

Distribution hybride pour la modélisation de données à deux queues lourdes: Application sur les données neuronales

Nehla DEBBABI^{1,2}, Marie KRATZ³, Sadok EL ASMI², Mamadou MBOUP^{1,4}

¹CRESTIC, Faculté des Sciences Exactes et Naturelles, Université de Reims Champagne Ardenne, France

²Laboratoire COSIM, École Supérieure des Communications de Tunis, Université de Carthage, Tunisie

³ESSEC Business School, CREAR risk research center, Cergy-Pontoise, France

⁴Non-A, Inria Lille Nord Europe

Nehla.Debbabi@supcom.tn, Kratz@essec.edu

Elasmi@supcom.rnu.tn, Mamadou.Mboup@univ-reims.fr

Résumé – Ce travail propose un modèle hybride pour modéliser des données à deux queues lourdes. Le modèle proposé est une distribution à trois composantes pondérées : une distribution Gaussienne, pour modéliser le comportement moyen des données, liée à deux distributions de Pareto généralisées pour modéliser les comportements extrêmes. Un algorithme itératif et non supervisé est ensuite proposé pour une estimation fiable des points de jonctions entre les trois distributions, les paramètres de ces dernières ainsi que les poids affectés à chaque composante du modèle hybride. Une application sur des données neuronales réelles issues d'un enregistrement extracellulaire, est développée pour évaluer les performances du modèle proposé, comparé à la distribution normale.

Abstract – A new hybrid model for two heavy tailed data modelling is proposed in this study. The proposed model is a weighted three-components distribution: a Gaussian distribution, to model the mean behavior of the data, linked to two generalized Pareto distributions, modelling the extreme ones. An unsupervised iterative algorithm is then developed to estimate accurately the junction points between the three distributions, the parameters of these latter as well as the weights of the hybrid model. An application on real extracellular neural recordings is developed to evaluate the performance of the proposed hybrid model, compared to the normal distribution.

1 Introduction

Les données d'observation de phénomènes complexes sont généralement non-homogènes et multi-composantes [1, 2, 3]. Trouver un modèle adéquat offrant le meilleur ajustement de ce type de données est un sujet d'importance majeure, qui intéresse les chercheurs depuis des décennies [4, 5]. Dans le présent travail, nous abordons ce sujet pour un cas particulier de données à deux queues lourdes. Ce type de données est rencontré dans différents domaines où les queues de distribution, contenant des valeurs extrêmes, portent souvent une information importante. À titre d'exemple, on cite le cas des rendements de l'indice S&P 500 en finance [6], où les valeurs extrêmes indiquent un risque du marché, ou encore le cas des données neuronales issues d'un enregistrement extracellulaire en neuroscience [7], où les extrêmes indiquent la manifestation de Potentiels d'Action (ou spikes). Modéliser ce type de données directement par une gaussienne ne suffit pas : les deux queues sont sous-estimées [6, 8]. Les méthodes non paramétriques représentent une bonne alternative mais ne donnent aucune information sur les extrêmes. Nous privilégions pour cette raison les méthodes paramétriques comme le mélange de noyaux gaussiens, la distribution q-gaussienne ou encore la distribution de

Pareto généralisée (GPD). Notons que notre objectif n'est pas seulement d'ajuster les données par une distribution adéquate mais c'est aussi, et surtout, de pouvoir interpréter les extrêmes. Une façon de le faire est d'exploiter le théorème de Pickands [9] de la Théorie des Valeurs Extrêmes (TVE) en modélisant les extrêmes au delà d'un seuil bien choisi par une GPD [10, 11]. Cependant, et contrairement aux autres méthodes, la TVE ne permet de bien estimer que les extrêmes mais pas toute la distribution. Pour garder l'avantage de la TVE mais avec une modélisation complète, nous proposons dans ce travail de séparer le comportement moyen des extrêmes. En effet, comme dans [6], nous proposons de modéliser le comportement moyen des données par une gaussienne *via* le Théorème Central Limite (TCL), et les comportements extrêmes par deux GPD différentes *via* le théorème de Pickands [9]. Pour un meilleur ajustement des données, et pour plus de souplesse au niveau des queues, contrairement à [6], nous attribuons un poids différent à chaque morceau. Les trois morceaux sont connectés entre eux en deux points de jonctions, où on impose des conditions de classe \mathcal{C}^1 . Ces conditions permettent non seulement d'avoir une fonction de densité de probabilité (pdf) lisse mais aussi de réduire le nombre de paramètres à estimer. Pour une estimation fiable des paramètres du modèle hybride ainsi obtenu, nous

proposons un algorithme itératif non supervisé, basé sur le principe de l'algorithme introduit dans [12]. Pour chaque itération, l'algorithme proposé estime les paramètres du modèle hybride en résolvant des problèmes d'optimisation au sens des moindres carrés en utilisant la méthode de Levenberg Marquardt [13, 14]. Nous mentionnons qu'à part l'estimation fiable des paramètres du modèle, l'algorithme est non supervisé permettant d'éviter les problèmes rencontrés par les méthodes existantes de détermination du seuil satisfaisant le théorème de Pickands, qui sont graphiques et non automatiques [11, 10, 15]. Pour évaluer les performances du modèle proposé, une application sur des signaux neuronaux issus d'un enregistrement extracellulaire est développée. Les résultats obtenus montrent que le modèle hybride fournit un bon ajustement des données neuronales, comparé à la distribution normale.

Le reste du papier est organisé comme suit : La première section décrit le modèle hybride proposé. Le pseudo-code de l'algorithme itératif est donné dans la section 3. Les résultats de simulations sont discutés dans la section 4 et la conclusion de ce travail est donnée à la fin.

2 Modèle hybride proposé

La distribution hybride proposée dans ce travail est une distribution gaussienne dont les queues, gauche et droite, ont été remplacées par deux GPD. La densité de probabilité correspondante est ainsi la fonction par morceaux définie par :

$$h(x; \theta) = \gamma_1 g(u_1 - x; \xi_1, \beta_1) \mathbb{1}_{\{x \leq u_1\}} + \gamma_2 f(x; \mu, \sigma) \mathbb{1}_{\{x \in]u_1, u_2\}} + \gamma_3 g(x - u_2; \xi_2, \beta_2) \mathbb{1}_{\{x > u_2\}}. \quad (1)$$

où $\mathbb{1}_{\{\cdot\}}$ est la fonction indicatrice, u_1 et u_2 représentent les points de jonction ou encore les seuils à partir desquels la distribution change de comportement et $\theta = [u_1, u_2, \mu, \sigma, \xi_1, \xi_2]$. Les paramètres γ_1 , γ_2 et γ_3 désignent les poids et g est la pdf de la GPD, donnée par (2) de paramètres ξ_1 et β_1 pour la queue gauche (resp. ξ_2 et β_2 pour la queue droite). Enfin, $f(\cdot; \mu, \sigma)$ est la densité de la loi normale de moyenne μ et d'écart type σ .

$$g(x; \xi, \beta) = \begin{cases} \frac{1}{\beta} (1 + \frac{\xi}{\beta} x)^{-1 - \frac{1}{\xi}}, & \text{si } \xi \neq 0 \\ \frac{1}{\beta} e^{-\frac{x}{\beta}}, & \text{si } \xi = 0 \end{cases}, \quad (2)$$

$$x \in \begin{cases} [0, \infty), & \text{si } \xi \geq 0 \\ [0, -\frac{\beta}{\xi}], & \text{si } \xi < 0 \end{cases}, \text{ et } (\xi, \beta) \in \mathbb{R} \times \mathbb{R}_+^*.$$

où ξ désigne l'indice de queue et β le paramètre d'échelle. Puisque h est une pdf, sa masse totale doit être égale à 1. Cette contrainte se traduit par la relation

$$\gamma_1 + \gamma_2 [F(u_2; \mu, \sigma) - F(u_1; \mu, \sigma)] + \gamma_3 = 1, \quad (3)$$

où F désigne la fonction de répartition (cdf) de la loi normale. Dans le but d'avoir une pdf lisse, on impose des conditions de continuité et de dérivabilité aux points de jonction. Ces conditions permettent non seulement de réduire le nombre de paramètres à estimer mais aussi, une fois les seuils sont déterminés, de pouvoir les affiner empiriquement (voir section 4.2).

Les relations ainsi obtenues aux points u_1 et u_2 avec (3) permettent d'exprimer explicitement les paramètres β_1 , β_2 , γ_1 , γ_2 et γ_3 en fonction du reste des paramètres comme suit

$$(E) \begin{cases} \beta_1 = -\frac{(1 + \xi_1)\sigma^2}{u_1 - \mu}; & \beta_2 = \frac{(1 + \xi_2)\sigma^2}{u_2 - \mu} \\ \gamma_2 = [\beta_1 f(u_1; \mu, \sigma) + \beta_2 f(u_2; \mu, \sigma) + \int_{u_1}^{u_2} f(x; \mu, \sigma) dx]^{-1} \\ \gamma_1 = \beta_1 \gamma_2 f(u_1; \mu, \sigma); & \gamma_3 = \beta_2 \gamma_2 f(u_2; \mu, \sigma) \end{cases}$$

La cdf de la loi hybride est donnée par :

$$H(x; \theta) = \gamma_1 [1 - G(u_1 - x; \xi_1, \beta_1)] \mathbb{1}_{\{x \leq u_1\}} + [\gamma_1 + \gamma_2 [F(x; \mu, \sigma) - F(u_1; \mu, \sigma)]] \mathbb{1}_{\{x \in]u_1, u_2\}} + [1 - \gamma_3 [1 - G(x - u_2; \xi_2, \beta_2)]] \mathbb{1}_{\{x > u_2\}} \quad (4)$$

où G représente la cdf de la GPD.

Nous mentionnons qu'en se basant sur (3) et (4), il est facile de démontrer que $(\gamma_1, \gamma_3) \in [0, 1]^2$ et que $\gamma_2 \geq 0$.

Un exemple de la pdf hybride est donné par la figure 1. Cette figure illustre bien la différence entre les queues de la pdf hybride (en continu) et celles de sa composante normale (en pointillé) ainsi que la signification physique des coefficients de pondération. Comme il est montré dans cette dernière, le paramètre γ_1 (resp. γ_3), représentant l'aire gris-clair (resp. l'aire gris-foncé), n'est autre que la probabilité que la population soit en dessous du seuil u_1 (resp. au delà du seuil u_2). En revanche, γ_2 , représentant deux fois l'aire blanche, est un paramètre régulateur assurant que h est une pdf.

Notons que le modèle hybride proposé généralise celui introduit dans [6]. En effet, ce dernier utilise la même pondération γ pour les trois distributions c -à- d $\gamma_1 = \gamma_2 = \gamma_3 = \gamma$. Dans ce cas, $H(u_1; \theta) = \gamma$ et $H(u_2; \theta) = 1 - \gamma$ ce qui implique que les deux seuils doivent être symétriques par rapport à la médiane pour pouvoir appliquer ce modèle. Une autre limitation de la pondération uniforme est que d'après (3), $\frac{1}{3} \leq \gamma \leq \frac{1}{2}$ c -à- d la probabilité de se trouver dans les queues doit appartenir à $[\frac{1}{3}, \frac{1}{2}]$, sinon le modèle ne s'applique plus. Toutes ces limitations sont surmontées avec la pondérations non uniforme, considérée dans ce travail, où les seuils sont sans contraintes.

Afin d'évaluer le modèle hybride, nous proposons dans la section suivante un algorithme itératif non supervisé qui donne une estimation de tous les paramètres du modèle.

3 Algorithme itératif pour l'évaluation du modèle hybride proposé

L'algorithme proposé est inspiré de celui décrit dans [12]. Il est non supervisé et itératif. Pour chaque itération les paramètres du modèle sont estimés en deux étapes : 1) estimer au sens des moindres carrés les paramètres de la gaussienne et des deux indices de queue des GPD en considérant les deux seuils

trouvés dans l'itération précédente, 2) fixer les paramètres trouvés dans 1) et estimer au sens des moindres carrés les deux seuils. L'algorithme proposé, dont le pseudo-code est donné par l'algorithme 1, commence par un choix arbitraire de u_1 et u_2 et itère jusqu'à ce qu'une précision satisfaisante ou un nombre maximal k_{max} d'itérations soit atteint. Une fois θ est estimée, le reste des paramètres est déterminé via le système (E).

Algorithm 1 Pseudo-code de l'algorithme itératif proposé

1: Détermination de la fonction de répartition empirique H_n du n-échantillon $X = (X_i)_{1 \leq i \leq n}$

$$H_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}, \quad \forall x \in \mathbb{R}.$$

2: Fixer les paramètres initiaux $\tilde{u}_1^{(0)}, \tilde{u}_2^{(0)}$, et ϵ

3: Procédure itérative

$k \leftarrow 1$

tant que $((\tilde{u}_2^{(k)} - \tilde{u}_2^{(k-1)})^2 + (\tilde{u}_1^{(k)} - \tilde{u}_1^{(k-1)})^2) \geq \epsilon$ et $k < k_{max}$

a. Détermination de $\tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}, \tilde{\xi}_1^{(k)}$ et $\tilde{\xi}_2^{(k)}$ les estimateurs, respectivement, de μ, σ, ξ_1 et ξ_2

$$(\tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}, \tilde{\xi}_1^{(k)}, \tilde{\xi}_2^{(k)}) \leftarrow \underset{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^2, (\xi_1, \xi_2) \in \mathbb{R}^2}{\operatorname{argmin}} \|H(X; \theta_1^{(k)}) - H_n(X)\|_2^2$$

$$\text{où } \theta_1^{(k)} = [u_1^{(k-1)}, u_2^{(k-1)}, \mu, \sigma, \xi_1, \xi_2].$$

b. Détermination de $\tilde{u}_1^{(k)}$ et $\tilde{u}_2^{(k)}$ les estimateurs, respectivement, de u_1 et u_2

$$(\tilde{u}_1^{(k)}, \tilde{u}_2^{(k)}) \leftarrow \underset{(u_1, u_2) \in \mathbb{R} \times \mathbb{R}_+}{\operatorname{argmin}} \|H(X; \theta_2^{(k)}) - H_n(X)\|_2^2$$

$$\text{où } \theta_2^{(k)} = [u_1, u_2, \tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}, \tilde{\xi}_1^{(k)}, \tilde{\xi}_2^{(k)}]$$

$k \leftarrow k + 1$

fin

retour $\theta^{(k)} = [\tilde{u}_1^{(k)}, \tilde{u}_2^{(k)}, \tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}, \tilde{\xi}_1^{(k)}, \tilde{\xi}_2^{(k)}]$

Notons que tous les problèmes d'optimisation de cet algorithme sont résolus par la méthode de Levenberg Marquardt [13, 14].

4 Résultats numériques et discussions

Dans cette section, nous étudions les performances de l'algorithme et du modèle hybride proposés, en terme de fiabilité de l'estimation des paramètres, du bon ajustement des données ainsi que la séparation du comportement moyen des extrêmes.

4.1 Étude de performances de l'algorithme itératif

Les performances de l'algorithme sont étudiées à partir de simulations de Monte Carlo. En effet, nous simulons $p = 100$

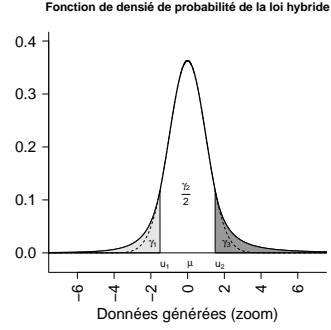


FIGURE 1 – Fonction de densité de probabilité hybride pour $\theta = [-1.5, 1.5, 0, 1, 0.2, 0.5]$.

séquences de test $\{X^{(k)} = (X_i^{(k)})_{1 \leq i \leq n}\}_{k=1}^p$ (resp. d'apprentissage $\{Y^{(k)} = (Y_i^{(k)})_{1 \leq i \leq n}\}_{k=1}^p$) de taille $n = 10^5$ et suivant la loi hybride de paramètre θ fixé. Pour tout $k \in \{1, \dots, p\}$, nous déterminons $\hat{\theta}^{(k)}$, l'estimateur de θ par l'algorithme itératif sur $\{Y^{(k)}\}$. Ensuite, nous déterminons sur les séquences de test le logarithme du rapport de vraisemblance moyen, \mathcal{C} :

$$\mathcal{C} = \frac{1}{np} \sum_{k=1}^p \sum_{i=1}^n \log \left(\frac{h(X_i^k; \theta)}{h(X_i^k; \hat{\theta}^{(k)})} \right) \quad (5)$$

Pour un exemple de $\theta = [-1, 1, 0, 1, 0.2, 0.2]$, les résultats obtenus pour $n = 10^5$ sont consignés dans le tableau 1. Ces résultats montrent une absence de biais asymptotique. La très faible valeur correspondante de $\mathcal{C} = 6.11 \cdot 10^{-5}$ suggère des performances intéressantes.

TABLE 1 – Estimation des paramètres du modèle hybride

θ	-1	1	0	1	0.2	0.2
μ_{θ_j}	-1.01	1.01	-0.00	1.00	0.20	0.20
$\sigma_{\theta_j}^2 (\times 10^{-4})$	9.82	7.45	0.53	0.82	1.91	1.56

où μ_{θ_j} and $\sigma_{\theta_j}^2$ désignent respectivement la moyenne et la variance empirique des estimations de θ_j , la composante j de θ , $j \in \{1, \dots, 6\}$.

4.2 Étude de performance du modèle hybride

Motivés par le problème de détection de spikes de neurones, nous avons choisi d'appliquer le modèle hybride sur des données neuronales réelles issues d'un enregistrement extracellulaire. Ces données représentent 4 enregistrements de 20 s chacun, filtrés entre 300 Hz et 5000 Hz, et mesurés à partir du lobe antennaire d'un criquet [16]. Nous présentons dans la figure 2, graphe A, 1 s du premier enregistrement où les pics représentent des spikes. L'histogramme relatif à ces données est tracé dans le graphe B, avec la pdf hybride ajustée (ligne rouge continue), dont l'estimée de θ est donnée par $\hat{\theta} = [-0.7792, 0.7241, -0.0288, 0.4636, 0.5082, 0.6251]$, ainsi

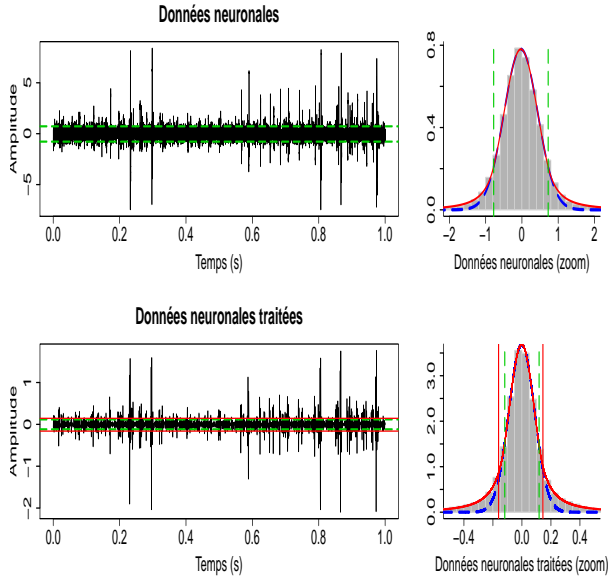


FIGURE 2 – Estimation de la pdf par la loi hybride et la loi normale sur des données neuronales réelles

que sa composante normale (ligne bleue discontinue). Nous remarquons une différence significative entre les deux pdfs au niveau des queues de distributions montrant que les données sont mieux ajustées avec le modèle hybride qu’avec la normale. De plus, les deux seuils du modèle hybride trouvés (représentés par des droites vertes discontinues dans tous les graphes de la figure 2) montrent bien qu’ils séparent les spikes du bruit. Comme il est montré dans le graphe A, le signal neuronal est bruité. Afin de réduire la puissance de ce bruit et rendre les spikes plus visibles, nous le filtrons par un filtre dérivateur, qui est un filtre algébrique linéaire [7]. La linéarité du filtre conserve la gaussianité du comportement moyen ce qui permettra d’ajuster aussi les signaux filtrés. Le graphe C montre le résultat du filtrage du signal représenté dans A. L’ajustement de ce dernier est illustré dans le graphe D. À travers les graphes C et D, nous remarquons un réhaussement des spikes par rapport au bruit ainsi qu’un bon ajustement de données (la différence entre la pdf hybride et la normale est devenue plus nette).

La figure 2 montre que le modèle hybride ajuste bien les données et sépare bien les spikes du bruit. Par contre, nous remarquons dans les graphes A et C la présence d’une zone intermédiaire entre les spikes et le bruit, où la distinction entre ces deux derniers est difficile. Afin d’avoir plus d’extrêmes que du bruit, nous proposons d’affiner les seuils d’une façon empirique, en tolérant une marge d’erreur d définie à partir de (6). Ceci est possible puisque h est lisse.

$$Err(x) = |h(x; \theta) - \gamma_2 f(x; \mu, \sigma)|, \quad \forall x \leq u_1 \text{ et } x \geq u_2. \quad (6)$$

Les seuils affinés, \tilde{u}_1 et \tilde{u}_2 , relatifs respectivement à u_1 et u_2 , sont déterminés par :

$$\tilde{u}_1 = \min_{x \leq u_1} |u_1 - \{x/Err(x) \geq d\}|$$

$$\tilde{u}_2 = \min_{x \geq u_2} |u_2 - \{x/Err(x) \geq d\}|$$

Pour $d = \frac{1}{4} \max(Err)$, nous traçons dans le graphe C et D les deux seuils \tilde{u}_1 et \tilde{u}_2 (droites rouges continues) relatifs aux estimateurs de u_1 et u_2 . Nous voyons bien que les deux seuils affinés ne changent pas trop l’allure de la pdf par contre ils permettent d’éliminer plus de bruit dans le signal neuronal.

5 Conclusion

Dans ce travail, nous avons proposé un modèle hybride pour des données à deux queues lourdes. Pour évaluer ce modèle, un algorithme itératif non supervisé a été développé et testé sur des données simulées. Les performances du modèle hybride ont été étudiées sur des données neuronales réelles, brutes et filtrées. Le modèle proposé a montré des résultats satisfaisants en terme d’ajustement et de séparation du comportement moyen, représentant le bruit, des extrêmes. Pour séparer encore davantage les deux comportements, un affinement empirique des deux points de jonction a été proposé pour réduire la zone intermédiaire, transitoire entre les comportements moyen et extrêmes. Une étude théorique de ce comportement intermédiaire ainsi qu’une étude de la convergence de l’algorithme feront l’objet de développements futurs.

Références

- [1] A. K. Mandava, L. Shahram, and E. E. Regentova, “Reliability assessment of microarray data using fuzzy classification methods: A comparative study,” in *Advances in Computing and Communications*, 2011, vol. 190.
- [2] P. Sermpezis and T. Spyropoulos, “Modelling and analysis of communication traffic heterogeneity in opportunistic networks,” *Mobile Computing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.
- [3] V. P. Singh C. Li and A. K. Mishra, “Simulation of the entire range of daily precipitation using a hybrid probability distribution,” *Water resources research*, vol. 48, pp. 1–17, 2012.
- [4] M.S. Sambridge and W. Compston, “Mixture modeling of multi-component data sets with application to ion-probe zircon ages,” *Earth and Planetary Science Letters*, vol. 128, no. 3–4, pp. 373–390, 1994.
- [5] M. Rangaswamy, J. H. Michels, and B. Himed, “Statistical analysis of the non-homogeneity detector for {STAP} applications,” *Digital Signal Processing*, vol. 14, no. 3, pp. 253–267, 2004.
- [6] W. Gehin, “Modélisation des queues de distribution des rendements des actif financiers,” M.S. thesis, EURIA, Brest, 2011.
- [7] M. Mboup, “Neural spike detection and localisation via volterra filtering,” in *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, Sept 2012, pp. 1–5.
- [8] Zoran Tiganj and Mamadou Mboup, “A non-parametric method for automatic neural spike clustering based on the non-uniform distribution of the data,” *Journal of neural engineering*, vol. 8, no. 6, pp. 066014, 2011.
- [9] J. Pickands, “Statistical inference using extreme order statistics,” *Annals of Statistics*, vol. 3, pp. 119–131, 1975.
- [10] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, 1997.
- [11] N. Debbabi, M. Kratz, M. Mboup, and S. El Asmi, “Combining algebraic approach with extreme value theory for spike detection,” in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, pp. 1836–1840.
- [12] N. Debbabi and M. Kratz, “A new unsupervised threshold determination for hybrid models,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 3440–3444.
- [13] K. Levenberg, “A method for the solution of certain nonlinear problems in least squares quart,” *Applied Math*, vol. 2, pp. 164–168, 1944.
- [14] D. W. Marquardt, “An algorithm for least-squares estimation of nonlinear parameters,” *SIAM Journal on Applied Mathematics*, vol. 11, pp. 431–441, 1963.
- [15] Y. Demichel, A. Estrade, M. Kratz, and G. Samorodnitsky, “How fast can the chord length distribution decay ?,” *Advances in Applied Probability*, vol. 43, pp. 504–523, 2011.
- [16] O. Mazor C. Pouzat and G. Laurent, “Using noise signature to optimize spike-sorting and to assess neuronal classification quality,” *Journal of neuroscience methods*, vol. 122, no. 1, pp. 43–57, 2002.