Estimation non-paramétrique de la loi des marques d'un shot-noise exponentiel

Paul Ilhe[1,2], Eric Moulines[1], François Roueff[1], Antoine Souloumiac[2]

[1]LTCI, Telecom ParisTech
CNRS, 46 rue Barrault, 75634 Paris Cedex 13, France

[2]CEA, LIST, 91191 Gif-sur-Yvette Cedex, France

paul.ilhe@telecom-paristech.fr,
eric.moulines@telecom-paristech.fr,francois.roueff@telecom-paristech.fr,antoine.souloumiac@cea.fr

Résumé − Nous proposons une méthode d'estimation non-paramétrique rapide et efficace pour estimer la distribution des marques d'un processus de shot-noise exponentiel dans le contexte d'un fort taux d'empilement. A partir d'une formule reliant la fonction caractéristique des marques à celle du shot-noise et sa dérivée, nous construisons un estimateur de type ≪ plug-in ≫ qui converge en norme uniforme vers la densité des marques à une vitesse logarithmique. Cette méthode est ensuite validée sur deux jeux de données simulées.

Abstract − We propose an efficient method to estimate in a nonparametric fashion the marks' density of a shot-noise process subject to a high *pile-up* effect. Based on a formula linking the characteristic function of the mark density to a function involving the shot-noise characteristic function and its derivative, we construct a "plug-in" estimator which converges to the mark density in uniform norm at a logarithmic speed. Two limited Monte-Carlo experiments are provided to support our findings.

1 Introduction

1.1 Dispositif expérimental et empilement

Nous nous intéressons à un problème inverse non-linéaire qui apparait en physique nucléaire, et plus particulièrement en spectrométrie Gamma ou X. Des particules (un faisceau de photons) frappent un détecteur. Chaque interaction est ensuite convertie par le détecteur en une impulsion électrique proportionnelle à l'énergie déposée par le photon correspondant. La forme de l'impulsion originelle -i.e. l'impulsion observée lorsque l'énergie déposée vaut 1 eV- dépend du détecteur et de l'électronique associée. Même si le dépôt d'énergie associé à un photon s'avère parfois partiel (effet Compton, voir par exemple [5]), la distribution des énergies déposées reflète très étroitement celle des énergies des photons : l'objectif de la spectrométrie gamma ou X consiste à mesurer précisement cette dernière quantité, caractéristique capitale du faisceau.

Les réactions nucléaires générant les photons étant indépendantes, on considère que le processus ponctuel des instants d'interaction faisceau/détecteur est bien modélisé par un processus de Poisson homogène d'intensité λ sur la droite réelle. Le signal analogique mesuré par le détecteur correspond alors à un processus de Poisson filtré par la réponse impulsionnelle- supposée connue- et marqué multiplicativement par une amplitude aléatoire de distribution inconnue : en d'autres termes, un processus shot-noise. Dans ce contexte, nous souhaitons retrouver la distribution des énergies photoniques à partir du courant électrique mesuré. Or, les techniques d'instrumentation nucléaire actuelles ne permettent pas de mesurer cette distribution lorsque l'intensité λ est suffisamment forte pour que les impulsions, de durée courte mais non nulle, s'empilent : ce phénomène est connu sous le nom d'empilement ou de *pile-up*. Cette superposition de courants électriques empêche donc d'établir une correspondance bijective entre une impulsion et l'énergie déposée associée. En pratique, comme le signal observé est échantillonné à une fréquence, bien que de l'ordre de 10 MHz, inférieure λ, l'estimation de la loi des marques ne peut se faire qu'à travers les distributions marginales du shot-noise.

1.2 Modèle

Afin de résoudre ce problème inverse, nous modélisons le courant électrique par un processus stationnaire de shot-noise $\mathbf{X} = (X_t)_{t\in\mathbb{R}}$ défini par :

$$X_t = \sum_{k:T_k\le t} Y_k\, h(t - T_k)\ , \tag{1}$$

où h représente la réponse impulsionnelle du détecteur et $\sum_k \delta_{(T_k,Y_k)}$ correspond à un processus de Poisson homogène à valeurs dans $\mathbb{R}$ de mesure d'intensité $\lambda > 0$ et

de marques i.i.d. à valeurs réelles admettant une densité θ par rapport à la mesure de Lebesgue.

Le processus (1) est bien défini si, et seulement si, la fonction h et la densité θ vérifient

$$\int \min(1, |y\,h(s)|)\theta(y)\mathrm{d}y\mathrm{d}s < \infty\,. \tag{2}$$

Lorque la réponse impulsionnelle h est une fonction exponentielle de taux de décroissance $\alpha > 0$, i.e.

$$h(t) \triangleq \mathrm{e}^{-\alpha\, t}\mathbb{1}_{\mathbb{R}_+}(t)\,, \tag{3}$$

le shot-noise est dit *exponentiel* et la condition (2) est alors équivalente à $\mathbb{E}\left[\log_+(|Y_1|)\right] < \infty$. La figure ci-dessous illustre la trajectoire d'un shot-noise et du processus ponctuel marqué associé.

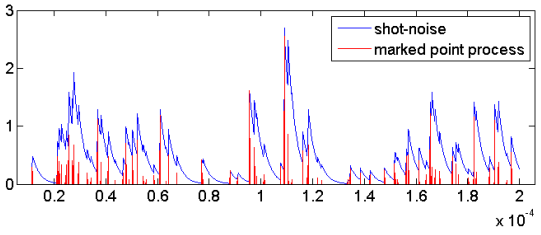

FIGURE 1 – Trajectoire d'un shot-noise *exponentiel*

Désormais, notre problème s'apparente à un *problème inverse* non-linéaire et se formule de la manière suivante : à partir d'observations $X_1, \cdots, X_n$ d'un shot-noise *exponentiel* d'intensité λ élevée, notre objectif consiste à identifier la densité θ des marques. Nous allons montrer que cette densité peut être estimée de manière non-paramétrique pour un certain espace de fonctions contenant θ. À notre connaissance, l'article qui se rapproche le plus de ce problème est [4]. Les auteurs proposent un estimateur non-paramétrique de la fonction k qu'ils nomment fonction canonique et qui est reliée à la densité θ par la formule

$$k(x) = \lambda\left(1 - \int_{-\infty}^{x}\theta(u)\mathrm{d}u\right) \quad , \quad x \in \mathbb{R}\,.$$

Néanmoins, aucune vitesse d'estimation n'est fournie dans [4].

2 Formule d'inversion

Notons Y une v.a. de même loi que les variables aléatoires $(Y_k)_{k\geq 0}$. La densité de la marginale du shot-noise *exponentiel* défini par (1) ne s'écrit pas sous forme analytique. En revanche, étant donné qu'un shot-noise peut être regardé comme un processus de Poisson filtré, la fonction caractéristique de X_0 s'écrit, pour tout réel u

$$\begin{aligned}\varphi_{X_0}(u) &= \exp\left(\lambda\int_0^\infty\int_{\mathbb{R}}(e^{iuxh(v)} - 1)\,\theta(x)\mathrm{d}x\mathrm{d}v\right)\\ &= \exp\left(\lambda\int_0^\infty\int_{\mathbb{R}}(e^{iux\mathrm{e}^{-\alpha v}} - 1)\,\theta(x)\mathrm{d}x\mathrm{d}v\right)\,. \end{aligned}\tag{4}$$

De plus, si on suppose que $\mathbb{E}\left[|Y|\right] = \int_{\mathbb{R}}|x|\theta(x)\mathrm{d}x < \infty$, la fonction φ_{X_0} est de classe $\mathcal{C}^1$ sur $\mathbb{R}$ et on a pour tout réel u

$$\varphi'_{X_0}(u) = \varphi_{X_0}(u)\frac{\lambda}{\alpha u}\left(\varphi_Y(u) - 1\right)\,. \tag{5}$$

Comme la loi de la variable X_0 est infiniment divisible, la fonction φ_{X_0} ne s'annule pas. Par conséquent, et sous l'hypothèse que φ_Y est intégrable, l'équation (5) permet d'exprimer la densité θ sous la forme

$$\theta(x) = \int_{\mathbb{R}}\mathrm{e}^{-\mathrm{i}xu}\varphi_Y(u)\mathrm{d}u = \int_{\mathbb{R}}\mathrm{e}^{-\mathrm{i}xu}\left(1 + \frac{\alpha u}{\lambda}\frac{\varphi'_{X_0}(u)}{\varphi_{X_0}(u)}\right)\mathrm{d}u\,. \tag{6}$$

Dans la suite du papier, nous supposerons que le rapport α/λ est une constante connue en tant que paramètres de l'appareil de mesure. Cependant, dans le cas où une incertitude surviendrait, il est possible d'estimer ce rapport en utilisant la propriété ([3][Theorème 1]) que la distribution G de X_0 est à variation régulière d'indice λ/α en 0, i.e.

$$G(x) \sim x^{\lambda/\alpha}L(x) \quad , \quad x \to 0 \tag{7}$$

où L est une fonction à variation lente en 0. L'estimateur de Hill appliqué à l'échantillon $X_1^{-1}, \cdots, X_n^{-1}$ fournit alors un estimateur consistant de α/λ.

3 Estimation non-paramétrique

Pour u réel, on définit $\hat{\varphi}_n(u) \triangleq n^{-1}\sum_{i=1}^n \mathrm{e}^{\mathrm{i}uX_j}$ comme la fonction caractéristique empirique basé sur les observations $X_1, \cdots, X_n$ et $\hat{\varphi}'_n$ sa dérivee. A partir de la formule d'inversion (6), nous sommes tentés de définir un estimateur $\hat{\theta}_n$ en insérant la fonction caractéristique empirique dans la formule (6), comme réalisé dans [2] .

Soient $(h_n)_{n\geq 0}$ et $(\kappa_n)_{n\geq 0}$ deux suites de nombres positifs telles que

$$\lim_{n\to\infty} h_n = \lim_{n\to\infty}\kappa_n = 0\,.$$

On définit alors l'estimateur $\hat{\theta}_n$ par

$$\hat{\theta}_n(x) \triangleq \frac{1}{2\pi}\int_{-h_n^{-1}}^{h_n^{-1}}\mathrm{e}^{-\mathrm{i}ux}\left(1 + \frac{\alpha u}{\lambda}\frac{\hat{\varphi}'_n(u)}{\hat{\varphi}_n(u)}\mathbb{1}_{|\hat{\varphi}_n(u)|>\kappa_n}\right)\mathrm{d}u\,. \tag{8}$$

Le choix de cet estimateur est motivé par des raisons de stabilité numérique : d'une part, à u fixé, on estime $1/\varphi(u)$ par $\mathbb{1}_{\{|\hat{\varphi}_n(u)|>\kappa_n\}}/\hat{\varphi}_n$ pour une suite $(\kappa_n)_{n\geq 0}$ strictement positive. Une méthode semblable est utilisée dans [6] pour des observations i.i.d. : les fluctuations de $\sqrt{n}\left(\hat{\varphi}_n(u) - \varphi(u)\right)$ étant bornées en probabilité, les auteurs proposent de choisir $\kappa_n = \kappa n^{-1/2}$ pour un certain $\kappa > 0$. D'autre part, nous tronquons l'intervalle d'intégration $\mathbb{R}$ par $\left[-h_n^{-1}, h_n^{-1}\right]$, permettant ainsi de contrôler l'erreur d'estimation $\theta - \hat{\theta}_n$ en norme uniforme. Comme les déviations de $\sqrt{n}\left(\hat{\varphi}_n(u) - \varphi(u)\right)$ sur $\left[-h_n^{-1}, h_n^{-1}\right]$ dépendent de la bande passante h_n, le paramètre κ_n sera choisi légérement plus grand que $n^{-1/2}$.

Afin d'évaluer la performance de notre estimateur en norme uniforme, nous définissons un ensemble de classes de fonctions auxquelles la densité est susceptible d'appartenir. Pour K, L, m des nombres strictement positifs et $s > 1/2$, on définit

$$\Theta(K,L,m,s) = \Big\{ \theta \text{ est une densité de probabilité t.q.} \int_{\mathbb{R}} |y|^{4+m}\theta(y)\mathrm{d}y \le K, \int_{\mathbb{R}} (1+|u|^2)^s ||\mathcal{F}\theta(u)|^2 \mathrm{d}u \le L \Big\} .$$

où $\mathcal{F}\theta$ représente la transformée de Fourier de θ.

4 Résultats principaux

Nous ajoutons dans cette section un indice θ aux symboles d'espérance et de probabilité afin de rendre explicite la dépendance en la densité inconnue θ. Le théorème ci-dessous garantit que l'estimateur proposé converge à une vitesse logarithmique dès lors que la densité θ appartient à l'une des classes $\Theta(K, L, s, m)$.

Theorème 1. *Soient $\boldsymbol{X}$ le processus de shot-noise exponentiel défini par (1), K, L, m des nombres strictement positifs et $s > 1/2$. Pour tout choix de constantes strictement positives γ, δ satisfaisant*

$$\gamma < 1/2 - \delta \frac{\lambda K^{2/(4+m)}}{4\alpha} ,$$

on pose $h_n \triangleq (\delta \log(n))^{-1/2}$, $\kappa_n = n^{\gamma - 1/2}$ et on définit $\hat{\theta}_n$ par (8). Alors, pour tout $M \ge \sqrt{L}/\pi$,

$$\limsup_n \sup_{\theta \in \Theta(K,L,m,s)} \mathbb{P}_\theta \left(\|\theta - \hat{\theta}_n\| > M h_n^s \right) = 0 .$$

Ce résultat établit une vitesse de convergence uniforme sur les espaces de Sobolev pour l'estimateur proposé. Des résultats de ce type ont été obtenus pour des processus de Lévy ([6], [2]) dont les accroissements sont indépendants et suivent des lois infiniment divisibles. Une difficulté spécifique au cas du shot-noise est que les variables de loi infiniment divisible observées ne sont pas indépendantes.

On voit d'après (8) que l'estimateur $\hat{\theta}_n$ dépend seulement des paramètres h_n et κ_n qui à leur tour sont choisis en fonction de K et m. Il est d'ailleurs intéressant de noter que l'exposant de Sobolev s n'intervient pas dans la définition de $\hat{\theta}_n$. En pratique, la connaissance de K et m sont des hypothèses relativement faibles : dans les applications en physique nucléaire, Y est borné presque sûrement par une constante connue. Cependant, une estimation trop élevée de la borne supérieure de $K^{2/(4+m)}$ entraîne nécessairement un choix de δ plus petit, et donc un plus grand h_n : autrement dit, une vitesse de convergence plus faible. Ainsi, il serait intéressant de pouvoir choisir ces paramètres en fonction des observations $X_1, \cdots, X_n$. Pour y parvenir, nous exploitons la formule connue des cumulants d'un shot-noise (cf. [1] par exemple) qui donne

$$Var(X_0) \triangleq \sigma_X^2 = \lambda \mathbb{E}_\theta \left[Y^2\right] \int_0^\infty \mathrm{e}^{-2\alpha t}\mathrm{d}t = \frac{\lambda}{2\alpha} \mathbb{E}_\theta \left[Y^2\right] .$$

Grâce à cette relation, nous pouvons formuler une version adaptée du théorème précédent avec $\gamma = 1/8$, $m = 1$ et un choix de δ en fonction des données de sorte que l'estimateur ne dépende plus de la connaissance de K.

Theorème 2. *Soient $\boldsymbol{X}$ le processus de shot-noise exponentiel défini par (1), K, L, m des nombres strictement positifs et $s > 1/2$. On définit*

$$\hat{\mu}_n^{(2)} \triangleq \frac{2\alpha}{\lambda} \hat{\sigma}_{X,n}^2$$

où $\hat{\sigma}_{X,n}^2$ est la variance empirique des observations $X_1, \cdots, X_n$. Posons

$$\hat{\delta}_n \triangleq \frac{3\alpha}{2\lambda(\hat{\mu}_n^{(2)} + 1)} ,$$

puis

$$h_n = \left(\hat{\delta}_n \log(n) \right)^{-1/2} \quad et \quad \kappa_n = n^{-3/8} .$$

Alors, pour tout $M \ge \sqrt{L}/\pi$, on a

$$\limsup_n \sup_{\theta \in \Theta(K,L,s,1)} \mathbb{P}_\theta \left(\|\theta - \hat{\theta}_n\| > M h_n^s \right) = 0 . \qquad (9)$$

Il est intéressant de noter que la finitude du moment d'ordre 5 de Y est une condition suffisante au contrôle uniforme de la variance empirique $\hat{\sigma}_{X,n}^2$ sur la classe de fonctions $\Theta(K, L, s, 1)$. Comme dans le théorème 1, la vitesse de convergence obtenue est logarithmique mais cette fois, la « constante » $\hat{\delta}_n$ (qui converge vers $2\alpha/(2\lambda(1+\mathbb{E}_\theta[Y2]))$ est obtenue à partir des données $X_1, \cdots, X_n$. Comme nous le verrons, la vitesse lente ne pose pas trop de problèmes dans les conditions pratiques où un très grand nombre d'observations sont disponibles.

5 Algorithme

Dans le cadre des applications en physique nucléaire, la taille de l'échantillon disponible atteint souvent plusieurs millions. Afin d'estimer d'une manière rapide et efficace la loi des marques, nous avons développé une procédure d'éstimation semblable à celle proposée en (8) dans laquelle l' évaluation de la fonction caractéristique empirique est substituée par la transformation discrète de Fourier d'un histogramme basé sur les observations $X_1, \cdots, X_n$. L'algorithme utilisée est présenté ci-dessous.

Algorithm 1: Estimation de la loi des marques

Données : Observations $X_1, \cdots, X_n$, λ, α.

1. Choix de $\Delta > 0$.
2. Calcul de l'histogramme $(H_i)_{i\in\mathbb{Z}}$ d'intervalle Δ basé sur les observations $X_1, \cdots, X_n$:
$$H_i = \{X_k \in [\Delta i; \Delta(i+1)]\}/n$$
3. Soit $(dH_i)_{i\in\mathbb{Z}}$ définie par $dH_i = \Delta i H_i$
4. Pour $N \triangleq 2^{16}$, on calcule $X\phi = \mathrm{FFT}(H, N)$, $dX\phi = \mathrm{FFT}(dH, N)$ Pour $k = 1, \cdots, N$, on a $(X\phi)_k \simeq \hat{\varphi}_n\left(-\frac{2\pi(k-1)}{N\Delta}\right)$ et $(dX\phi)_k \simeq -i\hat{\varphi}'_n\left(-\frac{2\pi(k-1)}{N\Delta}\right)$
5. Calcul de $\hat{\varphi}_{Y,n}$:
$\hat{\varphi}_{Y,n}\left(\frac{2\pi(k-1)}{N\Delta}\right) = 1 + \frac{2\pi\alpha(k-1)}{N\Delta\lambda}\frac{\overline{dX\phi_k}}{dX\phi_k}$ puis
$\hat{\theta}_n = \frac{2}{N\Delta} FFT(\hat{\varphi_Y}_n, N)$

6 Résultats numériques

Pour donner quelques ordres de grandeur en instrumentation nucléaire, les détecteurs possèdent une résolution temporelle d'environ 10 MHz tandis que leur réponse impulsionnelle associée devient négligeable en au plus quelques microsecondes. Dans un contexte de fort empilement, cette atténuation rapide du courant électrique n'est pas perceptible du fait d'un taux d'interarrivée entre deux photons relativement haut, d'environ 10^6 arrivées par seconde. Afin d'illustrer la performance de l'estimateur $\hat{\theta}_n$ défini en (8), nous fournissons ci-dessous deux exemples d'estimation de la loi des marques, à $\lambda = 5.10^5$ et $\alpha = 3.10^4$ fixés.

Cas d'un mélange de gaussiennes : Nous considérons le cas où la loi des marques suit une mélange de trois gaussiennes définies par

$$p = [0.3\ 0.5\ 0.2] \quad , \quad \mu = [4\ 12\ 22] \quad , \quad \sigma = [1\ 1\ 0.5]$$

où p, μ, σ représentent respectivement le poids, la moyenne et la variance de chaque gaussienne.

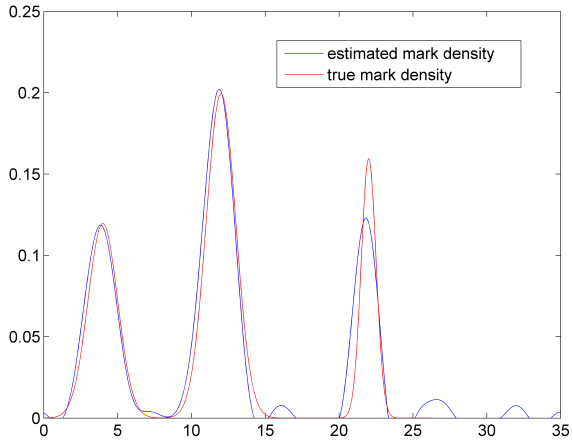

FIGURE 2 – Estimation d'un mélange de gaussiennes

Cas d'un mélange de lois Gamma(a, β) **:** la loi des marques suit un mélange de trois lois gammas indépendantes définies par

$$a = 2 \quad , \quad p = [0.2\ 0.3\ 0.5] \quad , \quad \beta = [3\ 6\ 12]$$

où a est le paramètre de forme commun et p, β le vecteur des poids et des paramètres d'échelle de chaque loi gamma.

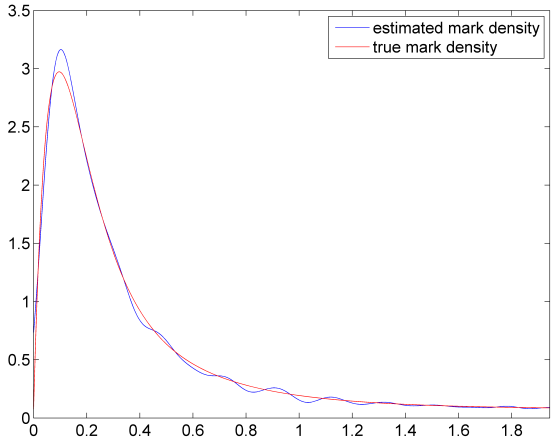

FIGURE 3 – Estimation d'un mélange de lois gamma

Références

[1] Daley, Daryl J and Vere-Jones, David, An introduction to the theory of point processes, *Springer*, (1988)

[2] Gugushvili, Shota, Nonparametric estimation of the characteristic triplet of a discretely observed Lévy process, *Journal of Nonparametric Statistics*, (2009)

[3] Iksanov, Aleksander M and Jurek, Zbigniew J, Shot noise distributions and selfdecomposability, *Taylor & Francis*, (2003)

[4] Jongbloed, Geurt and Van Der Meulen, Frank H and Van Der Vaart, Aad W and others, Nonparametric inference for Lévy-driven Ornstein-Uhlenbeck processes, *Bernoulli*, (2005)

[5] Knoll, Glenn F, Radiation detection and measurement, *Wiley New York*, (1989)

[6] Neumann Michael H and Reiß, Markus, Nonparametric estimation for Lévy processes from low-frequency observations, *Bernoulli*, (2009).