

Année scolaire: 2021-2022

Processus de Poisson comme outils dans la compréhension de la connectivité fonctionnelle au sein du cerveau

Sahi Gonsangbeu & Yassine Obeid

Janvier 2022

MAM4

Table des matières

Introduction	3
I. Première partie	3
1. Finalité poursuivie dans l'article intitulé MTGAUE	3
2. Explication des grands outils utilisés	3
3. La notion de coïncidence par "Multiple shift"	4
3.1. Train de spikes	4
3.2. Notion de coïncidence par multiple shift (cas symétrique)	4
4. Tests multiples - Procédure de Benjamini-Höcherberg	4
4.1. C'est quoi un test statistique?	4
4.2. Risque de type I	5
4.3. p-value	5
4.4 Tests multiples	5
4.5 La procédure de Benjamini-Höcherberg	6
5. Processus de Poisson homogènes	6
5.1. Définition préliminaires	6
5.2. Processus de Poisson homogènes	7
5.3. Définition du processus de Poisson à l'aide des temps inter-arrivées	8
5.4. Somme de deux processus de Poisson indépendants	8
II. Deuxième partie	8
1. Simulation de processus de Poisson homogènes	8
2. Test pour savoir si des données peuvent être modélisées par des processus de Poisson homogènes	10
2.1. Principe du test	11
2.2. Mise en oeuvre du test sur nos données	11
Conclusion	12
Références	12

Introduction

L'un des enjeux majeurs en neurosciences est d'acquérir une meilleure compréhension de la dynamique du cerveau et notamment d'identifier les connectivités fonctionnelles pouvant exister entre neurones au cours du temps. Dans cette quête, il est très fréquent de modéliser l'activité des neurones par des processus de Poisson.

Les processus de Poisson sont des processus stochastiques ponctuels, les plus simples à étudier. Ils permettent de modéliser types de phénomènes aléatoires que ce soit dans le domaine de la physique nucléaire, de la biologie cellulaire ou encore dans des situations de la vie courante, comme l'étude du congestionnement d'une centrale téléphonique.

Les événements modélisés peuvent être temporels comme dans le cas de l'activité des neurones ou bien spatiaux comme la position des gènes sur une chaîne d'ADN.

Dans ce projet, après avoir expliqué brièvement la finalité et les grands outils de l'article MTGAUE, nous définissons ce qu'est un processus de Poisson homogène (on ne se concentre que sur ce type de processus de Poisson par opposition avec les processus de Poisson inhomogènes), en simulons avec le logiciel R et testons si des données peuvent être modélisées au moyen de ses processus.

I. Première partie

1. Finalité poursuivie dans l'article intitulé MTGAUE

Dans cet article, le but est de proposer (en supposant que l'activité des neurones se fait au moyen de processus de Poisson homogène) une méthodologie permettant de tester l'indépendance entre deux neurones, localement dans le temps.

2. Explication des grands outils utilisés

Les outils les plus importants que nous avons identifiés dans l'article sont:

- La notion de coïncidence par "Multiple shift"
- Tests d'hypothèses, Tests multiples (procédure de Benjamini et Hochberg), False Discovery Rate (FDR)
- Processus de Poisson homogène

3. La notion de coincidence par “Multiple shift”

3.1. Train de spikes

On note h résolution d’acquisition des données (de l’ordre de $10^{-3}, 10^{-4}$)

Un train de spikes est une séquence de 0 et 1 notée $(H_n)_n$ où $H_i = 1$ traduit la présence d’un spike sur l’intervalle $\left[ih - \frac{h}{2}, ih + \frac{h}{2}\right]$

A une telle séquence, il est possible d’associer un processus ponctuel N défini comme étant l’ensemble des ponts ih pour lesquels $H_i = 1$.

On note N_1 et N_2 les processus ponctuels, et $(H_n^1)_n$ et $(H_n^2)_n$ les séquences, associés respectivement au train de spikes du neurone 1 et du neurone 2 enregistrés simultanément.

On se limite à une fenêtre temporelle W de longueur $T = nh$.

3.2. Notion de coincidence par multiple shift (cas symétrique)

Une coincidence est observée au temps ih sur la fenêtre W s’il existe un “shift” j dans $\{-d, \dots, d\}$ où d est un entier supérieur ou égal à 1, tel que $H_i^1 = H_{i+j}^2 = 1$

$$1 \leq i + j \leq n$$

Ainsi le nombre de coïncidences est donné par :

$$X = \sum_{i=1}^n \sum_{|j| \leq d, 1 \leq i+j \leq n} \mathbb{1}_{H_i^1 = H_{i+j}^2 = 1}$$

et donc par un changement d’indice $k = i + j$

on obtient:

$$X = \sum_{i=1}^n \sum_{k=1}^n \mathbb{1}_{|k-i| \leq d} \mathbb{1}_{H_i^1 = 1} \mathbb{1}_{H_k^2 = 1}$$

4. Tests multiples - Procédure de Benjamini-Höcherberg

Avant de parler de tests multiples, intéressons nous brièvement à la théorie des tests.

4.1. C’est quoi un test statistique?

Un test statistique ou test d’hypothèse, est une procédure de décision entre deux hypothèses. Il s’agit d’une démarche consistant à rejeter ou ne pas rejeter une hypothèse, appelée hypothèse nulle en fonction d’un échantillon de données.

L’hypothèse nulle souvent notée H_0 est celle que l’on considère à priori vraie. Le but est de décider si cette hypothèse est à priori crédible. L’hypothèse alternative notée H_1 est l’hypothèse complémentaire à l’hypothèse nulle.

4.2. Risque de type I

C'est la probabilité α de rejeter H_0 à tort (alors que celle-ci est vraie). Cette erreur est souvent appelée risque α .

La valeur seuil de α communément admise pour rejeter H_0 est de 0.05 (5%). Cette valeur de 5% est arbitraire.

Il existe aussi le risque de type II qui correspond à la probabilité de ne pas rejeter H_0 alors que celle-ci est fautive (et donc H_1 est vraie). cette probabilité est souvent notée β

4.3. p-value

La p-value souvent notée p est la probabilité pour un modèle statistique donnée sous l'hypothèse nulle H_0 , d'obtenir la même valeur ou une valeur encore plus extrême que celle observée.

Elle s'interprète aussi comme la probabilité d'un résultat au moins aussi extrême que le résultat observé, sachant l'hypothèse nulle H_0

Le grand principe d'un test est donc de calculer la probabilité de rejeter l'hypothèse nulle H_0 , tout en minimisant les chances de se tromper. Ainsi, quand nous comparons la p-valeur, associée à une statistique de test, à la loi de probabilité qu'elle suit et que nous rejetons H_0 au risque de 5%, dans la forme, nous prenons une décision mais, sur le fond, nous prenons également le risque de nous tromper avec une probabilité de 5%.

Imaginons désormais que pour une longue série de variables nous appliquons le même test au risque α de 5%. Il paraît évident que nous augmentons les chances de conclure à tort au rejet de H_0 au moins pour un certain nombre de variables testées.

C'est là que la théorie des tests multiples (ou comparaison multiples) intervient.

4.4 Tests multiples

Le grand principe des tests multiples est donc de corriger la liste des p-valeurs établies afin de rattraper les hypothèses H_0 rejetées à tort.

Définissons K le nombre d'hypothèses nulles H_0^1, \dots, H_0^K que nous pourrions également assimiler au test de l'hypothèse H_0 pour K variables X^1, \dots, X^K .

Enfin, posons R le nombre d'hypothèses nulles rejetées et m_0 le nombre d'hypothèses nulles vraies.

$K - m_0$ représente donc le nombre d'hypothèses alternatives vraies.

Il existe deux grandes familles de tests multiples qui sont basées sur les deux taux d'erreur de type I. Ce sont:

- Family Wise Error Rate : $FWER = \mathbb{P}(FP > 0) = 1 - \mathbb{P}(FP = 0)$: Estimation de la probabilité d'avoir au moins un faux positif
- False Discovery Rate : $FDR = \mathbb{E}\left(\frac{FP}{R} | R > 0\right)$: Taux moyen de faux positifs

NB: FP représente le nombre de faux positifs (le nombre de fois où on rejette H_0 alors que celle-ci est vraie)

De plus, le taux d'erreur de type II est donnée par la Non Discovery Rate : $NDR = \mathbb{E} \left(\frac{FN}{K-R} \right)$, FN : nombre de faux négatifs (le nombre de fois où on ne rejette pas H_0 alors que celle-ci est fautive)

4.5 La procédure de Benjamini-Höcherberg

La procédure de Benjamini-Höcherberg a été inventée par Yoav Benjamini et Yosef Höcherberg en 1995. Elle porte également le nom de méthode linéaire step-up de Benjamini-Höcherberg.

Principe: Rejet de H_0^1, \dots, H_0^k pour le plus grand k tel que $p_k \leq \frac{k}{K} \cdot \alpha \Rightarrow \alpha \geq \frac{K}{k} \cdot p_k$.

Où p_k représente la p-valeur associée à la statistique de test calculée. Enfin, notons α le seuil de significativité fixé au préalable.

Algorithme associé:

- Ranger les p par ordre croissant
- $p_K^* = p_K$
- $\forall k \in [K-1, \dots, 1], p_k^* = \min \left(p_{k+1}^*, \frac{K}{k} \cdot p_k \right)$
- $p^* = [p_1^*, \dots, p_K^*]$

5. Processus de Poisson homogènes

5.1. Définition préliminaires

Définition

Un processus stochastique est une fonction aléatoire $t \rightarrow X_t$.

Définition

Désignons par $N(t)$ le nombre d'événements se produisant dans l'intervalle de temps $[0, t]$, et supposons que $N(0) = 0$. Le processus $\{N(t); t \geq 0\}$, est appelé processus de comptage et vérifie :

- $\forall t \geq 0, N(t) \in \mathbb{N}$;
- $t \rightarrow N(t)$ est croissante;
- $\forall 0 < a < b, N(b) - N(a)$ représente le nombre d'événement se produisant dans l'intervalle de temps $]a, b]$.

Propriété

Soit $(N(t))_{t \geq 0}$ un processus de comptage.

On a :

- Si $a < b, N(a) \leq N(b)$

Définition

Un processus de comptage est dit à accroissements stationnaires si la loi de probabilité du nombre d'événements se produisant dans un intervalle de temps donné ne dépend que de la longueur de celui-ci.

Définition

Un processus de Poisson est dit à accroissements indépendants si les nombres d'événements se produisant dans des intervalles de temps disjoints sont indépendants.

5.2. Processus de Poisson homogènes

Définition 1

Un processus de comptage $\{N(t), t \geq 0\}$ est appelé processus de Poisson d'intensité $\lambda > 0$ si:

- $N(0) = 0$
- Le processus est à accroissements indépendants ie : $\forall 0 \leq t_1 \leq t_2 \cdots < t_n$, les variables aléatoires $N(t_i) - N(t_{i-1}), i = 1, \dots, n$ sont globalement indépendantes;
- Le nombre de tops se produisant dans un intervalle de temps de longueur $t \geq 0$ suit une loi de Poisson de paramètre λt , ie: $\forall h \geq 0, \forall t \geq 0, \forall n \in \mathbb{N}, \mathbb{P}(N(t+h) - N(t) = n) = \exp(-\lambda t) \frac{(\lambda t)^n}{n!}$

Définition 2

Un processus de comptage $\{N(t), t \geq 0\}$ est appelé processus de Poisson d'intensité $\lambda > 0$ si:

- $N(0) = 0$;
- Le processus est à accroissements indépendants et stationnaires (ie. $\forall t, h > 0$, la loi de $N(t+h) - N(t)$ ne dépend que de h);
- $P(N(h) = 1) = \lambda h + o(h)$ pour $h \rightarrow 0$
- $P(N(h) \geq 2) = o(h)$ pour $h \rightarrow 0$

Ces deux définitions sont équivalentes.

Remarque:

- Le processus de Poisson est un processus de Markov car comme les accroissements sont indépendants, la valeur du processus après l'instant t ne dépend que de la valeur du processus à l'instant t mais pas de tout ce qui s'est passé avant.
- $N(t)$ suit la loi de Poisson de paramètre λt pour $t > 0$
- L'homogénéité d'un processus de Poisson vient du fait que son intensité λ soit constante (ne dépend pas de t)

Considérons un processus de Poisson $(N(t))_t$. Le premier saut arrive à un instant aléatoire T_1 , puis il faut attendre un temps T_2 avant le second saut, puis un temps T_3 ainsi de suite ... On note T_n le temps écoulé entre le $(n-1)$ -ième saut et le n -ième saut.

La suite des instants $(T_n)_{n \geq 1}$ est appelée suite des instants inter-arrivées.

Proposition

La suite des instants inter-arrivées $(T_n)_{n \geq 1}$ est une suite de variables aléatoires iid de loi exponentielle de paramètre λ .

5.3. Définition du processus de Poisson à l'aide des temps inter-arrivées

On considère $(T_n)_{n \geq 1}$ une suite de variables aléatoires iid de loi exponentielle de paramètre λ . On définit $S_0 = 0$ et on pose

$$S_n = \sum_{i=1}^n T_i$$

Alors le processus $(N(t))_{t \geq 0}$ définit par

$$N(t) = \sum_{n=1}^{\infty} \mathbb{1}_{S_n \leq t} = \max\{n \geq 0 : S_n \leq t\}$$

est un processus de Poisson d'intensité λ .

5.4. Somme de deux processus de Poisson indépendants

Considérons deux processus de Poisson $(N_1(t))_{t \geq 0}$ et $(N_2(t))_{t \geq 0}$ indépendants d'intensité respective λ_1 et λ_2 . Alors le processus $N(t) = N_1(t) + N_2(t)$ est un processus de Poisson d'intensité $\lambda_1 + \lambda_2$.

La probabilité que le premier saute avant le second est égale à

$$P(T_1^1 < T_1^2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

T_1^1 est le premier instant de saut de N_1 et T_1^2 est celui de N_2

II. Deuxième partie

1. Simulation de processus de Poisson homogènes

Pour simuler un processus de Poisson homogène de paramètre λ , on génère un vecteur de variables aléatoires indépendantes de loi exponentielle de paramètre λ . Il s'agit des vecteurs des inter-arrivées $(T_n)_{n \geq 1}$ puis on génère le vecteur des occurrences $(t_k)_{k \geq 1}$ par la formule :

$$\begin{aligned} t_1 &= T_1 \\ t_k &= t_{k-1} + T_k, \forall k > 1 \end{aligned}$$

```
# Simulation en fonction du nombre d'évènements
simul_poisson <- function(n, lambda){
  T <- cumsum(rexp(n, lambda))
  t <- c(0, T)
```



```

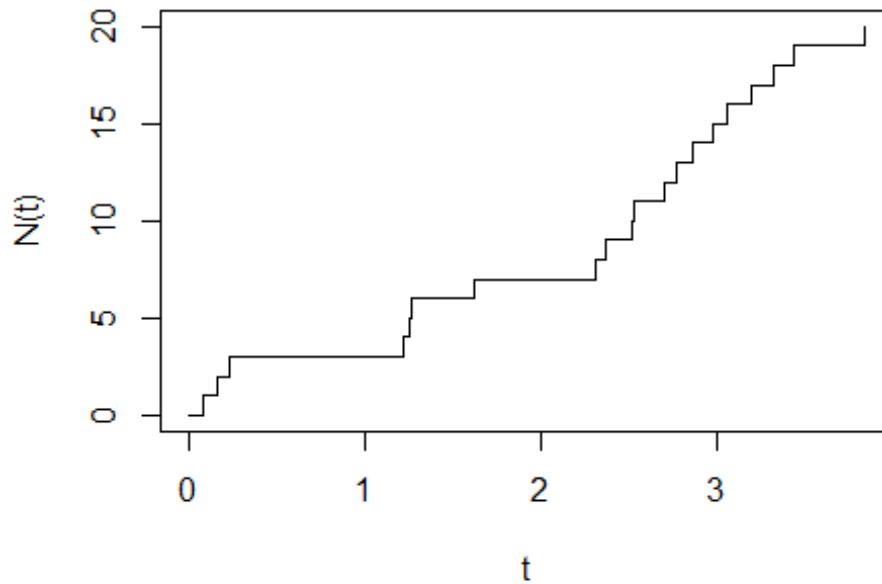
N <- c(0:n)
plot(t,N,type="s",ylab = "N(t)")
return(t)
}

```

```

# simulation d'un processus de Poisson de paramètre 5 et de taille 20
simul_poisson(20,5)

```



```

## [1] 0.00000000 0.08604891 0.16999046 0.23664134 1.21913778 1.25357475
## [7] 1.27031549 1.62405701 2.31185831 2.37060997 2.51404212 2.52286330
## [13] 2.70114284 2.77006781 2.86051854 2.96959306 3.05892606 3.18617895
## [19] 3.31335978 3.43647993 3.83174251

```

```

# Simulation jusqu'à une date T

```

```

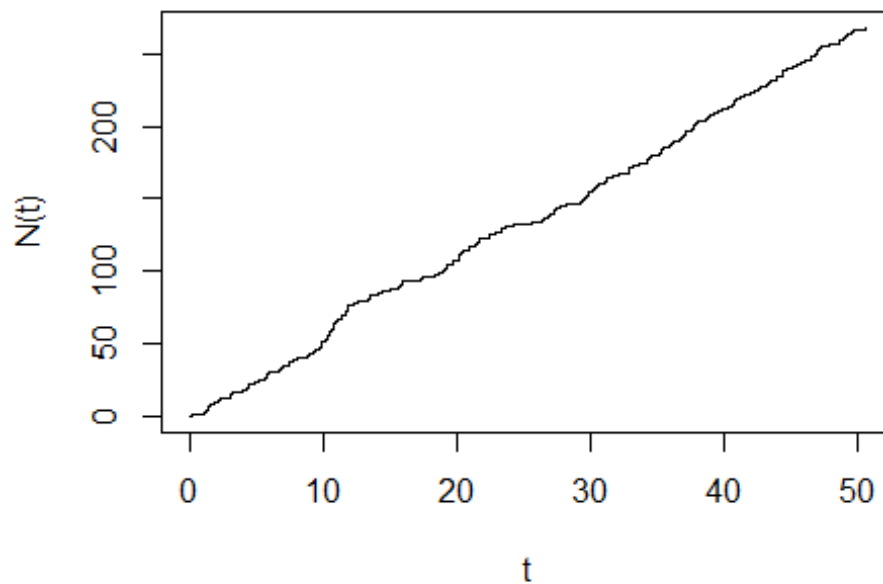
simul_poisson2 <- function(T,lambda){
  i <- 2
  t <- 0
  while(t[length(t)] < T){
    w <- rexp(1,lambda)
    t[i] <- t[i-1] + w
    i <- i + 1
  }
  return(t)
}

```

```

lambda = 5
T = 50
t <- simul_poisson2(T,lambda)
n <- length(t) - 1
N <- c(0:n)
plot(t,N,type="s", ylab = "N(t)")

```



2. Test pour savoir si des données peuvent être modélisées par des processus de Poisson homogènes

Notre but ici est de savoir si oui ou non des données peuvent être modélisées par des processus de Poisson homogènes. On va donc tester si les écarts entre les données sont indépendants et suivent une loi exponentielle de paramètre λ .

Pour cela, nous utilisons le test de Kolmogorov-Smirnov qui a pour but de vérifier si un échantillon suit une loi donnée connue par sa fonction de répartition continue ou si deux échantillons suivent la même loi.

Ce test repose sur les propriétés des fonctions de répartition empiriques.

Soit (X_1, \dots, X_n) un échantillon de variables aléatoires de même loi et donc de même fonction de répartition. On appelle fonction de répartition empirique la quantité suivante :

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq t\}}$$

Si F est continue alors $D(F, F_n) = \sup_{t \in \mathbb{R}} |F(t) - F_n(t)|$ ne dépend pas de F .

2.1. Principe du test

On teste l'hypothèse nulle $H_0 : "F = F_0"$ contre l'hypothèse alternative " $F \neq F_0$ " au niveau α .

On accepte H_0 si :

$$\forall t \in \mathbb{R}, |F_n(t) - F_0(t)| \leq \xi_{n,1-\alpha}$$

où $\xi_{n,1-\alpha}$ est le quantile d'ordre $1 - \alpha$ extrait de la table de Kolmogorov-Smirnov.

2.2. Mise en oeuvre du test sur nos données

Nous utilisons ce test pour voir si nos données peuvent être modélisées par un processus de Poisson.

comme décrit plus haut, on teste si les écarts entre les valeurs de nos données sont indépendantes et suivent une même loi exponentielle de paramètre λ .

Vu qu'on ne connaît pas λ , on va l'approximer par la moyenne (puisque on sait que la moyenne d'une loi exponentielle de paramètre λ est $\frac{1}{\lambda}$)

```
# Importation de nos données
data1 <- read.table("donnees/data1.txt")
data2 <- read.table("donnees/data2.txt")
data3 <- read.table("donnees/data3.txt")

data <- rbind(data1, data2)

# Concaténation de toutes les données ensemble
data <- rbind(data, data3)

data <- data$V1

# Calcul des écarts
t <- diff(data)

lambda_t <- 1./mean(t)

# test de Kolmogorov-Smirnov
ks.test(t, "pexp", lambda_t) # ici on teste si les écarts entre nos données s
uivent bien une loi exponentielle de paramètre lambda_t

##
## One-sample Kolmogorov-Smirnov test
##
## data:  t
## D = 0.37095, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

On obtient une p-value négligeable ce qui nous permet de rejeter l'hypothèse nulle H_0 . Du coup, les écarts entre nos données ne suivent pas une loi exponentielle. Donc nos données ne suivent pas un processus de Poisson homogène.

Conclusion

En conclusion, nous avons pu constater que les processus de Poisson représentent un outil très important dans la modélisation de nombreux phénomènes aléatoires.

Ce projet nous a permis de voir comment cet outil s'utilisait dans le domaine des neurosciences. Nous avons donné quelques caractéristiques des processus de Poisson homogène et en avons simulé. Nous avons aussi pu mettre en oeuvre le test de Kolmogorov-Smirnov sur des données pour voir si elles suivent un processus de Poisson ce qui nous a permis de nous initier à la réalisation d'un test d'hypothèse.

Références

Dominique FOATA, Aimé FUCHS, *Processus stochastiques*, DUNOD, 2002.

https://fr.wikipedia.org/wiki/Test_statistique

https://fr.wikipedia.org/wiki/Processus_de_Poisson

https://fr.wikipedia.org/wiki/Test_de_Kolmogorov-Smirnov

Hélène Guérin, *Notes de cours MAT-3071 Processus stochastiques*, Université du Québec Montréal (UQAM)

C. Tuleau-Malot, A. Rouis, F. Grammont, P. Reynaud-Bouret, *Multiple Tests based on a Gaussian Approximation of the Unitary Events method with delayed coincidence count*, <http://hal.archives-ouvertes.fr/hal-00757323>.