Classification contrainte de signaux, application à l'étude de la protéine neuronale tau

Vincent Brault, Emilie Lebrarbier, Amélie Rosier, Virginie Stoppin-Mellet

Classification contrainte de signaux, application à l'étude de la protéine neuronale tau

Vincent Brault [1] & Emilie Lebarbier [2] & Amélie Rosier [2,3] & Virginie Stoppin-Mellet [4]

[1] *Univ. Grenoble Alpes, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France*
[2] *MODAL'X, Université Paris Nanterre, Nanterre, France*
[3] *ESME Sudria, Paris, France*
amelie.rosier@esme.fr
[4] *Equipe "Dynamique et structure du cytosquelette", Institut Neuroscience, Grenoble, France*

Résumé. Ce travail est motivé par une application en neuroscience, en particulier par l'étude du (dys)fonctionnement d'une protéine appelée Tau. L'objectif est d'établir une classification de profils d'intensité, selon la présence ou pas de la protéine et sa proportion monomère ou dimère. Pour cela, nous proposons ici un modèle de mélange gaussienne en un nombre fixé de groupes dont les paramètres de moyennes sont contraints et partagés par les groupes. L'inférence de ce modèle est faite via l'algorithme classique EM. La méthode proposée sera évaluée via des études de simulations et une application sur des données réelles sera effectuée.

Mots-clés. Données fonctionnelles, Classification automatique, Modèle de mélange, paramètres contraints, algorithme ECM.

Abstract. This work is motivated by an application in neuroscience, in particular by the study of the (dys)functioning of a protein called Tau. The objective is to establish a classification of intensity profiles, according to the presence or absence of the protein and its monomer or dimer proportion. For this, we propose here a Gaussian mixture model with a fixed number of groups whose mean parameters are constrained and shared by the groups. The inference of this model is done via the classical EM algorithm. The performance of the method will be evaluated via simulation studies and an application on real data will be done.

Keywords. Functional data, Model-based clustering, Constrained parameters, ECM algorithm.

1 Introduction

Ce travail est motivé par une application en neuroscience que nous détaillerons dans un premier de temps avant de présenter la modélisation proposée pour répondre aux problématiques sous-jacentes.

∗. Institute of Engineering Univ. Grenoble Alpes

Contexte biologique. La protéine Tau est une protéine présente dans le cerveau qui participe au contrôle de l'architecture et la stabilité des réseaux de microtubules, qui sont essentiels à la propagation des messages nerveux. Des dysfonctionnements de la protéine Tau sont responsables de pathologies graves, comme la maladie d'Alzheimer. Pour étudier la façon dont la protéine Tau contrôle la dynamique des microtubules, l'équipe d'Isabelle Arnal de l'institut Neurosciences de Grenoble a mis au point une méthode permettant de reconstituer in vitro des réseaux de microtubules dynamiques en présence de la protéine Tau, et de les observer en temps réel par microscopie (Elie et al. (2015)). A l'issu de ces expériences sont obtenus des profils d'intensité lumineuses au cours du temps. En théorie, ces profils sont des fonctions en escalier décroissantes, comme l'illustre la figure 1, où chaque saut correspond à l'extinction d'une protéine et donc où le nombre de sauts (de même taille) indique le nombre de protéines présentes.

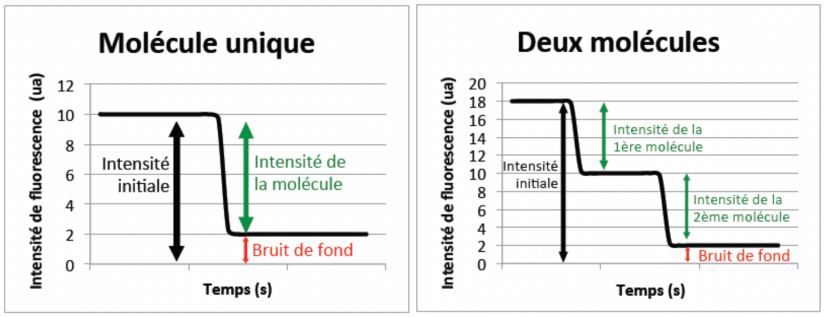

FIGURE 1 – Deux profils d'intensité lumineuse en théorie.

Classification contrainte. Etant donné un ensemble de profils, l'objectif statistique est d'établir une classification de ces profils selon le nombre de sauts et leurs tailles. Pour cela, on propose ici un modèle de mélange de distributions gaussiennes dont la moyenne est une fonction constante par morceaux décroissante et différente selon 4 groupes : elle est soit constante (groupe 1), i.e. sans présence de protéine, soit avec 1 ou 2 sauts de taille contrainte : un saut de taille δ pour le groupe 2, deux sauts de taille δ chacun pour le groupe 3 et un saut de taille 2δ pour le groupe 4. La taille des sauts est partagée par tous les profils/signaux tandis que les instants de sauts sont signal-spécifique. Un algorithme EM est utilisé pour estimer les paramètres du modèle.

2 Modèle

On considère S signaux de taille n^s chacun, notés $y^1, \ldots, y^S$ avec $y^s = (y^s_1, \ldots, y^s_{n^s})'$. On suppose que chaque signal y^s est une réalisation d'un processus gaussien Y^s de taille $[n^s \times 1]$ dont la moyenne dépend du groupe auquel il appartient. De plus, on suppose l'indépendance inter-signaux et intra-signal. Formellement, on introduit une suite de variables aléatoires indépendantes $Z^s = (Z^s_1, Z^s_2, Z^s_3, Z^s_4)$ telle que, pour tout $k \in [\![1, 4]\!]$, $Z^s_k = 1$ si le signal s

appartient au groupe k et $Z_k^s = 0$ sinon. Chaque signal a une probabilité $\pi_k = \mathbb{P}(Z_k^s = 1)$ ($\sum_{k=1}^4 \pi_k = 1$) d'appartenir au groupe k. Conditionnellement au groupe auquel il appartient, le signal d'intensité s a la distribution suivante

$$Y^s \mid Z_k^s = 1 \sim \mathcal{N}(m_k^s, \sigma_k^{2,s}\mathbf{I}_{n^s}) \quad \text{for } s \in [\![1\,, S]\!],$$

où le vecteur de la moyenne de taille $[n^s \times 1]$ est pour $t \in [\![1, n^s]\!]$

$$m_k^s(t) = \begin{cases} \mu_1^s & \text{if } k = 1 \\ \mu_2^s + \delta\ \mathbb{1}_{t>t_{21}^s} & \text{if } k = 2 \\ \mu_3^s + \delta\ \mathbb{1}_{t_{31}^s<t\leq t_{32}^s} + 2\delta\ \mathbb{1}_{t>t_{32}^s} & \text{if } k = 3 \\ \mu_4^s + 2\ \delta\ \mathbb{1}_{t>t_{41}^s} & \text{if } k = 4 \end{cases}$$

ou de façon équivalente

$$m_k^s = \mu_k^s \mathbb{1}_{n^s} + T_k^s \delta, \tag{1}$$

où T_k^s est la matrice d'incidence des instants de sauts, appelés instants de ruptures, pour le signal y^s dans le groupe k de taille $[n^s \times 1]$:

$$T_1^s = \left[0_{n^s}\right], \quad T_2^s = \begin{bmatrix} 0_{n_{21}^s} \\ \mathbb{1}_{n_{22}^s} \end{bmatrix}, \quad T_3^s = \begin{bmatrix} 0_{n_{31}^s} \\ \mathbb{1}_{n_{32}^s} \\ 2\mathbb{1}_{n_{33}^s} \end{bmatrix}, \quad T_4^s = \begin{bmatrix} 0_{n_{41}^s} \\ 2\mathbb{1}_{n_{42}^s} \end{bmatrix}, \tag{2}$$

On note t_{ki}^s la $i$ème rupture, $n_{ki}^s = t_{ki}^s - t_{k(i-1)}^s$ la longueur du $i$ème segment associé $[\![t_{k(i-1)}^s + 1, t_{ki}^s]\!]$, avec la convention $0 = t_{k0}^s < t_{k1}^s < t_{k2}^s < t_{kI_k}^s = n^s$ où I_k est le nombre de segments dans le groupe k ($I_1 = 1, I_2 = 2, I_3 = 3, I_4 = 2$). La moyenne pour les 4 groupes sont représentées en Figure 2.

Les paramètres du modèle sont $\mathbf{\Psi} = (\pi, \delta)$ avec $\pi = (\pi_k)_{k=1,\dots,4}$, qui implique tous les signaux, et $\mathbf{\Phi} = (\mathbf{\Phi}_k^s)_{k,s}$, où $\mathbf{\Phi}_1^s = (\mu_1^s, \sigma_1^{2,s})$ et pour $k \geq 2$ $\mathbf{\Phi}_k^s = (\mu_k^s, \sigma_k^{2,s}, T_k^s)$, qui est signal-spécifique et de taille $I_k + 1$ par signal y^s dans le groupe k. On note $\theta = (\mathbf{\Psi}, \mathbf{\Phi})$.

3 Inférence

On propose d'estimer les paramètres du modèle par maximum de vraisemblance. On utlise pour cela l'algorithme classique EM (Dempster et al. (1977)), qui consiste à maximiser l'espérance conditionnelle $Q(\theta, \theta')$ de la log-vraisemblance des données complète, $\log V(Y, Z; \theta)$ avec $Z = (Z^s)_s$, étant donné Y :

$$Q(\theta, \theta') = \mathbb{E}_{\theta'}[\log V(Y, Z; \theta) \mid Y],$$

où $\mathbb{E}_{\theta'}[.]$ est l'opérateur espérance utilisant θ' comme paramètre et

$$\log V(Y, Z; \theta) \quad = \quad \sum_{s=1}^{S} \log \left(\sum_{k=1}^{4} (\pi_k f(Y^s; \mathbf{\Phi}_k^s, \delta))^{Z_k^s} \right) = \sum_{s=1}^{S} \sum_{k=1}^{4} Z_k^s \log(\pi_k f(Y^s; \mathbf{\Phi}_k^s, \delta)),$$

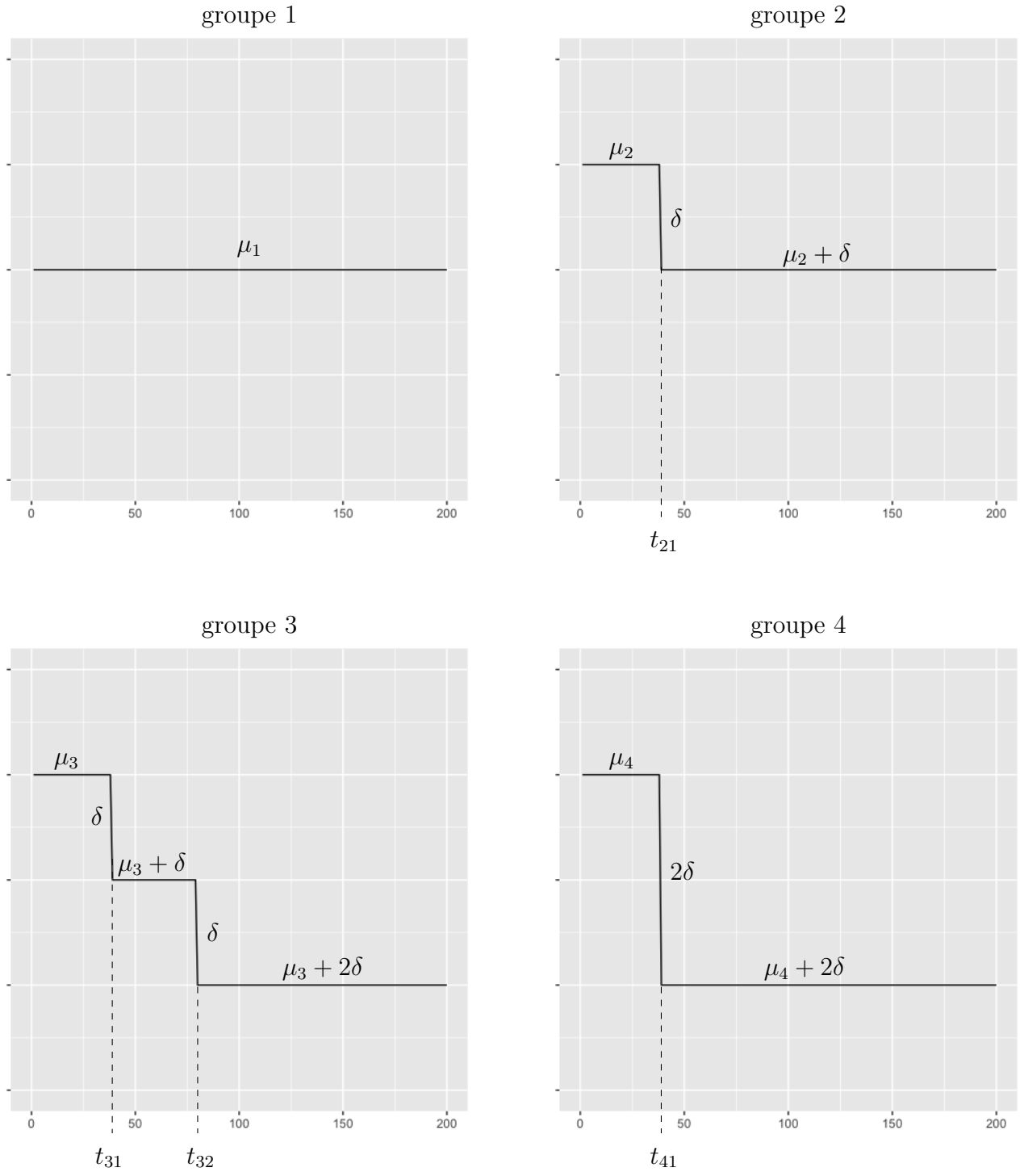

FIGURE 2 – La moyenne des 4 groupes.

avec $f(Y^s; \mathbf{\Phi}_k^s, \delta)$ la densité conditionnelle de Y^s donnée par

$$f(Y^s; \mathbf{\Phi}_k^s, \delta) = \frac{n^s}{\sigma_k^s \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_k^{2,s}} \|Y^s - m_k^s\|^2\right) = \frac{n^s}{\sigma_k^s \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_k^{2,s}} \|Y^s - \mu_k^s \mathbb{1}_{n^s} - T_k^s \delta\|^2\right).$$

L'algorithme EM est un algorithme itératif combinant, à chaque itération, deux étapes : l'étape E et l'étape M décrites respectivement dans les deux sous-sections suivantes.

3.1 Etape E.

Elle consiste à calculer $Q(\theta, \theta^{(h)})$ en utilisant le paramètre courant $\theta^{(h)}$:

$$Q(\theta, \theta^{(h)}) = \mathbb{E}_{\theta^{(h)}}\left[\sum_{s=1}^{S}\sum_{k=1}^{4} Z_k^s \log(\pi_k f(Y^s, \mathbf{\Phi}_k^s, \delta))|Y\right] = \sum_{s=1}^{S}\sum_{k=1}^{4} \tau_k^{s(h+1)} \log(\pi_k f(Y^s, \mathbf{\Phi}_k^s, \delta)),$$

avec

$$\tau_k^{s(h+1)} = = \mathbb{P}_{\theta^{(h)}}(Z_k^s = 1|Y^s) = \frac{\pi_k^{(h)} f(Y^s, \mathbf{\Phi}_k^{s,(h)}, \delta^{(h)})}{\sum_{l=1}^{4} \pi_l^{(h)} f_l(Y^s, \mathbf{\Phi}_l^{s,(h)}, \delta^{(h)})},$$

la probabilité a posteriori d'appartenance du signal s au groupe k.

3.2 CM-steps.

L'étape M globale consiste à mettre à jour les paramètres en maximisant l'espérance conditionnelle obtenue à l'étape précédente :

$$\theta^{(h+1)} = \arg\max_{\theta} \sum_{s=1}^{S}\sum_{k=1}^{4} \tau_k^{s(h+1)} \log(\pi_k \ f(Y^s, \mathbf{\Phi}_k^s, \delta)).$$

On utilise ici l'algorithme ECM qui découpe la maximisation de $Q(\theta, \theta^{(h)})$ selon θ en étapes CM qui se focalisent sur un paramètre, les autres étant fixés. Les propriétés de convergence de ECM ont été étudiées par (Meng and Rubin (1993)).

Estimation de π. Les proportions π_k sont estimés sous la contrainte $\sum_{k=1}^{4} \pi_k = 1$. On obtient

$$\pi_k^{(h+1)} = \arg\max_{\pi} Q(\theta, \pi, \delta^{(h)}, \mathbf{\Phi}^{(h)}) = \frac{\sum_{s=1}^{S} \tau_k^{s(h+1)}}{S} \quad \text{for } k = 1, \ldots, 4.$$

Estimation de δ. On obtient

$$\begin{aligned}\delta^{(h+1)} &= \arg\max_{\delta} Q(\theta, \pi^{(h+1)}, \delta, \Phi^{(h)}) = \arg\min_{\delta} \sum_{s=1}^{S}\sum_{k=1}^{4} \frac{\tau_k^{s(h+1)}}{\sigma_k^{2,(h)}} \|Y^s - \mu_k^{s(h)} \mathbb{1}_{n^s} - T_k^{s(h)} \delta\|^2 \\ &= \frac{\sum_{s=1}^{S}\sum_{k=2}^{4} \frac{\tau_k^{s(h+1)}}{\sigma_k^{2,(h)}} {}^t T_k^{s(h)} (Y^s - \mu_k^{s(h)} \mathbb{1}_{n^s})}{\sum_{s=1}^{S}\sum_{k=2}^{4} \frac{\tau_k^{s(h+1)}}{\sigma_k^{2,(h)}} {}^t T_k^{s(h)} T_k^{s(h)}}.\end{aligned}$$

Estimation de $\mathbf{\Phi}$. Cette étape se réduit à un problème de segmentation contrainte pour laquelle chaque signal y^s est segmenté selon chaque groupe, i.e. en accord avec la contrainte de décroissance associée à chaque groupe. Pour chaque $s = 1, \ldots, S$ et $k = 1, \ldots, 4$,

$$\mathbf{\Phi}_k^{s(h+1)} = \arg\max_{\mathbf{\Phi}_k^s} Q(\theta, \pi^{(h+1)}, \delta^{(h+1)}, \mathbf{\Phi})$$

i.e.

$$\begin{aligned}(\mu_k^{s(h+1)}, \sigma_k^{2,s(h+1)}, T_k^{s(h+1)}) &= \arg\max_{\mu_k^s, \sigma_k^{2,s}, T_k^s} Q(\theta, \pi^{(h+1)}, \delta^{(h+1)}, \mu_k^s, \sigma_k^{2,s}, T_k^s) \\ &= \arg\max_{\mu_k^s, \sigma_k^{2,s}, T_k^s} \left[-\frac{n^s}{2} \log(2\pi\sigma_k^{2,s}) - \frac{1}{2\sigma_k^{2,s}} \|Y^s - \mu_k^s \mathbb{1}_{n^s} - T_k^s \delta^{(h+1)}\|^2 \right]\end{aligned}$$

On différencie le cas du groupe 1 (pas de saut) des autres :

— pour $k = 1$,

$$(\mu_1^{s(h+1)}, \sigma_1^{2,s(h+1)}) = \arg\max_{\mu_1^s, \sigma_1^{2,s}} -\frac{n^s}{2} \log(2\pi\sigma_1^{2,s}) - \frac{1}{2\sigma_1^{2,s}} \|Y^s - \mu_1^s \mathbb{1}_{n^s}\|^2,$$

et on obtient les estimateurs des moindres carrés suivants

$$\mu_1^{s(h+1)} = \overline{Y}^s = \frac{1}{n^s} \sum_{t=1}^{n^s} Y_t^s \quad \text{et} \quad \sigma_1^{2,s(h+1)} = \frac{1}{n^s} \sum_{t=1}^{n^s} (Y_t^s - \overline{Y}^s)^2.$$

— pour $k \geq 2$, si les positions des ruptures sont connues, les estimateurs de la moyenne de base et de la variance dans chaque groupe sont simplement :

$$\widehat{\mu}_k^s(T_k^s) = \frac{1}{n^s} \|Y^s - T_k^s \delta^{(h+1)}\|^2 \quad \text{and} \quad \widehat{\sigma}_k^{2,s}(T_k^s) = \frac{1}{n^s} \|Y^s - \widehat{\mu}_k^s(T_k^s) \mathbb{1}_{n^s} - T_k^s \delta^{(h+1)}\|^2.$$

L'estimation des ruptures est obtenu par

$$\begin{aligned}T_k^{s(h+1)} &= \arg\max_{T_k^s} \max_{\mu_k^s} \max_{\sigma_k^{2,s}} Q(\theta, \pi^{(h+1)}, \delta^{(h+1)}, \mu_k^s, \sigma_k^{2,s}, T_k^s) \\ &= \arg\max_{T_k^s} Q(\theta, \pi^{(h+1)}, \delta^{(h+1)}, \widehat{\mu}_k^s(T_k^s), \widehat{\sigma}_k^{2,s}(T_k^s), T_k^s) \\ &= \arg\max_{T_k^s} \left[-\frac{n^s}{2} \log(2\pi\widehat{\sigma}_k^{2,s}(T_k^s)) - \frac{1}{2\widehat{\sigma}_k^{2,s}(T_k^s)} \|Y^s - \widehat{\mu}_k^s(T_k^s) \mathbb{1}_{n^s} - T_k^s \delta^{(h+1)}\|^2 \right] \\ &= \arg\max_{T_k^s} -\frac{n^s}{2} \left[\log(2\pi\widehat{\sigma}_k^{2,s}(T_k^s)) + 1 \right] \\ &= \arg\min_{T_k^s} \|Y^s - \widehat{\mu}_k^s(T_k^s) \mathbb{1}_{n^s} - T_k^s \delta^{(h+1)}\|^2,\end{aligned}$$

et les estimateurs finaux de la moyenne et de la variance sont

$$\mu_k^{s(h+1)} = \widehat{\mu}_k^s(T_k^{s(h+1)}) \quad \text{et} \quad \sigma_k^{2,s(h+1)} = \widehat{\sigma}_k^{2,s}(T_k^{s(h+1)}).$$

4 Simulations

Pour les simulations [1], nous avons choisi de prendre $S = 100$ signaux avec pour chaque signal s une longueur de $n^s = 100$, une probabilité d'appartenance à chacune des classes de $1/4$ et une variance de 1. Pour les sauts, nous supposons que $\delta \in \{-5, -2, -1, -0.5\}$ et les emplacements des sauts sont choisis au hasard entre 10% et 90% de la série et avec une distance de $\{0, 0.3, 0.6\} \times n^s + 1$ entre les deux sauts du cluster 3. Enfin, les moyennes μ^s sont choisies de telles sortes que la moyenne des dernières observations valent 2.

Nous estimons les paramètres à l'aide de l'algorithme EM défini en Section 3 et en prenant 1 ou 10 itérations pour la maximisation et en faisant 1 ou 10 initialisations aléatoires (l'estimation avec la meilleure vaisemblance est alors conservée dans ce dernier cas). Pour évaluer la qualité des résultats, nous regardons l'estimation $\widehat{\delta}$ par rapport à sa vraie valeur (voir la figure 3), la distance de Hausdorff des positions des ruptures connaissant les bons clusters (voir figure 5) et le pourcentage d'erreurs de classification (voir figure 4).

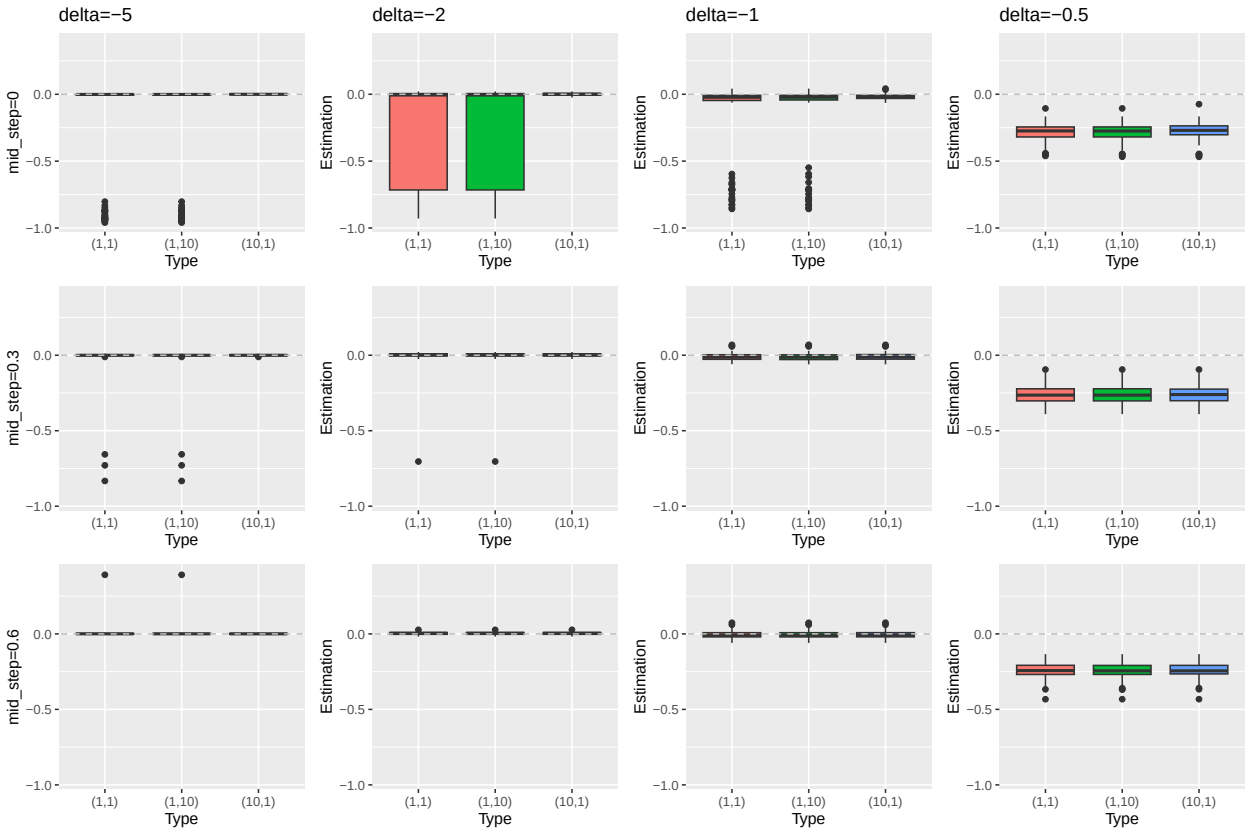

FIGURE 3 – Boxplot des $\frac{\widehat{\delta}-\delta^\star}{|\delta^\star|}$ en fonction de la valeur de $\delta^\star$ (colonnes) et de la longueur du plateau central (ligne). Pour chaque graphique, nous décomposons suivant s'il y a 1 ou 10 initialisations et 1 ou 10 itérations de l'étape M.

1. Toutes les simulations présentées dans cet article ont été réalisées sur les infrastructures de GRICAD (https://gricad.univ-grenoble-alpes.fr), qui sont supportées par la communauté scientifique de Grenoble.

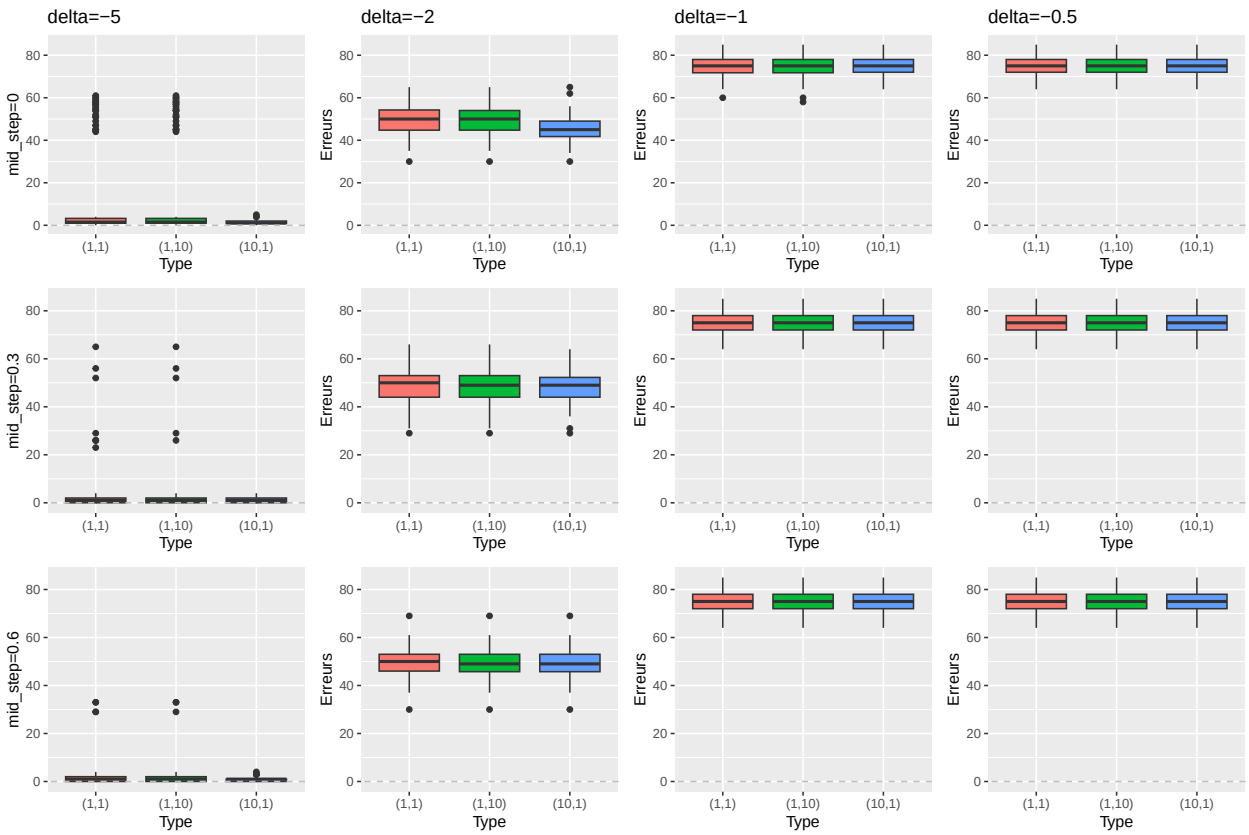

FIGURE 4 – Boxplot des erreurs de classifications en fonction de la valeur de $\delta^\star$ (colonnes) et de la longueur du plateau central (ligne). Pour chaque graphique, nous décomposons suivant s'il y a 1 ou 10 initialisations et 1 ou 10 itérations de l'étape M.

Les résultats s'améliorent avec le nombre d'initialisations testées, comme attendu, mais pas avec le nombre d'itérations de l'étape M.

On observe que quand la détection de ruptures est facile (δ grand), le saut est bien estimé et les ruptures sont bien positionnées alors que lorsque la détection est difficile (δ petit), le saut est sous-estimé et les ruptures sont moins bien positionnées en particulier pour le cluster 3 (voir figures 3 et 5). En terme de classification, le taux d'erreur augmente quand la taille du saut diminue (figure 4). Comme on peut l'observer en figure 6, seul le cluster 3 est correctement prédit. Pour les séries issues des autres clusters, si la détection est très difficile ($\delta = -1$), toutes les séries (ou quasiment) sont classées dans le cluster 3 et si la détection est un petit peu moins difficile, les séries du cluster 4 sont essentiellement classées en clusters 3 et 4, les séries du cluster 2 dans les clusters 2 et 3, et celles du cluster 1 dans les trois premiers clusters.
Dans la suite, nous chercherons à limiter ce surapprentissage.

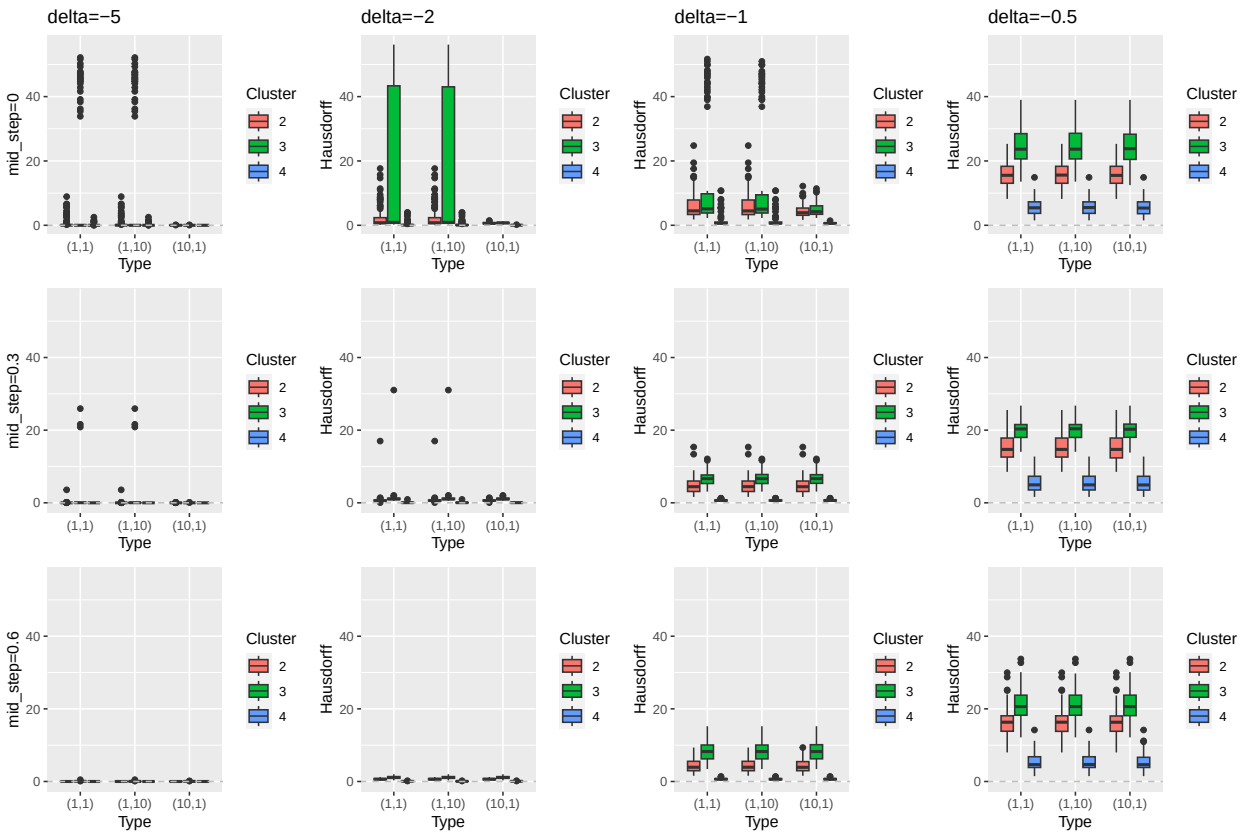

FIGURE 5 – Boxplot des distances de Hausdorff des estimations des ruptures connaissant la bonne classe en fonction de la valeur de $\delta^\star$ (colonnes) et de la longueur du plateau central (ligne). Pour chaque graphique, nous décomposons suivant s'il y a 1 ou 10 initialisations et 1 ou 10 itérations de l'étape M, et la couleur représente les distances pour les clusters 2, 3 et 4.

Références

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39 :1–38.

Elie, A., Prezel, E., Guérin, C., Denarier, E., Ramirez-Rios, S., Serre, L., Andrieux, A., Fourest-Lieuvin, A., Blanchoin, L., and Arnal, I. (2015). Tau co-organizes dynamic microtubule and actin networks. *Scientific reports*, 5(1) :1–10.

Meng, X.-L. and Rubin, D. (1993). Maximum likelihood estimation via the ecm algorithm : a general framework. *Biometrika*, 80(2) :267–278.

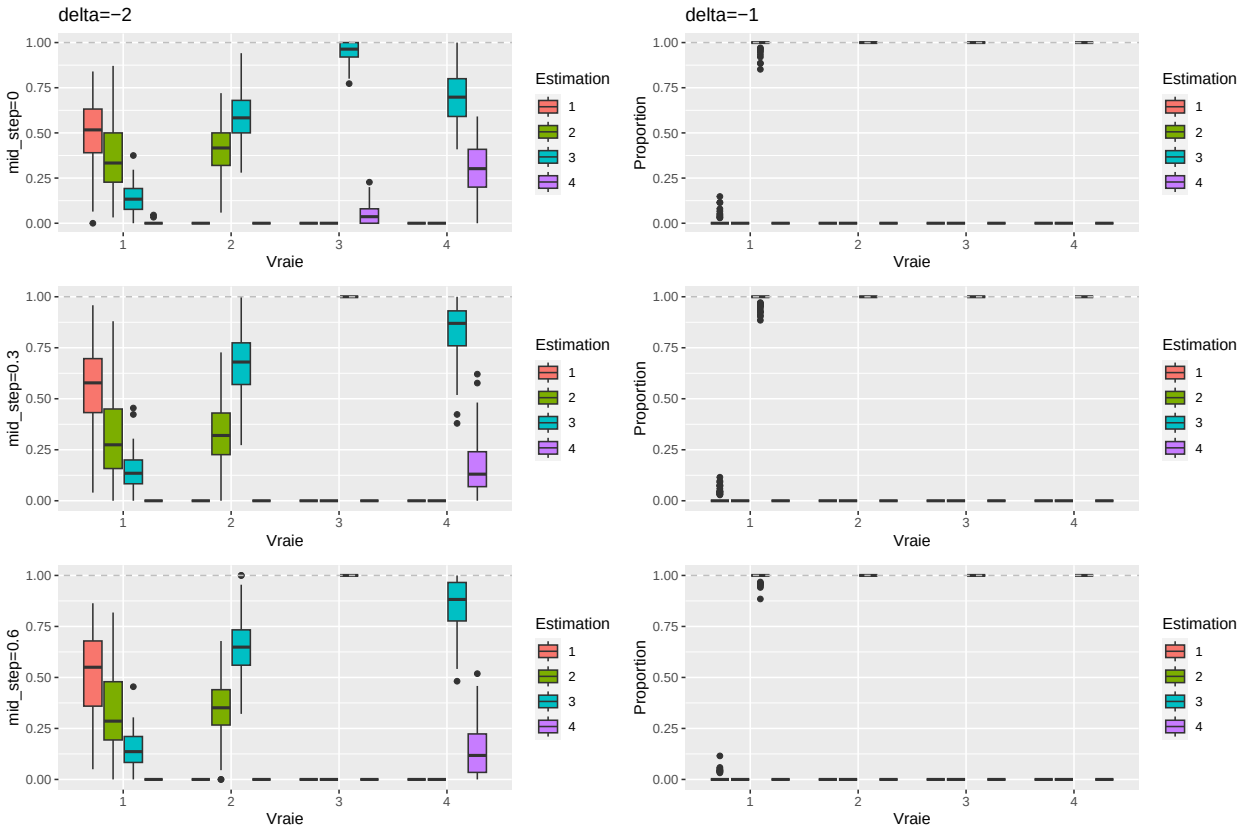

FIGURE 6 – Boxplot des proportions des estimations des clusters d'appartenances (couleur) en fonction du cluster d'origine (abscisse), de la valeur de $\delta^\star$ (colonnes) et de la longueur du plateau central (ligne).