Tests d'adéquation pour les processus de Poisson et les processus de Hawkes

Christine Malot, Patricia Reynaud-Bouret, Vincent Rivoirard, Franck Grammont

Tests d'adéquation pour les processus de Poisson et les processus de Hawkes

Christine Tuleau-Malot [1] & Patricia Reynaud-Bouret [1] & Vincent Rivoirard [2] & Franck Grammont [1]

[1] *Laboratoire J.-A. Dieudonné, UMR7351, UNS/CNRS, Université de Nice Sophia Antipolis, Parc Valrose, 06108 Nice Cedex 2, France. Email : malot@unice.fr, reynaudb@unice.fr, grammont@unice.fr*

[2] *CEREMADE UMR CNRS 7534, Université Paris Dauphine, Place du Maréchal De Lattre De Tassigny, 75775 PARIS Cedex 16, France. Email : rivoirard@ceremade.dauphine.fr*

Résumé. En neurosciences, le principal objet d'étude est le train de spike car il est considéré comme le vecteur principal de transmission de l'information de l'activité cérébrale. Au fil des différentes études, plusieurs modélisations pour les trains de spikes ont été proposées, plus pour des raisons biologiques que mathématiques. Nous proposons ici des procédures statistiques permettant de tester les diverses modélisations.

Keywords. Processus de Poisson, Processus de Hawkes, Tests.

Abstract. In neurosciences, the main studied object is the spike train since it is considered as the main vector for the transmission of information of the brain activity. Throughout the studies, several models for spikes trains have been proposed for biological reasons more than for mathematical ones. We propose several statistical procedures for testing the various models.

Keywords. Poissons processes, Hawkes Processes, Tests.

1 Introduction

En Neurosciences, un des intérêts majeurs est une meilleure compréhension de la dynamique du cerveau, et plus précisément de comment les réseaux de neurones transmettent l'information dans le cerveau. S'il existe différents aspects dans l'activité cérébrale, beaucoup de biologistes s'accordent à dire que le potentiel d'action (ou spike) est l'élément fondamental dans le traitement de l'information au niveau cérébral. Aujourd'hui, grâce aux avancées technologiques notamment en termes d'électrodes, il est possible d'enregistrer in vivo l'activité associée à différents neurones à savoir pour un neurone. En effet, on dipose de l'enregistrement simultané de plusieurs trains de spikes, chaque train étant associé à un neurone. En raison de la nature même de ces enregistrements, une modélisation par un processus ponctuel est évidente. Cependant, dans la plupart des études considérant de tels objets, des hypothèses supplémentaires ont dû être faites sur le modèle. Ainsi, le premier modèle considéré a été le Processus de Poisson homogène, à l'image de ce qui a été fait dans le livre de Dayan et Abbott (2005). Ce modèle très simple a été largement utilisé puisque pendant longtemps, les personnes pensaient que la seule information pertinente était le taux de décharge (firing rate) ou autrement dit l'intensité constante du processus. Cependant, une telle modélisation ne permet pas de prendre en considération d'éventuels effets de non-stationnarité ni d'interactions. Or, plusieurs études parmi lesquelles les articles de Grammont et Riehle (2003) et Grün (1996) montrent qu'omettre de tels phénomènes est difficilement concevable en toute généralité. Ainsi, afin de pallier les aspects non stationnaires, une modélisation par un processus de Poisson non-homogène a été envisagée. Cependant, dans la pratique, une telle modélisation n'a pas réellement été exploitée puisque généralement, des techniques de type bining ou multiple shift, à l'image de ce qui a été fait dans Grün (1996), étaient utilisées pour se ramener sur des zones de stationnarité. Pour ce qui est de prendre en considération les interactions possibles, qu'elles soient temporelles ou avec d'autres neurones, le processus de Hawkes a été introduit dans le cadre des neurosciences d'un point de vue théorique notamment par Brémaud et Massoulié (1996) et d'un point de vue pratique par Krumin et al. (2010). Pour autant, si ces différentes modélisations apparaissent déjà dans la littérature, il est difficile de trouver dans cette même littérature des outils statistiques permettant de tester les modélisations proposées ou d'estimer les intensités des processus mis en oeuvre. Nous montrons ici dans la théorie et dans la pratique comment tester les deux modélisations considérées. Ainsi, après cette introduction, nous allons considérer en Section 2 le test associé à un processus de Poisson tandis que la Section 3 sera consacrée au processus de Hawkes. La dernière partie quant à elle sera dévolue à la pratique.

2 Processus de Poisson inhomogène

Soit N un processus de Poisson inhomogène sur $[0,T]$ d'intensité $\lambda(.)$. Notons, pour tout intervalle I de $[0,T]$, $N(I)$ le nombre de points de N dans I. Une caractérisation du processus N est alors : pour tous les intervalles disjoints I et J de $[0,T]$, $N(I)$ et $N(J)$ sont indépendants et $N(I)$ est une variable de Poisson de paramètre $\int_I \lambda(x).dx$.
On suppose disposer de n réalisations indépendantes de N, notées respectivement $N^{(1)},\dots,N^{(n)}$. Voici une procédure permettant de tester si N est un procéssus de Poisson ou non :

1. On estime l'intensité cumulée $\Lambda(t) = \int_0^t \lambda(x)dx$ à partir des n réalisations par : $\hat{\Lambda}(t) = \int_0^t \hat{\lambda}(x)dx$ où $\hat{\lambda}$ est un estimateur de λ.

2. On tire un sous-échantillon $\mathcal{S}$ de $\{1,\dots,n\}$ de taille $m = n^{(2/3)}$.

3. Pour chaque élément i de $\mathcal{S}$, on considère $\hat{\mathcal{N}}^{(i)} = \{\hat{\Lambda}(U), U \in N^{(i)}\}$ qui est un processus ponctuel sur $[0,\hat{\Lambda}(T)]$. Cette étape revient simplement à changer de temps.

4. On calcule alors le processus cumulé associé noté $\hat{\mathcal{N}}^{c,\mathcal{S}}$ qui consiste à mettre bout à bout les différents processus $\hat{\mathcal{N}}^{(i)}$, $i \in \mathcal{S}$. Le processus cumulé est sur $[0, m.\hat{\Lambda}(T)]$.

5. Soit θ un réel positf plus petit que $\hat{\Lambda}(T)$, on calcule alors la statistique

$$Z^c := \sqrt{\hat{\mathcal{N}}^{c,\mathcal{S}}([0, m\theta])} \sup_{0 \leq u \leq 1} \left| \frac{1}{\hat{\mathcal{N}}^{c,\mathcal{S}}([0, m\theta])} \sum_{X \in \hat{\mathcal{N}}^{c,\mathcal{S}}} \mathbf{1}_{X \leq u.m.\theta} - u \right|.$$

6. On rejette si $Z^c > \tilde{k}_{1-\alpha}$ ou si $Z^c > \tilde{k}_{\alpha}$, avec $\tilde{k}_{1-\alpha}$ qui représente le quantile d'ordre $1-\alpha$ de la loi tabulée de Kolmogorov Smirnov.

Cette procédure repose sur une caractéristique fondamentale des processus de Poisson qui est que si N est un processus de Poisson inhomogène d'intensité $\lambda(.)$ sur $[0, T]$ (un processus de Poisson homogène étant un processus inhomogène particulier), alors la transformation temporelle $\mathcal{N} = \{\Lambda(U), U \in N\}$ permet de se ramener à un processus de Poisson homogène sur $[0, \Lambda(T)]$ d'intensité 1. De ce fait, lorsque l'intensité $\lambda(.)$ est connue, un simple test d'uniformité permet de conclure. Lorsque $\lambda(.)$ est inconnue, l'idée est alors de procéder à un plug-in. Cependant, contrairement au cadre habituel qui consiste à estimer λ puis à tester sur deux échantillons indépendants, ici tout l'échantillon sert à l'estimation de λ et un sous-échantillon est utilisé pour la phase de test.

Théorème 2.1
Le test précédemment défini est asymptotiquement de niveau α si $\mathbb{E}(\int_0^T (\hat{\lambda}(u) - \lambda(u))^2 du = O(n^{-\eta})$ avec $2/3 < \eta$ et si $\Lambda(T) > \theta$.

Pour obtenir un estimateur $\hat{\lambda}$ de λ, diverses méthodes existent. On peut par exemple citer les méthodes des noyaux qui permettent d'obtenir des estimateurs plus ou moins lisses selon le noyau considéré. Pour calibrer le choix de la fenêtre, on peut par exemple utiliser la méthode de Goldenshluger and Lepski (cf. Goldenshluger et Lepski (2011)). Parmi les autres procédures d'estimation, on peut citer les travaux de Reynaud-Bouret et al. (2011) qui permet d'obtenir un estimateur constant par morceaux ou encore la procédure de Willet et Nowak (cf. Willet et Nowak (2007)) qui produit un estimateur polynomial par morceaux avec une adaptation de la partition et du degré.

3 Processus de Hawkes

Soit $N_1, \ldots, N_p$ les trains de spikes associés à p neurones. Une modélisation qui permet de tenir compte d'une éventuelle interaction entre les p neurones est le Processus de Hawkes multivarié. Une façon de définir un tel processus est de donner son intensité conditionnelle qui représente le taux de décharge instantané. Une telle intensité conditionnelle est donnée par : $\forall i \in \{1,\dots,p\}, \forall t > 0, \lambda^{(i)}(t) = \left(\nu^{(i)} + \sum_{j=1}^{p} h_j^{(i)}(t-u)dN_j(u)\right)_+$, où $\nu^{(i)}$ représente le phénomène d'apparition spontanée sur le neurone i, et $h_j^{(i)}$ est la fonction d'interaction du neurone j sur le neurone i, fonction bornée de support $[0, A]$.

On souhaite, à l'image de ce qui a été fait ci-dessus, proposer une procédure permettant de tester si les données peuvent se modéliser via un processus de Hawkes. Il nous faut commencer par obtenir un estimateur de l'intensité conditionelle λ, à savoir, un estimateur de $f^* = ((\nu^{(j)})_{j=1\dots p}, (h_i^{(j)})_{i,j=1\dots p})$. On peut par exemple s'inspirer de travaux de Hansen et al. (2012) qui consistent à considérer un estimateur Lasso.

3.1 Estimateur Lasso

Soit $f = (((\mu^{(j)})_{j=1\dots p}, (g_i^{(j)})_{i,j=1\dots p})$, posons $\psi_t^{(j)}(f) = \mu^{(j)} + \sum_{l=1}^{p} g_l^{(j)}(t-u)dN_l(u)$. $\psi^{(j)}(f)$ sera un candidat estimateur de $\lambda^{(j)}$ si ces deux quantités sont proches.
Afin de pouvoir estimer f on va utiliser une procédure basée sur les moindres carré. On considère alors $(N_1^{(i)}, \dots, N_p^{(i)})_{i=1,\dots,n}$ n réalisations indépendantes de $(N_1,\dots,N_p)$, et une paramétrisation de $g_l^{(j)}$ par des fonctions constantes par morceaux, à savoir $g_l^{(j)} = \sum_{k=1}^{K} a_{j,l,k}\delta^{-1/2}1_{[(k-1)\delta,k\delta]}$, avec $\delta > 0$ la longueur des bins et $K\delta = A$. Du fait de la motivation neurobiologique, nous allons par ailleurs nous restreindre à des représentations parcimonieuses des fonctions d'interaction, ce qui conduit à non pas considérer le contraste des moindres carrés simple, mais une version pénalisée. Pour les neurones j et l et l'essai i, notons :

$$
\begin{aligned}
(\mathbf{a}^{(j)})' &= (\mu^{(j)}, a_{j,1,1}, \dots, a_{j,1,K}, a_{j,2,1}, \dots, a_{j,p,K}),\\
(\mathbf{c}_t^{(j,i)})' &= \left(N_j^{(i)}([t-\delta,t)), \dots, N_j^{(i)}([t-K\delta, t-(K-1)\delta))\right),\\
(\mathbf{Rc}_t^{(i)})' &= \left(1, \delta^{-1/2}(\mathbf{c}_t^{(1,i)})', \dots, \delta^{-1/2}(\mathbf{c}_t^{(p,i)})'\right),\\
\mathbf{n}_{j,l}^{(i)} &= \left(\int_0^T N_l^{(i)}([t-k\delta, t-(k-1)\delta])dN_j^{(i)}(t)\right)_{k=1,\dots,K}\\
(\mathbf{b}^{(j,i)})' &= \left(N_j^{(i)}]([0,T]), \delta^{-1/2}\mathbf{n}'_{j,1}, \dots, \delta^{-1/2}\mathbf{n}'_{j,p}\right)\\
\mathbf{G^{(i)}} &= \int_0^T \mathbf{Rc}_t^{(i)}(\mathbf{Rc}_t^{(i)})'dt
\end{aligned}
$$

et si l'on se réfère à l'article de Hansen et al. (2012), on peut montrer que les estimateurs minimisant le contraste pénalisé sont :

$$
\tilde{a}^{(j)} \in \underset{\mathbf{a}^{(j)}}{argmin} \left(-2(\mathbf{a}^{(j)})'\left(\sum_{i=1}^{n}\mathbf{b}^{(j,i)}\right) + (\mathbf{a}^{(j)})'\left(\sum_{i=1}^{n}\mathbf{G}^{(i)}\right)\mathbf{a}^{(j)} + 2(\mathbf{d}^{(j)})'|\mathbf{a}^{(j)}|\right)
$$

où $d^{(j)}$ est un vecteur de poids spécifique

3.2 Procédure de test avec hypotèse de Hawkes

A présent que l'on dispose d'un estimateur de λ, on peut définir la procédure de test suivante qui permet de tester si (N_j) est un processus de Hawkes

1. On tire un sous-échantillon $\mathcal{S}$ de $\{1,\ldots,n\}$ de taille $m = n^{(2/3)}$.

2. Pour chaque élément i de $\mathcal{S}$, on considère $\hat{\Lambda}^{(i)}(t) = \int_0^t (\mathbf{Rc}_t^{(i)})'\tilde{\mathbf{a}}^{(j)}dt$ et $\hat{\mathcal{N}}^{(i)} = \{\hat{\Lambda}^{(i)}(U), U \in N^{(i)}\}$ qui est un processus ponctuel sur $[0, \hat{\Lambda}^{(i)}(T)]$.

3. On calcule alors le processus cumulé associé noté $\hat{\mathcal{N}}^{c,\mathcal{S}}$. Le processus cumulé est sur $[0, \sum_{i\in\mathcal{S}} \hat{\Lambda}^{(i)}(T)]$.

4. Soit θ un réel plus petit que $\sum_{i\in\mathcal{S}} \hat{\Lambda}^{(i)}(T)/m$, on calcule alors la statistique

$$Z^c := \sqrt{\hat{\mathcal{N}}^{c,\mathcal{S}}([0,m\theta])} \sup_{0\le u\le 1} \left| \frac{1}{\hat{\mathcal{N}}^{c,\mathcal{S}}([0,m\theta])} \sum_{X\in\hat{\mathcal{N}}^{c,\mathcal{S}}} \mathbf{1}_{X\le u.m.\theta} - u \right|.$$

5. On rejette si $Z^c > \tilde{k}_{1-\alpha}$ ou si $Z^c > \tilde{k}_\alpha$

Théorème 3.1
Le test précédemment défini est asymptotiquement de niveau α sous certaines hypothèses liées à θ et à l'estimation.

4 Application

Afin d'illustrer ces deux procédures, nous les avons appliquées sur des données simulées et réelles. Pour ce qui est des données simulées, nous avons simulé, sur $[0, 2]$, 40 réalisations d'un processus de Poisson homogène d'intensité 20 ainsi que 40 réalisations d'un processus inhomogène d'intensité continue par morceaux. Pour ce qui est des données réelles nous avons considéré des enregistrements neuronaux mesurés chez le singe lors de la réalisation d'une tâche sensorimotrice. Pour chacun des deux enregistrements considérés, nous disposons d'une trentaine de réalisations. Plus de détails sur ces données sont disponibles dans l'article de Grammont et Riehle (2003).

En regardant les tables, on constate que les tests semblent permettent, pratiquement, de distinguer les deux modélisations. En effet, le test sous hypothèse poissonienne est cohérent avec les simulations, et le second test l'est tout autant. Pour ce qui est des données réelles, on constate qu'une même modélisation n'est pas possible pour les deux enregistrements considérés. En effet, pour le premier enregistrement, une modélisation par un processus de Hawkes semble envisageable tandis que pour le second enregistrement,

Poisson homogène	0.15	Poisson inhomogène	0.11
Hawkes N_1	10^{-6}	Hawkes N_2	10^{-6}
données réelles 1	10^{-16}	données réelles 2	0.15

Figure 1: application de la procédure de test Poisson pour des données simulées et réelles.

Hawkes N_1	0.08	Hawkes N_2	0.59
données réelles 1	0.21	données réelles 2	0.0025

Figure 2: application de la procédure de test Hawkes pour des données simulées et réelles.

il s'agirait plutôt d'une modélisation poissonienne.
Par conséquent, une hypothèse a priori unique sur les trains de spikes semble difficile à concevoir. D'autres modèles encore plus ambitieux pourraient être considérés de façon à certes prendre en considération les interactions, mais aussi des effets de non-stationnarité.

Bibliographie

[1] Dayan, P. et Abbott, LF. (2005), *Theoretical Neuroscience Computational and Mathematical Modeling of Neural Systems*, MIT Press.
[6] Grammont, F. et Riehle, A. (2003), Spike synchronisation and firing rate in a population of motor cortical neurons in relation to movement direction and reaction time, *Biological Cybernetics*, 88, 360–373.
[5] Grün, S. (1996), Unitary joint-events in multiple-neuron spiking activity: Detection, significance and interpretation, *PhD thesis*, Thun; Verlag Harri Deutsch.
[2] Brémaud, P. et Massoulié, L. (1996), Stability of nonlinear Hawkes processes, *Ann. Probab.*, 24(3), 1563–1588.
[7] Krumin, M. et Reutsky, I. et Shoham, S. (2010), Correlation-based analysis and generation of multiple spike trains using Haxkes models with exogenous input, *Frontiers in Computational Neuroscience*, 4, article 147.
[8] Goldenshluger, A. et Lepski, O. (2011), Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality, *Ann. Statist.*, 39(3), 1608–1632.
[9] Reynaud-Bouret, P. et Rivoirard, V. et Tuleau-Malot, C. (2011), Adaptive density estimation: a curse of support?, *J. Statist. Plann. Inference*, 141, 115–139.
[4] Willet, R. et Nowak, RD. (2007), Multiscale Poisson intensity and density estimation, *IEEE Trans. Inform. Theory*, 53(9), 115–139.
[3] Hansen, N. et Reynaud-Bouret, P. et Rivoirard, V. (2012), Lasso and probabilistic inequalities for multivariate point processes, *Tech. rep., http://arxiv.org/abs/1208.0570.*