Tests d'indépendance entre processus ponctuels, et application en neurosciences

Mélisande Albert [1] & Yann Bouret [2] & Magalie Fromont[3] & Patricia Reynaud-Bouret[1]

[1] *Univ. Nice Sophia Antipolis, CNRS, LJAD, UMR 7351, 06100 Nice, France.*
Emails : Melisande.Albert@unice.fr et Patricia.Reynaud-Bouret@unice.fr.
[2] *Univ. Nice Sophia Antipolis, CNRS, LPMC, UMR 7336, 06100 Nice, France.*
Email : Yann.Bouret@unice.fr.
[3] *Univ. Européene de Bretagne, CNRS, IRMAR, UMR 6625, 35042 Rennes, France.*
Email : magalie.fromont@univ-rennes2.fr

Résumé. Considérant un échantillon de couples i.i.d. de processus ponctuels observés sur $[0, 1]$, on se pose la question de la détection de dépendances entre les processus marginaux sous-jacents. Cette question est motivée par l'étude des synchronisations entre trains de spikes en neurosciences. Nous proposons des tests d'indépendance non-paramétriques entre deux processus ponctuels, basés sur des méthodes de bootstrap et de permutation.

Mots-clés. Test d'indépendance, U-statistiques, processus ponctuels, bootstrap, randomisation, permutation, neurosciences, analyse de trains de spikes.

Abstract. Considering an i.i.d. sample from the joint distribution of a pair of point processes observed on $[0, 1]$, we address the question of detecting dependences between the two underlying marginal point processes. This question is motivated by the study of the synchrony phenomenon between spike trains in neuroscience. We propose distribution-free tests of independence between two point processes, that are based on bootstrap and permutation approaches.

Keywords. Independence test, U-statistics, point processes, bootstrap, randomization, permutation, neuroscience, spike train analysis.

1 Introduction

Savoir détecter les dépendances est un point fondamental en Neurosciences. Récemment, les neurobiologistes ont découvert que les neurones communiquent entre eux, non seulement en émettant un grand nombre de potentiels d'action (aussi appelés spikes) par unité de temps, mais aussi en se synchronisant. Ainsi, déterminer les éventuelles connectivités fonctionnelles (interactions entre les neurones) permet une meilleure compréhension de la transmission du message nerveux. Ayant accès aux temps d'occurrence des potentiels d'actions (aussi appelés trains de spikes) de plusieurs neurones enregistrés simultanément, la question qu'on se pose est donc de déterminer si ces trains de spikes sont indépendants ou non. En raison du degré de résolution, les enregistrements des trains de spikes sont discretisés dans le temps, mais leur dimension est si grande qu'il n'est ni réaliste, ni raisonnable d'appliquer des tests d'indépendance classiques entre variables ou vecteurs aléatoires réels.

Plusieurs méthodes ont été proposées pour détecter le phénomène de synchronisation entre trains de spikes en neurosciences. Parmi elles, la méthode des Unitary Events de Grün et al. (2010) a été très largement appliquée cette dernière décennie. Elle est basée sur une méthode dite de *binning* qui permet une réduction de la dimension des données, mais cela au prix d'une perte d'information (jusqu'à 60% dans certains cas). Il est donc préférable de modéliser les trains de spikes par des processus ponctuels, chaque point correspondant à un spike lui-même et d'utiliser des tests d'indépendance spécifiquement dédiés à de tels processus. Des tests asymptotiques entre processus ponctuels ont déjà été introduits dans la littérature neuroscientifique, notamment celui de Tuleau-Malot et al. (2014), mais dans le cas particulier de processus de Poisson homogènes. Un tel cadre paramétrique est nécessairement restrictif, d'autant qu'aucun modèle, et encore moins le modèle poissonnien, n'est à ce jour reconnu en neurosciences. Nous nous focalisons donc sur la construction de tests d'indépendance non-paramétriques. Dans cet esprit, des méthodes dites de *trial shuffling*, basées sur des approches par bootstrap, ont été introduites par Pipa et Grün (2003), mais pour des données binnées. Nous construisons ici des tests d'indépendance non-paramétriques entre deux processus ponctuels, à partir de l'observation de n copies indépendantes de ces processus, nécessitant un minimum d'hypothèses sur leur distribution.

Considérons l'ensemble des processus ponctuels finis à valeurs dans $[0,1]$ et définis sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$. Notons $\mathcal{X}$ l'ensemble de leurs valeurs possibles, à savoir l'ensemble des parties dénombrables de $[0,1]$, que l'on munit d'une distance $d_{\mathcal{X}}$ issue de la distance de Skorokhod, rendant l'espace $\mathcal{X}$ séparable. On peut donc considérer la tribu borélienne de $\mathcal{X}$ et donc par extension, la tribu borélienne de $\mathcal{X}^2$. Ainsi, on dira qu'un couple de processus ponctuels finis $X = (X^1, X^2)$ défini sur $(\Omega, \mathcal{A}, \mathbb{P})$ a pour loi jointe P, de marginales P^1 et P^2, si pour tout borélien $\mathcal{B}$ de $\mathcal{X}^2$, et tous boréliens $\mathcal{B}^1$ et $\mathcal{B}^2$ de $\mathcal{X}$, $P(\mathcal{B}) = \mathbb{P}(X \in \mathcal{B})$, $P^1(\mathcal{B}^1) = \mathbb{P}(X^1 \in \mathcal{B}^1)$ et $P^2(\mathcal{B}^2) = \mathbb{P}(X^2 \in \mathcal{B}^2)$. Dans toute la suite, on se donne un n-échantillon $\mathbb{X}_n = (X_1, \dots, X_n)$, de loi P sur $\mathcal{X}^2$, de marginales P^1 et P^2. On cherche ici à tester

$$(H_0)\ P = P^1 \otimes P^2 \quad \text{contre} \quad (H_1)\ P \neq P^1 \otimes P^2.$$

2 Des Neurosciences à notre statistique de test

Les neurobiologistes cherchent à détecter des dépendances bien particulières entre neurones correspondant à des synchronisations dans le temps, appelées coïncidences. Plus précisément, ils cherchent à savoir si ces coïncidences apparaissent de façon significative.

La notion de nombre de coïncidences entre deux processus ponctuels X^1 et X^2, avec délai δ, introduit par Malot-Tuleau et al. (2014) est définie par

$$\varphi_\delta^{coinc}(X^1, X^2) = \int_0^1 \mathbf{1}_{|u-v|\leq\delta} dN_{X^1}(u) dN_{X^2}(v) = \sum_{u\in X^1,\ v\in X^2} \mathbf{1}_{|u-v|\leq\delta}.$$

Dans le cadre poissonnien de Tuleau-Malot et al. (2014), l'espérance de $\varphi_\delta^{coinc}(X^1, X^2)$ s'exprime facilement en fonction de δ et des intensités des processus. Ainsi, en estimant ces dernières, et en appliquant le principe du plug-in, on peut aisément construire une statistique de test asymptotiquement centrée sous (H_0). Dans le cadre non-paramétrique considéré ici, cette étape de recentrage n'est pas applicable. Nous proposons donc une nouvelle statistique, basée sur une astuce particulière de recentrage, définie par

$$\frac{1}{n(n-1)} \sum_{i\neq i'\in\{1,2,\ldots,n\}} (\varphi_\delta^{coinc}(X_i^1, X_i^2) - \varphi_\delta^{coinc}(X_i^1, X_{i'}^2)). \tag{1}$$

Remarquons d'une part que cette statistique, construite principalement pour les neurobiologistes, peut être généralisée pour détecter d'autres types de dépendances en remplaçant φ_δ^{coinc} par n'importe quelle fonction φ intégrable. D'autre part, remarquons que, dans ce cas plus général, la nouvelle statistique est un estimateur sans biais de $\int\int \varphi(x^1, x^2)(dP(x^1, x^2) - dP^1(x^1)dP^2(x^2))$, et semble donc être une statistique de test d'indépendance adéquate, centrée sous (H_0). Il est possible d'écrire cette nouvelle statistique sous forme de U-statistique, ce qui mène à la définition de notre statistique de test la plus générale :

$$U_{n,h}(\mathbb{X}_n) = \frac{1}{n(n-1)} \sum_{i\neq i'\in\{1,2,\ldots,n\}} h(X_i, X_{i'}),$$

où $h : (\mathcal{X}^2)^2 \to \mathbb{R}$ est un noyau symétrique, centré sous (H_0), i.e. vérifiant l'hypothèse :

$(\mathcal{A}_{Cent})$ | Si $P = P^1 \otimes P^2$, $\mathbb{E}[h(X_1, X_2)] = 0$.

Dans toute la suite, on supposera par ailleurs que la U-statistique $U_{n,h}(\mathbb{X}_n)$ est non-dégénérée sous (H_0), ce qui revient à supposer :

$(\mathcal{A}_{non-deg})$ | Si $P = P^1 \otimes P^2$, $\mathrm{Var}(\mathbb{E}[h(X_1, X_2)|X_1]) > 0$.

On s'intéresse notamment à deux cas particuliers :

- au *cas linéaire* où $h = h_\varphi$ avec

  $$h_\varphi(x, y) = \frac{1}{2}\left(\varphi(x^1, x^2) + \varphi(y^1, y^2) - \varphi(x^1, y^2) - \varphi(y^1, x^2)\right),$$

  pour lequel l'hypothèse $(\mathcal{A}_{Cent})$ est toujours vérifiée,

- au *cas des coïncidences*, sous-cas du *cas linéaire* avec $\varphi = \varphi_\delta^{coinc}$, pour lequel $U_{n,h_{\varphi_\delta^{coinc}}}(\mathbb{X}_n)$ n'est autre que la statistique définie en (1). Nous montrons que dans ce cas, $(\mathcal{A}_{non-deg})$ est vérifiée dès que $P^1(\emptyset) > 0$, $P^2(\emptyset) > 0$, et φ_δ^{coinc} n'est pas identiquement nulle, ce qui est facilement vérifiable en pratique.

Rappelons d'abord des résultats classiques pour les U-statistiques. Le théorème de la limite centrale est exprimé ici en termes de distance de Wassertein, définie pour deux mesures de probabilité de carré intégrable Q et Q' par

$$d_2(Q, Q') = \inf\left\{ \left(\mathbb{E}[(Z - Z')^2]\right)^{1/2} \ ; \ Z \sim Q, Z' \sim Q' \right\}.$$

Théoreme 1. *Soit $\mathbb{X}_n$ un n-échantillon de loi P.*

- Loi des grands nombres : *si* $\mathbb{E}[h(X_1, X_2)] < +\infty$, *alors* $U_{n,h}(\mathbb{X}_n) \xrightarrow[n\to+\infty]{p.s.} \mathbb{E}[h(X_1, X_2)]$.
- Théorème de la limite centrale : *Si* $P = P^1 \otimes P^2$, *et si h vérifie* $(\mathcal{A}_{Cent})$, $(\mathcal{A}_{non-deg})$ *et* $\mathbb{E}[h^2(X_1, X_2)] < +\infty$, *alors*
$$d_2\left(\mathcal{L}(\sqrt{n}U_{n,h}(\mathbb{X}_n)), \mathcal{N}(0, \sigma^2_{P^1\otimes P^2})\right) \underset{n\to+\infty}{\longrightarrow} 0,$$
où $\sigma^2_{P^1\otimes P^2} = 4\,\mathrm{Var}(\mathbb{E}[h(X_1, X_2)|X^1])$.

Remarque : En se basant sur ce théorème, on peut construire un test purement asymptotique, en estimant la variance $\sigma^2_{P^1\otimes P^2}$. Cependant, il pourrait souffrir d'un manque de puissance lorsque le nombre d'observations n est petit, ce qui est souvent le cas en pratique pour des raisons biologiques et économiques. L'idée naturelle est donc de se tourner vers des approches par bootstrap ou par permutation.

3 Différentes approches

3.1 Approche par bootstrap

S'inspirant de Romano (1989), on définit l'*échantillon bootstap* associé à $\mathbb{X}_n$ par $\mathbb{X}^*_n = (X^*_1, \ldots, X^*_n)$ où les X^*_i sont i.i.d. de loi $P^1_n \otimes P^2_n$, avec pour $j = 1, 2$ $P^j_n = \frac{1}{n}\sum_{i=1}^n \delta_{X^j_i}$. Cela revient à tirer indépendamment avec remise chaque coordonnée selon les lois marginales empiriques. Afin de justifier cette approche, on a besoin des hypothèses suivantes :

- une hypothèse de centrage empirique :
$$(\mathcal{A}^*_{Cent}) \ \Big|\ \forall (x^1_1, x^2_1), \ldots (x^1_n, x^2_n) \in \mathcal{X}^2, \quad \textstyle\sum_{i,i',j,j'=1}^n h((x^1_i, x^2_{i'}), (x^1_j, x^2_{j'})) = 0,$$
- une hypothèse générale de moments :
$$(\mathcal{A}^*_{Mmt}) \ \left|\ \begin{array}{l} \forall X_1, X_2, X_3, X_4 \text{ i.i.d. } \sim P^1 \otimes P^2, \forall i, i', j, j' \in \{1, 2, 3, 4\}, \\ \mathbb{E}[h^2((X^1_i, X^2_{i'}), (X^1_j, X^2_{j'}))] = 0. \end{array}\right.$$
- une hypothèse de continuité :
$$(\mathcal{A}_{Cont}) \ \Big|\ \text{il existe } \mathcal{C} \subset (\mathcal{X}^2)^2 \text{ tel que } P^1 \otimes P^2(\mathcal{C}) = 1, \text{ sur lequel } h \text{ est continue.}$$

Théoreme 2. *Soit $\mathbb{X}_n$ et $\mathbb{X}^{\perp\!\!\!\perp}_n$ des n-échantillons de loi respectives P et $P^1 \otimes P^2$, et $\mathbb{X}^*_n$ un* échantillon bootstrap *associé à $\mathbb{X}_n$. Supposons que h vérifie* $(\mathcal{A}_{Cent})$, $(\mathcal{A}^*_{Cent})$, $(\mathcal{A}^*_{Mmt})$ *et* $(\mathcal{A}_{Cont})$. *Alors*

$$d_2\left(\mathcal{L}(\sqrt{n}U_{n,h}(\mathbb{X}^*_n)|\mathbb{X}_n), \mathcal{L}(\sqrt{n}U_{n,h}(\mathbb{X}^{\perp\!\!\!\perp}_n))\right) \xrightarrow[n\to+\infty]{p.s.} 0.$$

Remarque : Par conséquent, la loi conditionnelle de la statistique bootstrappée sachant les données initiales mime la loi de la statistique sous l'hypothèse d'indépendance, et ce même si l'échantillon de départ ne vérifie pas (H_0). Intuitivement, cela vient du fait qu'on impose l'indépendance entre les coordonnées dans l'*échantillon bootstrap*.

3.2 Approche par permutation

S'inspirant encore une fois de Romano (1989), nous nous sommes aussi intéressés à l'approche par permutation. Un *échantillon permuté* associé à $\mathbb{X}_n$ est alors défini par $\mathbb{X}_n^\star = (X_1^\star, \dots, X_n^\star)$ où $X_i^\star = (X_i^1, X_{\pi(i)}^2)$, π étant une permutation aléatoire uniforme de $\{1, 2, \dots, n\}$. Cette nouvelle approche n'est ici justifiée que dans le *cas linéaire* et sous l'hypothèse suivante :

$(\mathcal{A}_{Mmt,\varphi})$ | pour tout (X^1, X^2) de loi P ou $P^1 \otimes P^2$ sur $\mathcal{X}^2$, $\mathbb{E}[\varphi^4(X^1, X^2)] = 0$.

Théoreme 3. *Soit $\mathbb{X}_n$ un n-échantillon de la loi P et $\mathbb{X}_n^\star$ un* échantillon permuté *associé à $\mathbb{X}_n$. Dans le* cas linéaire*, si h vérifie $(\mathcal{A}_{Mmt,\varphi})$ et $(\mathcal{A}_{non-deg})$, alors*

$$d_2\left(\mathcal{L}(\sqrt{n}U_{n,h}(\mathbb{X}_n^\star)|\mathbb{X}_n), \mathcal{N}(0, \sigma^2_{P^1\otimes P^2})\right) \xrightarrow[n\to+\infty]{\mathbb{P}} 0.$$

Remarque : Ainsi, la loi conditionnelle de la *statistique permutée* converge en probabilité vers la même limite gaussienne que la loi de la statistique initiale sous (H_0), et ce, encore une fois, même si les données initiales ne vérifient pas (H_0). Intuitivement, cela vient du peu de points fixes pour les permutations uniformes (une seule en moyenne). Ce résultat sous (H_1) est, à notre connaissance, complètement nouveau même dans des cadres plus généraux que celui des processus ponctuels. En particulier, il répond à une question ouverte de Van der Vaart et Wellner (1996).

4 Tests d'indépendance

Combinant les théorèmes 1, 2 et 3, on obtient les convergences faibles de la loi de la statistique sous (H_0), et des lois conditionnelles des statistiques bootstrappée et permutée vers la même limite gaussienne. Grâce à cela, on peut construire des tests d'indépendance unilatéraux à droite :

$$\Delta^+_{n,\alpha}(\mathbb{X}_n) = \mathbf{1}_{\sqrt{n}U_{n,h}(\mathbb{X}_n) > c^+_{\alpha,n}(\mathbb{X}_n)},$$

en prenant comme valeurs critiques :

- *(bootstrap)* $c^+_{\alpha,n}(\mathbb{X}_n) = q^*_{1-\alpha,n}(\mathbb{X}_n)$ le $(1-\alpha)$-quantile de $\mathcal{L}(\sqrt{n}U_{n,h}(\mathbb{X}_n^*)|\mathbb{X}_n)$,
- *(permutation dans le cas linéaire uniquement)* $c^+_{\alpha,n}(\mathbb{X}_n) = q^\star_{1-\alpha,n}(\mathbb{X}_n)$ le $(1-\alpha)$-quantile de $\mathcal{L}(\sqrt{n}U_{n,h}(\mathbb{X}_n^\star)|\mathbb{X}_n)$.

Il est bien connu que les tests par permutation ont le grand avantage d'être de niveau non-asymptotique voulu, à savoir :

$$\text{sous } (H_0),\ \forall n \geq 2, \quad \mathbb{P}(\sqrt{n}U_{n,h}(\mathbb{X}_n) > q^{\star}_{1-\alpha,n}(\mathbb{X}_n)) \leq \alpha.$$

Ainsi, lorsque les deux approches sont applicables, celle par permutation doit être privilégiée par rapport à celle par bootstrap, comme le recommandent Efron et Tibshirani (1993).

Par ailleurs, nous montrons que les tests d'indépendance introduits ci-dessus vérifient les propriétés asymptotiques suivantes.

Théoreme 4. *Soit* $(\Delta^{+}_{n,\alpha}(\mathbb{X}_n))_{n\geq 2}$ *une suite de tests par bootstrap ou par permutation. Sous les hypothèses du théorème 1 et de celui correspondant à l'approche choisie, on a :*

- (taille asymptotique) *si* $P = P^1 \otimes P^2$, *alors* $\mathbb{P}(\Delta^{+}_{n,\alpha}(\mathbb{X}_n) = 1) \underset{n\to+\infty}{\longrightarrow} \alpha$,
- (consistance) *si* P *est tel que* $\mathbb{E}[h(X_1, X_2)] > 0$, *alors* $\mathbb{P}(\Delta^{+}_{n,\alpha}(\mathbb{X}_n) = 1) \underset{n\to+\infty}{\longrightarrow} 1$.

Nous montrons également que lorsque les quantiles sont approchés par des méthodes de Monte Carlo, les tests vérifient les mêmes propriétés (asymptotiques ou non).

Cela nous a permis d'implémenter ces tests d'indépendance afin de vérifier leur validité d'un point de vue pratique. Les résultats des études par simulation montrent une réelle amélioration par rapport aux tests de Tuleau-Malot et al (2014) et de *trial-shuffling* discutés dans l'introduction, ainsi qu'au test "naïf" basé sur le théorème 1. Ces études, ainsi que toute l'analyse théorique, sont disponibles sur HAL à l'adresse éléctronique `http://hal.archives-ouvertes.fr/hal-01001984`.

Bibliographie

[1] Efron, B. et Tibshirani, R.J. (1993) An introduction to the bootstrap, *Monographs on Statistics and Applied Probability*, Chapman and Hall, New York.

[2] Grün, S., Diesmann, M et Aertsen, A.M. (2010), Unitary Events Analysis, chapitre 10 dans *Unitary Events Analysis*, Springer Series in Computational Neurosciences.

[3] Pipa, G. et Grün, S. (2003) Non-parametric significance estimation of joint-spike events by shuffling and resampling, *Neurocomputing*, 52–54:31–37.

[4] Romano, J.P. (1989) Bootstrap and randomization tests of some nonparametric hypotheses, *Annals of Statistics*, 17(1):141–159.

[5] Tuleau-Malot, C., Rouis, A., Grammont, F. et Reynaud-Bouret, P. (2014), Multiple tests based on Gaussian approximation of the Unitary Events method, *Neural Computation*, 26(7):1408–1454.

[6] Van der Vaart, A.W., Wellner, J.A. (1996) *Weak Convergence and Empirical Processes*, Springer, New-York.