Procédure LASSO pour la reconstruction du support d'un processus de Hawkes multivarié en grande dimension

Christophe Denis[1,2] & Charlotte Dion-Blanc[2] & Romain E. Lacoste[1] & Laure Sansonnet[2,3]

[1] *LAMA, Université Gustave-Eiffel, France romain.lacoste@polytechnique.edu*
[2] *LPSM, Sorbonne Université, France*
[3] *MIA Paris-Saclay, Université Paris-Saclay, France*

Résumé. Dans cette étude, on s'intéresse au problème de la reconstruction du support de la matrice d'interaction d'un processus de Hawkes multivarié en grande dimension. Afin de composer avec la grande dimension, on impose des hypothèses de parcimonie sur la matrice d'interaction. On suppose que l'on a accès à des répétitions de trajectoires de processus de Hawkes multivariés en temps court. La stratégie proposée consiste à minimiser le contraste des moindres carrés moyenné sur les répétitions, couplé à une pénalité de type LASSO. On établit un résultat de consistance du support et de convergence de l'estimateur associé lorsque le nombre d'observations tend vers l'infini. Pour résoudre le problème de minimisation de la fonction objective, incluant un terme non-différentiable, on utilise des algorithmes de descente de gradient proximal. En présence d'une pénalité ℓ_1, l'opérateur proximal associé s'écrit comme le seuillage doux et on utilise l'algorithme FISTA. L'implémentation de la procédure est réalisée en `C++` et bénéficie de propriétés computationelles compétitives. Enfin on propose une étude numérique sur données simulées pour valider la procédure.

Mots-clés. Processus de Hawkes, Grande Dimension, Parcimonie, LASSO, FISTA.

Abstract. In this study, we focus on the problem of support recovery of the interaction matrix of a high-dimensional multivariate Hawkes process. In order to deal with the high dimensionality, we impose sparsity assumptions on the interaction matrix. We assume that we have access to short-time path repetitions of multivariate Hawkes process. Our strategy consists in minimizing the least-squared contrast averaged over the repetitions, coupled with a LASSO-type penalty. We establish a result of support consistency and convergence of the associated estimator when the number of observations tends to infinity. To solve the problem of minimizing the objective function, which includes a non-differentiable term, we use proximal gradient descent algorithms. In the presence of a ℓ_1 penalty, the associated proximal operator is written as a soft thresholding and the FISTA algorithm is used. The procedure is implemented in `C++` and benefits from competitive computational properties. Finally, we propose a numerical study on simulated data to validate the procedure.

Keywords. Hawkes Process, High Dimension, Sparsity, LASSO, FISTA.

1 Introduction

Les processus de Hawkes, qui furent introduits dans [10], permettent de modéliser des séries temporelles où l'occurrence d'un évènement augmente la probabilité d'en observer un nouveau dans un futur proche. En vertu de leur nature intrinsèque à modéliser des dynamiques auto-excitantes, les processus de Hawkes sont extrêmement versatiles quant à leur domaine d'applicabilité. En effet, bien qu'ils aient historiquement été utilisés pour la modélisation de secousses sismiques [11, 13], leur champ d'application s'est rapidement étendu à de nombreux autres domaines, comme les neurosciences [15], l'étude de réseaux sociaux [14], le football [2] ou encore l'écologie [6], etc.

La version multidimensionnelle de ces processus, appelée processus de Hawkes multivariés (PHM), constitue une généralisation naturelle qui enrichit considérablement les possibilités de modélisation. En effet, en plus de modéliser les interactions auto-excitantes, un tel modèle prend en compte les interactions positives entre les individus connectés au sein d'un réseau. Par conséquent, l'activité future d'une composante ne dépend plus seulement de ses activités passées, mais également de l'activité passée des individus qui interagissent avec elle et qui sont donc susceptibles d'avoir eu une influence sur elle. Là encore, les potentielles applications sont légions, comme la modélisation de potentiels d'actions au sein d'un réseau de neurone [4], la finance [8], etc.

D'un point de vue théorique, de nombreuses méthodes statistiques pour l'inférence ont été proposées pour les PHM linéaires ou non linéaires. On peut citer par exemple [1, 7, 17, 5]. En particulier dans [1], les auteurs proposent une procédure d'inférence reposant sur la minimisation du contraste des moindres carrés avec des pénalités ℓ_1 et de faible rang, laquelle est basée sur l'observation d'une seule trajectoire d'un PHM en temps long sur l'intervalle $[0, T]$ avec T qui tend vers l'infini. Dans cette étude, on propose d'adapter cette procédure à un cadre asymptotique différent. En effet, on suppose qu'on observe n trajectoires d'un PHM sur l'intervalle $[0, T]$ avec T fixé. Notre stratégie consiste à étudier le contraste des moindres carrés moyenné sur les répétitions avec une pénalité de LASSO. Sous des hypothèses classiques, un résultat de consistance du support et de convergence de l'estimateur associé en norme infini est établi lorsque le nombre d'observations tend vers l'infini. Enfin, pour valider la procédure, on illustre ses performances sur des données simulés.

2 Processus de Hawkes multivarié

Dans cette étude, on s'intéresse aux processus de Hawkes linéaires exponentiels M-multivariés (PHM), où $M > 1$ est la dimension du réseau. Un tel processus, noté $N = (N_1, \dots, N_M)$ est la donnée de M processus ponctuels sur $\mathbb{R}_+^*$. Le processus de comptage N_j est le processus $(N_j(t))_{t\in\mathbb{R}_+}$ tel que $N_j(t) = \sum_{\ell\geq 1} \mathbb{1}_{T_{j,\ell}\leq t}$, où $\{\{T_{j,\ell}\}_{\ell\geq 1}, 1 \leq j \leq M\}$ sont les temps de sauts du processus N. Chaque processus N_j peut se caractériser par son intensité conditionnelle,

fonction définie pour tout $t \geq 0$ par :

$$\lambda_j(t) := \mu_j + \sum_{j'=1}^{M} a_{j,j'} \int_0^t \beta e^{-\beta(t-s)} \, dN_{j'}(s) = \mu_j + \sum_{j'=1}^{M} \sum_{T_{j',\ell}<t} a_{j,j'} \beta e^{-\beta(t-T_{j',\ell})}, \tag{1}$$

où

- $\mu = (\mu_1, \ldots, \mu_M) \in (\mathbb{R}_+^*)^M$ appelé vecteur d'intensité de base, régit la dynamique de chaque processus N_j avant l'occurrence d'un saut et modélise l'arrivée d'évènements spontanés ;
- $A = (a_{j,j'})_{j,j'} \in \mathbb{R}_+^{M \times M}$ appelé matrice d'interaction, modélise l'influence positive de la composante j' sur la composante j ;
- $\beta \geq 0$ appelé taux de décroissance, contrôle les interactions entres composantes et donne la vitesse avec laquelle ces influences disparaissent avec le temps.

On dit qu'un individu j' interagit avec un individu j dès lors qu'il exerce une influence non-nulle sur j, c'est-à-dire lorsque $a_{j,j'} > 0$. Du fait de l'hypothèse de positivité des coefficients de A, chaque occurrence d'un évènement $T_{j',\ell}$ augmente temporairement la probabilité d'observer un nouveau saut dans un futur proche chez tous les individus avec lequel j' interagit. Ainsi, un PHM permet de modéliser des effets d'excitation mutuels entre les composantes d'un réseau, lesquels dépendent des interactions survenues dans le passé.

D'après la définition (1) ci-dessus, on considère que les intensités conditionnelles des composantes du PHM N dépendent d'un paramètre inconnu θ^* qui appartient à la famille de paramètres suivante :

$$\Theta := \left\{ \mu \in (\mathbb{R}_+^*)^M, \ A \in \mathcal{M}_M(\mathbb{R}_+), \ \rho(A) < 1 \right\} \in \mathbb{R}_+^{M \times M+1}$$

avec $\rho(A)$ le rayon spectral de A. Cette hypothèse de stabilité sous-critique assure la non-explosion du processus N. Dans la suite, on suppose le paramètre β connu et commun à toutes les composantes de N. On note $\theta^* = (\mu^*, A^*) \in \Theta$ le paramètre à estimer et pour tout $j \in \{1, \ldots, M\}$, on note λ_{j,θ^*} l'intensité conditionnelle de la j-ème composante associée à ce paramètre. Notre objectif est de reconstruire le support de θ^*, qu'on note $\text{supp}(\theta^*)$.

3 Cadre statistique

Soit $T > 0$ fixé. On note $\mathcal{T}_T := \{\{T_{j,\ell}\}_{1 \leq \ell \leq N_j(T)}, 1 \leq j \leq M\}$ les temps de saut d'un PHM $N = (N_1, \ldots, N_M)$ d'intensité conditionnelle $(\lambda_{1,\theta^*}, \ldots, \lambda_{M,\theta^*})$ observés en temps court sur l'intervalle $[0, T]$. On dispose d'un n-échantillon $D_n := \{\mathcal{T}_T^{(1)}, \ldots, \mathcal{T}_T^{(n)}\}$ qui consiste en copies indépendantes de $\mathcal{T}_T$. De plus, on se place dans le cadre de la grande dimension, à savoir qu'on considère des cas où la dimension du réseau M peut être très grande (en particulier le nombre total de paramètre $M(M+1)$ peut être plus grand que n). Afin de composer avec la grande dimension, on impose des hypothèses de parcimonie sur la matrice d'interaction A^*. D'un point de vue concret, supposer A^* creuse traduit que chaque individu du réseau n'est impacté que par une faible portion d'autres individus. Bien que cette hypothèse soit motivée par la réduction de la dimension du problème et pour faciliter l'interprétation, elle s'avère

très souvent naturelle du point de vue de la modélisation. À titre d'exemple, si l'on cherche à modéliser les interactions au sein de réseaux sociaux, il est vraisemblable de supposer qu'il existe des individus n'interagissant que très peu en dehors de leur cercle de connaissances proche ainsi que l'existence d'individus très peu connectés. A des fins de modélisation, on ne fait pas d'hypothèses de parcimonie sur le vecteur μ^*, chaque μ_j^* étant par définition pris strictement positif. Le support de θ^*, supposé de faible cardinalité, est noté $\mathrm{supp}(\theta^*)$. Pour $j \in \{1, \dots, M\}$, on note $S_j^* := \{\theta_{j,j'}^* \neq 0,\ 1 \leq j' \leq M+1\}$ le support de la j-ème ligne de θ. En particulier, en raison du caractère non creux de μ^*, chaque S_j^* contient donc au moins un élément.

On considère, à titre de fonction d'ajustement au modèle, le contraste des moindres carrés moyenné sur les répétitions défini comme suit :

$$R_{T,n}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{T} \sum_{j=1}^{M} \int_0^T \lambda_{j,\theta}^{(i)}(t)^2 \, dt - 2 \int_0^T \lambda_{j,\theta}^{(i)}(t) \, dN_j(t) \right), \tag{2}$$

où $\lambda_{j,\theta}^{(i)}(t)$ donné en équation (1) est définie à partir de la i-ème répétition.

Notre objectif est de reconstruire le support de θ^*, $\mathrm{supp}(\theta^*)$. A cette fin, étant donné l'échantillon D_n, notre stratégie consiste à minimiser le contraste des moindres carrés couplé à une pénalité de type LASSO :

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^{M \times M+1}}{\mathrm{argmin}} \left\{ R_{T,n}(\theta) + \kappa \sum_{i=1}^{M} \sum_{j=1}^{M} |\theta_{i,j}| \right\}, \tag{3}$$

où κ est la constante de régularisation à calibrer. A noter que l'on n'impose aucune contraintes de régularisation sur μ.

4 Résultats théoriques

On commence par donner quelques notations supplémentaires. Pour une matrice B, on note B' sa transposée. Pour tout $t \in (0, T]$, on définit la matrice aléatoire $\mathbb{H}_t \in \mathbb{R}^{n \times M+1}$ comme suit :

$$(\mathbb{H}_t)_{i,j} = H_j^{(i)}(t), \ \text{ avec } \ H_j^{(i)}(t) := \int_0^t \beta e^{-\beta(t-s)} \, dN_j^{(i)}(s), \ j \in \{1, \dots, M\},$$

avec $H_0^{(i)} \equiv 1$. On définit aussi :

$$\mathbb{H} = \frac{1}{T} \int_0^T \mathbb{H}_t' \mathbb{H}_t \, dt.$$

Pour $j \in \{1, \dots, M\}$, on note $\mathbb{H}_{S_j^*, S_j^*} := (\mathbb{H}_{j',j'})_{j' \in S_j^*}$ la matrice extraite en supprimant les lignes et colonnes appartenant au complémentaire du support $S_j^{*c} := \{\theta_{j,j'}^* = 0,\ 1 \leq j' \leq M+1\}$.

On énonce les hypothèses suivantes qui portent sur la matrice $\mathbb{H}$.

Hypothèse 1 (Incohérence mutuelle)**.** *Il existe un* $\gamma > 0$ *tel que*

$$\max_{j\in\{1,\dots,M\}} \|\mathbb{H}_{S_j^{*c},S_j^*}\mathbb{H}_{S_j^*,S_j^*}^{-1}\|_\infty \leq 1-\gamma$$

Hypothèse 2 (Valeur propre minimale)**.** *Il existe* $\Lambda_0 > 0$ *tel que*

$$\min_{j\in\{1,\dots,M\}} \Lambda_{\min}\left(\frac{\mathbb{H}_{S_j^*,S_j^*}}{n}\right) \geq \Lambda_0$$

Hypothèse 3 (Condition de signal minimal)**.**

$$\min_{j,j'\in S^*} \left|\theta^*_{j,j'}\right| > \Lambda_0 \max_j \left|S_j^*\right|^2 \frac{\log^4(nM^2)}{\sqrt{n}}$$

A présent nous pouvons énoncer le théorème principal de cette étude, on l'on établit la consistance du support et la convergence de l'estimateur associé.

Théorème 4. *On se place sous les hypothèses 1, 2, et 3. Soit* $\kappa = \dfrac{\log^4(nM^2)}{\sqrt{n}}$ *Pour* n *assez grand, avec probabilité supérieure à* $1-\frac{C_0}{n}$ *avec* $C_0 > 0$*, le contraste des moindres carrés pénalisé défini dans l'équation* (3) *admet une unique solution* $\hat{\theta}$ *qui satisfait les propriétés suivantes :*

1. $\hat{\theta}_{j,j'} \geq 0$
2. $\mathrm{supp}(\hat{\theta}) = \mathrm{supp}(\theta^*)$ *;*
3. $\left\|\hat{\theta}-\theta^*\right\|_\infty \leq \dfrac{\Lambda_0 \max_j |S_j^*|^2 \log^4(nM^2)}{\sqrt{n}}$

En particulier, une conséquence de ce résultat est que l'on a :

$$\mathbb{P}\left(\mathrm{supp}(\hat{\theta}) = \mathrm{supp}(\theta^*)\right) \xrightarrow[n\to\infty]{} 1$$

La preuve de ce résultat suit la méthode de construction *primal-dual witness* décrite dans le Chapitre 11 de [9].

5 Étude numérique

On termine cette étude en proposant une étude de la procédure sur des données simulées. La simulation des trajectoires est effectuée en utilisant l'algorithme de représentation par grappe (voir [12]), qui utilise les propriétés de branchement des PHM. Pour résoudre le problème de minimisation de la fonction objective définie dans l'équation (3), incluant un terme non-différentiable, on utilise des algorithmes de descente de gradient proximal. En présence d'une pénalité ℓ_1, l'opérateur proximal associé s'écrit comme le seuillage doux et on utilise l'algorithme FISTA (voir [3]). Le choix du pas de la descente de gradient préconisé

est $1/L$ avec L la constante de Lipschitz du gradient et on peut prendre la plus grande valeur propre de la matrice hessienne. Les calculs de gradient, hessienne et d'évaluation de la fonction objective sont implémentés en `C++`. Ces derniers sont optimisés de façon à réduire considérablement le coût computationel associé. Enfin, la calibration de la constante de régularisation κ est effectué via le critère BIC (voir [16]). Ce critère, en plus de donner de meilleurs résultats en terme de reconstruction du support, induit des temps de calculs bien moindre qu'une procédure par validation croisée. Pour toute ces raisons, prise dans son ensemble, la procédure bénéficie donc de propriétés computationelles compétitives ce qui la rend applicable à des réseaux de grande dimension.

Pour évaluer la qualité de la reconstruction du support, on utilise les deux métriques d'évaluation suivantes :

- Distance de Hamming : pour A^*, $\hat{A} \in \mathbb{R}_+^{M\times M}$,

$$d_H\left(A^*, \hat{A}\right) = \frac{1}{M^2}\sum_{j,j'=1}^{M} \mathbb{1}_{\left\{A^*_{j,j'} \neq \widehat{A}_{j,j'}\right\}}$$

- Distance ℓ_2 : pour A^*, $\hat{A} \in \mathbb{R}_+^{M\times M}$,

$$d_{\ell_2}\left(A^*, \hat{A}\right) = \sqrt{\sum_{j,j'=1}^{M} |A^*_{j,j'} - \widehat{A}_{j,j'}|^2}$$

Afin de rendre l'étude numérique la plus complète possible, on fait varier le taux de parcimonie de la matrice d'interaction A^* ainsi que sa structure, ces deux facteurs ayant un impact direct sur la difficulté de la tâche de reconstruction du support. Dans chacun de ces scénarios, on choisit $M = 25$ pour la dimension du réseau, $T = 20$ comme borne supérieure de l'intervalle d'observation et $\beta = 3$ pour le taux de décroissance. Dans les deux scénarios, on choisit $\mu^*_j = 0.5$ commun à toutes les composantes et on prend les paramètres non nuls $a^*_{j,j'}$ dans l'intervalle $[0.1, 0.5]$. Dans la Figure 1, on affiche le paramètre θ^* (où la première colonne correspond à μ^*) associé à chacun des deux scénarios considérés. Dans la Figure 2, on affiche le vrai support $\text{supp}(\theta^*)$ ainsi que le support $\text{supp}(\hat{\theta})$ reconstruit par la procédure pour $n \in \{100, 1000, 5000\}$. Enfin dans la Table 1, on affiche les valeurs des métriques d'évaluation moyennées sur 10 répétitions Monte-Carlo dans les deux scénarios et pour les trois valeurs de n.

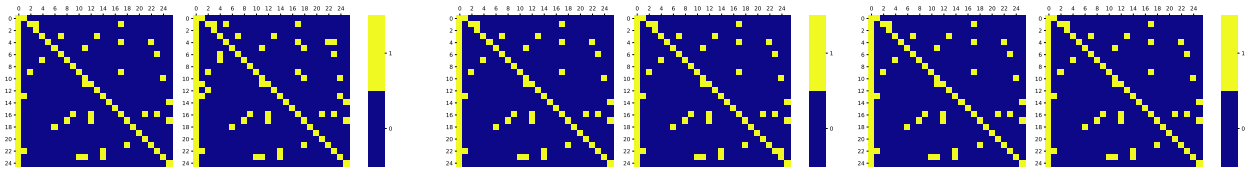
	d_H			d_{l_2}		
	$n = 100$	$n = 1000$	$n = 5000$	$n = 100$	$n = 1000$	$n = 5000$
Scénario 1	0.05 (0.05)	0.00 (0.00)	0.00 (0.00)	0.54 (0.11)	0.29 (0.01)	0.28 (0.01)
Scénario 2	0.09 (0.05)	0.04 (0.01)	0.01 (0.00)	0.53 (0.06)	0.22 (0.01)	0.22 (0.00)

TABLE 1 – Valeurs des métriques obtenues moyennées sur 10 répétitions Monte-Carlo avec écart-type dans les deux scénarios et pour les trois valeurs de n.

Dans les deux scénarios on remarque que le support est d'autant bien reconstruit que n est grand, que ce soit en terme de distance de Hamming et ℓ_2. On peut mettre en lumière que dans le cas du scénario 2 le carré inférieur droit, ayant des valeurs $a^*_{j,j'} = 0.1$ petites, est bien détecté par la procédure.

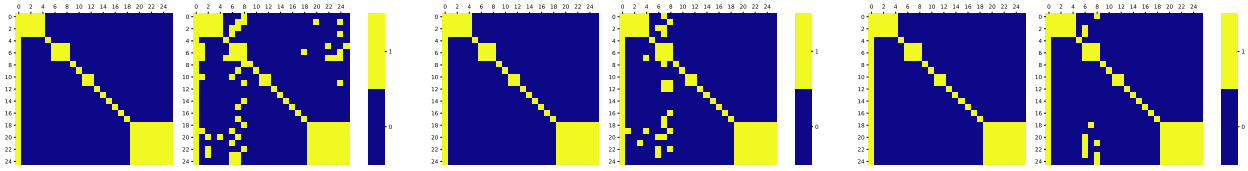
(a) *Scénario 1* (b) *Scénario 2*

FIGURE 1 – Représentation de la matrice $\theta^* = (\mu^*, A^*)$ dans les deux scénarios pour $M = 25$. Taux de parcimonie de A^* dans le *Scénario 1* : 8.64%, dans le *Scénario 2* : 13.92%



(a) $n = 100$ (b) $n = 1000$ (c) $n = 5000$

FIGURE 2 – Vrai support $\text{supp}(\theta^*)$ et support reconstruit $\text{supp}(\hat{\theta})$ pour $n = 100$, pour $n = 1000$ et pour $n = 5000$ dans le *Scénario 1*.



(a) $n = 100$ (b) $n = 1000$ (c) $n = 5000$

FIGURE 3 – Vrai support $\text{supp}(\theta^*)$ et support reconstruit $\text{supp}(\hat{\theta})$ pour $n = 100$, pour $n = 1000$ et pour $n = 5000$ dans le *Scénario 2*.

6 Remerciements

Ce travail a été soutenu par la Chaire « Modélisation Mathématique et Biodiversité » de Veolia-École polytechnique-Museum national d'Histoire naturelle-Fondation X.

Références

[1] E. Bacry, M. Bompaire, S. Gaïffas, and J-F Muzy. Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, 21(50) :1–32, 2020.

[2] A. Baouan, S. Coustou, M. Lacome, S. Pulido, and M. Rosenbaum. Crediting football players for creating dangerous actions in an unbiased way : the generation of threat (got) indices. *arXiv preprint arXiv :2304.05242*, 2023.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1) :183–202, 2009.

[4] A. Bonnet, C. Dion-Blanc, F. Gindraud, and S. Lemler. Neuronal network inference and membrane potential model using multivariate Hawkes processes. *Journal of Neuroscience Methods*, 372 :109550, 2022.

[5] A. Bonnet, M. Martinez Herrera, and M. Sangnier. Inference of multivariate exponential Hawkes processes with inhibition and application to neuronal activity. *Statistics and Computing*, 33(4) :91, 2023.

[6] C. Denis, C. Dion-Blanc, R.E. Lacoste, L. Sansonnet, and Y. Bas. Bats monitoring : A classification procedure of bats behaviors based on Hawkes processes. *hal preprint hal-04345822*, 2023.

[7] S. Donnet, V. Rivoirard, and J. Rousseau. Nonparametric Bayesian estimation for multivariate Hawkes processes. *Annals of Statistics*, 48(5) :2698–2727, 2020.

[8] P. Embrechts, T. Liniger, and L. Lin. Multivariate Hawkes processes : an application to financial data. *Journal of Applied Probability*, 48(A) :367–378, 2011.

[9] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

[10] A.G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1) :83–90, 1971.

[11] A.G. Hawkes. Cluster models for earthquakes-regional comparisons. *Bulletin of the International Statistical Institute*, 45(3) :454–461, 1973.

[12] J. Møller and J.G. Rasmussen. Perfect simulation of Hawkes processes. *Advances in Applied Probability*, 37(3) :629–646, 2005.

[13] Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401) :9–27, 1988.

[14] Z. Qu, C. Lyu, and C. Chi. Mush : Multi-stimuli Hawkes process based sybil attacker detector for user-review social networks. *IEEE Transactions on Network and Service Management*, 2022.

[15] P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot. Inference of functional connectivity in neurosciences via Hawkes processes. In *2013 IEEE global conference on signal and information processing*, pages 317–320. IEEE, 2013.

[16] G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2) :461 – 464, 1978.

[17] J. Worrall, R. Browning, P. Wu, and K. Mengersen. Fifty years later : new directions in Hawkes processes. *SORT (Statistics and Operations Research Transactions)*, 46(1) :3–38, 2022.