Detection of onset of neuronal activity by allowing for heterogeneity in the change points

Y. Ritov[1,3] A. Raz[2,3] H. Bergman[2,3]

July, 2002

Correspondence:
Ya'acov Ritov
Department of Statistics
The Hebrew University of Jerusalem
91905 Jerusalem, Israel
Tel.: (972-2) 652-9929
Fax: (972-2) 652-9996
E-mail: yaacov@mscc.huji.ac.il

[1] Department of Statistics, The Hebrew University of Jerusalem, 91905 Jerusalem, Israel. The research was support by part by a grant from the Israel Science Foundation.

[2] Department of Physiology, The Hebrew University of Jerusalem, P.O.Box 12272, 91120 Jerusalem, Israel

[3] The Center for Neural Computation, The Hebrew University of Jerusalem.

Abstract

We consider situations in which there is a change point in the activity of a cell, that is, some time after an external event the firing rate of the cell changes. The change can occur after a random delay. The distribution of the time to change is considered unknown.

Formally we deal with n i.i.d. random point processes, each of these is an inhomogeneous Poisson processes, with one intensity until a random time, and a different intensity thereafter. Thus, the change point is not explicitly observed. We present both a simple estimator and the non-parametric maximum likelihood estimator (NPMLE) of the change point distribution, both having the same rate of convergence. This rate is proved to be the best possible. The extension of the basic model to multiple processes per trial with different intensities and joint multiple change points is demonstrated using both simulated and neural data. We show that for realistic spike train data, trial by trial estimation of a change point may be misleading, while the distribution of the change point distribution can be well estimated.

Keywords: Peri-stimulus time histogram, neural data, semiparametric models, hidden Markov models, mixture models, Empirical Bayes.

1 Introduction

The synchrony between neural activity and external event is a major tool in neurophysiological studies of the brain. The data is typically analyzed using the Peri-Stimulus Time Histogram (PSTH). Different trials are aligned with respect to the time of the external event and the average change over many trials of the intensity of the neural activity is observed. However, using this technique, one cannot distinguish between a smooth transition in any single trial between two regimes on one hand, and a sharp transition at each trial, but with a jitter in the transition times between the trials on the other.

Technically, we consider a situation in which i.i.d. copies of a multivariate point process on a fix interval are observed. The intensity of the process is not fixed along the interval but changes once or more. The time of the change points may vary from copy to copy.

In this manuscript the problem is treated in two different ways: both as a formal statistical problem and as tool for the analysis of neural data. The biological question behind the statistical discussion is the extent to which the activity of a specific group of neurons in the monkey's brain is synchronized with the external behavior of the animal. Here the model is extended to more than one change point and to multivariate counting processes. The estimation procedure assumes that the change points actually exist, and that the different components of the multivariate process are, given the change points times, independent inhomogeneous Poisson processes. These assumptions are not necessarily valid in all situations, but we argue that they are reasonable for our examples. The different extensions of the model are applied to real and simulated data.

Mathematically, we analyze a problem in which it is assumed that there is only one change point per trial, whose time is a random variable distributed according to a distribution function G. For each trial, we assume an inhomogeneous Poisson process that has a constant intensity λ_0 until the change point, and a constant intensity λ_1 thereafter. The actual time of the change point is not observed explicitly. The parameters G, λ_0 and λ_1 are not known. The information bounds and the efficient score functions are given. In a nutshell, we have an explicit expression for the score function only in a relatively trivial case. The maximum likelihood estimator as well as simple estimators are presented. In particular, the distribution function of the time to change can be estimated using a simple monotone regression estimator. The rates of convergence of these simple estimators are optimal.

The change point model for a single Poisson process (and a single trial) was discussed in a few earlier papers, e.g., Matthews, Farewell and Pyke (1985); Akman and Raftery (1986) and the standard Bayesian analysis is discussed in Raftery and Akman (1986). The change point methodology was discussed in the context of neuron activity by Commenges and Seal (1985), where the change point was estimated for each trial separately, which may be difficult in some applications. A typical firing rate for a neuron is a spike every 20 to 200ms on the average. Hence there is an error of a few hundreds millisecond in the estimation of a single change point. This is too crude for a typical behavioral task. In our simulation, we present an example, where the change point distribution can be estimated, reasonably well, while it almost impossible to locate the individual change points. Note however, that the brain system observes many cells at the same time, and therefore can detect the change point exactly, even when it is not possible in the experimental setting where only one or at most a few cells can be observed simultaneously.

Our point of view is akin to hierarchical Bayes or, closer, to the empirical Bayes formulation of the problem. Previously, empirical Bayes models were employed by Joseph and Wolfson (1992); Bélisle, Joseph, MacGibbon, Wolfson and du Berger (1998) in the context of change point detection for spike data. See a relevant recent discussion of empirical Bayes procedures in Efron (1996). Leaving philosophy aside, we consider the problem as a semiparametric mixture model, Bickel, Klaassen, Ritov and Wellner (1993); Robins and Ritov (1997). The typical neuronal experiment in which the activity of single cells is recorded involves a repeated task in which the animal is reacting to external cues. The experimenter tries to understand how the cells activity is related to different sensory and motor events. The distribution of the change point time may be interesting in particular in situations where it is not known *a priori* with which external event the neuronal activity is synchronized. In a typical experiment, an observed change in the neurons activity may be related to the visual cue that the monkey receives, to the eye movement that follows, or to the movement of his arm. The times of these events are recorded, and we may try to test to which of them the activity is better synchronized (Seal, Commenges, Salamon and Bioulac (1983); Seal and Commenges (1985); Schwartz, Kettner and Georgopouls (1988); Montgomery (1989); Crutcher and Alexander (1990); Romo and Schultz (1990)). In this paper, a single change point was located for each neuron and task condition using a formal hierarchical Bayesian methods. Our method was applied to other data set as presented in Ritov, Raz and Bergman (1997). The model analyzed by Bélisle et al. (1998) is similar to ours, except that it was analyzed using Bayesian tools, both in the model formulation and in the algorithms, they used Gibbs sampler, while we used a non-iterative simple estimator and the EM algorithm to calculate the maximum-likelihood estimator. Moreover, we extend their model to examples of multiple change point and multiple cells. Finally, we give the theoretical justification to the technique used.

2 Methods

Our empirical data were recorded from two awake vervet (green) monkeys (*Cercopithecus aethiops aethiops*). The monkeys were trained to perform a visual-motor task with two behavioral paradigms, see details in Raz, Vaadia and Bergman (2000). Briefly, the trials were as follows. Four seconds after the end of the previous trial the program started checking if the central key is touched. In most cases, the monkey would have touched the key during the inter-trial period. If not, the program waited until the key has been touched. Immediately (less than 1ms) after it touched the key, the “get ready” LED was turned on. After a variable delay (3 - 6s), one of the two peripheral target keys was illuminated for 0.25s, and the monkey got a trigger signal after another random delay of 1, 2, 4, or 8s. At this point the monkey was supposed either to release the central key and touch the target key (the “GO” paradigm), or to keep touching the central key (the “NO-GO” paradigm). If the monkey did this, it was rewarded with 0.15ml of juice. After 4 correct trials, there was a 4s signal instructing the monkey to change paradigm from “GO” to “NO-GO” or vice versa. The monkey was fully trained before recording started. In each recording session, the activity of two to eight single cells in the basal ganglia was recorded. In a single recording session, a few hundreds trials were recorded. Typically the number of valid records from any single cell is between a few tens to a few hundreds trials.

The cells whose activity we analyze are from the external segment of the globus pallidus (GPe). These cells are characterized with a fast tonic rate (a few tens of spikes per second), and with unexplained short intervals in which they are silent, DeLong (1971); DeLong and Georgopoulos (1981)). It looks as if the cells behave independently, Nini, Feingold, Slovin and Bergman (1995); Raz et al. (2000). The exact function and the mode of activity of the basal ganglia are not known. We assume that a model in which some cells switch abruptly to a different mode is reasonable.

3 Results

3.1 The model

The data used for the analysis can be summarized as follows. We measure the activity of $K \geq 1$ cells during n trials. For each trial we record the activity of each of the cells during a window synchronized on an given activity of the monkey during the trial.

Formally, the observations are at discrete time, $a, a+1, \dots, b$ for some a and b. We assume that for each trial i, $i = 1, \dots, n$, there are multiple change points $a < T_{i1} < \cdots < T_{iM} < b$, for some $M \geq 1$. We observe for each trial K counting processes. The processes are independent homogeneous Poisson processes between the common change points. In other words, for all $i = 1, \dots, n$, $\mathbb{N}_{i1}, \dots \mathbb{N}_{iK}$

are independent given $T_{i1},\ldots,T_{iM}$. The values $\mathbb{N}_{ik}(t)$, $i=1,\ldots,n$, $k=1,\ldots,K$, $t=a,\ldots,b$, represent the total number of spikes fired by the k-th cell during the time interval from a to t: $\mathbb{N}_{ik}(t) \equiv \sum_{s=a}^{t} N_{ik}(s)$. We assume that the Bernoulli random variables $N_{ik}(a),\ldots,N_{ik}(b)$ are independent given $T_{i1},\ldots,T_{iM}$, and $P(N_{ik}(t)=1 \mid T_{i1},\ldots,T_{iM}) = 1-P(N_{ik}(t)=0 \mid T_{i1},\ldots,T_{iM}) = p_{km}$, $T_{im} \leq t < T_{i,m+1}$, m=0,...,M, where, formally, $T_{i0} \equiv a$ and $T_{i,M+1} \equiv b+1$. In other words, the data are a collection of independent Bernoulli random variables, whose probability of success depend on the cell and the random time interval to which they belong. The latter is defined in terms of the change point times. The Bernoulli model is typically not valid, as the cells have refractory periods, however, it can be a valid approximation if the refractory period is much shorter than then the mean inter spikes time.

We need to restrict the structure of the joint distribution of the change point times, because of statistical and computational considerations. We considered two alternative assumptions:

[A1] $T_{i1},\ldots,T_{iM}$ are independent with distribution functions $G_1,\ldots,G_M$ respectively. In particular, the supports of these distributions are mutually exclusive.

or

[A2] $T_{i1}, T_{i2}-T_{i1},\ldots,T_{iM}-T_{i,M-1}$ are independent with distribution functions $G_1,\ldots,G_M$ respectively.

A1 describes a situation in which all change point are relative to the synchronizing event, while A2, for $M=2$, describes an situation in which the cell react to the external event at time T_{i1} for a duration of $T_{i2}-T_{i1}$. Practically, the algorithm for the second case was useful only for $M=2$: a random length first period, then an intermediate interval with a random length, and thereafter a final period. The algorithm can be too slow for any larger M.

The likelihood function to be maximized under assumption A1 is:

$$\begin{aligned} & L(\{g_m(t)\}_{\substack{m=1,\dots,M \\ t=a,\dots,b}}, \{p_{km}\}_{\substack{k=1,\dots,K \\ m=1,\dots,M}}) \\ & = \prod_{i=1}^{n} \sum_{a<t_1<\cdots<t_M<b} \prod_{m}^{M} g_m(t_m) \prod_{k=1}^{K} \prod_{m}^{M+1} p_{km}^{\sum_{t=t_{m-1}}^{t_m} N_{ik}(t)} (1-p_{km})^{t_m - t_{m-1}+1-\sum_{t=t_{m-1}}^{t_m} N_{ik}(t)}, \end{aligned}$$

where $t_0 = a$ and $t_{M+1} = b$, and $g_m(t)$ is the point mass at t of the m-th distribution.

It is similar under assumption A2:

$$\begin{aligned} & L(\{g_m(t)\}_{\substack{m=1,\dots,M \\ t=a,\dots,b}}, \{p_{km}\}_{\substack{k=1,\dots,K \\ m=1,\dots,M}}) \\ & = \prod_{i=1}^{n} \sum_{a<t_1<\cdots<t_M<b} \prod_{m}^{M} g_m(t_m - t_{m-1}) \prod_{k=1}^{K} \prod_{m}^{M+1} p_{km}^{\sum_{t=t_{m-1}}^{t_m} N_{ik}(t)} (1-p_{km})^{t_m - t_{m-1}+1-\sum_{t=t_{m-1}}^{t_m} N_{ik}(t)}, \end{aligned}$$

where $t_0 = a$ and $t_{M+1} = b$.

3.2 The algorithms

Suppose the change point times, T_{ij}, $i = 1,\dots,n$, $j = 1,\dots,M$, were observed. Then, the distribution of the time to the change points could be estimated easily by the empirical distribution of the corresponding variables, and the probabilities, $p_1,\dots,p_M$ could be estimated by the corresponding means in the sample. This makes the model a typical missing data model. Note that by missing we don't necessarily mean that data was lost. It may that the model can be derived as a simpler model in which some of the variables are unobserved. A standard method to maximize the likelihood in models with missing data is the EM algorithm, Dempster, Laird and Rubin (1977). Generally speaking, when the EM algorithm is used, it is assumed that besides the observed data there are unobserved data, and we iterate between computing expectation of the log-likelihood of the complete data over the conditional distribution of the unobserved random variables giving the observed ones (the E-steps), and maximizing this expectation (the M-steps).

We used two versions of the EM algorithm for computing the (approximate) maximum likelihood estimators (MLE) of the different parameters. The following notation is used. Hat above a parameter denote an estimator. The distributions are approximated by discrete distributions have r_m, $m = 1,\dots,M$ support points. Note that both M and the r_m's are prescribed by the user. The discrete distributions of the change point time are denoted by G, and their probability functions are denoted by g. The EM algorithm for multiple independent change points was as follows:

Algorithm 1:

1. Initial step Set $l = 0$, $\hat{g}_m^{(l)}(j) = 1/r_m$, $j = 1,\dots,r_m$, $m = 1,\dots,M$ and $\hat{p}_{km}^{(l)} = (n(b-a+1))^{-1}\sum_{i=1}^n \mathbb{N}_{ik}(b)$, $k = 1,\dots,K$, $m = 0,\dots,M$. Let $z_m(1),\dots,z_m(r_m)$ be the support point of the distribution of the m-th change point, $m = 1,\dots,M$.

2. E-step Compute the likelihood function that the changes of the i-th trial happened at $j_1,\dots,j_M$:

$$L_i(j_1,\dots,j_M) = \prod_{k=1}^K \prod_{m=1}^M \left(\frac{\hat{p}_{k,m-1}^{(l)}(1-\hat{p}_{km}^{(l)})}{(1-\hat{p}_{k,m-1}^{(l)})\hat{p}_{km}^{(l)}} \right)^{\mathbb{N}_{ik}(z_m(j_m))} \left(\frac{1-\hat{p}_{k,m-1}^{(l)}}{1-\hat{p}_{km}^{(l)}} \right)^{z_m(j_m)},$$

for $i = 1,\dots,n$ $1 \le j_m \le r_m$, $m = 1,\dots,M$. Compute the *a posteriori* probabilities for the vector of the i-th change point times:

$$P_i(j_1,\dots,j_m) = \frac{L_i(j_1,\dots,j_M)\prod_{m=1}^M \hat{g}_m^{(l)}(j_m)}{\sum_{j_1'=1}^{r_1}\cdots\sum_{j_M'=1}^{r_M} L_i(j_1',\dots,j_M')\prod_{m=1}^M \hat{g}_m^{(l)}(j_m')}$$

3. M step Set $l = l+1$. Update $\hat{g}_1^{(l)},\dots,\hat{g}_M^{(l)}$ to be the marginal distributions of $n^{-1}\sum_{i=1}^n P_i(\cdot)$. Update for $k = 1,\dots,K$ and $m = 0,\dots,M$:

$$\hat{p}_{km}^{(l)} = \frac{\sum_{i=1}^n \sum_{j_1=1}^{r_1}\cdots\sum_{j_M=1}^{r_M} P_i(j_1,\dots,j_M)(\mathbb{N}_{ik}(z_{m+1}(j_{m+1})) - \mathbb{N}_{ik}(z_m(j_m)))}{\sum_{i=1}^n \sum_{j_1=1}^{r_1}\cdots\sum_{j_M=1}^{r_M} P_i(j_1,\dots,j_M)(z_{m+1}(j_{m+1}) - z_m(j_m))} \tag{3.1}$$

4. Convergence check Stop if the number of iterations exceeds the pre-decided tolerable number or the convergence criterion (3.2) below was less than the pre-decided value. Other wise return to the E-step.

The algorithm for dependent change points (or independent start and duration of an intermediate period) was as follows:

Algorithm 2:

1. Initial step: Like the initial step of algorithm 1 with $M = 2$.

2. E-step Let M=2. Let $z_1(1),\dots,z_1(r_1)$ be the support of the distribution of the first change point, and let $t_2(1),\dots,t_2(r_2)$ be the support of the distribution of the time between the two change points. Compute the likelihood function for the i-th trial:

$$L_i(j_1,j_2) = \left(\frac{\hat{p}_{k0}^{(l)}(1-\hat{p}_{k1}^{(l)})}{(1-\hat{p}_{k0}^{(l)})\hat{p}_{k1}^{(l)}}\right)^{\mathbb{N}_{ik}(z_1(j_1))} \left(\frac{\hat{p}_{k1}^{(l)}(1-\hat{p}_{k2}^{(l)})}{(1-\hat{p}_{k1}^{(l)})\hat{p}_{k2}^{(l)}}\right)^{\mathbb{N}_{ik}(z_1(j_1)+z_2(j_2))}$$
$$\left(\frac{1-\hat{p}_{k0}^{(l)}}{1-\hat{p}_{k1}^{(l)}}\right)^{z_1(j_1)} \left(\frac{1-\hat{p}_{k1}^{(l)}}{1-\hat{p}_{k2}^{(l)}}\right)^{z_1(j_1)+z_2(j_2)}$$

Define P_i as in (3.1).

3. M-step Set $l = l+1$ Update $\hat{g}_1^{(l)},\hat{g}_2^{(l)}$ to be the two marginals of $n^{-1}\sum_{i=1}^n P_i(\cdot)$, Update for $k = 1,\dots K$:

$$\hat{p}_{k0}^{(l)} = \frac{\sum_{i=1}^n \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} P_i(j_1,j_1)\mathbb{N}_{ik}(z_1(j_1))}{\sum_{i=1}^n \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} P_i(j_1,j_1)(z_1(j_1)-a)}$$
$$\hat{p}_{k1}^{(l)} = \frac{\sum_{i=1}^n \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} P_i(j_1,j_1)\left(\mathbb{N}_{ik}(z_1(j_1)+z_2(j_1)) - \mathbb{N}_{ik}(z_1(j_1))\right)}{\sum_{i=1}^n \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} P_i(j_1,j_1)z_2(j_2)}$$
$$\hat{p}_{k2}^{(l)} = \frac{\sum_{i=1}^n \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} P_i(j_1,j_1)\left(\mathbb{N}_{ik}(b) - \mathbb{N}_{ik}(z_1(j_1)+z_2(j_2))\right)}{\sum_{i=1}^n \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} P_i(j_1,j_1)\left(b - z_1(j_1) - z_2(j_2)\right)}$$

4. Convergence check Stop if the number of iterations exceeds the pre-decided tolerable number or the convergence criterion (3.2) below was less than the pre-decided value. Other wise return to the E-step.

In general, the EM algorithm may be very slow. In our simulation it was reasonably fast for the single change points examples. It took around 1 minute on 133MHz PC. It was quite slow for some of our extensions where there were more than one change point.

The stopping time was defined as the minimum between $l = 500$ and the first l such that first time the difference between the estimates in two consecutive iteration as measure by

$$\frac{\sum_{m=1}^{M}\sum_{k=1}^{K}\left|\hat{p}_{km}^{(l)} - \hat{p}_{km}^{(l-1)}\right|}{\sum_{m=1}^{M}\sum_{k=1}^{K} p_{km}} + \sum_{j=1}^{r}\left|\hat{g}_j^{(l)} - \hat{g}_j^{(l-1)}\right| < 4 \times 10^{-6}, \tag{3.2}$$

where $\hat{p}_{km}^{(l)}$ and $\hat{g}_j^{(l)}$ are the estimators after l cycles of the algorithm.

In the first example the algorithm converged after 10 iterations. It converges after 9 in the case of the third example. On the other hand, it stopped in the second example after 500 iterations, when the first term of (3.2) was equal to $2.5\ 10^{-7}$ and the second term was equal to $1.4\ 10^{-4}$.

3.3 Examples

3.3.1 One cell and one change point

The first record we discuss is of a GPe cell. We consider the interval starting 600ms before the time the monkey released the central key (RELEASE) and ending 500ms after this event. The raster plot is given in Figure 1a, where each horizontal line of dots represents a single trial, and each dot denotes a spike at the trial and time relative to the synchronizing event, as given by its coordinates. The same data is summarized in Figure 1b by the PSTH (Peri Stimulus Time Histogram). Here we plot the total number of spikes (over all the trials) in each 1ms interval, relative to the RELEASE time. The vertical axis is scaled to denote the intensity (in spikes per second). The average intensity is 57.8, or, on the average, a spike every 17.3ms The PSTH is smoothed in Figure 1c with a Gaussian kernel with bandwidth of 4ms. On the same graph, the PSTH is smoothed also by a monotone regression estimator. We can observe from these figures that the intensity in decreasing. The transition from high to low average intensity, as can be observed from the PSTH, is smooth. A more detailed observation of the raster plot shows that in each trial the transition between the two periods is quite abrupt, but the change time varies from trial to trial. One possible interpretation of Figure 1b or Figure 1c, is that a change is happening before or at time -200ms, and there after intensity is decreasing during approximately 200ms before it stabilized again. However, this interpretation of the PSTH is wrong in view of the raster plot. Our analysis will model this exactly. In this example and in the other two examples, there is no need for a sophisticated test to verify that the intensity is not constant. Formally, we considered a t-test for comparing the total count in the first 550ms to the total count of the second half. The t-statistics has a value of 14.3 ($P < 0.001$).

The estimated p.d.f. and c.d.f. of the change point distribution are shown in Figure 2. It can be observed that the lower intensity period starts in most trials before the actual release, but the actual time varies between trial to trial.

3.3.2 Two change points

We consider now a second GPe cell from the same recording session as the cell analyzed above. The raster plot and the PSTH are given in Figure 3. As can clearly be seen from the PSTH, a simple change point model cannot fit the data. However, we can try to fit a model with two change points. The fact that there is an intermediate period with higher intensity can be verified simply by considering the interval of length 1200ms around RELEASE. We divided the interval into three equal parts and counted the number of spikes in each sub-interval. The P-value of the t-test that compare the two extreme sub-intervals is 0.7, while the t-test that compares the first and second sub-interval has an apparent P-value of 10^{-8}.

A model in which the width of the interval is independent of its initial time was fitted to the data. That is, we assumed that there are i.i.d. pairs (T_{i1}, T_{i2}),

$i = 1,\dots,n$, such that T_{i1} and $T_{i2} - T_{i1}$ are independent, and the intensity of the process is λ_1, λ_2 and λ_3, for $t \leq T_{i1}$, $T_{i1} < t \leq T_{i2}$, and $t > T_{i2}$ respectively. The support of the distribution was fitted by eye, to be as wide as possible. Thus the support of the first change point was in the range of -300ms to -50ms, while the width of the the second period was restricted to be in the range of 50ms to 450ms. The results obtained from applying Algorithm 2 to these data are given in Figure 4. Note that the p.d.f. is 0 at most points. The first change point is mainly distributed between 250ms to 150ms before RELEASE. The width of the high intensity period was found to be mostly around 240ms, but with probability of approximately 0.25 it got the maximal value that was permitted, as if some in a quarter of the trial the second change point is missing. Note that the probability mass assigned by the estimator is quite negligible on most of the permitted support points. In fact, only in 6 out of the 26 support points of the first distribution, and in 4 out of the 41 support points of the second distribution, the algorithm assigned a probability larger than 0.001.

To check the reliability of the estimation procedure, we introduced a 100ms jitter. That is, each spike train was shifted by a random time distributed uniformly between -50ms to 50ms. The spread of the distribution of the first change point was increased (although, less than could be expected), while the distribution of the second change point remained almost the same as expected (since the time between the two change points was not expected to change by the random shift). One can judge from the shape of the estimated distribution and the effect of introducing the jitter, that either the change point model is not appropriate to this cell, or a much larger sample is needed for a stable estimator.

3.3.3 Two cells with one change point

We consider now a record of two GPe cells of the second monkey. (This monkey was trained for somewhat simplified experiment with only the "GO" paradigm). We observe two cells around the RELEASE time. The data is exhibited in Figure 5. There are 44 valid trials.

It seems that the two cells behave similarly. Clearly, the processes are not homogeneous. Formally, we calculated a t-statistics that compared the number of spikes in the first half to the number in the second half of the segment. We obtained the values of 9.1 and 7.2 ($P < 0.001$) respectively for the two cells. Marginally, for each cell we assume the same model as above. We assume, however, that the two processes have the same change point, this change point may vary from trial to trial, and the cells are independent given this change point. This assumption seems to be plausible: We applied the algorithm to the two neurons independently. The root-mean-squares distance between the two estimated p.d.f.'s (one for each cell) was 0.07. The correlation coefficient between the two vectors of *a posteriori* expected values of the change point times was 0.46 .

The estimator is given in Figure 6. In Figures 6b and 6c we compare the MLE estimate of the distribution function of the joint change point to the separate estimates based on the monotone regression of the corresponding PSTHs. These graphs show that the assumption of the existence of a joint change point is reasonable.

3.3.4 Simulation: multiple processes and change points

We continue in our generalization. This time we simulated 100 trials in which two neurons are observed. The two cells have the same change points, but are independent otherwise. Three change points were simulated. The change point time were independent and with different supports. The distributions of the change points were gamma with a scale parameter 2 and shape parameters

125, 250 and 375 respectively (and hence the mean times were 250, 500 and 750, while the standard deviations are 22.36, 31.63, 38.73, respectively). The time scale was chosen to be similar to the biological data, so the whole interval was considered as having 1000ms length. The distribution was truncated to the intervals (125,375), (375,625), and (625,875) respectively. The intensities of the two processes were (40,10)(in the units of spikes/s) before the first change point, (60,50) between the first and the second change points, (40,50) after the second and (40,30) after the last change point. The raster plot is given in Figure 7a, and the PSTH of these data is given in Figures 7b and 7c.

The first change can be observed nicely, the other change points can be observed but less clearly. We looked for a single change point in each of the intervals (125,375), (375,620), and (625,875). The starting point was a homogeneous Poisson process and uniform change point distribution on the grid of 5ms in each of the intervals. In Figure 8 the estimated densities of the change points are plotted together with the histogram of the actual "unobserved" times. As can be expected from the raster plot, the distribution of the first change point was well estimated. The two other distributions were estimated better than we expected, but not as good as the first. The estimates of the intensities are given in Table 1.

In Figure 9 we ordered the trials according to the time of the first change point, and plotted the a posteriori expectation and the actual time of the first change point as against trial number. That is, for each trial we computed the *a-posteriori* distribution of the change point, as in the E-step of the algorithm, and calculated the expectation of this distribution. It can be observed that although the distribution was estimated quite well the individual times were not. Of course, it can be expected that the Bayes estimator will shrink towards the mean. The estimator seems to depend mainly on the *a priori* distribution. In the introduction we argued that estimators which are based on the trial by trial estimation of the change point may yield a poor estimator of the change point distribution. Figure 9 proves our case.

3.4 Mathematical background

In the appendix we give a rigorous analysis of a mathematical model of the problem. Unlike the model discussed above, the theoretical model considered have one change point and the observed process is of inhomogeneous Poisson process. We discuss in the extended version the information bound for estimating the intensities, and show that they can be estimated in the parametric $\sqrt{n}$ rate. The distribution function, on the other hand, can be estimated only in the rate of $n^{1/3}$, a much slower rate than the $n^{1/2}$ which is attainable with direct observations on the change point times. A bound on the achievable rate is established by presenting pairs of distributions which are $n^{-1/3}$ apart but the Neyman-Pearson tests between them have sum of errors bounded away from 0. That this bound is actually achievable is proved by exhibiting a simple non-iterative estimator that actually achieves the optimal rate. This estimator is based on monotone smoothing of the PSTH which was used above. See Figures 1(c), 2(b), 5(c), 6 (b) and (c), and 7(c). We also show that under some conditions the maximum-likelihood is rate optimal.

4 Discussion

We applied the empirical Bayes change point methodology to neural data. Using empirical examples and simulated data we showed that this technique can be used to obtain a sound understanding of the nature of the synchronization between an external event and the cells activity.

The theoretical statistical discussion was restricted to point processes in continuous time, while the algorithms were restricted to 0-1 processes in discrete time. Both are approximations of reality. In practice, the cells operate in continuous time while the output of the experimental system is in discrete one.

Moreover, the spikes are not points in time, but have a duration of the order of 1ms. So we preferred to use the convenient model for the given discussion.

Another statistical method that was used for similar data is that of the hidden Markov model (HMM), Radons et al. (1994); Abeles et al. (1995); Gat, Tishby and Abeles (1997); Ver Hoef and Cressie (1997). This model presumes that the recorded cells are behaving as a Markov process with a finite state space. These states are not observed directly. Instead, each state is characterized by a different vector of cell intensities — the hidden mechanism. The above papers suggest different algorithms to estimate the parameters of the model, and show that the inferred states may have a biological meaning.

The change point model suggested in this paper may seem more restricted than the HMM. Practically, the number of possible states was restricted to two or three, with a prescribed transition order. However, by definition, the hidden Markov model assumes that the brain stays at each state an exponential time. In theory, this can bypassed by assuming many pseudo-states. Actually, any stationary process can be weakly approximated by (not necessarily simple) HMM, see Kunsch, Geman and Kchagias (1995). However, for this we may need many more states than it would be practical to assume.

The change point model does not suffer from this problem. Any distribution function for the time of change can be assumed. Therefore, the change point model is preferred to the HMM, whenever we assume that the number of change points is small, and the distribution of the time to the changes is of interest.

In this paper we consider a non-parametric model for the time to the change in the intensity. We could assume a parametric model, such as a gamma distribution with one or two unknown parameters. However, such a parametric assumption is restricting and without the usual benefits in terms of speed of convergence and simplicity of the estimation procedure. The algorithm will be much very much the same, and the rate of convergence will be not much different.

We considered the testing of the existence of a change point versus the hypothesis that no change occurs. We intend to discuss elsewhere the more difficult problem of the existence of a relatively sharp change at a random time versus graduate change.

In this paper we considered a mathematical models of neurons which react to external event after a random delay. The reaction is an abrupt change in the firing intensity, but whose time is different in different trials. The main take home message of our analysis is that in "real cases" of neuronal data, the distribution of the change point can be estimated, while single trial estimating of the specific value of the change point is prone to a large error. However, proper estimation of the distribution of the change point can reliably help in the discrimination between two plausible physiological scenarios. In the first scenario there is a smooth transition of the discharge rate in all single trials, whereas in the second scenario there are sharp transitions with jitter of their timings. This discrimination can't be done by the classical PSTH analysis, and the present manuscript provides a quantitative (rather than the subjective raster plot display) method for the discrimination between these two possible physiological settings.

Acknowledgement This study was supported in part by the Israeli Academy of Science. We like to thank the reviewers for their comments which improved the quality of the paper.

References

Abeles, M., Bergman, H., Gat, I., Meilijson, I., Seidemann, E., and Tishby, N. Cortical activity flips among quasi-stationary states. Proc Natl Acad Sci USA 1995; 92:8616–8620.

Akman, V. E. and Raftery, A. E. Asymptotic inference for a change-point poisson process. Ann Statist 1986; 14:1583–1590.

Alexander, K. S. Probability inequalities for empirical processes and a law of iterated logarithm. Ann Probab 1984; 12:1041–1067.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. *Statistical Models Based on Counting Processes.* Springer New York 1993.

Bélisle, P., Joseph, L., MacGibbon, B., Wolfson, D. B., and du Berger, R. Change-point analysis of neuron spike train data. Biometrics 1998; 54:113–123.

Bickel, P., Klaassen, C., Ritov, Y., and Wellner, J. *Efficient and Adaptive Estimation for Semiparametric Models.* John Hopkins University Press, Baltimore 1993.

Commenges, D. and Seal, J. The analysis of neuronal discharge sequence: change-point estimation and comparison of variances. Stat Med 1985; 4:91–104.

Crutcher, M. D. and Alexander, G. E. Movement-related neuronal activity selectively coding either direction or muscle pattern in three motor areas of the monkey. J Neurophysiol 1990; 64:151–163.

DeLong, M. and Georgopoulos, A. 1981;. Motor functions of the basal ganglia. In J. M. Brookhart, V. B. Mountcastle, V. B. Brooks, and S. R. Geiger (Eds.), *The Nervous System. Motor Control.* volume II, Pt. 2 of *Handbook of Physiology* (pp. 1017–1061). American Physiological Society Bethesda.

DeLong, M. R. Activity of pallidal neurons during movement. J Neurophysiol 1971; 34:414–427.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm (with comments). JRSS-B 1977; 39:1–37.

Efron, B. Empirical bayes methods for combining likelihoods. J Amer Statist Assoc 1996; 91:538–550.

Gat, I., Tishby, N., and Abeles, M. Hidden markov modelling of simultaneously recorded cells in the associative cortex of behaving monkeys. Network-Comp Neura 1997; 8:297–322.

Groeneboom, P. 1985;. Estimating a monotone density. In M., L. C. L. and A., O. R. (Eds.), *Proceedings of the Berkeley Symposium in Honor of Jerzy Neyman and Jack Kiefer* volume II (pp. 539–555). Wadsworth, Belmont.

Joseph, L. and Wolfson, D. B. Estimation in multi-path change-point problems. Comm Statist A—Theory Methods 1992; 21:897–913.

Kunsch, H., Geman, S., and Kchagias, A. Hidden markov random fields. Ann Appl Probab 1995; 5:577–602.

Le Cam, L. and Yang, G. L. *Asymptotics in Statistics*. Springer Verlag, New York 1990.

Matthews, D. E., Farewell, V. T., and Pyke, R. Asymptotic score-statistics processes and test for constant hazard against a change-point alternative. Ann. Stat. 1985; 13:583–591.

Montgomery, E. B. J. A new method for relating behavior to neronal activity in performing monkeys. J Neurosci Methods 1989; 28:197–204.

Nini, A., Feingold, A., Slovin, H., and Bergman, H. Neurons in the globus pallidus do not show correlated activity in the normal monkey, but phase-locked oscillations appear in the mptp model of parkinsonism. J Neurophysiol 1995; 74:1800–1805.

Prakasa Rao, B. L. S. Estimation for distribution with monotone failure rate. Ann Math Statist 1970; 41:507–519.

Prakasa Rao, B. L. S. Estimation of a unimodal density. Sankhyā Ser. A 1983; 31:23–36.

Radons, G., Becker, J. D., Dulfer, B., and Kruger, J. Analysis, classification, and coding of multielectrode spike trains with hidden markov models. Biol Cybern 1994; 71:359–373.

Raftery, A. E. and Akman, V. E. Bayesian analysis of a poisson process with a change-point. Biometrika 1986; 73:85–89.

Raz, A., Vaadia, E., and Bergman, H. Firing patterns and correlations of spontaneous discharge of pallidal neurons in the normal and the tremulous 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine vervet model of parkinsonism. J Neurosci 2000; 20:8559–8571.

Ritov, Y., Raz, A., and Bergman, H. 1997;. An empirical bayes change point problem with application to neurological data. In *The International Symposium on Contemporary Multivariate Analysis and Its Applications, Hong Kong*.

Ritov, Y., Raz, A., and Bergman, H. 2002;. Detection of onset of neuronal activity by allowing for heterogeneity in the change points. Technical Report 622 Dept. of Stat., UCB http://www.stat.berkeley.edu/tech-reports/index.html.

Robins, J. M. and Ritov, Y. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semiparametric models. Stat Med 1997; 17:285–319.

Romo, R. and Schultz, W. Dopamine neurons of the monkey midbrain: Contingencies of responses to active tough during self-initiated are movements. J Neurophysiol 1990; 63:592–606.

Schwartz, A. B., Kettner, R. E., and Georgopouls, A. P. Primate motor cortex and free arm movemnets to visual tragets in three-dimensional space. i. relations betwen single cell discharge and direction of movement. J Neurosci 1988; 8:2913–2927.

Seal, J. and Commenges, D. A quantitative analysis of stimulus- and movement-related responses in the posterior parietal cortex of the monkey. Exp Brain Res 1985; 58:144–153.

Seal, J., Commenges, D., Salamon, R., and Bioulac, B. A statisitical method for the estiamtion of neuronal response latency and its functional interpretation. Brain Res 1983; 278:382–386.

Ver Hoef, J. M. and Cressie, N. Using hidden markov chains and empirical bayes change-point estimation for transect data. Environ Ecol Stat 1997; 4:247–264.

A Appendix: Mathematical background

A.1 Model and information bounds

Let $T_1,\dots,T_n$ be *unobserved* i.i.d. random variables with a distribution function G supported on the interval $[0,1]$. We observe the i.i.d. processes $\mathbb{N}_1,\dots,\mathbb{N}_n$, such that given $T_1,\dots,T_n$, $\mathbb{N}_1,\dots,\mathbb{N}_n$ are independent, $\mathbb{N}_i$ is an inhomogeneous Poisson process on $[0,1]$, with a constant intensity λ_0 on the interval $[0,T_i)$ and another constant intensity λ_1 on $[T_i,1]$, $i=1,\dots,n$. The parameters λ_0, λ_1 and G are unknown.

Let $\mathbb{N}^*$ be a Poisson process with (known) intensity λ^*. Reparametrize the model by writing $\lambda_0=\lambda^*+\nu-\delta$ and $\lambda_1=\lambda^*+\nu+\delta$. Suppose $G\in G$, a family dominated by a σ-finite measure μ_G. With some abuse of notation, we interchange the two parametrizations. The process $\mathbb{N}$ has a density with respect to the distribution of $\mathbb{N}^*$ given by

$$\begin{aligned} f(\mathbb{N};\nu,\delta,g) &= c_1(\mathbb{N})\int e^{-\lambda_0 t}\lambda_0^{\mathbb{N}(t)}e^{-\lambda_1(1-t)}\lambda_1^{\mathbb{N}(1)-\mathbb{N}(t)}g(t)\,d\mu_G(t)\\ &= c_2(\mathbb{N})\int e^{-\nu+\delta(2t-1)+\mathbb{N}(t)\log(\lambda^*+\nu-\delta)+(\mathbb{N}(1)-\mathbb{N}(t))\log(\lambda^*+\nu+\delta)}g(t)\,d\mu_G(t), \end{aligned}$$

where c_1 and c_2 do not depend on the unknown parameters. See, for example, Andersen, Borgan, Gill and Keiding (1993) page 98.

Let $\{G_\eta : |\eta|<\varepsilon\}\subseteq G$ be a one dimensional regular parametric sub-family of G. Let G_η have density g_η with respect to μ_G, and let $h=\frac{\partial}{\partial\eta}g_\eta/g_\eta|_{\eta=0}$. The score functions for ν, δ, and η, each with the other parameters known, are

$$\begin{aligned} \ell^*_{\nu\,|\,\delta,g}(\mathbb{N};\lambda_0,\lambda_1,g) &= \mathrm{E}_g\left(\lambda_o^{-1}\mathbb{N}(T)+\lambda_1^{-1}(\mathbb{N}(1)-\mathbb{N}(T))-1\,\middle|\,\mathbb{N}\right)\\ \ell^*_{\delta\,|\,\nu,g}(\mathbb{N};\lambda_0,\lambda_1,g) &= \mathrm{E}_g\left(-\lambda_o^{-1}\mathbb{N}(T)+\lambda_1^{-1}(\mathbb{N}(1)-\mathbb{N}(T))+2T-1\,\middle|\,\mathbb{N}\right)\\ \ell^*_{h\,|\,\nu,\delta}(\mathbb{N};\lambda_0,\lambda_1,g) &= \mathrm{E}_g\left(h(T)\,|\,\mathbb{N}\right). \end{aligned}$$

To find the efficient score function for λ_0 and λ_1 (or, equivalently, ν and δ) we have to find h_ν and h_δ such that

$$\mathrm{E}\left(\ell^*_{\nu\,|\,\delta,g}(\mathbb{N};\lambda_0,\lambda_1,g)+\ell^*_{h_\nu\,|\,\nu,\delta}(\mathbb{N};\lambda_0,\lambda_1,g)\,\middle|\,T=t\right)\equiv 0$$

$$\mathrm{E}\left(\ell^*_{\delta\,|\,\nu,g}(\mathbb{N};\lambda_0,\lambda_1,g)+\ell^*_{h_\delta\,|\,\nu,\delta}(\mathbb{N};\lambda_0,\lambda_1,g)\,\middle|\, T=t\right)\equiv 0,$$

see Bickel, Klaassen, Ritov and Wellner (1993) Section 3.4 for details. We are not able to find h_δ and h_ν explicitly in the general case. There is, however, an exception, which is very important for testing. When $\lambda_0=\lambda_1$, $\mathbb{N}$ and T are independent. Therefore

$$\begin{aligned}\ell^*_{\nu|\delta,g}(\mathbb{N};\lambda_0,\lambda_0,g)&=\lambda_0^{-1}(\mathbb{N}(1)-1)\\ \ell^*_{\delta|\nu,g}(\mathbb{N};\lambda_0,\lambda_0,g)&=\lambda_0^{-1}(\mathbb{N}(1)-2\mathrm{E}_g\left(\mathbb{N}(T)\,|\,\mathbb{N}\right))+2E_gT-1,\end{aligned}\tag{A.1}$$

where, explicitly, $\mathrm{E}_g\left(\mathbb{N}(T)\,|\,\mathbb{N}\right)=\int\mathbb{N}(t)g(t)\,d\mu_G(t)$. Both of the expressions in (**??**) have mean 0 given T (since $\mathbb{N}$ and T are independent and the expressions have, unconditionally, mean 0). Hence $h_\nu\equiv 0$ and $h_\delta\equiv 0$ and adaptation is possible.

Moreover, the efficient score function for δ when both ν and G are unknown, calculated under $\delta=0$, is given by the linear combination of the form $\ell^*_{\delta\,|\,\nu,g}+c\ell^*_{\nu\,|\,\delta,g}$ which has mean 0 even if ν is misspecified, Bickel et al. (1993). It can easily checked that it is given by

$$\ell^*_\delta(\mathbb{N};\lambda_0,\lambda_1,g)=2\lambda_0^{-1}\big(\mathrm{E}_g(T)\mathbb{N}(1)-\mathrm{E}_g\left(\mathbb{N}(T)\,|\,\mathbb{N}\right)\big)\tag{A.2}$$

A.2 Simple estimators and the rate of convergence

We consider now a simple non-iterative estimators of the parameters, and prove that these estimators actually achieve the best rate of convergence. That is, the parameters λ_0 and λ_1 can be estimated at a rate of $n^{1/2}$, and G can be estimated at the slower rate of $n^{1/3}$. However, some functions of G, notably its mean, can be estimated at the $n^{1/2}$ rate.

Theorem A.1 *Suppose $\lambda_0\neq\lambda_1$ and there are $t_1,t_2\in(0,1)$ with $0<G(t_1)<G(t_2)<1$. Then λ_0 and λ_1 can be estimated at $n^{1/2}$ rate. Moreover, $\mathrm{E}(T)$ can be estimated in the same rate. Finally let $\tilde{G}(\cdot)=\int_0^{\cdot}G(s)\,ds$. Then there is an estimator $\hat{\tilde{G}}(\cdot)$ such that* *the process* $\{\sqrt{n}(\hat{\tilde{G}}(t)-\tilde{G}(t)) : t\in[0,1]\}$ *is tight and converges to a limiting Gaussian process.*

Proof. The random variable $\mathbb{N}(t)$ (for a given $t\in(0,1)$) is a mixture of Poisson random variables: Given the unobserved T, $\mathbb{N}(t)$ is Poisson with mean $\lambda_0 + 2\delta(t-T)^+$, where $\delta=(\lambda_1-\lambda_0)/2$, where x^+ is x for $x>0$ and 0 otherwise. Hence

$$\begin{aligned} e_t \equiv \mathrm{E}\,\mathbb{N}(t) &= \lambda_0 t + 2\delta\int_0^t (t-s)\,dG(s) \\ &= \lambda_0 t + 2\delta\int_0^t G(s)\,ds \\ &\equiv \lambda_0 t + 2\delta\tilde{G}(t), \quad \text{say.} \end{aligned} \tag{A.3}$$

Similarly,

$$\begin{aligned} v_t &\equiv \operatorname{Var}\mathbb{N}(t) \\ &= \mathrm{E}\operatorname{Var}\bigl(\mathbb{N}(t)\bigm| T\bigr) + \operatorname{Var}\mathrm{E}\bigl(\mathbb{N}(t)\bigm| T\bigr) \\ &= \mathrm{E}\bigl(\lambda_0 t + 2\delta(t-T)\bigr) + \operatorname{Var}\bigl(\lambda_0 t + 2\delta(T-t)\bigr) \\ &= e_t + 4\delta^2\int_0^t (t-s)^2\,dG(s) - 4\delta^2\Bigl(\int_0^t (t-s)\,dG(s)\Bigr)^2 \\ &= e_t + 8\delta^2\int_0^t \tilde{G}(s)\,ds - 4\delta^2\tilde{G}^2(t), \quad \text{integration by parts} \\ &= e_t + 4\delta\int_0^t e_s\,ds - 2\lambda_0\delta t^2 - e_t^2 + 2\lambda_0 e_t t - \lambda_0^2 t^2, \quad \text{by } \textbf{(??)} \end{aligned} \tag{A.4}$$

Let $\hat{e}_t$ and $\hat{v}_t$ be estimates of v_t and e_t based on the i.i.d. sample, $\mathbb{N}_1(t),\dots,\mathbb{N}_n(t)$. These estimators are $n^{1/2}$ consistent uniformly in $t\in[0,1]$. In the simple case, the functions $t\mapsto te_t$, $\int_0^t e_s\,ds\,t^2$, $t\in(0,1)$ are linearly independent. We can then isolate $J\ge 3$ points, $t_1,\dots,t_J$, and find the least squares solution for the system

$$\hat{v}_{t_j} - \hat{e}_{t_j} + \hat{e}_{t_j}^2 = 4\hat{\delta}\int_0^{t_i}\hat{e}_s\,ds - (\hat{\lambda}_0^2 + 2\hat{\lambda}_0\hat{\delta})t_j^2 + 2\lambda_0\hat{e}_{t_j}t_j, \quad j=1,\dots,J \tag{A.5}$$

with the three unknowns, $\hat{\delta}$, $\hat{\lambda}_0$ and $\hat{\lambda}_0^2+2\hat{\lambda}_0\hat{\delta}$. The solution for the first two unknowns can serve as estimates of δ and λ_0.

The abovementioned three functions can be linearly dependent. Some tedious argument shows that they are linearly dependent iff $G(t)=t^\nu$, $\nu>0$,

$t \in (0,1)$. In particular they are linearly independent if the support of G is a (necessarily proper) subset of $(0,1)$. If the functions are linearly dependent, we can solve the non-linear system (**??**), only for t_1 and t_2 and the unknown $\hat{\delta}$ and $\hat{\lambda}_0$. It can be verified that this system has at most 3 isolated solutions. The correct solution can be found since, say, λ_0 is the derivative of e_t at 0. In practice, we can either start with the non-linear system, or consider the linear system, and retreat to the non-linear system if the condition number is smaller than $n^{-1/4}$.

Since $e_1 = \lambda_0 + 2\delta(1 - \mathrm{E}T)$, the expectation of T can be estimated in the $n^{1/2}$ rate, as well as the smooth function $\tilde{G}(\cdot)$ of G, which can be estimated by $\hat{\tilde{G}}(t) = (\hat{e}(t) - \hat{\lambda}_0 t)/2\hat{\delta}$. Moreover, the process $\sqrt{n}(\hat{\tilde{G}}(\cdot) - \tilde{G}(\cdot))$ is tight, since $\sqrt{n}(\hat{e}_\cdot - e_\cdot)$ is tight.

□

Finally, G itself can be estimated using a simple monotone intensity estimator as proved next. Explicitly, let $\mathbb{N}_+(\cdot) = \sum_{i=1}^n \mathbb{N}_i(\cdot)$. That is, if, without any lose of generality, $\hat{\lambda}_1 < \hat{\lambda}_0$, then let $\mathbb{C}_n$ be the piece-wise constant derivative of the least concave function larger than $\mathbb{N}_+$. Then G can be estimated by $\hat{G} = (\hat{\lambda}_0 - \hat{\lambda}_1)^{-1}(\hat{\lambda}_0 - \mathbb{C}_n)$.

Theorem A.2 *Suppose that G has a bounded density and there is an open interval on which G has a density bounded away from 0. Then the optimal rate of convergence of an estimator of G to G is $n^{1/3}$ in the sense that there is an estimator $\hat{G}_n$ of G such that*

1. $\hat{G}(t) = G(t) + \mathrm{O_p}\left(n^{-1/3}\right)$ *for any* $t \in (0,1)$.
2. $n^\alpha \left\|\hat{G} - G\right\|_\infty \xrightarrow{\mathrm{p}} 0$ *for any* $\alpha < 1/3$.
3. $n^\alpha \left|\tilde{G}(t) - G(t)\right| \xrightarrow{\mathrm{p}} \infty$ *for any* $\alpha > 1/3$, $t \in (0,1)$, $G'(t) > 0$ *and any estimator* $\tilde{G}$ *of G.*

Proof. We prove first that the $n^{1/3}$ rate is attainable. Assume, w.l.o.g., that $\lambda_1 > \lambda_0$. Ignore the trial information, and let $\mathbb{N}_+(\cdot) = \sum_{i=1}^n \mathbb{N}_i(\cdot)$. Then $\mathbb{N}_+$ (given the set $\{ T_1,\ldots,T_n\}$) is an inhomogeneous Poisson process with intensity $\lambda^*(t) = n\Big[\lambda_1\mathbb{G}_n(t)+\lambda_0\big(1-\mathbb{G}_n(t)\big)\Big]$, where $\mathbb{G}_n$ is the empirical distribution function of $T_1,\ldots,T_n$. In particular, $\lambda^*(\cdot)$ is monotone non-decreasing. It is well known, (Prakasa Rao (1970); Prakasa Rao (1983); Groeneboom (1985)), that λ^* can be estimated at the $n^{1/3}$ rate. Since λ_j, $j=0,1$ can be estimated in a much faster rate and $\|\mathbb{G}_n - G\|_\infty = \mathrm{O}_\mathrm{p}\left(n^{-1/2}\right)$, we conclude that $\hat{G}$ satisfies conditions 1. and 2.

We prove now that this is the best possible rate. It suffices to show that there exists a sequence of distribution functions $\{G_n\}$ with densities $\{g_n\}$, such that $n^{1/3}(G_n(t)-G(t)) \to c_1 > 0$ while

$$\sqrt{n}H\left(F_n,F\right) \to c_2 < \infty, \tag{A.6}$$

where

$$H^2(F_n,F) = \mathrm{E}\left(\left(\frac{f(\mathbb{N};\lambda_0,\lambda_1,g_n)}{f(\mathbb{N};\lambda_0,\lambda_1,g)}\right)^{1/2} - 1\right)^2,$$

the square Hellinger distance. If (**??**) is satisfied, then the variational distance between the distributions of $\mathbb{N}_1,\ldots,\mathbb{N}_n$ under G and G_n is bounded away from 2 (Cf. Le Cam and Yang (1990) page 29).

Fix any t_0. Assume that G has a density g, bounded from below by b^* on some small interval $(t_0-\zeta,t_0+\zeta)$, $\zeta>0$, and $G(t_0-\zeta)>0$. Let $0<b<b^*$. Let $a_n>0$ be a solution of

$$\big(2e^{-(\lambda_1-\lambda_0)a_n n^{-1/3}} - e^{-2(\lambda_1-\lambda_0)a_n n^{-1/3}}\big) - \big(2e^{(\lambda_1-\lambda_0)n^{-1/3}} - e^{2(\lambda_1-\lambda_0)n^{-1/3}}\big) = 0.$$

Note that the left hand side is monotone increasing as a function of a_n on $(0,\infty)$ and it is equal to

$$(a_n^2-1)(\lambda_1-\lambda_0)^2 n^{-2/3} + \mathrm{o}\left(n^{-1}\right)$$

It follows that a_n is well defined and $a_n = 1 + \mathrm{o}\left(n^{-1/3}\right)$. Let $g_n = g + h_n$, where

$$h_n(t) = \begin{cases} b & t_0 - 2a_n n^{-1/3} \leq t < t_0 - a_n n^{-1/3} \\ -b & t_0 - a_n n^{-1/3} \leq t < t_0 + n^{-1/3} \\ b & t_0 + n^{-1/3} \leq x < t_0 + 2n^{-1/3} \\ 0 & \text{elsewhere.} \end{cases}$$

Finally let $H_n(\cdot) = \int_0^{\cdot} h_n(t)\,dt$, $\bar{G}(\cdot) = \int_0^{\cdot} \exp((\lambda_1 - \lambda_0)t)\,dG(t)$. and $\bar{H}_n(\cdot) = \int_0^{\cdot} \exp((\lambda_1 - \lambda_0)t) h_n(t)\,dt$. Note that $H = \bar{H} = 0$ outside the interval $(t_0 - an^{-1/3}, t_0 + a_n n^{-1/3})$. In particular $G_n = G + H_n$ is a cdf, and $G_n(t_0 + n^{-1/3}) = G(t_0 + n^{-1/3}) - bn^{-1/3}$. Hence we should only argue that (**??**) is satisfied.

First note that the likelihood ratio between G_n and G satisfies

$$\begin{aligned} L_n &= \frac{f(\mathbb{N};\lambda_0,\lambda_1,g_n)}{f(\mathbb{N};\lambda_0,\lambda_1,g)} \\ &= 1 + \frac{\int_0^1 e^{(\lambda_1-\lambda_0)t-\mathbb{N}(t)\log(\lambda_1/\lambda_0)} h_n(t)\,dt}{\int_0^1 e^{(\lambda_1-\lambda_0)t-\mathbb{N}(t)\log(\lambda_1/\lambda_0)} g(t)\,dt} \\ &= 1 + \frac{\sum_{j=0}^{\mathbb{N}(1)} (\bar{H}_n(\tau_{j+1}) - \bar{H}_n(\tau_j)) e^{-j\log(\lambda_1/\lambda 0)}}{\sum_{j=0}^{\mathbb{N}(1)} (\bar{G}(\tau_{j+1}) - \bar{G}(\tau_j)) e^{-j\log(\lambda_1/\lambda_0)}}, \end{aligned} \tag{A.7}$$

where $\tau_1 < \cdots < \tau_{\mathbb{N}(1)}$ are the jump points of the point process $\mathbb{N}$, and $\tau_0 = 0$, $\tau_{\mathbb{N}(1)+1} = 1$. Now, by construction $(\bar{H}_n(t_{j+1}) - \bar{H}_n(t_j))/(\bar{G}(t_{j+1}) - \bar{G}(t_j))$ is positive and bounded. Hence

$$\left|L_n - 1\right| \leq b/b^*. \tag{A.8}$$

Note that the numerator in the right hand side of (**??**) is 0 unless there is an event in the interval $[t_0 - 2a_n n^{-1/3}, t_0 + 2n^{-1/3}]$. We can assume that the process was generated in the following way. We start with a Poisson process $\bar{\mathbb{N}}$ with intensity λ_1 and a random time T_i. The process $\mathbb{N}_i$ is obtained by deleting each of the events of $\bar{\mathbb{N}}$ in the interval $(0, T_i)$ with probability $1 - \lambda_0/\lambda_1$. Hence

$$P(L_n = 1) \geq P(A_n) = e^{-2\lambda_1(1+a_n)n^{-1/3}} \tag{A.9}$$

where A_n is the event that $\bar{\mathbb{N}}$ has no jumps in the interval $[t_0 - 2an^{-1/3}, t_0 + 2n^{-1/3}]$.

We can further bound the second line of (**??**):

$$\begin{aligned}\left|L_n-1\right| &\le \frac{\int_0^1 e^{(\lambda_1-\lambda_0)t}\left|h_n(t)\right|dt\,\mathbb{1}(L_n\neq 1)}{\int_0^{t_0-\zeta} e^{-\log(\lambda_1/\lambda_0)\bar{\mathbb{N}}(t_0-\zeta)}g(t)\,dt} \\ &\le cn^{-1/3}\mathbb{1}(L_n\neq 1)e^{\log(\lambda_1/\lambda_0)\bar{\mathbb{N}}(t_0-\zeta)},\end{aligned} \tag{A.10}$$

for some constant c. Now, $\bar{\mathbb{N}}(t_0-\zeta)$ is independent of A_n, and $\exp(\beta\bar{\mathbb{N}}(t))$ has a finite expectation. It follows from (**??**) and (**??**) that

$$\mathrm{E}(L_n-1)^2 \le c_1 n^{-2/3}\big(1-P(A_n)\big) = c_2 n^{-1} \tag{A.11}$$

for some finite c_1 and c_2. But then

$$\begin{aligned}H\,(F_n,F) &= \mathrm{E}(L_n^{1/2}-1)^2 \\ &= \mathrm{E}\left(\frac{L_n-1}{L_n^{1/2}+1}\right)^2 \le c_3 n^{-1}\end{aligned}$$

by (**??**) and (**??**). The proof is now complete. □

It may seem from the proof of Theorem **??** that $\mathbb{N}_+=\sum_{i=1}^n \mathbb{N}_i$ is almost a sufficient statistics. I.e., the estimator based on $\mathbb{N}_+$ is almost as good as the best estimator based on $\mathbb{N}_1,\dots,\mathbb{N}_n$. This is not true. To make this explicit we consider a different asymptotic. We compare now the following two models

Model M_1 Let $T_1,T_2,\dots$ be i.i.d. random variables with common distribution G. Given $T_1,\dots,T_n$, let $\mathbb{N}_1,\dots,\mathbb{N}_n$ be independent, $\{\mathbb{N}_i : t\in(0,T_i)\}$ is a Poisson counting process with intensity $m\lambda_0$, and $\{\mathbb{N}_i : t\in(T_i,1)\}$ is an independent Poisson counting process with intensity $m\lambda_1$.

Model M_2 Let $\mathbb{N}_1,\dots,\mathbb{N}_n$ be i.i.d. , counting processes with intensity $\lambda(t)=m[\lambda_0+(\lambda_1-\lambda_0)G(t)]$.

The following proposition is intended to demonstrate the differences between these two models. It is not the strongest possible result of its kind.

Proposition A.3 *Suppose G has a bounded density and λ_0 and λ_1 are known.*

i. Suppose $n \to \infty$, $m \to \infty$, $m^4/n \to 0$, and $m^5/n \to \infty$. Then there is an estimator $\hat{G}$ of G such that under model M_1, $\left\|\hat{G} - G\right\|_\infty = \mathrm{O_p}\left(n^{-1/2} + m^{-2}\right)$, while the best estimator of G under model M_2 satisfies $\hat{G} - G = \mathrm{O_p}\left((mn)^{-1/3}\right)$ (point-wise). That is, G can be estimated much faster under model M_1 than it can be estimated under model M_2.
ii. Suppose $n \to \infty$ and $m^2/n \to \infty$. Then G can be estimated at a rate of $n^{1/2}$ under model M_1. It can be estimated with the faster rate of $(mn)^{1/3}$ under model M_2 (point-wise, and almost at that rate uniformly).

Proof.
i. Since $m \to \infty$, T_i can be "estimated" based on $\mathbb{N}_i$. Let the estimator be $\hat{T}_i$. Then $\hat{T}_i = T_i + \varepsilon_i$, where $\varepsilon_i = \mathrm{O_p}\left(m^{-1}\right)$ and has, asymptotically, a symmetric distribution. For example, we can take $\hat{T}_i = \mathrm{argmin}\{\mathbb{N}_i(t) - (\lambda_0 + \lambda_1)t/2\}$. Then $m(\hat{T}_i - T_i)$ has the asymptotic distribution of $\mathrm{argmin}\{B(t) - |t|\}$, where B is a Brownian motion with zero drift. Let $\tilde{G}$ be the distribution of $\hat{T}_i$. Then $\hat{G}_n$, the empirical distribution function of $\hat{T}_1, \dots, \hat{T}_n$ satisfies $\left\|\hat{G} - \tilde{G}\right\|_\infty = \mathrm{O_p}\left(n^{-1/2}\right)$. While $\left\|\tilde{G} - G\right\|_\infty = \mathrm{O_p}\left(m^{-2}\right)$ (since the distribution of ε_1 is, essentially, symmetric). The convergence of the estimator of the monotone intensity under M_2 is standard.
ii. Even if $T_1, \dots, T_n$ are known, then under M_1, $n^{1/2}$ is the best rate of convergence for an estimator of G. □

A.3 MLE

The NPMLE is defined as the distribution $\hat{G}$ and parameters $\hat{\lambda}_0$ and $\hat{\lambda}_1$ (if they exist) that maximize the expression $\sum_{i=1}^n \log f(\mathbb{N}_i; \hat{\lambda}_0, \hat{\lambda}_1, \hat{G})$. Unfortunately, we have a minor technical problem, since, as would follow from the proof of the next proposition, if we define the processes $\mathbb{N}_i$ to be left continuous then the NPMLE does not exists if $\lambda_1 < \lambda_0$ and if we define it to be right continuous, the NPMLE does not exists if $\lambda_1 > \lambda_0$. So, we prefer to give up notational consistency, and define the processes $\mathbb{N}_1, \dots, \mathbb{N}_n$ such that $\log(\hat{\lambda}_0/\hat{\lambda}_1)\mathbb{N}_i$ is upper semi-continuous. Anyway, the NPMLE can be found using the EM algorithm given in Section 3.2 .

An immediate consequence of the form of the likelihood is the following proposition.

Proposition A.4 *The MLE $\hat{G}$ of G is supported on the observed events. It has a version with at most $n+1$ points.*

Proof. Suppose, w.l.o.g., that $\hat{\lambda}_1 > \hat{\lambda}_0$. Then the MLE, $\hat{G}$, maximizes the log-likelihood function given by

$$\sum_i \log \int e^{(\hat{\lambda}_1 - \hat{\lambda}_0)t - \mathbb{N}_i(t)\log(\hat{\lambda}_1/\hat{\lambda}_0)}\, d\hat{G}(t).$$

Suppose that there was a mass in a tiny interval centered at a point t between two spikes. Then by moving it to the right up to the next spike, the term $(\hat{\lambda}_1 - \hat{\lambda}_0)t$ increases, the term $\mathbb{N}_i(t)\log(\hat{\lambda}_0/\hat{\lambda}_1)$ does not change and hence the likelihood increases.

Let $t_1 < \cdots < t_K$ be the points in which at least one of the processes $\mathbb{N}_1, \ldots, \mathbb{N}_n$ has a jump, and let $\hat{g} = (\hat{g}_1, \ldots, \hat{g}_K)^T$ where $\hat{g}_j$ is the mass of $\hat{G}$ at t_j. Also, let M be the matrix with entries $M_{ij} = \exp\{(\hat{\lambda}_1 - \hat{\lambda}_0)t_j - \mathbb{N}_i(t_j)\log(\lambda_1/\lambda_0)\}$, $i = 1, \ldots, n$, $j = 1, \ldots, K$, and $M_{0,j} = 1$. Then the log-likelihood function is $\sum \log(f_i)$, where $(1, f_1, \ldots, f_n)^T = M\hat{g}$. In other words, the likelihood depends on g only through Mg. Hence, if $K > n+1$, $\hat{g}$ is not uniquely defined and there is a solution with at most $n+1$ entries different from 0. □

We now prove the consistency of the NPMLE of G. For simplicity we prove it for the unrealistic case of λ_0 and λ_1 known. We believe that the result is valid for the more general case where the intensities are unknown. The proof, however, would be more complicated since the convexity argument which is crucial to our proof is not valid for the general case.

Theorem A.5 *Suppose $\lambda_1 > \lambda_0$ and both are known and that G has a bounded density. The NPMLE $\hat{G}$ is $n^{1/3}$ consistent.*

Proof. To simplify the notation let $\phi = \lambda_1 - \lambda_0$ and $\psi = \log(\lambda_1/\lambda_0)$. Denote the log-likelihood of the single observation by

$$\ell_i(G) = \log\left(\frac{\int_0^1 e^{\phi t - \psi \mathbb{N}_i(t)}\, dG(t)}{\int_0^1 e^{\phi t - \psi \mathbb{N}_i(t)}\, dG^0(t)}\right),$$

where G^0 is the true distribution. Let $\ell(G)$ be the generic random variable. Recall the notation

$$\bar{G}(t) = \int_0^t e^{\phi s}\, dG(s) = e^{\phi t}G(t) - \phi \int_0^t e^{\phi s} G(s)\, ds. \tag{A.12}$$

Let $\tau_1, \ldots, \tau_{\mathbb{N}(1)}$ be the jump points (if any) of the process $\mathbb{N}$. The following expressions for the numerator and denominator of $\ell(G)$ will be useful:

$$\begin{aligned} \int_0^1 e^{\phi t - \psi \mathbb{N}(t)}\, dG(t) &= \sum_{j=0}^{\mathbb{N}(1)} \big(\bar{G}(\tau_{j+1}) - \bar{G}(\tau_j)\big) e^{-j\psi} \\ &= \bar{G}(1) e^{-j\mathbb{N}(1)} + (e^{\psi} - 1) \sum_{j=1}^{\mathbb{N}(1)} \bar{G}(\tau_j) e^{-j\psi} \end{aligned} \tag{A.13}$$

Let $G_n = \{G : \; \|G - G^0\|_\infty \le A_n n^{-1/3}\}$ for some $A_n \to \infty$ slowly (all distributions considered in this discussion are supported on $[0,1]$). Obviously, $\|\bar{G} - \bar{G}^0\|_\infty < (1+\phi)A_n e^{\phi} n^{-1/3}$ for $G \in G_n$. Let $\ell(G) \equiv \log\big(1 + X(G)\big)$ where

$$X(G) \equiv \frac{\int_0^1 e^{\phi t - \psi \mathbb{N}(t)}\, (dG(t) - dG^0(t))}{\int_0^1 e^{\phi t - \psi \mathbb{N}(t)}\, dG^0(t)}$$

It follows from (**??**) that For $G \in G_n$

$$\big|X(G)\big| \le 2\big\|G - G^0\big\|_\infty e^{\psi \mathbb{N}(1)} \le c_1 A_n n^{-1/3} e^{\psi \mathbb{N}(1)}. \tag{A.14}$$

and

$$\Big|\frac{\int_0^1 e^{\phi t - \psi \mathbb{N}(t)}\, dG(t)}{\int_0^1 e^{\phi t - \psi \mathbb{N}(t)}\, dG^0(t)}\Big| \le e^{\phi + \psi \mathbb{N}(1)}. \tag{A.15}$$

Hence

$$\ell(G) = X(G) - \frac{1}{2}X^2(G) + R;$$

where for any $\varepsilon > 0$:

$$\big|R\big| \le \varepsilon X^2(G) + \big(\phi + \psi \mathbb{N}(1) + \big|X(G) - X^2(G)/2\big|\big) \mathbb{1}\big(\big|X(G)\big| > \varepsilon\big). \tag{A.16}$$

The last term in (**??**) is negligible since by (**??**), $\left|X(G)\right| > \varepsilon$ for any $G \in G_n$ implies that $\mathbb{N}(1) \le \gamma \log n$ for some $\gamma > 0$, and for any $\alpha, \beta, \gamma, K > 0$

$$\begin{aligned} \mathrm{E}\left\{e^{\beta \mathbb{N}(1)} \mathbb{1}(\mathbb{N}(1) \ge \gamma \log n)\right\} &\le \left(\frac{\lambda_1 e}{\gamma \log n}\right)^{\gamma \log n} \\ &\le n^{-K}, \quad n > n_0, \end{aligned}$$

since $\mathbb{N}(1)$ is stochastically smaller than a Poisson mean λ_1 random variable.

Clearly $\mathrm{E}X(G) = 0$. Hence

$$\mathrm{E}\ell(g) = -(1 + \mathrm{o}(1))\frac{1}{2}\mathrm{E}X^2(G) + \mathrm{o}_\mathrm{p}\left(n^{-1}\right), \tag{A.17}$$

uniformly for $G \in G_n$. Similarly we could consider only one term in the Taylor expansion of $\log(1 + X(G))$, to obtain that

$$\ell(G) \le \frac{1}{1-\varepsilon}\left|X(G)\right| + \left(\phi + \psi \mathbb{N}(1)\right) \mathbb{1}(\left|X(G)\right| > \varepsilon).$$

Therefore

$$\mathrm{Var}\,\ell(G) = (1 + \mathrm{o}(1))\mathrm{E}X^2(G). \tag{A.18}$$

This family of random variables is not bounded, but for any $\gamma > 0$

$$P(\max_{1 \le i \le n} \mathbb{N}_i(1) < \gamma \log n) \to 1.$$

It follows from (**??**) that for any $G \in G_n$

$$P(\max_{G \in G_n} \max_{1 \le i \le n} \left|\ell_i(G)\right| \le c_1 n^{-\gamma}) \to 1, \quad 0 < \gamma < 1/3.$$

Hence, we tactically assume that $\mathbb{N}_i$ is bounded by $\gamma_1 \log n$ for some $\gamma_1 > 0$ and $\{\ell_i(G) : \ G \in G_n\}$ is bounded by $n^{-\gamma_2}$ for some $0 < \gamma_2 < 1/3$.

We want to prove that $\sup_{G \in G_n} n^{-1/2}\left|\sum_{i=1}^n (\ell_i(G) - \mathrm{E}\,\ell(G))\right|$ is bounded in probability. For that we want to bound the coverage number of the set $L_n = \{\ell(G) : \ G \in G_n\}$. We actually consider a somewhat larger family of random variables. Let

$$\bar{\ell}(H) = \log\left(\frac{H(1)e^{-j\mathbb{N}(1)} + (e^{\psi} - 1)\sum_{j=1}^{\mathbb{N}(1)} H(\tau_j)e^{-j\psi}}{\bar{G}^0(1)e^{-j\mathbb{N}(1)} + (e^{\psi} - 1)\sum_{j=1}^{\mathbb{N}(1)} \bar{G}^0(\tau_j)e^{-j\psi}}\right)$$

for $H \in H_n \equiv \{H : H$ is monotone non-decreasing. $\|H - \bar{G}^0\|_\infty < 2A_n n^{-1/3}\}$. Let $\bar{L}_n = \{\bar{\ell}(H) : H \in H_n\}$. Note that $L_n \subset \bar{L}_n$ since $\ell(G) = \bar{\ell}(\bar{G})$. However, $\bar{L}_n$ is strictly larger, since $\int e^{-\phi t}\, d\bar{G}(t) = 1$ for any $G \in G_n$, and there is no such restriction on members of H_n. Now, it is easy to approximate the family $\bar{L}_n$. Define for any $H \in H_n$

$$H^\eta(t) = \eta\lceil H(\eta\lceil t/\eta\rceil)/\eta\rceil, \quad t \in [0,1]$$

where for every real x, $\lceil x \rceil$ is the smallest integer not smaller than x. That is, H^η is a step-wise approximation of H, with jumps at multiples of η and values which are multiples of η. Clearly, $\|H^\eta - H\|_\infty \le \eta$. The number of such functions H^η is, at most η^{-2}. Recall that we consider $\mathbb{N}_i(1)$ to be bounded by $\gamma \log n$ for any γ. Hence

$$\|\bar{\ell}(H^\eta) - \bar{\ell}(H)\|_\infty \le c\eta n^\gamma.$$

Hereafter, c and γ are positive finite constants. Let $Y_i(H) = n^{1/3-\gamma}\bar{\ell}(H)$. Then $|Y_i| \le 1$. Let H_n^* be any subset H_n, and let $\sigma_n^{*2} = \sup_{H \in H_n^*} \operatorname{Var}(\bar{\ell}(H))$. Then $\nu_n^2 = \max_{H \in H_n^*} \operatorname{Var}(Y(H)) \le n^{2/3-2\gamma}\sigma_n^{*2}$. Let

$$I_n = \int_s^t E_n^{1/2}(u)\, du,$$

where $E_n(u) = \log N_n(u)$ is the log of the smallest number of random variables $Y(H_1), \dots, Y(H_{N_n(u)})$, such that $\sup_{H \in H_n} \min_k |Y(H) - Y(H_k)| < u$. In our case

$$E_n(u) = \log c + \gamma \log n - 2\log u.$$

The relevant range of u in the above integral is shrinking to 0, hence, $I_n \to 0$. It follows from Theorem 2.1 of Alexander (1984) that

$$P\left(\max_{H \in H_n^*} \Big|\sum_{i=1}^n \big(\bar{\ell}_i(H) - \mathrm{E}\,\bar{\ell}(H)\big)\Big| \ge M\sigma_n^*\sqrt{n}\right) \le K_1 e^{-K_2 M^2} \tag{A.19}$$

for all $\varepsilon < M < n^{1/3-\gamma}$ and some K_1 and K_2 which do not depend on M or n.

In Theorem **??** we proved that it is possible to estimate G at the $n^{1/3}$ rate. This implies that for any $M_n \to \infty$

$$n \sup\{\mathrm{E}\,\ell(G) : \|G - G^0\|_\infty > M_n n^{-1/3}\} \to -\infty, \tag{A.20}$$

since, otherwise, there was a sequence G_n, $n^{1/3}\left\|G_n - G^0\right\|_\infty \to \infty$ and the sum of errors of the Neyman-Pearson test between G_n and G^0 converges to 1, contradicting Theorem **??** .

Let now

$$G_n^* = \{G:\ \mathrm{E}\ell(G) \le -a_n/n\}.$$

for some $a_n \to \infty$. It follows from (**??**) that $G_n^* \subseteq G_n$ (for appropriate sequences A_n and a_n). Comparing (**??**) to (**??**), we obtain that

$$n\sigma_n^{*2}/2a_n \to 1,$$

where

$$\sigma_n^{*2} \equiv \sup\{\mathrm{Var}\,\ell(G):\ G \in G_n^*\}.$$

Finally,

$$\begin{aligned} P\left(\sup_{G\in\partial G_n^*}\sum_{i=1}^n \ell_i(G) \ge 0\right) &= P\left(\sup_{G\in\partial G_n^*}\sum_{i=1}^n (\ell_i(G) - \mathrm{E}\,\ell(G)) \ge a_n\right) \\ &\le P\left(\sup_{G\in\partial G_n^*}\sum_{i=1}^n (\ell_i(G) - \mathrm{E}\,\ell(G)) \ge \sqrt{a_n/4}\sigma_n^*\sqrt{n}\right) \\ &\le K_1 e^{-K_2 a_n/4} \to 0. \end{aligned}$$

Since $\ell(G^0) \equiv 0$ and $\sum_{i=1}^n \ell(G)$ is concave in G, we obtain that all its maxima are inside G_n^* and are $n^{1/3}$ continuous. □

A.4 Testing

In this section we discuss how a formal test for the existence of a change point versus no change occurs in the relevant interval can be devised. The purpose tests are not the tests used in the examples of this paper. In all the example simple minded, ad-hoc tests were powerful enough to reject the null assumption and establish the existence of change points. Consider the hypotheses H_0: the process $\{\mathbb{N}(t);\ t \in (0,1)\}$ is a homogeneous Poisson process, versus H_1:

for some random time $T \in (0,1)$ with cdf G, the intensity is λ_0 on $(0,T)$ and intensity λ_1 on $(T,1)$. We could base our test statistic directly on (**??**). We prefer, however, to motivate the test statistic from a different perspective. If we assume that $T = t$, then an asymptotic efficient test would be a t-test based on $(\mathbb{N}(1) - \mathbb{N}(t))/(1-t) - \mathbb{N}(t)/t$. Since the change point is not known, we consider a test statistic which is a weighted sum of such statistics:

$$\begin{aligned} W_\alpha &= \int_0^1 t(1-t)\left(\frac{N(1)-N(t)}{1-t} - \frac{N(t)}{t}\right) d\alpha(t) \\ &= \int_0^1 \big(tN(1) - N(t)\big)\, d\alpha(t), \end{aligned}$$

where α is a probability measure on $[0,1]$. Now,

$$\begin{aligned} \mathrm{E}(W_\alpha \mid T) = \int_0^1 t(1-t) \Bigg[& \left(\lambda_1 - \frac{\lambda_0 T + \lambda_1(T-s)}{t}\right) \mathbb{1}(T \le t) \\ & + \left(\frac{\lambda_0(T-t) + \lambda_1(1-t)}{1-t} - \lambda_0\right) \mathbb{1}(T > t)\Bigg]\, d\alpha(t) \\ = (\lambda_1 - \lambda_0) \int_0^1 & [(1-t)T\mathbb{1}(T \le t) + (1-t)(1-T)\mathbb{1}(T > t)]\, d\alpha(t) \end{aligned}$$

Therefore:

$$\mathrm{E}W_\alpha = (\lambda_1 - \lambda_0)\int_0^1 \left[(1-t)\int_0^t s\,dG(s) + t\int_t^1 (1-s)\,dG(s)\right] d\alpha(t). \tag{A.21}$$

In particular $\mathrm{E}(W_\alpha) = 0$ under H_0. Now, under H_0, if $s \le t$

$$\mathrm{cov}\big(sN(1) - N(s), tN(1) - N(t)\big) = \lambda_0 s(1-t).$$

Therefore,

$$\begin{aligned} \mathrm{Var}(W_\alpha) &= 2\lambda_0 \int_0^1 \int_0^{t-} s(1-t)\, d\alpha(s)\, d\alpha(t) + \sum t(1-t)\big(\alpha(t) - \alpha(t-)\big)^2 \\ &= \int_0^1 \left[(1-t)\int_0^{t+} s\, d\alpha(s) + t\int_{t+}^1 (1-s)\, d\alpha(s)\right] d\alpha(t). \end{aligned} \tag{A.22}$$

We look for the weight function α that maximizes the ratio of the expectation of W_α (under H_1) to its standard deviation (under H_0). It is straight forward to check that the optimal α equates the two terms within the square brackets in equations (**??**) and (**??**). That is, $\alpha = G$.

We reached the following test statistic

$$S_n(G) = \left(\sum_{i=1}^{n} \mathbb{N}_i(1)\right)^{-1/2} \sum_{i=1}^{n} \big(\int_0^1 \mathbb{N}_i(t)\, dG(t) - \mathbb{N}_i(1) \int_0^1 t\, dG(t)\big)$$

Let $\mathbb{N}_+(\cdot) = \sum_{i=1}^{n} \mathbb{N}_i(\cdot)$. Then

$$S_n(G) = \mathbb{N}_+^{-1/2}(1) \int_0^1 \big(\mathbb{N}_+(t) - t\mathbb{N}_+(1)\big)\, dG(t).$$

Denote $\mathbb{W}_n(t) \equiv \mathbb{N}_+^{-1/2}(1)\big(\mathbb{N}_+(t) - t\mathbb{N}_+(1)\big)$. Then $\mathbb{W}_n$ converges weakly under H_0 to a Brownian bridge, since $\mathbb{N}_+(\cdot)$ is a homogeneous Poisson process under H_0.

Typically, the distribution G is not known. We can proceed in a number of ways. The simplest is to maximize $S_n(G)$ over G. We arrive at the statistic $\max_G \big|S_n(G)\big| = \max_t \big|\mathbb{W}_n(t)\big|$ which is distributed asymptotically like the Kolmogorov-Smirnov statistic.

Alternatively, one may wish to maximize power against an alleged distribution. That is, he may consider $S_n(G_0)$, where G_0 may be an a priori specified distribution function. The following theorem shows that $S_n(G_0)$ has power against all alternatives G, whether G_0 was specified correctly or not. It is optimal when the right distribution of the change point is actually G_0. It is close to optimal if the right distribution is close to G_0. Hence, $S_n(G_0)$ is the reasonable test statistic, if one has a pre-specified direction towards which he wants to maximize the power of his test.

In the following theorem we are using the notation of Section **??** .

Theorem A.6 *Let H_0 be the hypothesis that $\delta = 0$ and H_n the alternative that $\delta = \delta_n$, where $n = 1, 2, \ldots$ and $\sqrt{n}\delta_n \to \delta_0$. Consider a sequence of models with fixed G, $\lambda = \lambda^* + \nu$ and δ as above. Then:*

1. *Suppose that H_0 holds. Then $S_n(G_0)$ has mean zero and variance $V(G_0) = 2\int_0^1 \int_s^t s(1-t)\, dG_0(s)\, dG_0(t) - \sum g_{0j}^2 t_j(1-t_j)$, where G_0 has point mass g_{0j} at t_j.*

2. *Suppose that hypothesis H_n holds for sample size n. Then $S_n(G_0)$ is asymptotically normal with the same variance as above and mean*

$$-2\lambda_0^{-1/2}\int_0^1\int_0^1 (s\wedge t)(1-s\vee t)\,dG_0(s)\,dG(t).$$

Proof. Assume H_0 holds. The first part follows since $\mathbb{N}_+(\cdot)$ is a simple Poisson process and for $s<t$:

$$\mathrm{E}\left(\frac{(\mathbb{N}_+(s)-s\mathbb{N}_+(1))\,(\mathbb{N}_+(t)-t\mathbb{N}_+(1))}{\mathbb{N}_+(1)}\,\middle|\,\mathbb{N}_+(1)\right)=s(1-t).$$

Assume now that H_n holds. The family of distributions defined by the sequence H_n is contiguous to H_0. Now, $S_n(G_0)$ is asymptotically equivalent to the statistic

$$\begin{aligned}\tilde{S}_n(G_0) &= (\lambda n)^{-1/2}\sum_{i=1}^n\int_0^1 (\mathbb{N}_i(t)-t\mathbb{N}_i(1))\,dG_0(t)\\ &\equiv n^{-1/2}\sum_{i=1}^n s_i(G_0).\end{aligned}$$

Since, under H_0, $(\tilde{S}_n(G_0),\sum_{i=1}^n \ell^*_{\delta|\nu,G}(\mathbb{N}_i))$ has asymptotically binormal distribution, we obtain that $\tilde{S}_n(G_0)$ has, under H_n, asymptotically normal distribution with the variance as under H_0 and mean given by

$$\begin{aligned}\delta_0\mathrm{E}\left(s_i(G_0)\ell_{\delta|\nu,G}(\mathbb{N}_1)\right) &= \delta_0\lambda^{-3/2}\int_0^1\int_0^1 (\mathbb{N}_1(s)-s\mathbb{N}_1(1))(\mathbb{N}_1(1)-2\mathbb{N}_1(t))\,dG_0(s)\,dG(t)\\ &= -2\delta_0\lambda^{-1/2}\int_0^1\int_0^1 (s\wedge t)(1-s\vee t)\,dG_0(s)\,dG(t).\end{aligned}$$

□

If we compare the two alternatives suggested so far, we conclude that although the process $\{S_n(G):\ G\in G\}$ is tight, whatever the set G is, the distribution of $S_n(G)$ depends on whether G is pre-specified or data dependent.

Another alternative is to try to estimate G. We consider a cross-validation version of the statistic. Let $\hat{G}_{(-i)}$ be an estimator of G based on all trial except for the i-th trial. Define s_i by

$$s_i=\int_0^1 \left(t\mathbb{N}_i(1)-\mathbb{N}_i(t)\right)d\hat{G}_{(-i)}(t).$$

Finally, let

$$S_n^{CV} = \sum_{i=1}^{n} s_i.$$

Note that in particular $\mathrm{E} S_n^{CV} = 0$ under H_0, and it is different from zero under H_1.

Captions list:

Figure 1 : The spiking activity of a GPe cell. The monkey released the central key at time 0.

(a) The raster plot;

(b) The PSTH;

(c) The PSTH smoothed and a monotone regression estimate of intensity.

Figure 2 : Change point distribution of the single cell described in Figure 1.

(a) The MLE p.d.f.;

(b) The MLE c.d.f. (stars) compared to the c.d.f. as estimated by the monotone regression of the PSTH (solid line).

Figure 3 : One GPe cell with seemingly two change points.

(a) Raster plot;

(b) PSTH.

Figure 4 : One GPe cell with two dependent change points.

(a) The estimated distribution of the first change point (0 is the RELEASE time);

(b) Estimated distribution of the time between the two change points.

Figure 5 : Two GPe cells.

(a) raster plot;

(b) The PSTH's for the two cells;

(c) Smoothed PSTH's.

Figure 6 : Two GPe cells with seemingly one change point.

(a) MLE of p.d.f. of the change point;

(b) Cell 9: MLE of the c.d.f, and the estimated based on the monotone regression;

(c) Cell 13: MLE of the c.d.f, and the estimated based on the monotone regression;

Figure 7 : Simulated data:

(a) The raster plot;

(b) PSTH;

(c) Smoothed PSTH.

Figure 8 : Simulated data: Histogram of the actual time of the change points and their estimates (stars).

(a) First change point;

(b) Second change point;

(c) Third change point;

Figure 9 : Simulated data: Actual times of the change point times and their *a posteriori* mean. The trials are ordered according to the time of the change point. The actual times are given by the solid time as function of the trial number. The opened circles are the *a-posteriori* expectation.

Figure 1 (a)

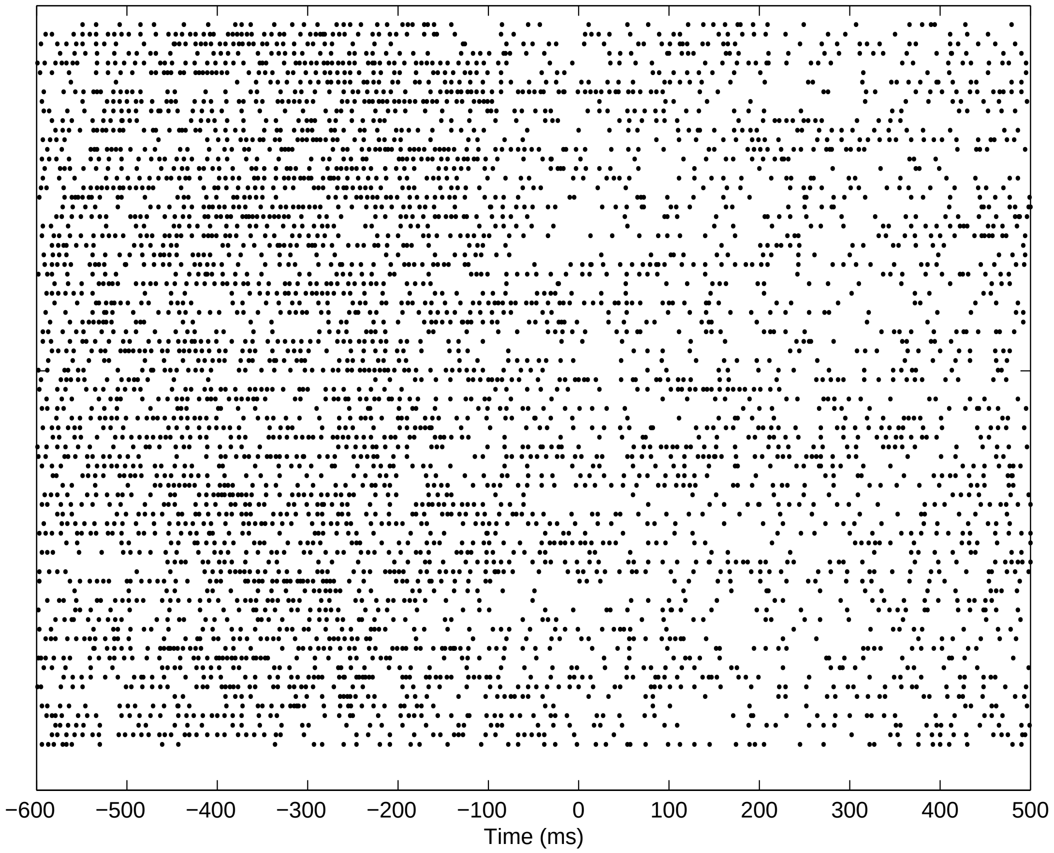
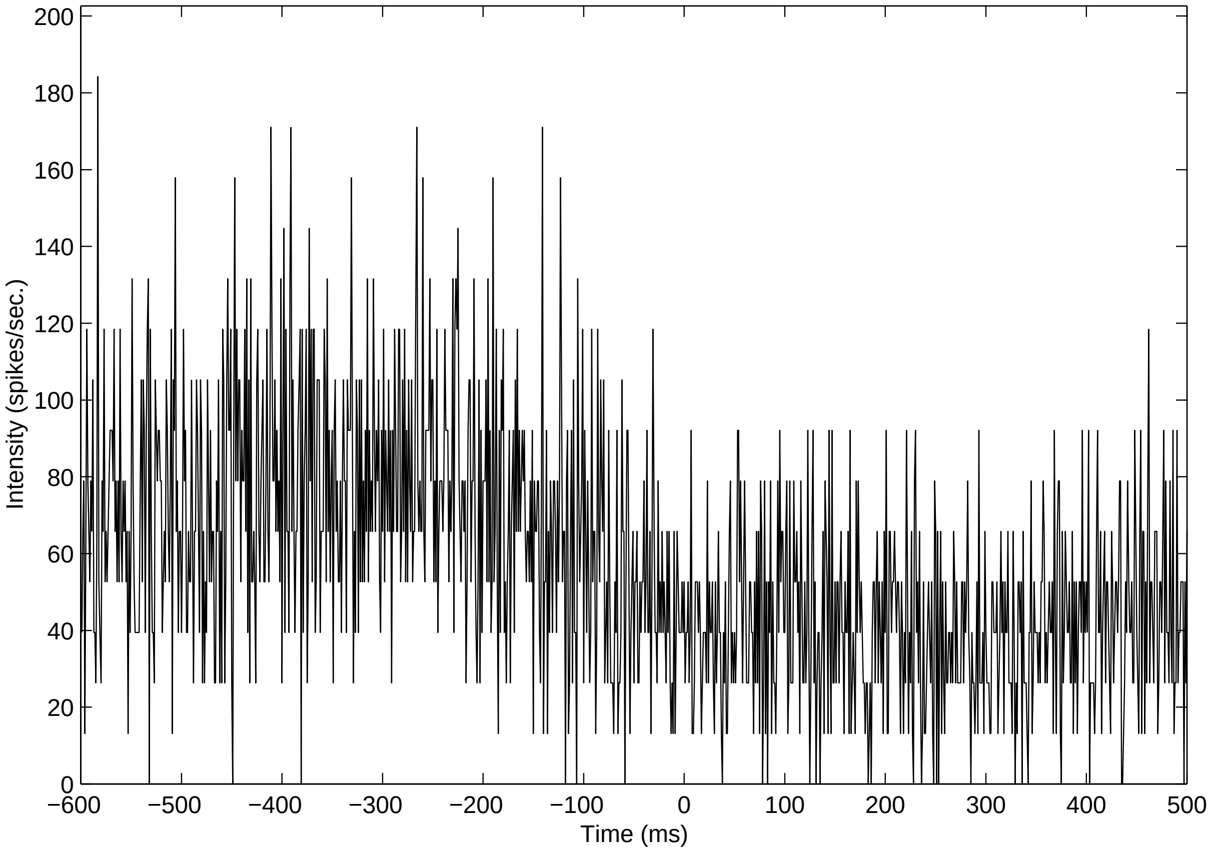


Figure 1 (b)



Intensity (spikes/sec.)
0
20
40
60
80
100
120
140
160
180
200
−600
−500
−400
−300
−200
−100
0
100
200
300
400
500
Time (ms)

Figure 1 (c)

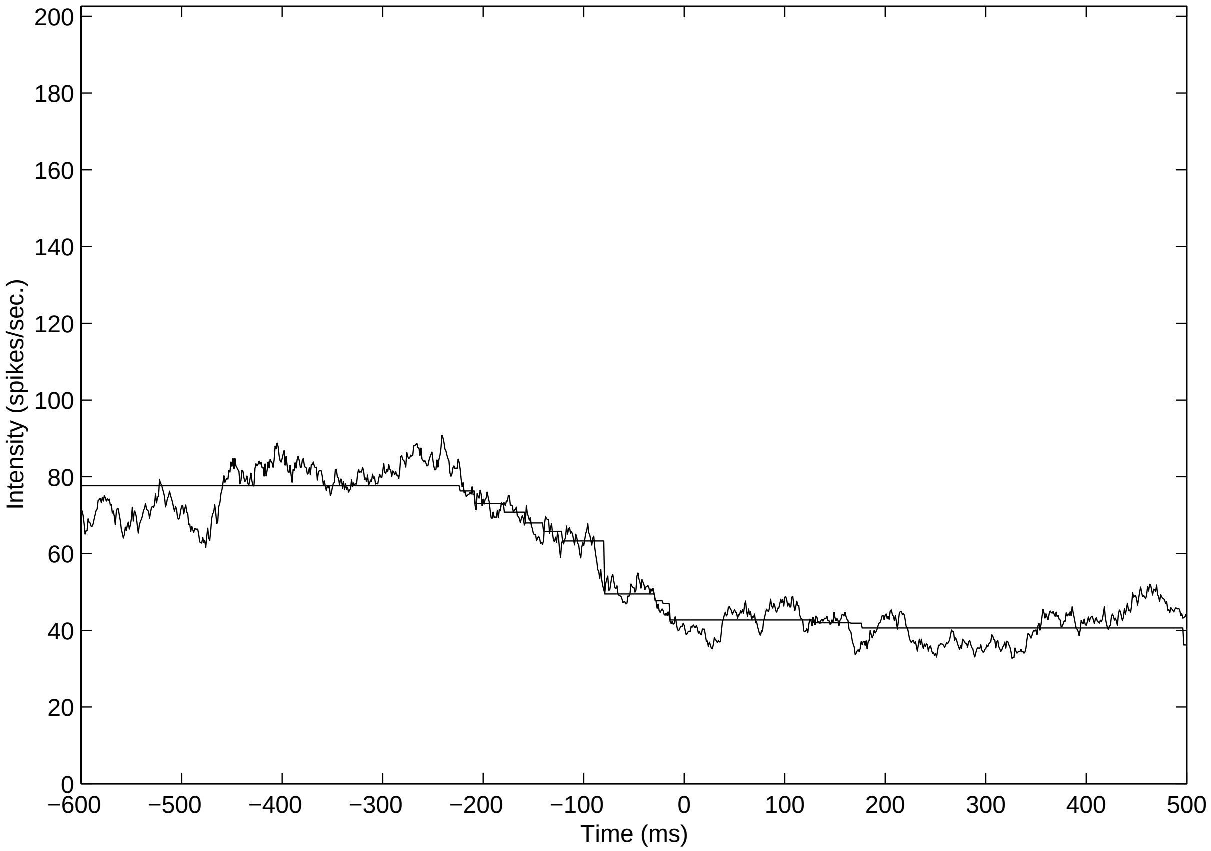

Intensity (spikes/sec.)
200
180
160
140
120
100
80
60
40
20
0
−600
−500
−400
−300
−200
−100
0
100
200
300
400
500
Time (ms)

Figure 2 (a)

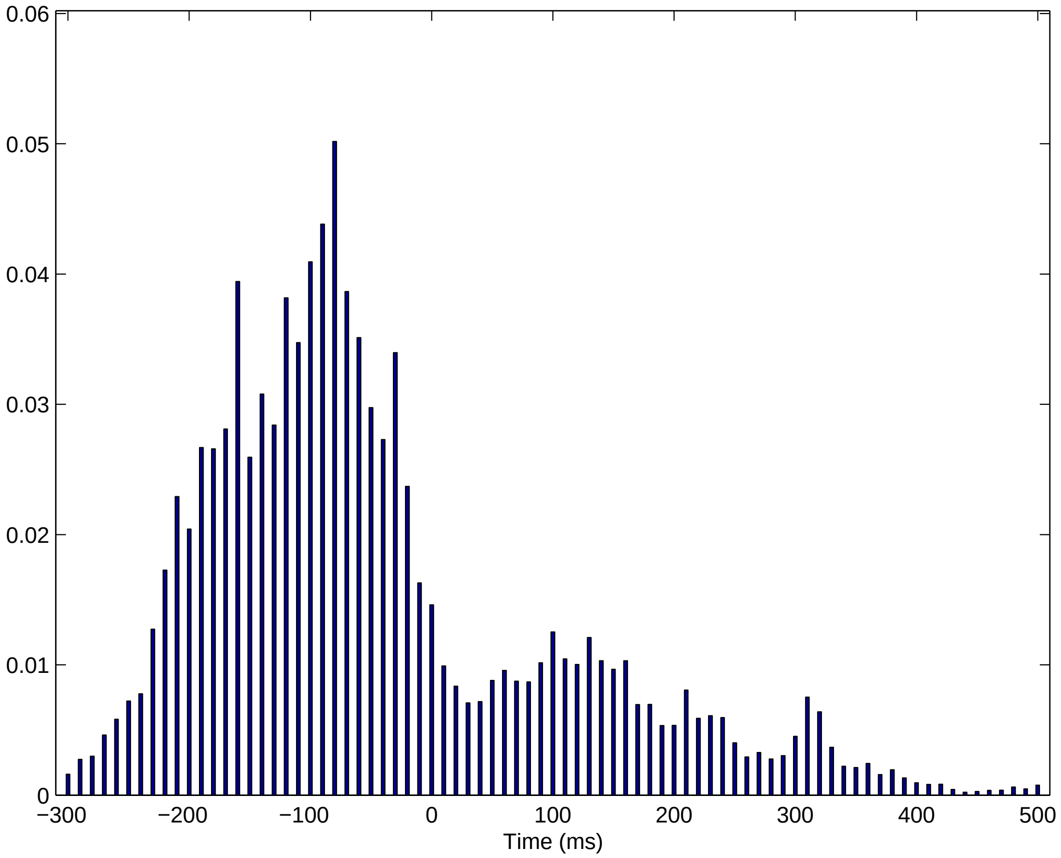

0.06
0.05
0.04
0.03
0.02
0.01
0
−300
−200
−100
0
100
200
300
400
500
Time (ms)

Figure 2 (b)

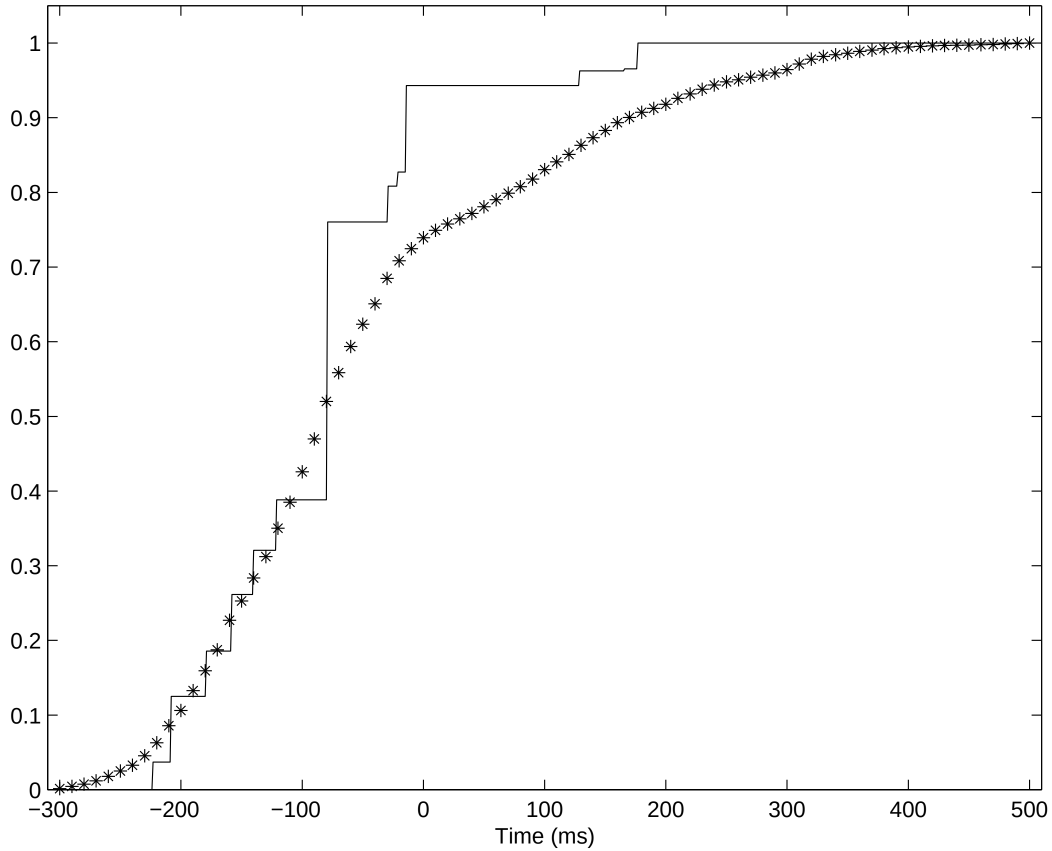

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−300
−200
−100
0
100
200
300
400
500
Time (ms)

Figure 3 (a)

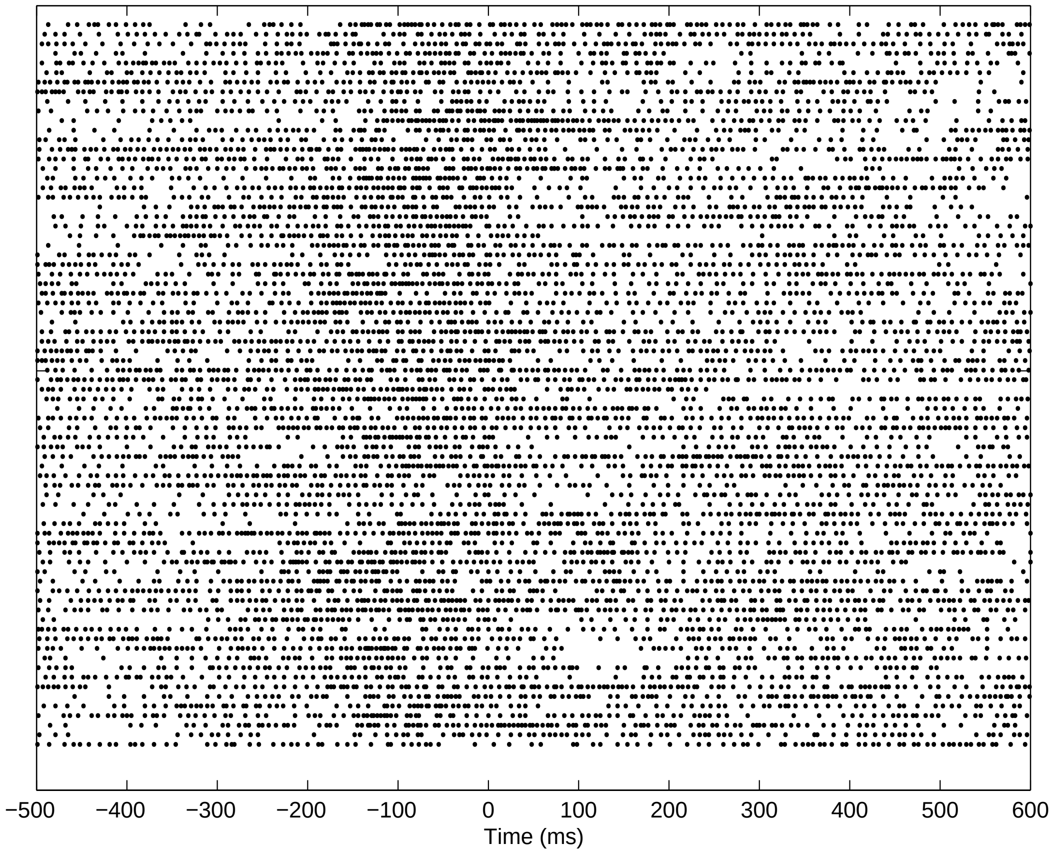


Figure 3 (b)

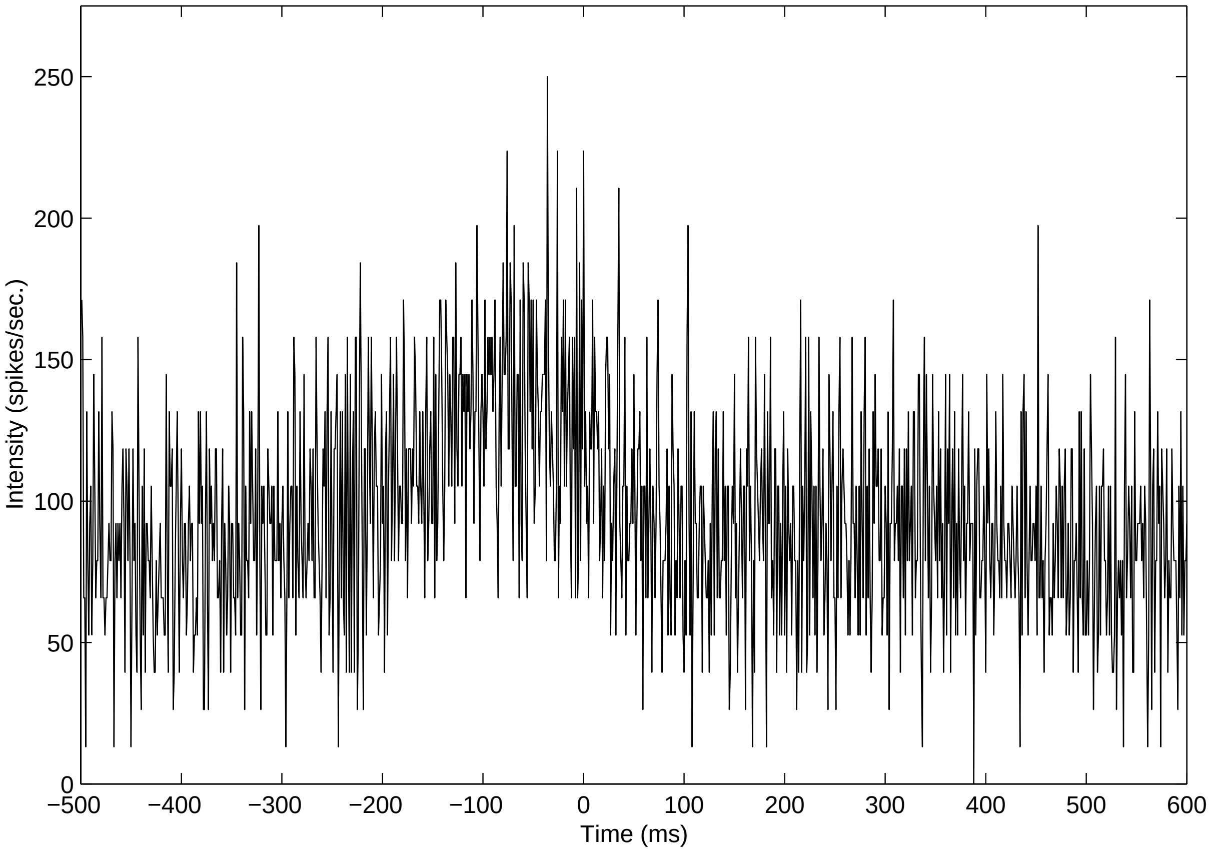

Intensity (spikes/sec.)
250
200
150
100
50
0
−500
−400
−300
−200
−100
0
100
200
300
400
500
600
Time (ms)

Figure 4 (a)

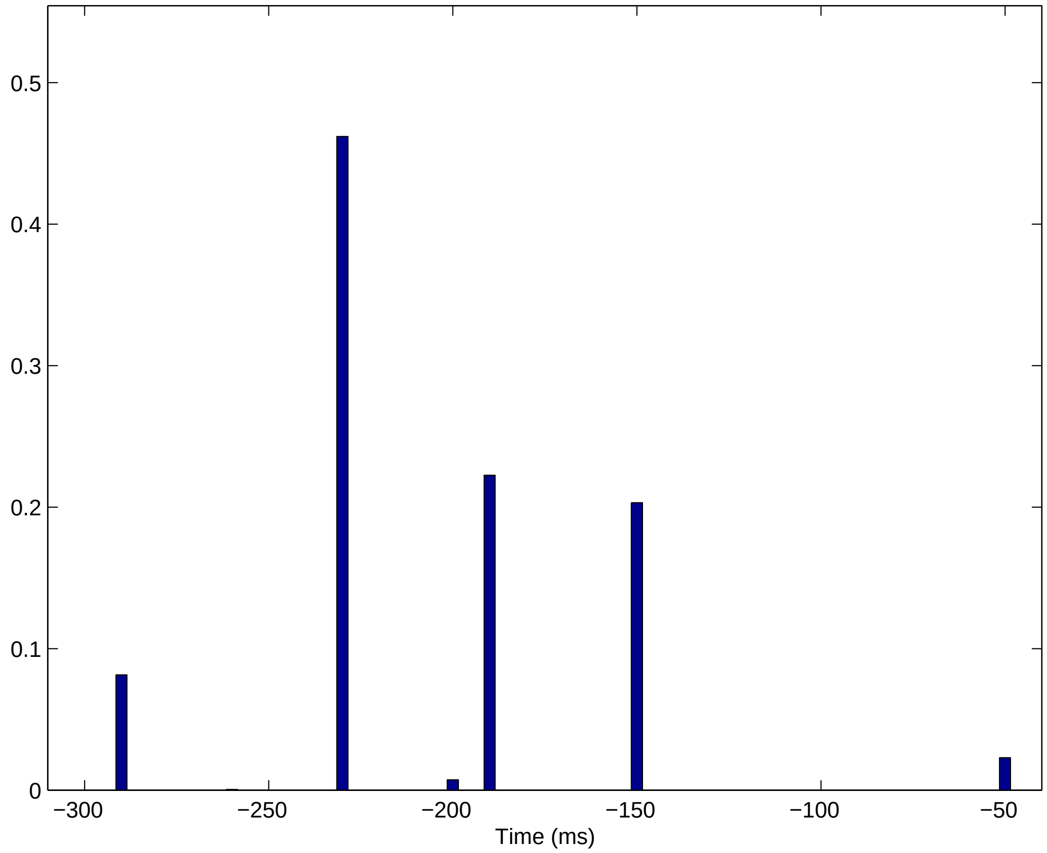


Figure 4 (b)

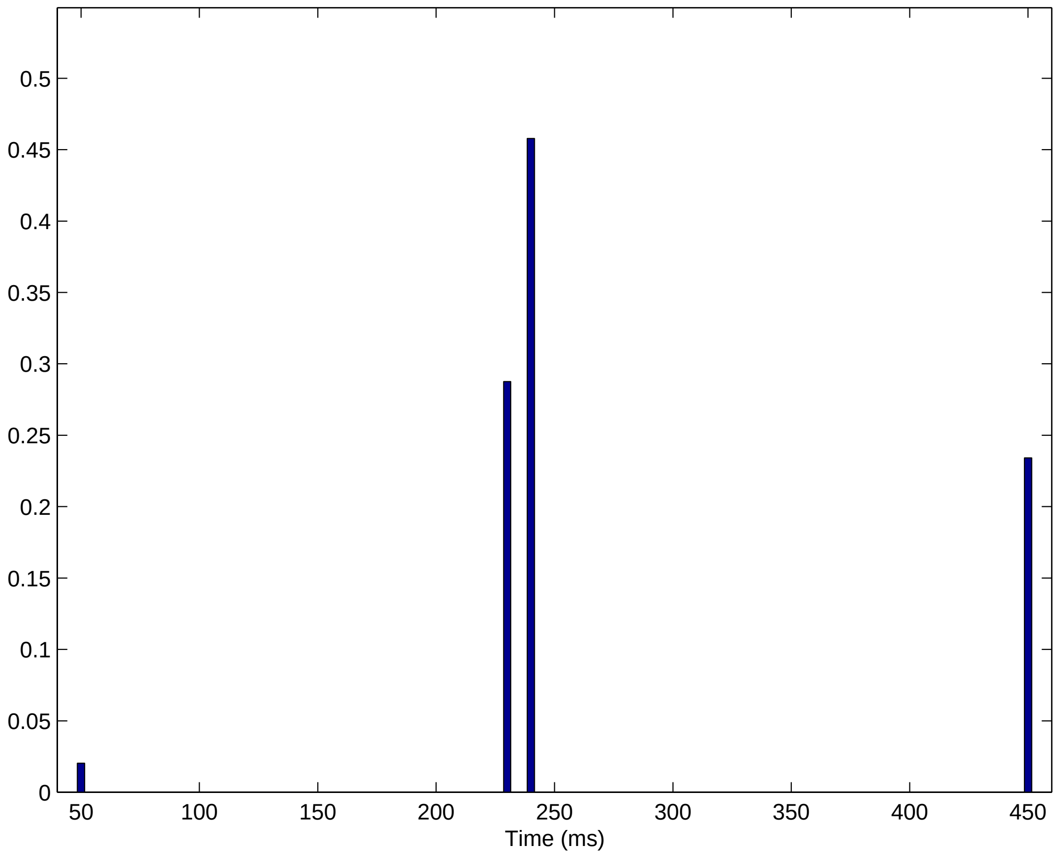

0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
50
100
150
200
250
300
350
400
450
Time (ms)

Figure 5 (a)

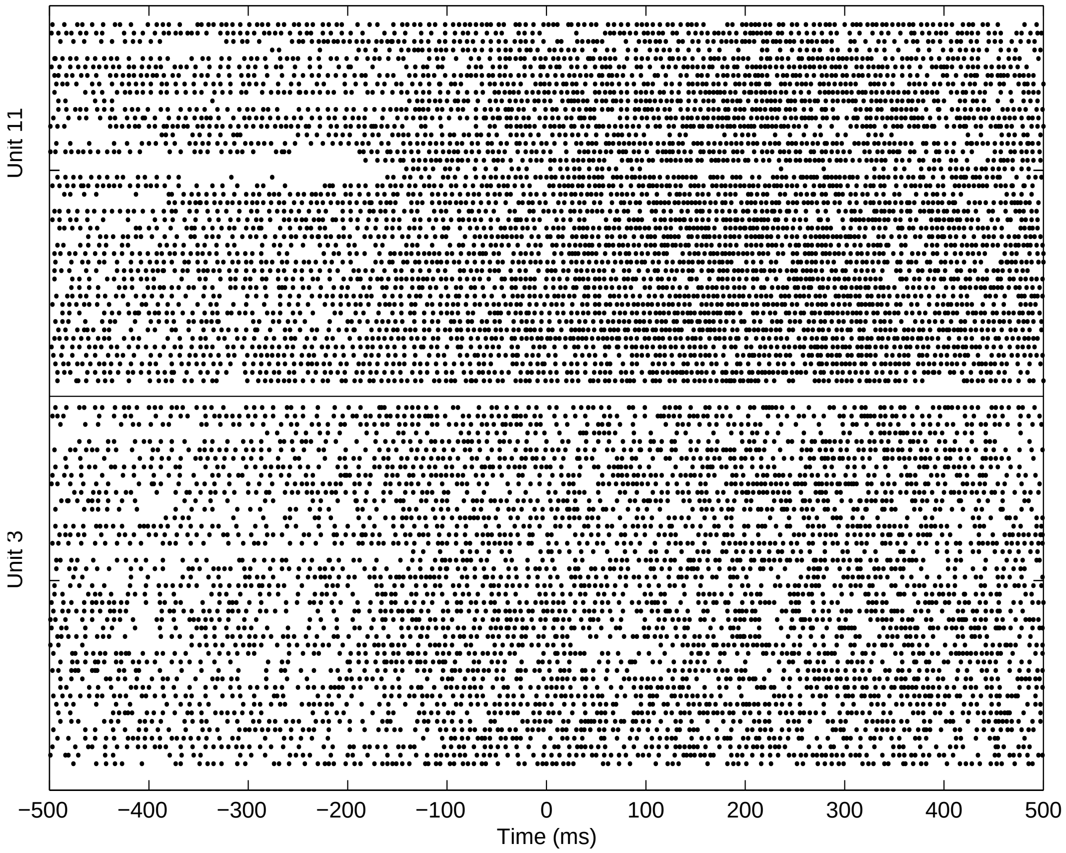
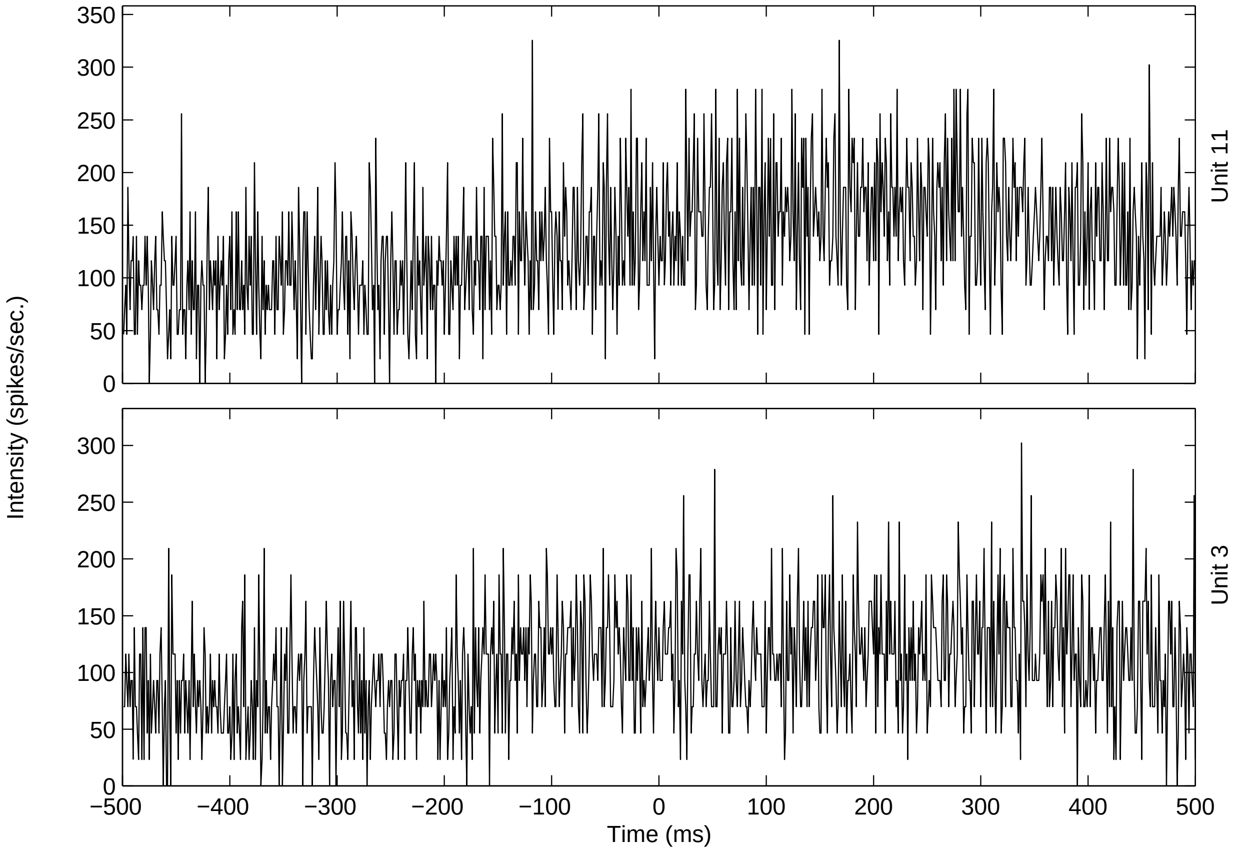


Figure 5 (b)



Intensity (spikes/sec.)
Unit 11
Unit 3
350
300
250
200
150
100
50
0
300
250
200
150
100
50
0
−500
−400
−300
−200
−100
0
100
200
300
400
500
Time (ms)

Figure 5 (c)

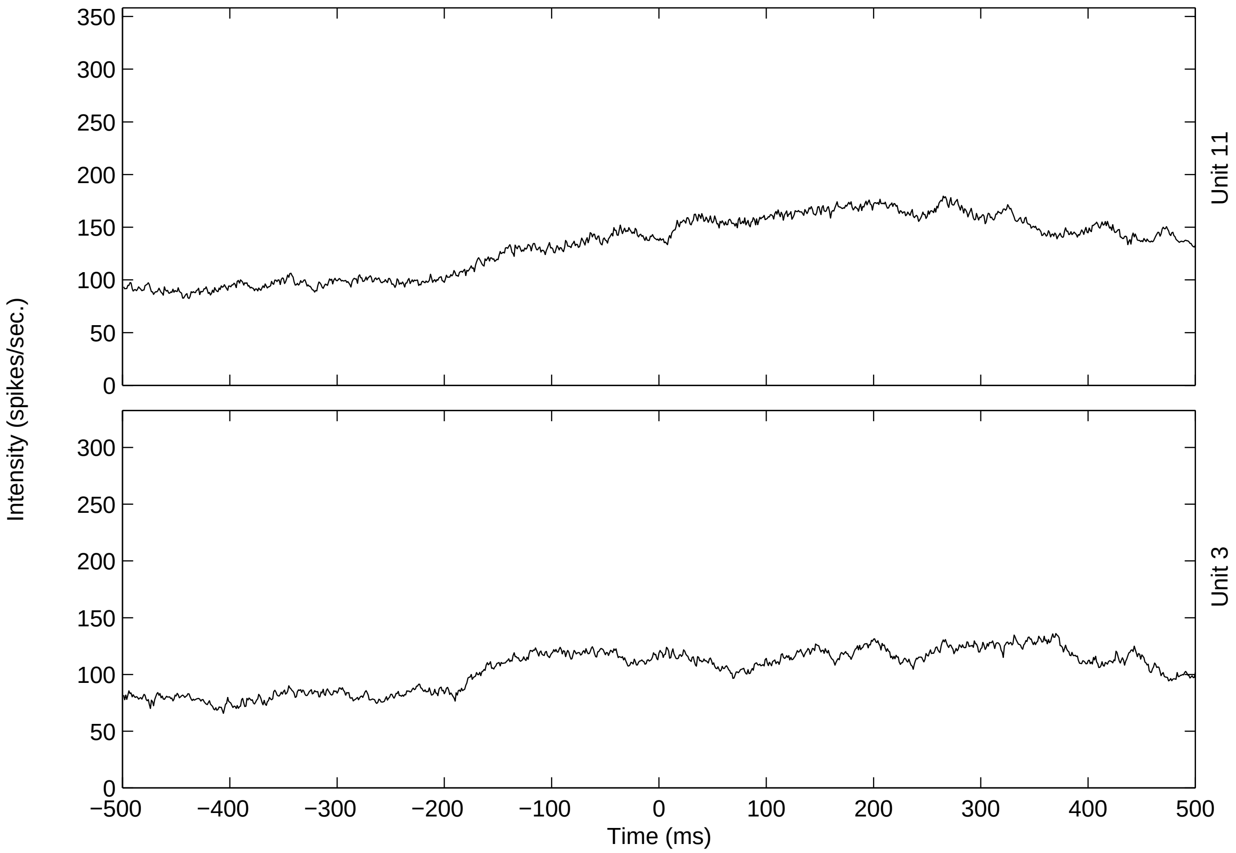

Intensity (spikes/sec.)
350
300
250
200
150
100
50
0
Unit 11
300
250
200
150
100
50
0
Unit 3
−500
−400
−300
−200
−100
0
100
200
300
400
500
Time (ms)

Figure 6 (a)

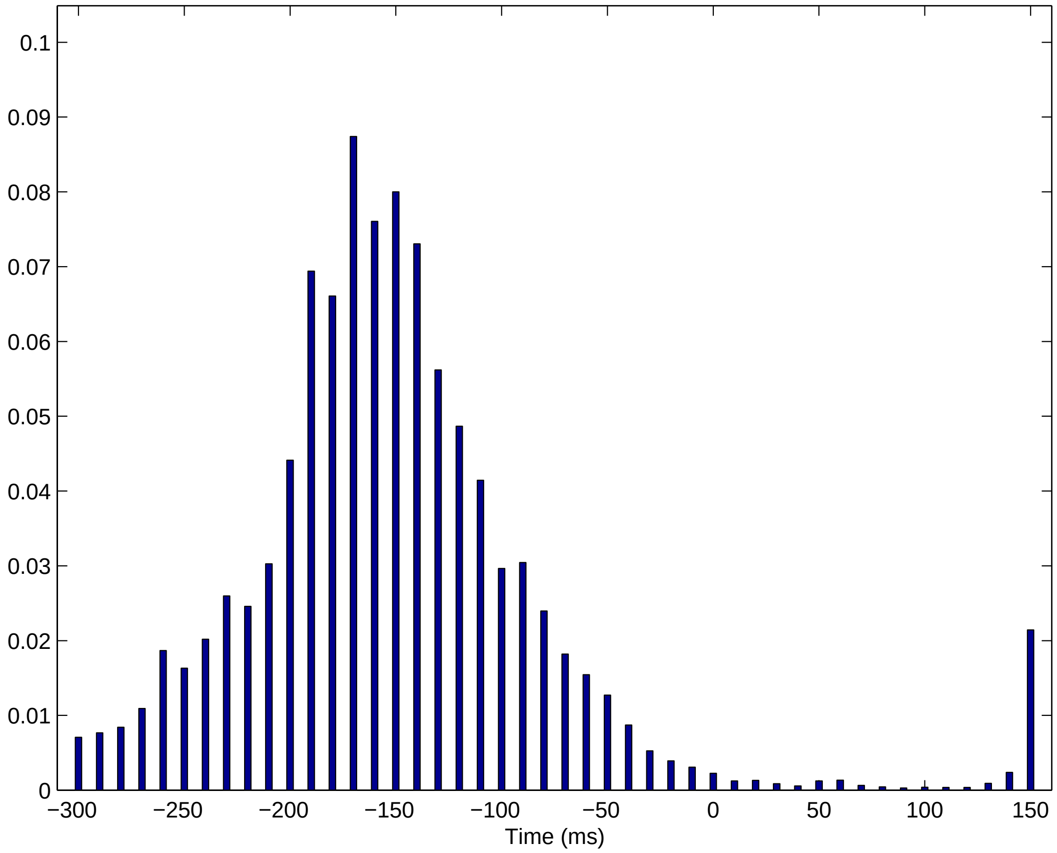
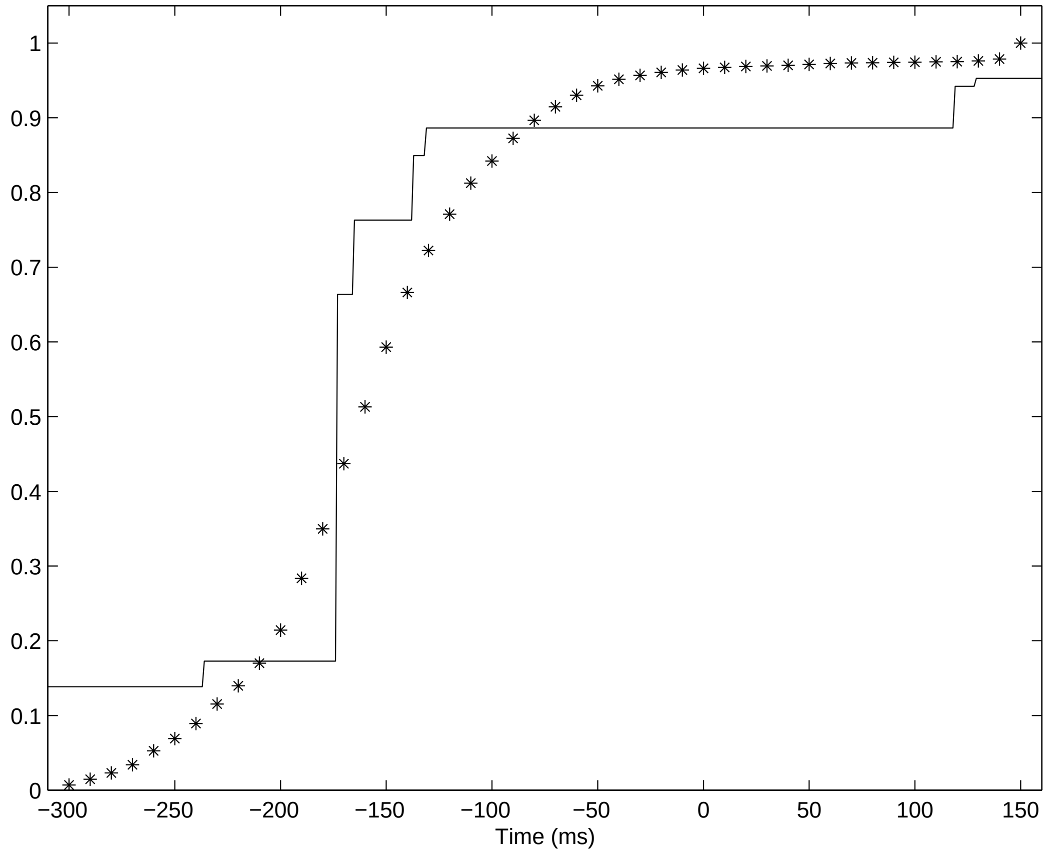

0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
−300
−250
−200
−150
−100
−50
0
50
100
150
Time (ms)

Figure 6 (b)



1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−300
−250
−200
−150
−100
−50
0
50
100
150
Time (ms)

Figure 6 (c)

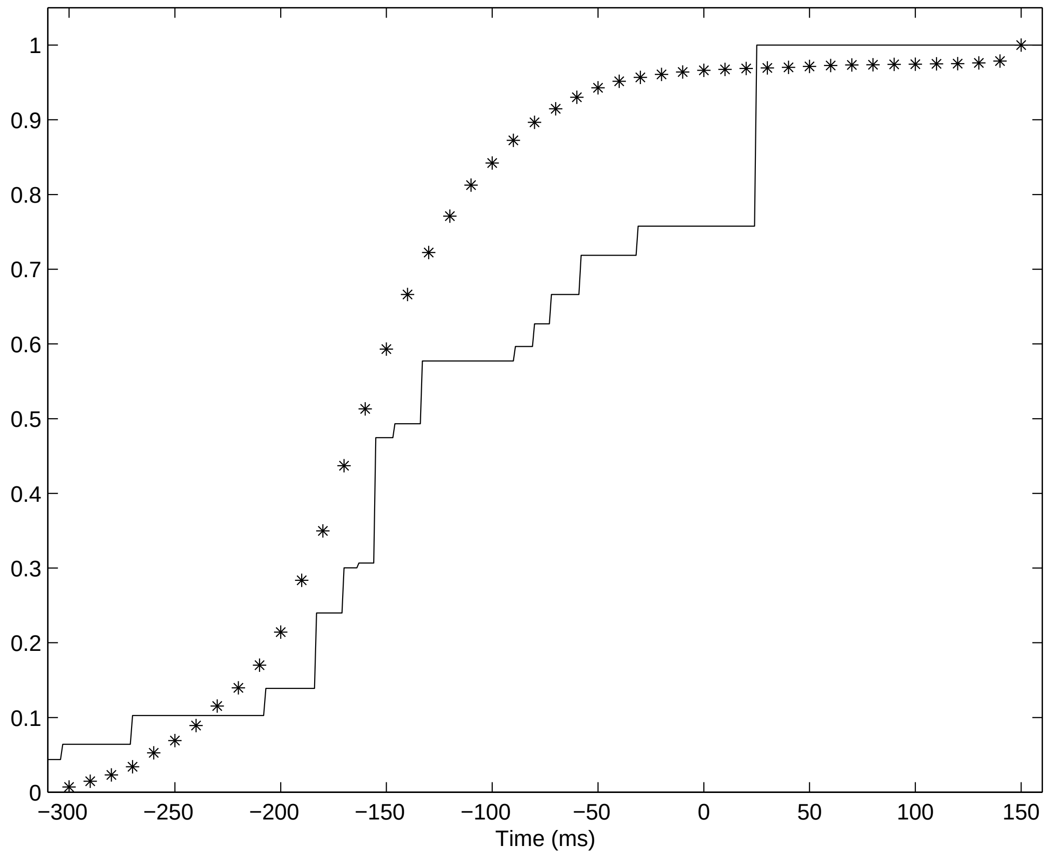

0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
−300
−250
−200
−150
−100
−50
0
50
100
150
Time (ms)

Figure 7 (a)

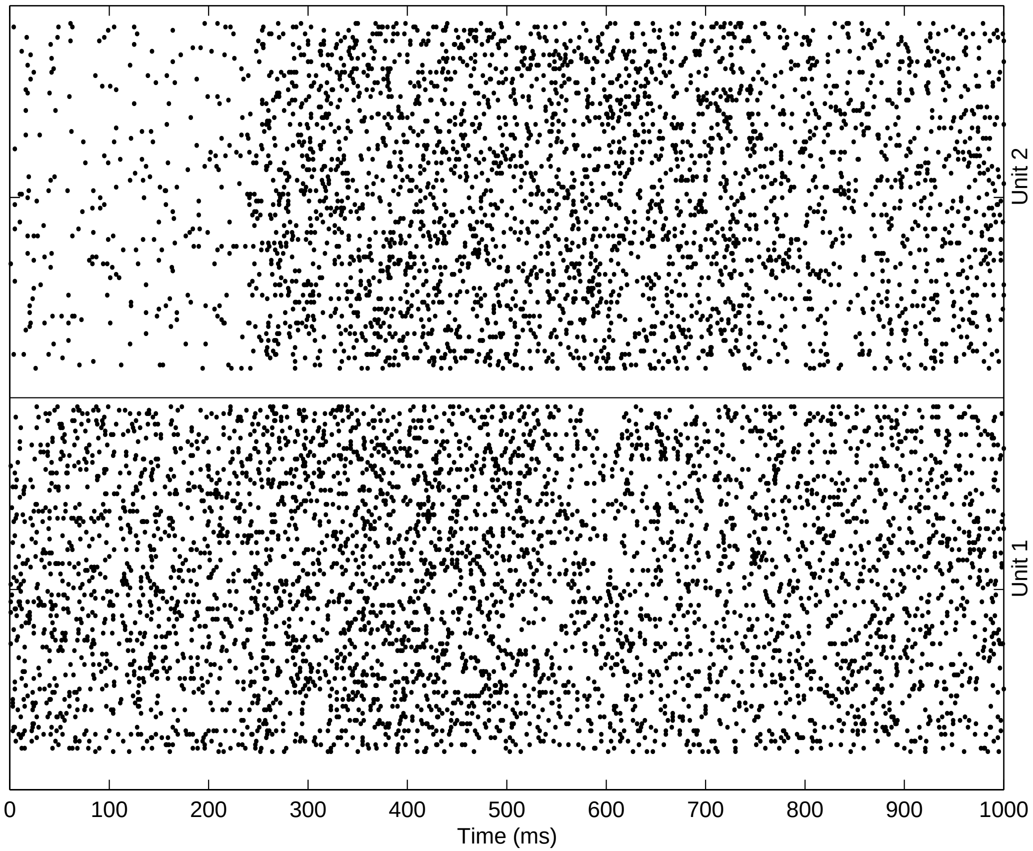


Figure 7 (b)

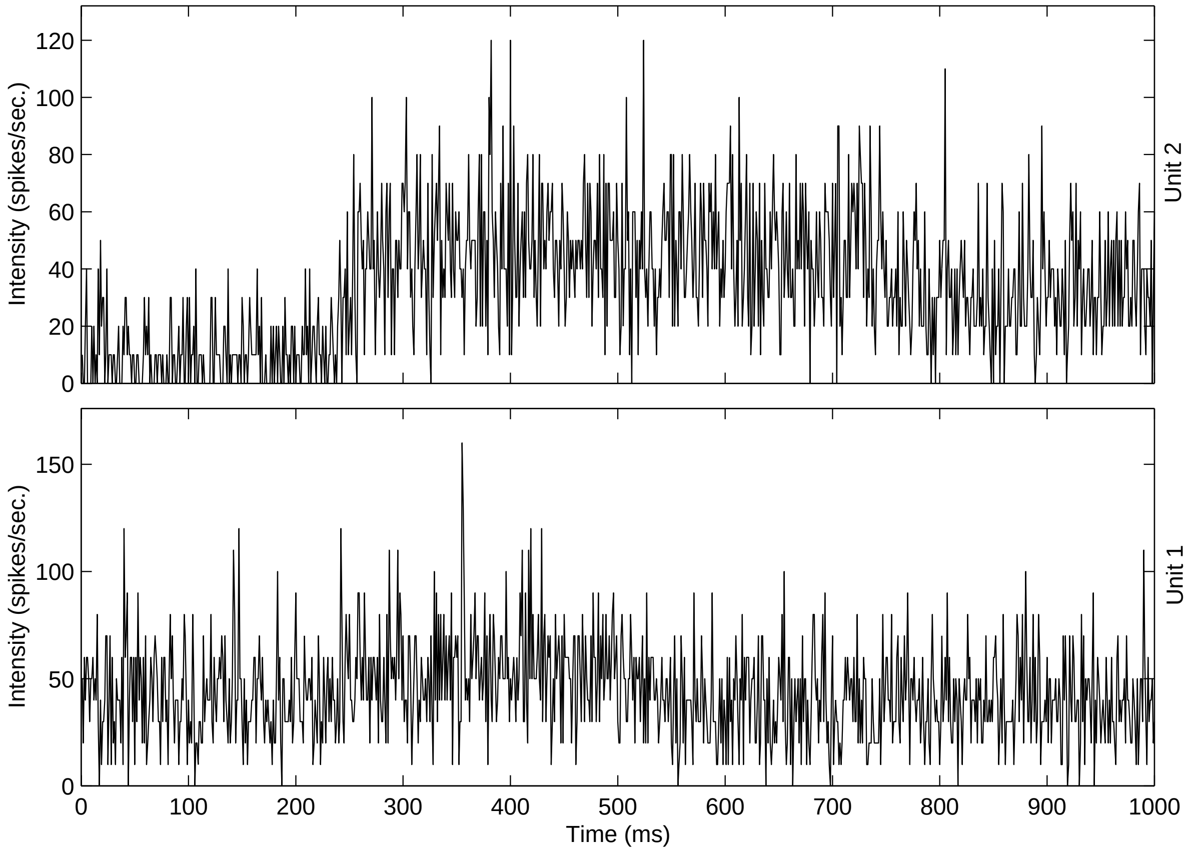

Intensity (spikes/sec.)
0
20
40
60
80
100
120
Unit 2
Intensity (spikes/sec.)
0
50
100
150
Unit 1
0
100
200
300
400
500
600
700
800
900
1000
Time (ms)

Figure 7 (c)

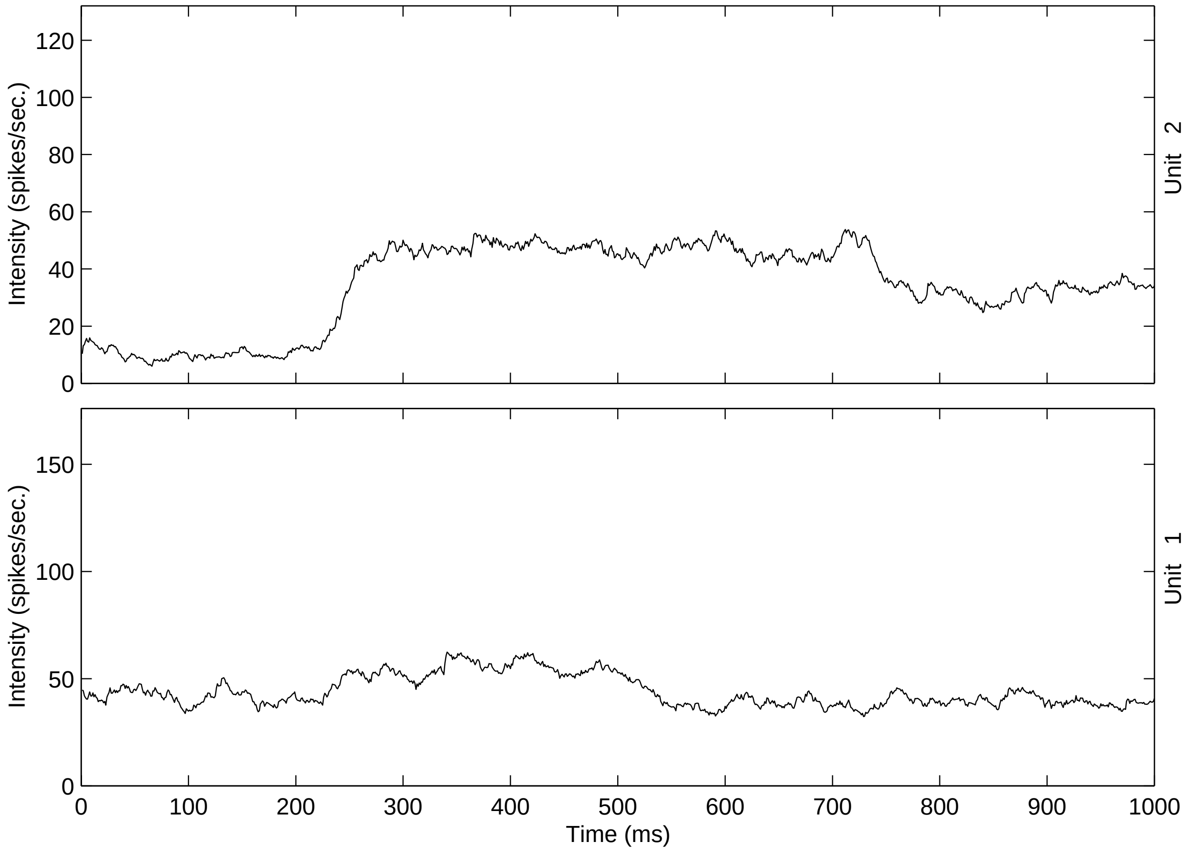

Intensity (spikes/sec.)
120
100
80
60
40
20
0
Unit 2
Intensity (spikes/sec.)
150
100
50
0
Unit 1
0
100
200
300
400
500
600
700
800
900
1000
Time (ms)

Figure 8 (a)

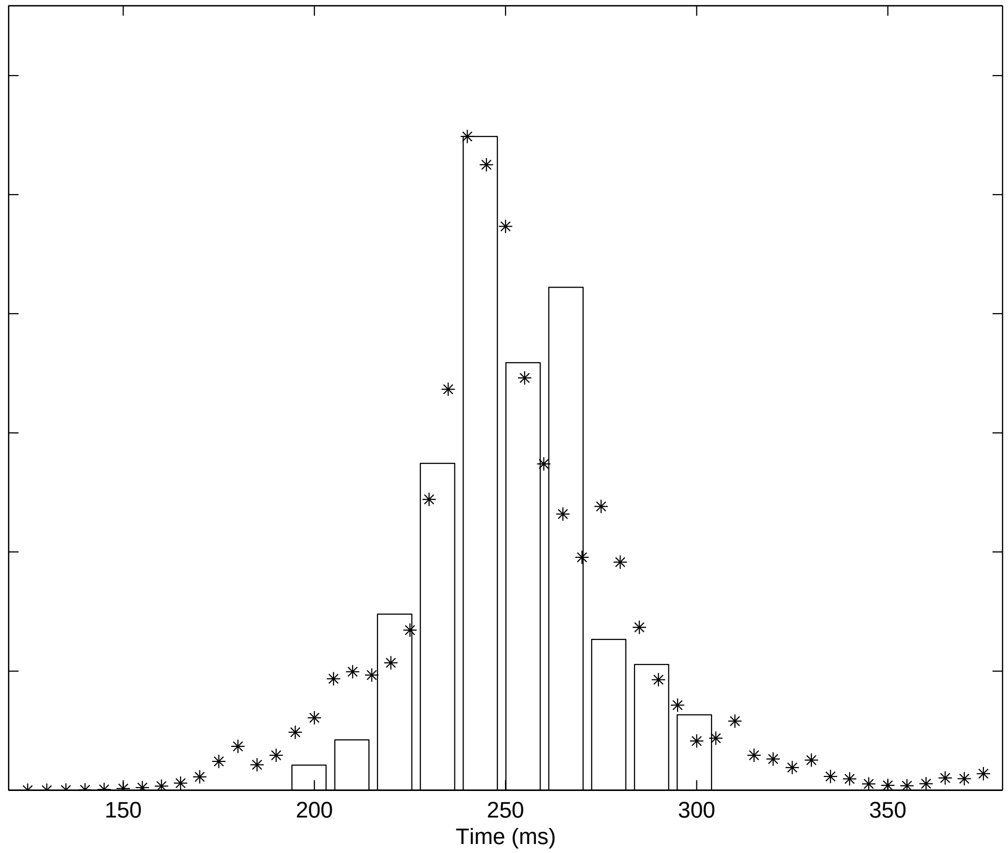


Figure 8 (b)

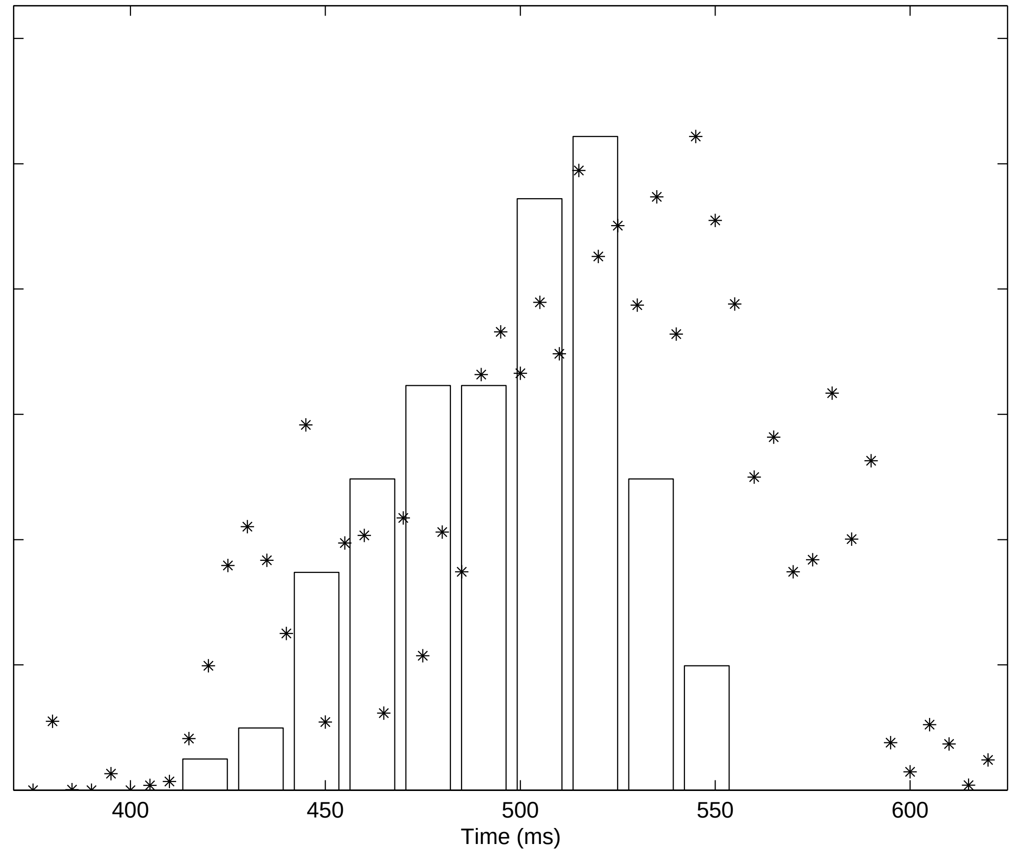


Figure 8 (c)

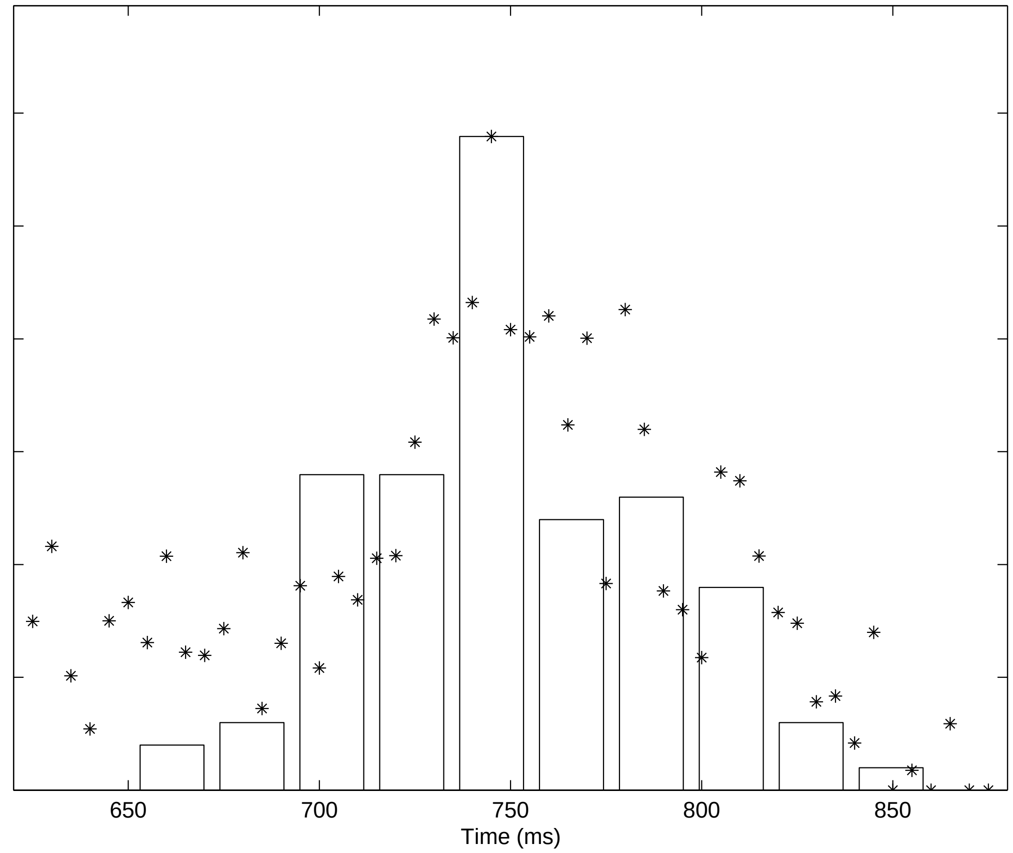
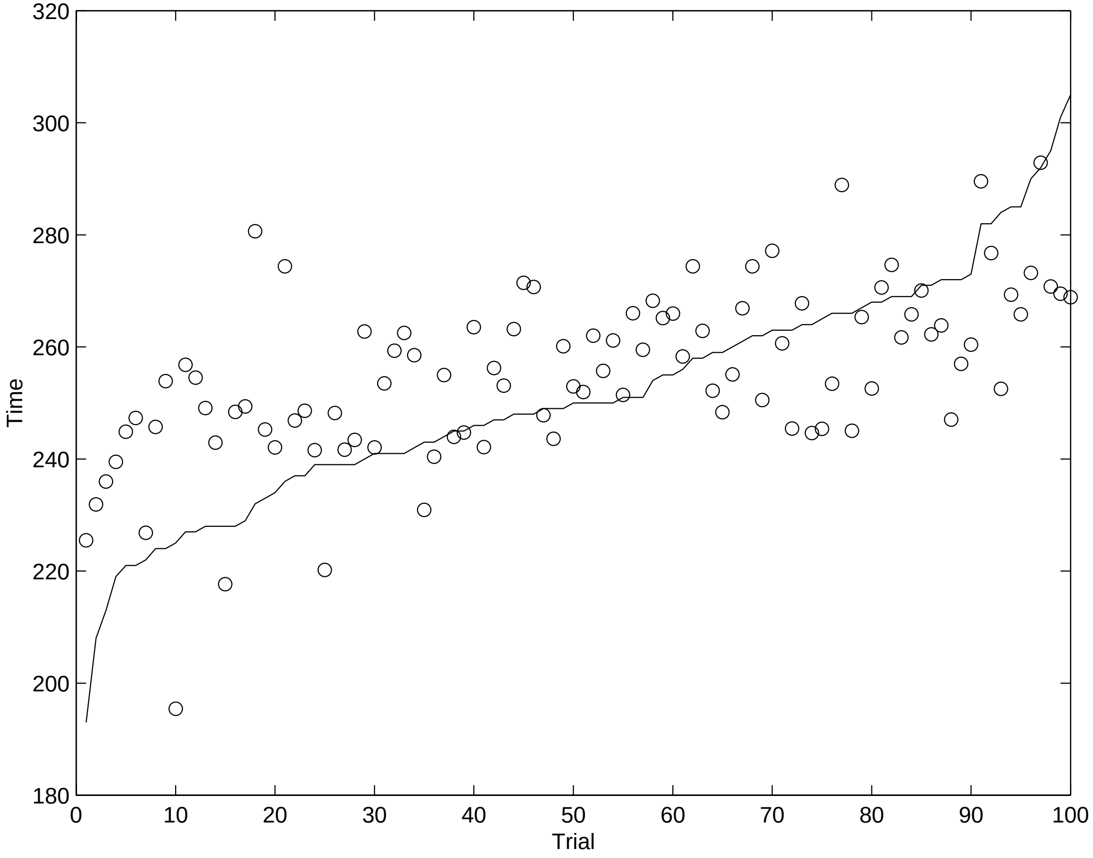


Figure 9.


Time
320
300
280
260
240
220
200
180
0
10
20
30
40
50
60
70
80
90
100
Trial

Interval:	$(0, \tau_1)$	(τ_1, τ_2)	(τ_2, τ_3)	$(\tau_3, 1)$
First process:	41.5 (40)	56.2 (60)	37.6 (40)	39.7 (40)
Second process:	9.9 (10)	47.9 (50)	48.0 (50)	31.4 (30)

Table 1: Simulation: estimates and true intensities (in spikes per second)