Désempilement non-paramétrique de la densité d'un processus shot-noise

Paul Ilhe, François Roueff, Eric Moulines, Antoine Souloumiac

Désempilement non-paramétrique de la densité d'un processus shot-noise

Paul Ilhe [1,3] , Francois Roueff [1] , Éric Moulines [2] & Antoine Souloumiac [3]

[1] *LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France prenom.nom@telecom-paristech.fr*

[2] *École Polytechnique, Centre de Mathématiques Appliquées- UMR 7641, Route de Saclay 91128 Palaiseau, eric.moulines@polytechnique.edu*

[3] *CEA, LIST, 91191 Gif-sur-Yvette Cedex, France, prenom.nom@cea.fr*

Résumé. Nous proposons une méthode d'estimation non-paramétrique rapide pour estimer la distribution des marques d'un processus de shot-noise en présence d'*empilement* à partir d'un nombre potentiellement important d'observations mais échantillonnées à basse fréquence. À partir d'une équation fonctionnelle liant la densité des marques à la fonction caractéristique des observations et sa dérivée, nous proposons un estimateur de cette densité en utilisant la base des B-splines. Nous discutons de l'implémentation pratique de l'algorithme et illustrons les performances de l'estimateur sur des données simulées.

Mots-clés. Shot-noise, B-spline, problème inverse, spectrométrie γ

Abstract. In this paper, we propose an efficient method to estimate in a nonparametric fashion the marks' density of a shot-noise process in presence of *pileup* from a sample of low-frequency observations. Based on a functional equation linking the marks' density to the characteristic function of the observations and its derivative, we propose a new time-efficient method using B-splines to estimate the density of the underlying γ-ray spectrum, which is able to handle large datasets used in nuclear physics. A discussion on the numerical computation of the algorithm and its performances on simulated data are provided to support our findings.

Keywords. Shot-noise, B-splines, inverse problem, γ spectrometry

Introduction

Dans ce papier, on s'intéresse à un problème inverse non-linéaire de science nucléaire. Dans le cadre de la spectrométrie γ, un faisceau de photons, provenant de la désintégration de noyaux radioactifs, frappe un détecteur. Chaque interaction se traduit par une impulsion électrique, dont la forme dépend de la chaîne d'instrumentation et du type de détecteur, et dont l'amplitude dépend de l'énergie, totale ou partielle [5], déposée par le photon. Le courant électrique observé est modélisé par un processus shot-noise $(X_t)_{t\geq 0}$ défini par

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

pour lequel on suppose que

(SN-1) $\sum_k \delta_{T_k,Y_k}$ correspond à un processus ponctuel de Poisson homogène d'intensité λ de temps d'arrivée $T_k \in \mathbb{R}$ et de marques Y_k indépendantes des temps d'arrivés et i.i.d. à valeurs positives admettant une densité f appartenant à $\mathcal{H}^2([0,M])$ défini pour $M>0$ par

$$\mathcal{H}^2([0,M]) := \{f : [0,M] \to \mathbb{R}, f, f' \text{ absolument continues et } \int_0^M |f''(t)|^2 \,\mathrm{d}t < \infty\} .$$

(SN-2) la réponse impulsionnelle h du détecteur est causale, intégrable et satisfait

$$h(t) \underset{t\to\infty}{=} O\left(\mathrm{e}^{-t}\right) .$$

Un tel processus est bien défini dès lors que la densité f et la réponse impulsionnelle h satisfont conjointement la condition $\int \min\left(1, |y\, h(s)|\right) f(y)\mathrm{d}y\mathrm{d}s < \infty$. La figure ci-dessous illustre la trajectoire d'un shot-noise d'intensité $\lambda = 3$, de réponse impulsionnelle $h : t \to 10t\mathrm{e}^{-10t}\mathbb{1}_{t\geq 0}$ et dont les marques sont distribuées suivant un mélange de gaussiennes.

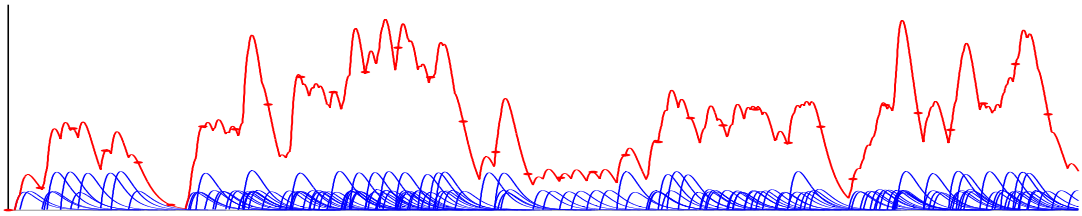

FIGURE 1 – illustration de la trajectoire d'un processus shot-noise (en rouge) et des contributions singulières (en bleu).

Les processus shot-noise constituent des modèles naturels dans l'étude des systèmes dynamiques pouvant être représentés comme la convolution d'un filtre et d'un processus ponctuel marqué. En spectrométrie γ, les marques $(Y_k)_k$ représentent des émissions d'énergie γ caractéristiques d'un ou plusieurs radionucléides. Une estimation du *flux* λf permet alors d'identifier le ou les radionucléides présents en le comparant à des spectres d'énergie connus ainsi que de quantifier l'intensité des émissions γ.

D'un point de vue statistique, nous étudions le problème inverse suivant : à partir d'un échantillon de pas fini $X_\delta, \ldots, X_{n\delta}$ de taille n du shot-noise défini par (1) et sous l'hypothèse que la réponse impulsionnelle est connue, notre objectif consiste à estimer la fonction

$$f_\lambda := \lambda f .$$

Le problème de l'identification statistique des processus shot-noise a été initié dans [6] et, plus récemment, dans [4]. De plus, les méthodes basées sur la vraisemblance ne sont pas applicables car la densité de la loi marginale de X_0, si elle existe [1], est trop complexe à calculer. En outre, l'échantillonnage a basse fréquence du signal rend les lois jointes de $(X_t)_t$ difficilement exploitables, de telle sorte que nous considérons uniquement des méthodes reposant sur la marginale du processus. Celle-ci étant infiniment divisible, il est tentant d'estimer le triplet caractéristique de Lévy associé (voir [8] pour les processus de Lévy). Cependant, dans notre cas, les observations $X_1, \ldots, X_n$ ne sont ni i.i.d., ni même Markoviennes comme dans [4] pour $h(t) = \mathrm{e}^{-\alpha t}\mathbb{1}_{t\geq 0}$.

Procédure d'estimation

Nous proposons ici une méthode plus générale que [4] permettant d'estimer la densité f_λ. Sans perte de généralité (quitte à modifier l'unité de temps et l'intensité), nous supposons que la fréquence d'échantillonnage $\delta^{-1} = 1$ et le support de f est inclus dans l'intervalle $[0, 1]$. Il est possible d'exprimer analytiquement la fonction caractéristique φ de la loi marginale de X_0. Pour tout réel u, on a

$$\varphi(u) := \mathbb{E}\left[\mathrm{e}^{\mathrm{i}uX_0}\right] = \exp\left(\int_0^\infty \int_0^1 \left[\mathrm{e}^{\mathrm{i}uxh(s)} - 1\right] f_\lambda(x)\mathrm{d}x\mathrm{d}s\right) . \tag{2}$$

Sous l'hypothèse que $\int |x| f_\lambda(x)\mathrm{d}x < \infty$, une dérivation de (2) conduit à l équation fonctionnelle

$$g(u) = K_h\left[f_\lambda\right](u) \quad , \quad u \in \mathbb{R} ,$$

où $g = \varphi'/\varphi$ et K_h correspond à un opérateur linéaire dépendant seulement de h défini par

$$K_h : f \to \left[u \to \int_0^\infty \int_0^1 \mathrm{i}xh(s)\mathrm{e}^{\mathrm{i}uxh(s)} f(x)\mathrm{d}x\mathrm{d}s\right] .$$

En pratique, on dispose seulement d'une version bruitée $\hat{g}_n$ de la fonction g obtenue par une méthode "plug-in" où φ et sa dérivée sont respectivement remplacées par la fonction caractéristique empirique associé aux observations $X_1, \ldots, X_n$ et sa dérivée. Le cadre général de ce type de problème est étudié dans [7] : les auteurs supposent qu'il est possible de construire des estimateurs sans biais de la fonction g ainsi que d'exprimer la décomposition en valeurs singulières de l'opérateur K_h. Ces deux hypothèses n'étant, en général, pas vérifiées, nous proposons de construire un estimateur de f_λ à partir d'une grille finie de N points d'évaluation $\{u_i, i \in 1, \ldots, N\}$ pour lesquels on considère les équations

$$\hat{g}_n(u_i) = K_h\left[f_\lambda\right](u_i) + \epsilon_n(u_i), \quad i \in \{1, \ldots, N\} ,$$

où $\epsilon(u_1), \ldots, \epsilon(u_N)$ sont les erreurs d'estimation de $g(u_i)$ par $\hat{\varphi}'_n(u_i)/\hat{\varphi}_n(u_i)$ pour $i = 1, \ldots, N$. En particulier, le résultat suivant assure que le vecteur normalisé des erreurs peut être approché par un vecteur gaussien centré et corrélé.

Proposition 1. *Soit $\lambda > 0$ et $X_1, \cdots, X_n$ des échantillons du processus shot-noise défini par* (1) *satisfaisant les hypothèses (*SN*-1), (*SN*-2) et $\mathbb{E}\left[|Y_1|^{4+\eta}\right] < \infty$ pour un certain $\eta > 0$. Soient φ et φ' respectivement la fonction caractéristique de X_1 et sa dérivée. Alors, pour tout entier N et tout N-uplet de réels $\underline{u} = (u_1, \cdots, u_N)$, on a*

$$\sqrt{n}\left(\epsilon_n(u_1), \ldots, \epsilon_n(u_N)\right) \Rightarrow \tilde{Z}(\underline{u}) \ ,$$

où $\tilde{Z}(\underline{u})$ est un vecteur gaussien centré à valeurs complexes et de matrice de covariance notée $W(\underline{u})$.

Par souci de concision, nous omettons ici l'expression exacte de $W(\underline{u})$. Il est possible de l'estimer de différentes manières : "plug-in", méthodes des moments généralisée ou bootstrap. Cette dernière est utilisée dans les applications numériques qui suivent.

Paramétrisation par des B-splines

Comme f_λ appartient à l'espace fonctionnel Hilbertien $\mathcal{H} = \mathcal{H}^2([0,1])$, le problème d'estimation se traduit de la manière suivante

$$\hat{f}_{n,\lambda}(\alpha) := \underset{f \in \mathcal{H}}{\operatorname{argmin}} \ \|\hat{g}_n(\underline{u}) - K_h[f](\underline{u})\|^2_{\hat{W}_n^{-1}} + \alpha \|f\|^2_{\mathcal{H}} \ , \tag{3}$$

où $\|f\|^2_{\mathcal{H}} := \int_0^1 |f''(t)|^2 \, \mathrm{d}t$. Nous proposons d'estimer f_λ en l'approchant par des fonctions appartenant à la base des B-splines, suggérée par une similarité de notre problème avec [2]. Étant donné deux entiers q, k et un ensemble de points $\underline{\mathbf{t}} = (t_1, \ldots, t_k)$ appelés *nœuds* tels que $t_1 < \cdots < t_k$, l'espace $\mathcal{S}^q_{\underline{\mathbf{t}}}$ *des B-splines de degré q associé aux nœuds* $\underline{\mathbf{t}}$ est l'ensemble des fonctions v qui s'écrivent sous la forme

$$v(t) = \sum_{j=1}^{k+q} \omega_j B^q_{j,k}(t) =: \omega^T \mathbf{B}(t) \ ,$$

pour $(\omega_1, \ldots, \omega_{k+q})$ un vecteur de réels et où les fonctions $(B^q_{1,k}, \ldots, B^q_{k+q,k})$ satisfont les relations

$$B_{j,l}(t) := \begin{cases} 1 & \text{si} \quad t_j \leqslant t < t_{j+1} \\ 0 & \text{sinon} \end{cases} \ , \quad \text{pour } l = 0 \ ,$$

$$B_{j,l}(t) := \frac{t - t_j}{t_{j+l} - t_j} B_{j,l-1}(t) + \frac{t_{j+l+1} - t}{t_{j+l+1} - t_{j+1}} B_{j+1,l-1}(t), \quad \text{pour } l \in \{1, \ldots, q\}.$$

Trois raisons motivent une telle paramétrisation : d'une part, les fonctions $(B_{j,q})_{1 \leq j \leq k+q}$ (pour k, q et $\mathbf{t}$ donnés) sont à support compact (ce qui permet dans les applications nucléaires de modéliser efficacement la résolution du détecteur, connue à priori), d'autre part il existe une relation linéaire entre la dérivée d'une B-spline d'un certain ordre et les B-splines de l'ordre inférieur. En outre, l'espace des B-splines permet d'approcher les

fonctions de $\mathcal{H}$ à une précision donnée. Ainsi, il est possible de construire un estimateur de f_λ capable de s'adapter à la régularité locale de la fonction cible via

$$\hat{f}_{n,\lambda}(\alpha) := \hat{\omega}_{n,\lambda}(\alpha)^T \mathbf{B} \ .$$

où $\hat{\omega}_{n,\lambda}(\alpha)$ est solution du problème de minimisation quadratique

$$\hat{\omega}_{n,\lambda}(\alpha) = \underset{\omega \in \mathbb{R}_+^{k+q}}{\operatorname{argmin}} \|\hat{g}_n(\underline{u}) - K_h[\mathbf{B}](\underline{u})^T \omega\|^2_{\hat{W}_n^{-1}} + \alpha \omega^T D_2^{k+q} \omega \ , \tag{4}$$

et où D_2^{k+q} est une matrice dépendant seulement de k et q. Afin de choisir la pénalité α, le problème de minimisation (4) appartient à la classe de ceux abordés dans [9]. Pour un nombre fini de pénalités $\alpha_1, \ldots, \alpha_p$, on choisit celle qui minimise la fonction de validation croisée généralisée

$$\mathrm{CV}(\alpha) := \frac{\|\hat{g}_n(\underline{u}) - A_n(\alpha)\hat{g}_n(\underline{u})\|^2_{\hat{W}_n^{-1}}}{(1-\mathrm{Tr}(A_n(\alpha))/N)^2} \text{ où } A_n(\alpha) := K_h[\mathbf{B}] \left(K_h[\mathbf{B}]^\star \hat{W}_n^{-1} K_h[\mathbf{B}] + \alpha D_2^{k+q} \right)^{-1} K_h[\mathbf{B}]^\star \hat{W}_n^{-1}.$$

Implémentation et résultats numériques

La procédure décrite dans la section précédente et résumée dans le pseudo-code ci-dessous comporte un certain nombre d'hyperparamètres pour lesquels nous proposons un choix efficace sur des données simulées.

Algorithm 1 Estimateur par B-splines

Entrées : Observations $X_1, \cdots, X_n$ du shot-noise,

1. Choix de $\Delta_n > 0$. Calcul de l'histogramme $(H_i)_{i\in\mathbb{Z}}$ d'intervalle Δ_n basé sur les observations $X_1, \cdots, X_n$: $H_i := \{X_k \in [\Delta_n i; \Delta_n(i+1)]\}/n$ et de $(dH_i)_{i\in\mathbb{Z}}$ définie par $dH_i = \Delta i H_i$.
2. Choix de $N := 2^{\log(n)}$ et calcul de $\phi = \mathrm{FFT}(H, N)$, $d\phi = \mathrm{FFT}(dH, N)$. Pour $k = 1, \cdots, N$, on définit $\hat{g}_n\left(\frac{2\pi(k-1)}{N\Delta}\right) := \frac{\overline{d\phi_k}}{\phi_k}$.
3. Choix du nombre de nœuds $k = N$, du degré $q = 3$ et calcul de la matrice de design $(K_h[B_{j,q}](2\pi(i-1)/N\Delta))_{i,j}$ pour $i \in \{1, \ldots, N\}$, $j \in \{1, \ldots k+q\}$.
4. Calcul de $\hat{W}_n$ par Bootstrap.
5. Estimation des $\hat{\omega}_n(\alpha)$ pour $\alpha \in \{2^{-15}, \ldots, 2^{15}\}$ via (4) et choix par validation croisée généralisée.

Nous illustrons la performance des estimateurs sur des données simulées pour un processus shot noise d'intensité $\lambda = 3$, de réponse impulsionnelle $h : t \to 10te^{-10t}1_{t\geq 0}$ et dont les marques possèdent une densité $f = 0{,}7\mathcal{N}(0{,}3; 2{,}5.10^{-3}) + 0{,}3\mathcal{N}(0{,}6; 2{,}5.10^{-3})$, où $\mathcal{N}(\mu; \sigma^2)$ correspond à la densité d'une variable aléatoire gaussienne de moyenne μ et de variance σ^2. Les figures ci-dessous illustrent les différentes densités obtenues ainsi qu'un diagramme en boîte de l'erreur quadratique moyenne intégrée entre $\hat{f}_{n,\lambda}$ et f_λ pour 100 itérations indépendantes de l'algorithme présenté et différentes tailles d'échantillons.

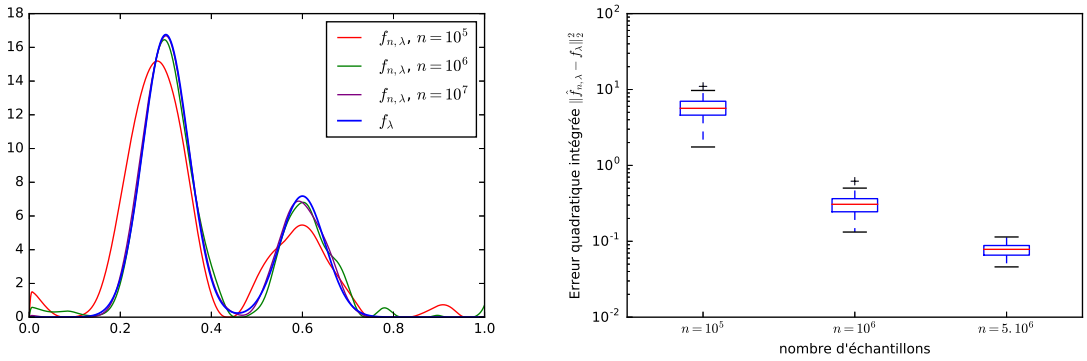

FIGURE 2 – De gauche à droite : estimation du flux f_λ pour $n \in \{10^5, 10^6, 10^7\}$ - boxplot de l'erreur $\|\hat{f}_{n,\lambda} - f_\lambda\|_2^2$ pour $n \in \{10^5, 10^6, 5.10^6\}$.

Outre que $\hat{f}_{n,\lambda}$ converge vers f_λ, nous observons que l'estimateur retrouve les modes de la fonction f_λ, même en présence d'un nombre restreint d'observations. Ceci présente un intêrét dans les applications de spectrométrie γ, car permet d'identifier les pics caractéristiques d'un radionucléide donné.

Bibliographie

[1] Biermé, H., Desolneux, A. (2012). A Fourier approach for the level crossings of shot noise processes with jumps. *Journal of Applied Probability*, 49(1), 100-113.
[2] Cardot, H. (2002). Spatially adaptive splines for statistical linear inverse problems. *Journal of Multivariate Analysis*, 81(1).
[3] Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied and computational harmonic analysis*, 2(2).
[4] Ilhe, P., Moulines, E., Roueff, F. et Souloumiac, A. (2015). Nonparametric estimation of mark's distribution of an exponential Shot-noise process. *Electronic Journal of Statistics*.
[5] Knoll, Glenn F. (1989). Radiation detection and measurement. *Wiley New York*
[6] Macchi, O., Picinbono, B. C. (1972). Estimation and detection of weak optical signals. *Information Theory, IEEE Transactions on*, 18(5), 562-573.
[7] Mair, B. A., Ruymgaart, F. H. (1996). Statistical inverse estimation in Hilbert scales. *SIAM Journal on Applied Mathematics*, 56(5).
[8] Neumann, M. H., Reiß, M. (2009). Nonparametric estimation for Lévy processes from low-frequency observations. *Bernoulli*, 15(1), 223-248.
[9] Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*.