Peri-Stimulus Time Histograms Estimation Through Poisson Regression Without Generalized Linear Models

Christophe Pouzat, Antoine Chaffiol and Avner Bar-Hen

Mathématiques Appliquées à Paris 5 (MAP5)

Université Paris-Descartes and CNRS UMR 8145

christophe.pouzat@parisdescartes.fr

LASCON, January 27 2018

Outline

The Data and the Questions

First tool

Confidence sets

Second Tool

Where are we ?

The Data and the Questions

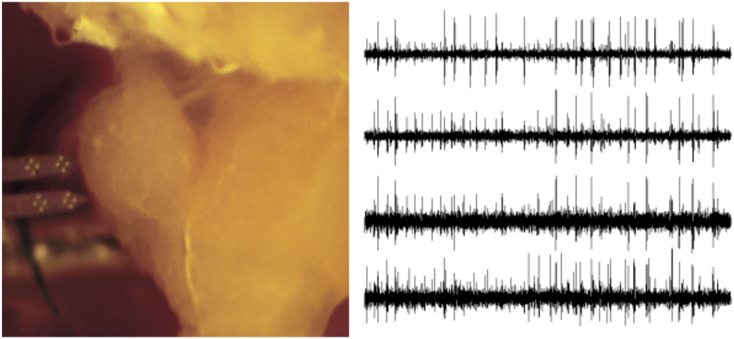
First tool

Confidence sets

Second Tool

Data's origin

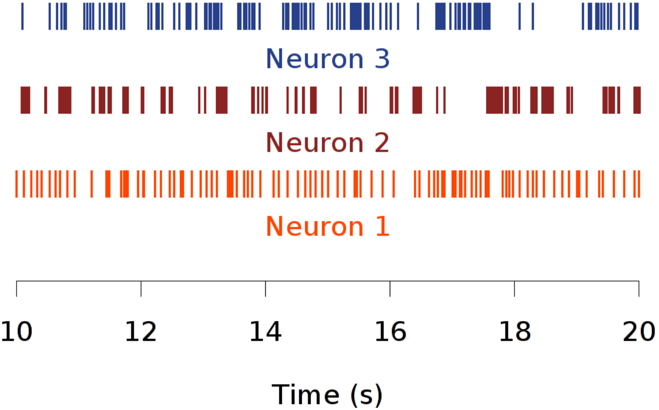
Viewed "from the outside", neurons generate brief electrical pulses: the action potentials



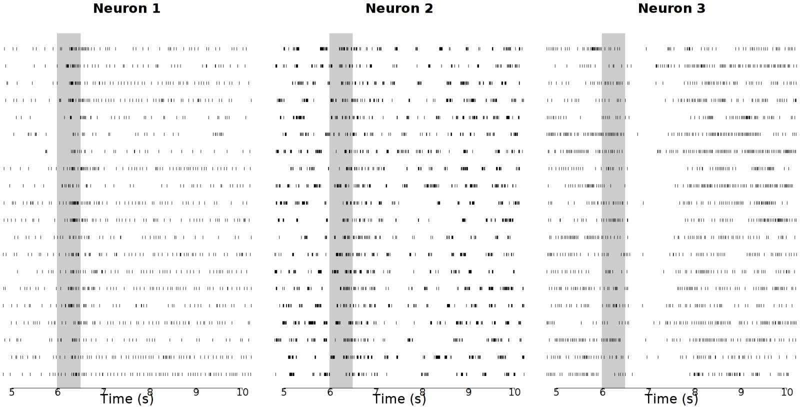
Left, the brain of an insect with the recording probe on which 16 electrodes (the bright spots) have been etched. Each probe's branch has a 80 μm width. Right, 1 sec of data from 4 electrodes. The spikes are the action potentials.

Spike trains

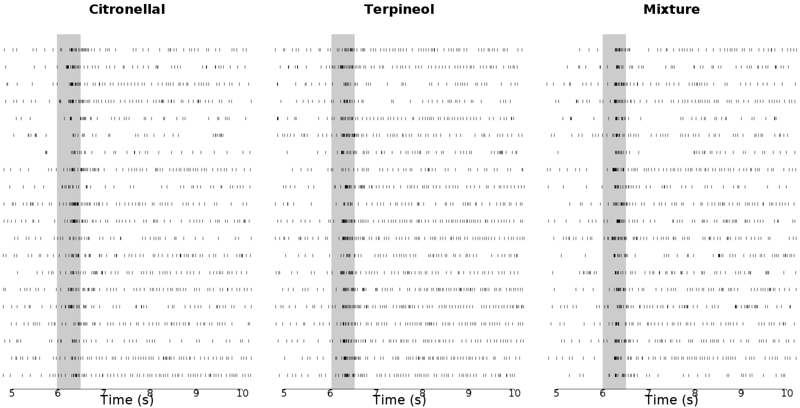
After a "rather heavy" pre-processing called spike sorting, the raster plot representing the spike trains can be built:



Non-stationary regime: odor responses



20 stimulation with citronellal. Stimulation are delivered during 500 ms (gray background). Is neuron 2 responding to the stimulation? Cockroach (*Periplaneta americana*) recordings and spike sorting by Antoine Chaffiol.

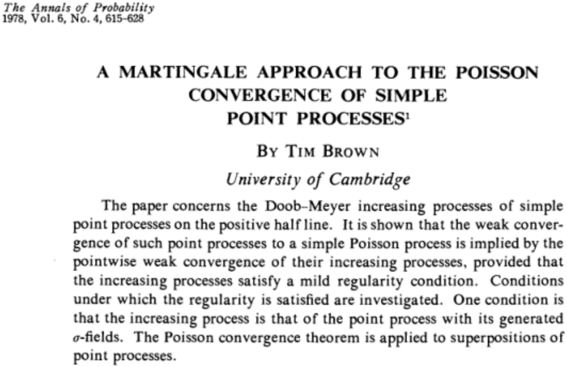


Neuron 1: 20 stimulation with citronellal, terpineol and a mixture of the two. Are the reponses any different?

What do we want?

- We want to estimate the peri-stimulus time histogram (PSTH) considered as an observation from an inhomogeneous Poisson process.
- In addition to estimation we want to:
    - Test if a neuron is responding to a given stimulation.
    - Test if the responses of a given neuron to two different stimulations are different.
- This implies building some sort of confidence bands around our best estimation.

Clarification on the convergence towards an IHP



The Annals of Probability
1978, Vol. 6, No. 4, 615-628

A MARTINGALE APPROACH TO THE POISSON CONVERGENCE OF SIMPLE POINT PROCESSES[1]

BY TIM BROWN

University of Cambridge

The paper concerns the Doob-Meyer increasing processes of simple point processes on the positive half line. It is shown that the weak convergence of such point processes to a simple Poisson process is implied by the pointwise weak convergence of their increasing processes, provided that the increasing processes satisfy a mild regularity condition. Conditions under which the regularity is satisfied are investigated. One condition is that the increasing process is that of the point process with its generated σ-fields. The Poisson convergence theorem is applied to superpositions of point processes.

In this paper (Corollary 2, p. 626), Tim Brown shows that aggregated uncorrelated point processes converge towards an inhomogenous Poisson Process (IHP).

Where are we ?

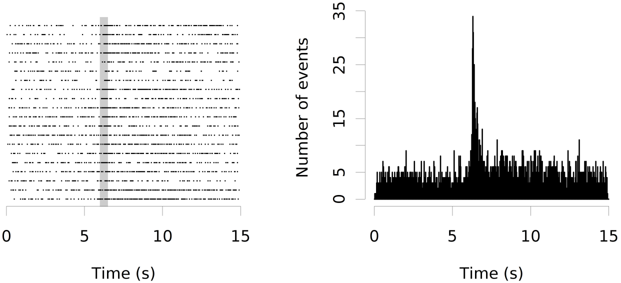
The Data and the Questions

First tool

Confidence sets

Second Tool

The PSTH



We go from the raw data to an histogram built with a tiny time step (25 ms), leading to an estimator with little bias and large variance.

- We model this "averaged process" as an inhomogeneous Poisson process with intensity λ(t).
- The histogram we just built can then be seen as the observation of a collection of Poisson random variables, $\{Y_1, \ldots, Y_k\}$, with parameters:

  $$n \int_{t_i-\delta/2}^{t_i+\delta/2} \lambda(u)\, du \;\approx\; n\, \lambda(t_i)\, \delta \; , \quad i = 1, \ldots, k \; ,$$

  where t_i is the center of a class (bin), δ is the bin width, n is the number of stimulations and k is the number of bins.
- A piecewise constant estimator of λ(t) is then obtained with:

  $$\hat{\lambda}(t) = y_i/(n\delta)\,, \quad \text{if} \quad t \in [t_i - \delta/2, t_i + \delta/2) \; .$$

  This is the "classical" PSTH.

- We are going to assume that $\lambda(\mathrm{t})$ is smooth—this is a very reasonable assumption given what we know about the insect olfactory system.
- We can then attempt to improve on the "classical" PSTH by trading a little bias increase for (an hopefully large) variance decrease.
- Many nonparametric methods are available to do that: kernel regression, local polynomials, smoothing splines, wavelets, etc.
- A problem in the case of the PSTH is that the observed counts ($\{y_1, \ldots, y_k\}$) follow Poisson distributions with different parameters implying that they have different variances.
- We have then at least two possibilities: i) use a generalized linear model (GLM); ii) transform the data to stabilize the variance.
- We are going to use the second approach.

Kernel smoothing

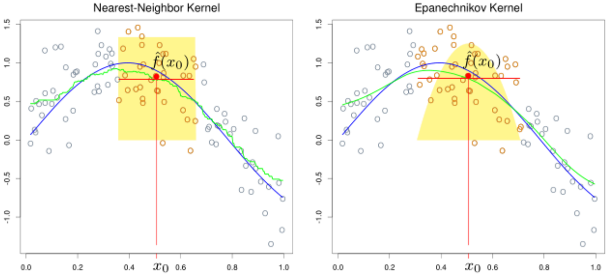


FIGURE 6.1. *In each panel* 100 *pairs* x_i, y_i *are generated at random from the blue curve with Gaussian errors:* $Y = \sin(4X) + \varepsilon$, $X \sim U[0,1]$, $\varepsilon \sim N(0, 1/3)$. *In the left panel the green curve is the result of a* 30*-nearest-neighbor running-mean smoother. The red point is the fitted constant* $\hat{f}(x_0)$, *and the red circles indicate those observations contributing to the fit at* x_0. *The solid yellow region indicates the weights assigned to observations. In the right panel, the green curve is the kernel-weighted average, using an Epanechnikov kernel with (half) window width* $\lambda = 0.2$.

From Hastie, Tibshirani & Friedman (2009) *The Elements of Statistical Learning*.

Error propagation

- Let us consider two random variables: X and Z such that:
- $X \approx \mathcal{N}(\mu_X, \sigma_X^2)$ or $X \approx \mu_X + \sigma_X\,\epsilon$
- $Z = G(X)$, with f continuous and differentiable.
- Using a first order Taylor expansion we then have:

$$\begin{aligned} Z &\approx G(\mu_X + \sigma_X\,\epsilon) \\ &\approx G(\mu_X) + \sigma_X\,\epsilon\,\tfrac{dG}{dX}(\mu_X) \end{aligned}$$

- $\mathrm{E}Z \approx G(\mu_X) = G(\mathrm{E}X)$
- $\mathrm{Var}Z \equiv \mathrm{E}[(Z - \mathrm{E}Z)^2] \approx \sigma_X^2\,\tfrac{dG}{dX}^2(\mu_X)$
- $Z \approx G(\mu_X) + \sigma_X\left|\tfrac{dG}{dX}(\mu_X)\right|\,\epsilon$

Variance stabilisation

- Following Brown, Cai and Zhou (2010), let's consider $X_1, \ldots, X_n$ IID from a Poisson distribution with parameter ν.
- Define $X = \sum_{j=1}^{n} X_j$, the CLT gives us:

$$\sqrt{n}\,(X/n - \nu) \stackrel{L}{\rightarrow} \mathcal{N}(0, \nu) \quad \text{as } n \rightarrow \infty\,.$$

- A variance stabilizing transformation is a function $G : \mathbb{R} \rightarrow \mathbb{R}$, such that:

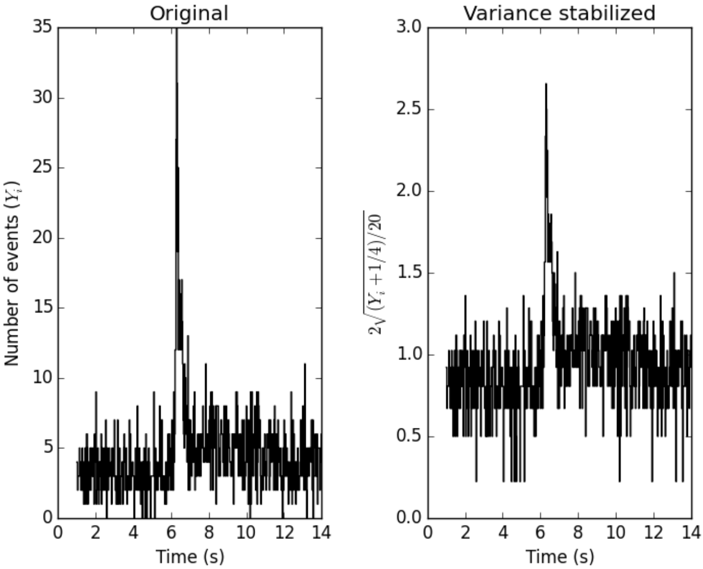
$$G'(x) = 1/\sqrt{x}\,.$$

- The delta method (or the error propagation method; a first order Taylor expansion) then yields:

$$\sqrt{n}\,(G(X/n) - G(\nu)) \stackrel{L}{\rightarrow} \mathcal{N}(0, 1)\,.$$

- It is known (Anscombe, 1948) that the variance stabilizing properties can be further improved by using transformation of the form:
$$H_n(X) = G\left(\frac{X+a}{n+b}\right)$$
for suitable choices of a and b.
- In nonparametric regression we want to set a and b such that $\mathrm{E}\left(H_n(X)\right)$ optimally matches $G(\nu)$.
- Brown, Cai and Zhou (2010) show that in all relevant PSTH estimation problems we have:
$$\mathrm{Var}\left(2\sqrt{(X+1/4)/n}\right) = \frac{1}{n} + O(n^{-2})\,.$$
- They also show that:
$$\mathrm{E}\left(2\sqrt{(X+1/4)/n}\right) - 2\sqrt{\nu} = O(n^{-2})\,.$$
- They get similar transformations for binomial and negative binomial random variables.

Example



Original
Number of events (Y_i)
0
5
10
15
20
25
30
35
0
2
4
6
8
10
12
14
Time (s)
Variance stabilized
$2\sqrt{(Y_i+1/4)/20}$
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
2
4
6
8
10
12
14
Time (s)

Nonparametric estimation

- Since our knowledge of the biophysics of these neurons and of the network they form is still in its infancy, we can hardly propose a reasonable parametric from for our PSTHs (or their variance stabilized versions).
- We therefore model our stabilized PSTH by:

$$Z_i \doteq 2\sqrt{(Y_i + 1/4)/n} = r(t_i) + \epsilon_i \sigma \,,$$

where the $\epsilon_i \overset{\text{IID}}{\sim} \mathcal{N}(0,1)$, r is assumed "smooth" and is estimated with a linear smoother (kernel regression, local polynomials, smoothing splines) or with wavelets (or with any nonparametric method you like).

- Following Larry Wasserman (*All of Nonparametric Statistics*, 2006) we define a linear smoother by a collection of functions $l(t) = (l_1(t), \ldots, l_k(t))^T$ such that:

$$\hat{r}(t) = \sum_{i=1}^{k} l_i(t) Z_i \,.$$

- The simplest smoother we are going to use is built from the tricube kernel:

$$K(t) = \frac{70}{81} \left(1 - |t|^3\right)^3 I(t) \,,$$

  where $I(t)$ is the indicator function of $[-1, 1]$.
- The functions l_i are then defined by:

$$l_i(t) = \frac{K\left(\frac{t - t_i}{h}\right)}{\sum_{j=1}^{k} K\left(\frac{t - t_j}{h}\right)} \,.$$

- When using this kind of approach the choice of the bandwidth h is clearly critical.
- Since after variance stabilization the variance is known we can set our bandwidth by minimizing Mallows' C_p criterion instead of using cross-validation. For (soft) wavelet thresholding we use the universal threshold that requires the knowledge (or an estimation) of the variance.
- More explicitly, with linear smoothers our estimations $(\widehat{r}(t_1), \ldots, \widehat{r}(t_k))^T$ can be written in matrix form as:

  $$\widehat{\mathbf{r}} = L(h)\,\mathbf{Z}\,,$$

  where $L(h)$ is the $k \times k$ symmetric matrix whose element (i, j) is given by $l_i(t_j)$.

- Ideally we would like to set $\widehat{h}$ as:

$$\arg\min_h (1/k) \sum_{i=1}^{k} \left(r(t_i) - \hat{r}(t_i)\right)^2 \ .$$

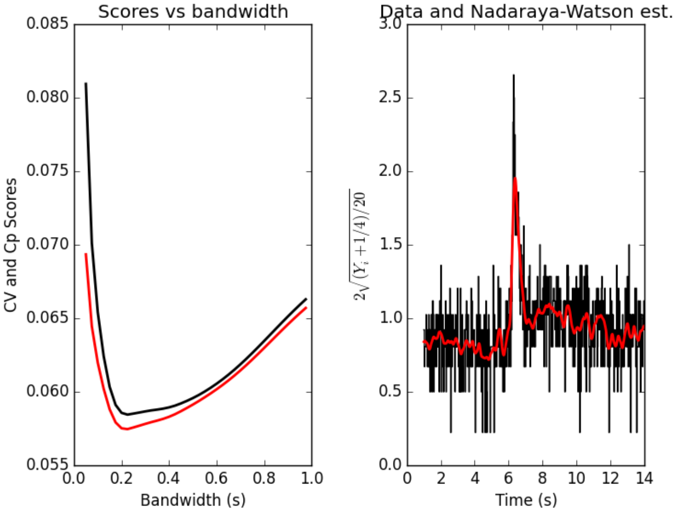
- But we don't know r (that's what we want to estimate!) so we minimize Mallows' C_p criterion:

$$(1/k) \sum_{i=1}^{k} \left(Z_i - \hat{r}(t_i)\right)^2 + 2\sigma^2 \mathrm{tr}\left(L(h)\right)/k \,,$$

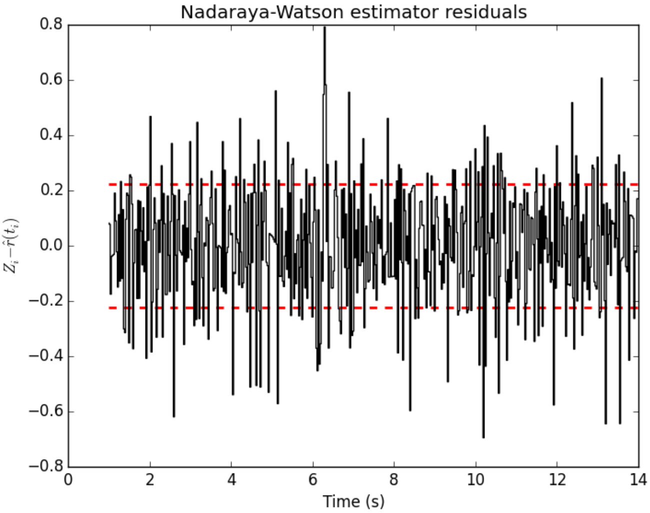
where $\mathrm{tr}\left(L(h)\right)$ stands for the trace of $L(h)$.

- If we don't know σ^2, we minimize the cross-validation criterion:

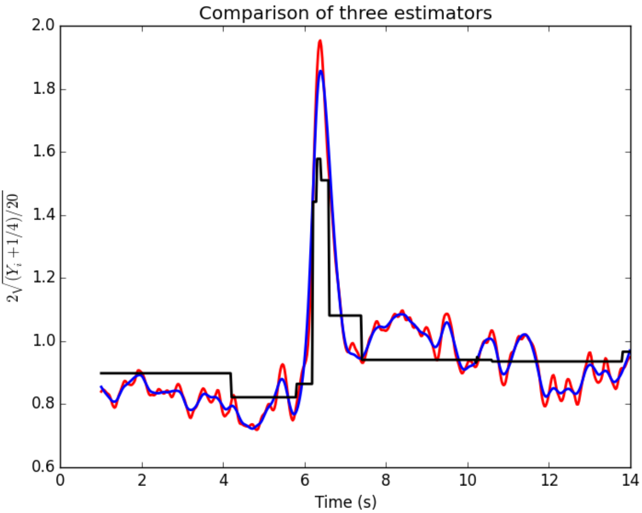
$$\frac{1}{k} \sum_{i=1}^{k} \frac{\left(Z_i - \hat{r}(t_i)\right)^2}{1 - L_{ii}(h)} \,.$$



Left: CV score in black, Cp score in red. Right: Variance stabilized data (black) with Nadaraya-Watson estimator (red) with "best" bandwidth.



Residuals obtained with the Nadaraya-Watson estimator. The red dashed lines correspond to $\pm\sigma$.



Nadaraya-Watson estimator (red), smoothing splines estimator (blue) and wavelet estimator (black; Haar wavelets, soft thresholding, universal threshold).

Where are we ?

The Data and the Questions

First tool

Confidence sets

Second Tool

Confidence sets

- Keeping in line with Wasserman (2006), we consider that providing an estimate $\hat{r}$ of a curve r is not sufficient for drawing scientific conclusions.
- We would like to provide a confidence set for r in the form of a band:

  $$\mathcal{B} = \{s : l(t) \leq s(t) \leq u(t),\ \forall t \in [a, b]\}$$

  based on a pair of functions $(l(t), u(t))$.
- We would like to have:

  $$\Pr\{r \in \mathcal{B}\} \geq 1 - \alpha$$

  for all $r \in \mathcal{R}$ where $\mathcal{R}$ is a large class of functions.

- When working with smoothers, our estimators exhibit a bias that does not disappear even with large sample sizes.
- We will therefore try to built sets around $\overline{r} = \mathrm{E}(\hat{r})$; that will be sufficient to address some of the questions we started with.
- For a linear smoother, $\hat{r}(t) = \sum_{i=1}^{k} l_i(t) Z_i$, we have:

$$\overline{r}(t) = \mathrm{E}\left(\hat{r}(t)\right) = \sum_{i=1}^{k} l_i(t) r(t_i)$$

  and

$$\mathrm{Var}\left(\hat{r}(t)\right) = \sigma^2 \sum_{i=1}^{k} l_i(t)^2 = (1/n) \|l(t)\|^2 \,.$$

  Remember that we stabilized the variance at $1/n$.
- We will consider a confidence band for $\overline{r}(t)$ of the form:

$$I(t) = \left(\hat{r}(t) - c\|l(t)\|/\sqrt{n}, \hat{r}(t) + c\|l(t)\|/\sqrt{n}\right) \,,$$

  for some $c > 0$ and $a \leq t \leq b$.

Following Sun and Loader (1994), we have:

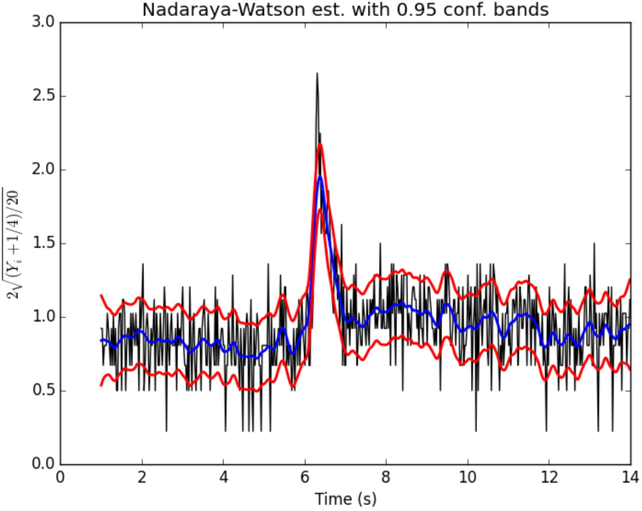
$$\begin{aligned}
\Pr\left\{\bar{r}(t) \notin I(t) \text{ for some } t \in [a,b]\right\} &= \Pr\left\{\max_{t\in[a,b]} \frac{|\hat{r}(t)-\bar{r}(t)|}{\|l(t)\|/\sqrt{n}} > c\right\} \\
&= \Pr\left\{\max_{t\in[a,b]} \frac{|\sum_{i=1}^{k}(\epsilon_i/\sqrt{n})l_i(t)|}{\|l(t)\|/\sqrt{n}}\right. \\
&= \Pr\left\{\max_{t\in[a,b]} |W(t)| > c\right\},
\end{aligned}$$

where $W(t) = \sum_{i=1}^{k} \epsilon_i l_i(t)/\|l(t)\|$ is a Gaussian process. To find c we need to know the distribution of the maximum of a Gaussian process. Sun and Loader (1994) showed the tube formula:

$$\Pr\left\{\max_{t\in[a,b]} |\sum_{i=1}^{k} \epsilon_i l_i(t)/\|l(t)\|| > c\right\} \approx 2\left(1-\Phi(c)\right) + \frac{\kappa_0}{\pi}\exp -\frac{c^2}{2},$$

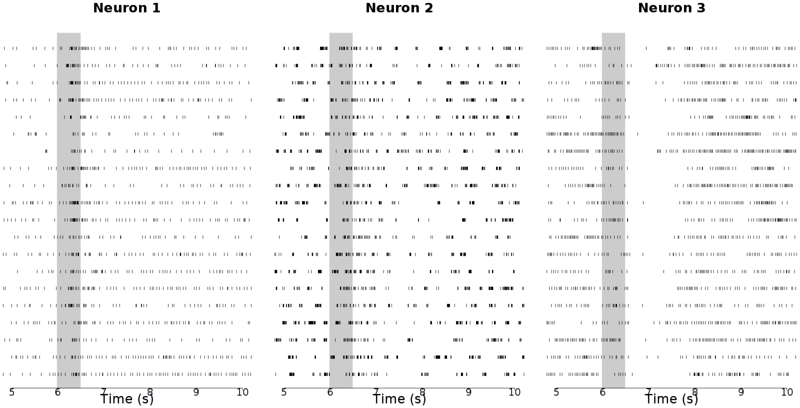
for large c, where, in our case, $\kappa_0 \approx (b-a)/h\left(\int_a^b K'(t)^2 dt\right)^{1/2}$. We get c by solving:

$$2\left(1-\Phi(c)\right) + \frac{\kappa_0}{\pi}\exp -\frac{c^2}{2} = \alpha\,.$$

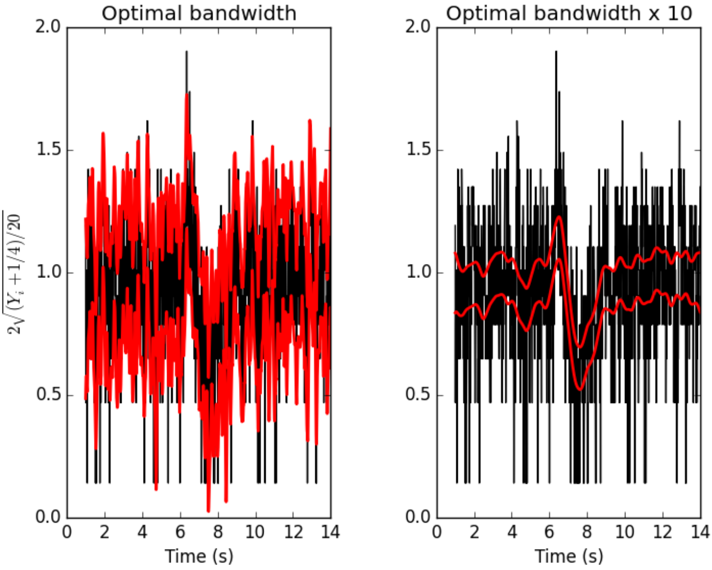


Variance stabilized data (black) Nadaraya-Watson estimator (blue) and 0.95 confidence band (red).

Do you remember this slide?



20 stimulation with citronellal. Stimulation are delivered during 500 ms (gray background). Is neuron 2 responding to the stimulation?



Since the null hypothesis is a constant, there is no bias and we can increase the bandwidth (right side) if necessary.

Where are we ?

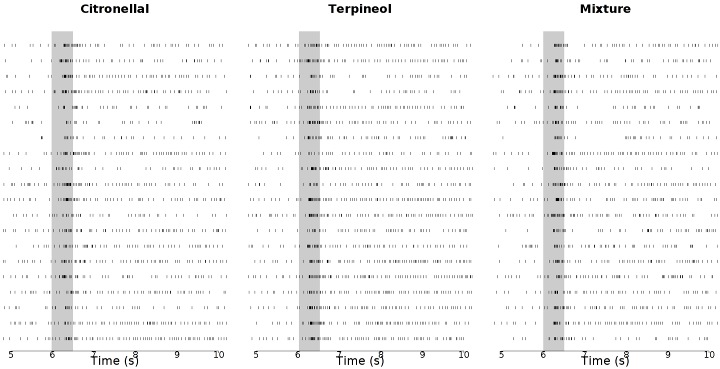
The Data and the Questions

First tool

Confidence sets

Second Tool

Remember again?



Neuron 1: 20 stimulation with citronellal, terpineol and a mixture of the two. Are the reponses any different?

Setting the test

- We start like previously by building a "classical" PSTH with very fine bins (25 ms) with the citronellal and terpineol trials to get: $\{y_1^{citron}, \ldots, y_k^{citron}\}$ and $\{y_1^{terpi}, \ldots, y_k^{terpi}\}$.
- We stabilize the variance as we did before ($z_i = 2\sqrt{(y_i + 0.25)/n}$) to get: $\{z_1^{citron}, \ldots, z_k^{citron}\}$ and $\{z_1^{terpi}, \ldots, z_k^{terpi}\}$.
- Our null hypothesis is that the two underlying inhomogeneous Poisson processes are the same, therefore:

  $$z_i^{citron} = r(t_i) + \epsilon_i^{citron}\sigma \quad \text{and} \quad z_i^{terpi} = r(t_i) + \epsilon_i^{terpi}\sigma \,,$$

  then

  $$z_i^{terpi} - z_i^{citron} = \sqrt{2}\epsilon_i\sigma \,.$$

- We then want to test if our collection of observed differences $\{z_1^{terpi} - z_1^{citron}, \ldots, z_k^{terpi} - z_k^{citron}\}$ is compatible with k IID draws from $\mathcal{N}(0, 2\sigma^2)$.

Invariance principle / Donsker theorem

Theorem

If $X_1, X_2, \ldots$ is a sequence of IID random variables such that $\mathrm{E}(X_i) = 0$ and $\mathrm{E}(X_i^2) = 1$, then the sequence of processes:

$$S_k(t) = \frac{1}{\sqrt{k}} \sum_{i=0}^{\lfloor kt \rfloor} X_i, \quad 0 \leq t \leq 1, \quad X_0 = 0$$

converges in law towards a canonical Brownian motion.

Proof

You can find a proof in:

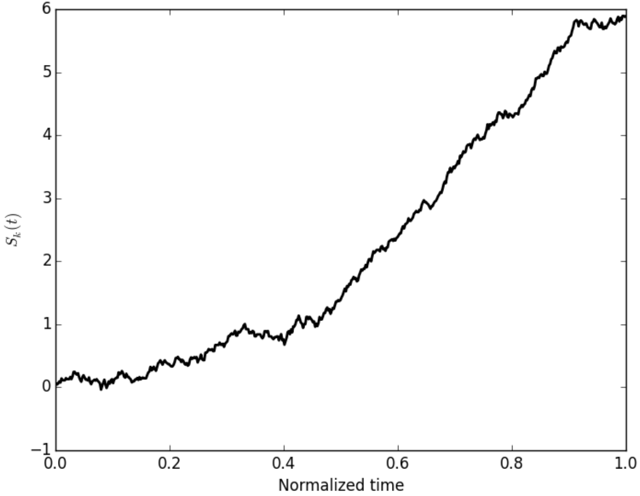
- R Durrett (2009) *Probability: Theory and Examples*. CUP. Sec. 7.6, pp 323-329 ;
- P Billingsley (1999) *Convergence of Probability Measures*. Wiley. p 121.

Recognizing a Brownian motion when we see one

- Under our null hypothesis (same inhomogeneous Poisson process for citronellal and terpineol), the random variables:

$$\frac{Z_i^{terpi} - Z_i^{citron}}{\sqrt{2}\sigma} \, ,$$

  should correspond to the X_i of Donsker's theorem.
- We can then construct $S_k(t)$ and check if the observed trajectory looks Brownian or not.
- Ideally, we would like to define a domain in $[0,1] \times \mathbb{R}$ containing the realizations of a canonical Brownian motion with a given probability.
- To have a reasonable power, we would like the surface of this domain to be minimal.

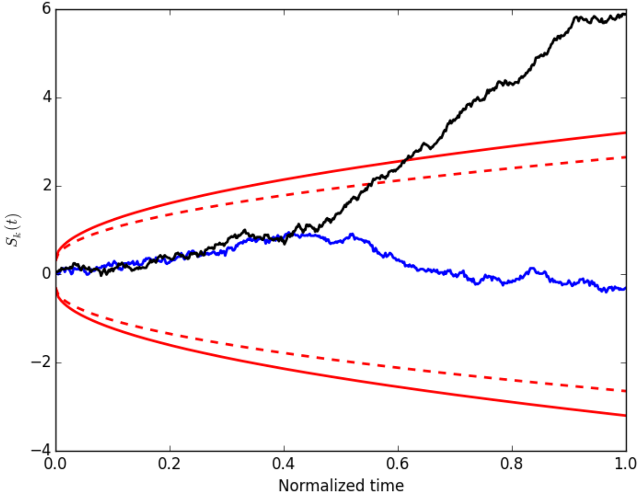


Does this look like the realization of a canonical Brownian motion?

- In a (non trivial) paper, Kendall, Marin et Robert (2007) showed that the upper boundary of this minimal surface domain is given by:

$$u^*(t) \equiv \sqrt{-W_{-1}\left(-(\kappa t)^2\right)}\,\sqrt{t}, \quad \text{for} \quad \kappa\, t \leq 1/\sqrt{e}$$

  where $\mathrm{W}_{\text{-}1}$ is the secondary real branch of the Lambert W function (defined as the solution of $W(z) \exp W(z) = z$); κ being adjusted to get the desired probability.
- They also showed that a domain whose upper boundary is given by: $u(t) = a + b\sqrt{t}$ is almost of minimal surface ($a > 0$ and $b > 0$ being adjusted to get the correct probability).
- Loader and Deely (1987) give a very efficient algorithm to adjust a and b or κ.
- The `R` package `STAR` (Spike Train Analysis with R) provides all that (and much more) out of the box.



Almost minimal surface domains with probabilities 0.95 (dashed red) and 0.99 (red) of containing an observed canonical Brownian motion. Black: terpineol - citronellal; blue: odd terpineol trials - even terpineol trials