delta)-1} \mathbf{1}\{T_i \le t\}, \quad t \le \Delta,$$

defined for $\Phi(\Delta) > 1$. Here, the number of summands heavily depends on the sequence $\{T_i\}$. Moreover, not all of the information contained in the data is used in forming this estimator. For example, the interval $[X_{\Phi(\Delta)}, \Delta]$ is not taken into account although it gives partial information about $T_{\Phi(\Delta)} = X_{\Phi(\Delta)+1} - X_{\Phi(\Delta)}$. We know that $T_{\Phi(\Delta)} > \Delta - X_{\Phi(\Delta)} = B_\Delta$ (known as backward recurrence time at Δ). Therefore, we define the modified estimator

$$\hat{F}(t) = \begin{cases} \frac{\Phi(\Delta)-1}{\Phi(\Delta)} \tilde{F}(t), & t \le B_\Delta, \\ \tilde{F}(t), & t > B_\Delta, \end{cases} \tag{1}$$

see also [5], p. 313. For $\Phi(\Delta) = 1$ we put $\hat{F}(t) = \mathbf{1}\{t > B_\Delta\}$. As it is also noted in [5], the shortcoming is the length bias in the estimators. Shorter intervals have greater chance to be taken into account. Conversely, the intervals of length larger than Δ are not observed at all.

We can also view the situation as survival data problem. If $\Phi(\Delta) \ge 2$, the interarrival times $T_1, \dots, T_{\Phi(\Delta)-1}$ are known exactly, while the interarrival time $T_{\Phi(\Delta)}$ is right censored by the end of the observation period. To each T_i, $i = 1, \dots, \Phi(\Delta)$ we can associate a censoring time $C_i = \Delta - X_i$. This resembles censoring schemes known from survival analysis. There are three types of censoring. If the censoring times are fixed for each T_i, then we have type I censoring. On the other hand, if each variable is censored at random by a censoring time which is independent of this variable, then we speak about type III (or random) censoring. In our setting, the censoring times C_i are random but they are not independent and depend on the interarrival times T_i. Therefore, we are not in the situation of either type I or type III censoring, see [2] for the related issue in the context of spatial point processes. Nevertheless, in analogy with the theory of random censoring, we propose the Kaplan-Meier estimator,

$$\hat{F}^{KM}(t) = 1 - \prod_{s \le t} \left(1 - \frac{D(s)}{S(s)}\right), \tag{2}$$

where $D(s) = \sum_{i=1}^{\Phi(\Delta)-1} \mathbf{1}\{T_i = s\}$ and $S(s) = \sum_{i=1}^{\Phi(\Delta)-1} \mathbf{1}\{T_i \ge s\} + \mathbf{1}\{B_\Delta \ge s\}$. Again, for $\Phi(\Delta) = 1$ we put $\hat{F}^{KM}(t) = \mathbf{1}\{t > B_\Delta\}$. Since the assumptions of random censoring are violated, the optimality of the Kaplan-Meier estimator is destroyed. However, we may hope that it can compete with (1). The Kaplan-Meier estimator is nondecreasing and piecewise constant but may not reach 1 (it is always less or equal to 1). As opposed to the estimator (1), it does not have a jump in B_Δ. Alternatively, one may try to use the analogy with type I censoring and consider the corresponding estimator. We decided not to follow this direction.

Another way to estimate the distribution function $F(t)$ is to use the so-called reduced-sample estimator. We restrict attention to the points lying in the interval

$[0, \Delta - t]$. For these points the interarrival intervals T_i shorter than t are observed in a bounded window $[0, \Delta]$. We define

$$\hat{F}^{RS}(t) = \frac{1}{\Phi(\Delta - t)} \sum_{i=1}^{\Phi(\Delta - t)} \mathbf{1}\{T_i \le t\}, \quad t \le \Delta - X_1, \tag{3}$$

and $\hat{F}^{RS}(t) = 1$ for $t > \Delta - X_1$. Estimators of this type are often used in spatial statistics in order to deal with edge effects caused by the bounded observation window, see e.g. [1]. This approach is also called border method. For larger t the estimator (3) discards a lot of information given by the data. Since it is not necessarily monotone, we take its upper envelope

$$\hat{F}^{RS\,m}(t) = \sup_{s \le t} \hat{F}^{RS}(s). \tag{4}$$

2.3 Mixed Poisson process

Next possibility how to generalize the Poisson process is obtained by supposing stochastic intensity. Such processes are called doubly stochastic Poisson processes or Cox processes, see e.g. [3] or [5]. If there exists a nonnegative random variable Λ (called driving intensity) such that conditionally on $\Lambda = \lambda$, Φ is a Poisson process with intensity λ, then the doubly stochastic Poisson process Φ is termed the mixed Poisson process. It is an example of a stationary point process. Random driving intensity Λ is unobserved. From a single realization, mixed Poisson process cannot be distinguished from Poisson process. The interarrival time distribution function is

$$F(t) = \mathbb{E}\,[\mathbb{P}(T \le t \mid \Lambda)] = \mathbb{E}(1 - \mathrm{e}^{-\Lambda t}) = 1 - \mathbb{E}\mathrm{e}^{-\Lambda t}, \quad t \ge 0. \tag{5}$$

3. Estimation from multiple replications

We consider the situation when n independent copies $\Phi_1, \ldots, \Phi_n$ of a stationary point process Φ are observed over the interval $[0, \Delta]$. The points of Φ_k will be denoted by $X_i^{(k)}$ and the interarrival intervals by $T_i^{(k)}$. Moreover, let $B_\Delta^{(k)} = \Delta - X_{\Phi_k(\Delta)}$ be the backward recurrence time associated with Φ_k.

For the Poisson process the maximum likelihood estimator of its intensity is given by

$$\hat{\lambda}_n = \frac{1}{n\Delta} \sum_{k=1}^{n} \Phi_k(\Delta)$$

and the plug-in estimator of the distribution function $F(t)$ is

$$\hat{F}_n^{PP}(t) = 1 - \mathrm{e}^{-\hat{\lambda}_n t}, \quad t > 0.$$

A first approach to the nonparametric estimation from n observations is to simply take the mean of the separate estimators for each window:

$$\bar{F}_n(t) = \frac{1}{n}\sum_{k=1}^{n} \hat{F}^{\Phi_k}(t),$$

where $\hat{F}^{\Phi_k}$ is (1) applied to the process Φ_k. Similarly, we can define $\bar{F}_n^{KM}(t)$, $\bar{F}_n^{RS}(t)$ and $\bar{F}_n^{RSm}(t)$ as the averages of (2), (3) and (4), respectively.

However, more efficient strategy might be to pool the information from all replicated observations. The pooled Kaplan-Meier estimator and reduced-sample estimator are obtained by analogues of (2) and (3) in which the numerators and denominators are replaced by the sums over all Φ_k,

$$\hat{F}_n^{KM}(t) = 1 - \prod_{s\le t}\left(1 - \frac{\sum_{k=1}^{n} D_k(s)}{\sum_{k=1}^{n} S_k(s)}\right),$$

where $D_k(s) = \sum_{i=1}^{\Phi_k(\Delta)-1} \mathbf{1}\{T_i^{(k)} = s\}$ and $S_k(s) = \sum_{i=1}^{\Phi_k(\Delta)-1} \mathbf{1}\{T_i^{(k)} \ge s\} + \mathbf{1}\{B_\Delta^{(k)} \ge s\}$, and

$$\hat{F}_n^{RS}(t) = \frac{\sum_{k=1}^{n}\sum_{i=1}^{\Phi_k(\Delta-t)} \mathbf{1}\{T_i^{(k)} \le t\}}{\sum_{k=1}^{n} \Phi_k(\Delta - t)}, \quad t \le \Delta - \min_{k=1,\dots,n} X_1^{(k)}.$$

When the data consist of replicated spatial point processes, the analogous pooled estimators of summary statistics were considered in [2]. We also define a monotone version of the pooled reduced-sample estimator

$$\hat{F}_n^{RSm}(t) = \sup_{s\le t} \hat{F}_n^{RS}(s).$$

In the case of mixed Poisson process, we will estimate $F(t)$ by estimating the Laplace functional, see e.g. [4] or [5]. The Laplace functional for mixed Poisson process is

$$L_\Phi(g) = \mathbb{E}\exp\left\{-\int g(x)\,\Phi(\mathrm{d}x)\right\} = \mathbb{E}\exp\left\{-\Lambda\int (1-\mathrm{e}^{-g(x)})\,\mathrm{d}x\right\}.$$

In particular, if $g(x) = -\log(1 - t/\Delta)$, $0 \le x \le \Delta$, then $L_\Phi(g) = \mathbb{E}\mathrm{e}^{-\Lambda t}$, $t < \Delta$. Using (5) we have $F(t) = 1 - L_\Phi(g)$. The maximum likelihood estimator of $L_\Phi(g)$ is the empirical Laplace functional (see [5])

$$\widehat{L_\Phi(g)} = \frac{1}{n}\sum_{k=1}^{n} \exp\left\{-\sum_{i=1}^{\Phi_k(\Delta)} g(X_i^{(k)})\right\}.$$

Therefore, we define

$$\hat{F}_n^{MP}(t) = 1 - \widehat{L_\Phi(g)} = 1 - \frac{1}{n}\sum_{k=1}^{n}\left(1 - \frac{t}{\Delta}\right)^{\Phi_k(\Delta)}.$$

Since $\mathbb{E}\widehat{L_\Phi(g)} = L_\Phi(g)$, the estimator $\hat{F}_n^{MP}(t)$ is unbiased. Note that similarly as in the Poisson case we don't need the information about exact arrival times. The distribution function is estimated just from the point counts observed in each window.

The numbers of events form a sufficient statistic for $F(t)$ and thus the arrival times can be disregarded.

4. Simulation study

In order to demonstrate the properties of the estimators introduced in Section 3 we perform a small simulation study. Since we cannot observe interarrival times greater than Δ in the data, our main interest will be in the distribution function of typical interarrival interval T conditional on $T \le \Delta$, i.e. $G(t) = F(t)/F(\Delta)$, $0 \le t \le \Delta$. Based on an arbitrary estimator $\hat{F}_n$ of $F(t)$, we estimate the truncated distribution function $G(t)$ by $\hat{G}_n(t) = \hat{F}_n(t)/\hat{F}_n(\Delta)$, $0 \le t \le \Delta$. For the comparison of estimators quality we use integrated square error

$$d(\hat{G}_n, G) = \int_0^\Delta (\hat{G}_n(t) - G(t))^2 \,\mathrm{d}t. \tag{6}$$

We consider the following stationary point process models:

(i) Poisson point process with intensity λ,
(ii) renewal process with gamma interarrival distribution $\Gamma(a, b)$,
(iii) renewal process with inverse Gaussian interarrival distribution $\mathrm{IG}(a, b)$,
(iv) mixed Poisson process with $\Gamma(a, b)$-distributed driving intensity Λ.

Let $\mu = \mathbb{E}T$ and $\gamma = \sqrt{\operatorname{var} T}/\mu$ be the mean and coefficient of variation of the typical interarrival time T, respectively. The theoretical distribution functions, means and coefficients of variation of T are ($t > 0$)

(i) $F(t) = 1 - \exp\{-\lambda t\}$, $\mu = 1/\lambda$, $\gamma = 1$ (exponential distribution),
(ii) $F(t) = \int_0^t \frac{b^a}{\Gamma(a)} \mathrm{e}^{-bs} s^{a-1} \,\mathrm{d}s$, $\mu = a/b$, $\gamma = 1/\sqrt{a}$ (gamma distribution),
(iii) $F(t) = H((t/a - 1)/\sqrt{bt}) + \exp\{\frac{2}{ab}\} H(-(t/a + 1)/\sqrt{bt})$, $\mu = a$, $\gamma = \sqrt{ab}$ (inverse Gaussian distribution),
(iv) $F(t) = 1 - \left(\frac{b}{b+t}\right)^a$, $\mu = b/(a-1)$, $\gamma = \sqrt{a/(a-2)}$ (Pareto distribution),

where H is a distribution function of the standard normal distribution. We choose two different values of μ: 0.5 and 2. The coefficient of variation is $\gamma = 1$ for the Poisson process and we choose $\gamma = 1.5$ for remaining processes. We have generated $n = 500$ independent realizations of the selected point processes in the time window $[0, \Delta]$ with $\Delta = 1$. From simulated data we estimate $G(t)$ and compute integrated square errors (6). This procedure is repeated 1000 times for each choice of model parameters. The sample means of computed errors over 1000 simulations are shown in Table 1 for various estimators which were defined in Section 3. The simulations and computations were carried out in R [8].

It is not surprising that in the case of Poisson or mixed Poisson process, $\hat{F}_n^{PP}$ and $\hat{F}_n^{MP}$ have the best performance because these estimators are constructed from the corresponding models. However, they may turn to be useless when the assumption

TABLE 1. Mean integrated square errors computed from 1000 repetitions of $n = 500$ realizations of given point process. For each model two different values of μ are considered: 0.5 (left) and 2 (right)

	$1000 \cdot d(\hat{G}_n, G)$							
	Poisson		renewal Γ		renewal IG		mixed Poisson	
$\hat{F}_n^{PP}$	0.023	0.008	17.321	36.134	6.562	4.576	0.009	0.007
$\hat{F}_n^{KM}$	0.740	8.097	0.724	4.655	0.580	5.720	2.049	4.760
$\hat{F}_n^{RS}$	3.157	51.335	4.873	43.015	3.238	32.500	1.456	25.118
$\hat{F}_n^{RSm}$	2.156	32.044	3.789	30.651	2.227	19.033	1.114	15.922
$\bar{F}_n$	2.518	1.218	10.409	22.542	5.639	1.795	1.911	1.425
$\bar{F}_n^{KM}$	1.670	1.196	6.468	19.888	3.394	2.120	1.167	1.308
$\bar{F}_n^{RS}$	4.087	1.282	13.879	23.969	7.833	1.763	3.668	1.677
$\bar{F}_n^{RSm}$	2.611	1.219	10.690	22.599	5.802	1.794	2.083	1.434
$\hat{F}_n^{MP}$	0.040	0.037	11.383	22.207	2.838	5.715	0.002	0.004

of Poisson or mixed Poisson process fails. For renewal processes the pooled Kaplan-Meier estimator $\hat{F}_n^{KM}$ seems to be the best choice. The reduced-sample estimators differ only slightly from the Kaplan-Meier estimators for smaller values t but they are becoming very inaccurate for t close to Δ where the method discards much information. This causes large overall deviations from the true distribution function. Monotone modification of the estimates reduces the error little bit. For larger μ, which means less observed points, the average estimators $\bar{F}_n$ and $\bar{F}_n^{KM}$ give quite satisfactory results.

References

[1] BADDELEY, A. J.: *Spatial sampling and censoring*. In Stochastic Geometry: Likelihood and Computation, Barndorff-Nielsen, O. E., Kendall, W. S. and van Lieshout, M. N. M. (eds.), Chapman and Hall, London (1999), 37–78.

[2] BADDELEY, A. J., GILL, R. D.: *Kaplan-Meier estimators of distance distributions for spatial point processes*, Ann. Statist. **25** (1997), 263–292.

[3] COX, D. R., ISHAM, V.: Point Processes. Chapman and Hall, London (1980).

[4] DALEY, D. J., VERE-JONES, D.: An Introduction to the Theory of Point Processes. Springer Verlag, New York (1988).

[5] KARR, A. F.: Point Processes and Their Statistical Inference, 2nd edition. Marcel Dekker, Inc., New York (1991).

[6] KLEBANOV, L. B.: *Unbiased parametric estimation of a probability distribution*, Math. Notes **25** (1979), 383–387.

[7] Pawlas, Z., Klebanov, L. B., Prokop, M., Lansky, P.: *Parameters of spike trains observed in a short time window*, Neural Comput. **20** (2008), 1325–1343.
[8] R Development Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria (2010). URL: `http://www.R-project.org`.