



# **Codage espace-échelle parcimonieux en présence de bruit non-gaussien. Application à l'analyse de la réplication de l'ADN en molécule unique**

Clara Lage, Nelly Pustelnik, Jean-Michel Arbona, Benjamin Audit

## **► To cite this version:**

Clara Lage, Nelly Pustelnik, Jean-Michel Arbona, Benjamin Audit. Codage espace-échelle parcimonieux en présence de bruit non-gaussien. Application à l'analyse de la réplication de l'ADN en molécule unique. XXIXème Colloque GRETSI Traitement du Signal et des Images, 2023. <hal-04274266>

**HAL Id: hal-04274266**

**<https://hal.science/hal-04274266v1>**

Submitted on 10 Nov 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

# Codage espace-échelle parcimonieux en présence de bruit non-gaussien. Application à l'analyse de la réplication de l'ADN en molécule unique.

Clara LAGE<sup>1</sup> Nelly PUSTELNIK<sup>1</sup> Jean-Michel ARBONA<sup>2</sup> Benjamin AUDIT<sup>1</sup>

<sup>1</sup>Univ Lyon, ENS de Lyon, CNRS, Laboratoire de Physique, F-69342 Lyon, France

<sup>2</sup>Laboratoire de Biologie et Modélisation de la Cellule, ENS de Lyon, Lyon, France

**Résumé** – La réplication de l'ADN est un processus hautement régulé dont la dynamique peut être analysée par des expériences en molécule unique. La vitesse de réplication i.e. le taux de synthèse local, un paramètre clé de cette dynamique, est alors estimée par la détection de motifs le long de profil d'incorporation de nucléotides marqués. Dans ce travail, nous explorons des méthodes issues du codage parcimonieux pour cette estimation : le LASSO, une des formulations les plus utilisées, et deux méthodes visant à augmenter la précision de cette estimation dans le cadre d'un bruit non gaussien. En particulier, nous proposons une méthode hybride discrète-continue dont les performances surpassent une approche discrète basée sur la divergence de Kullback-Leibler pourtant spécifiquement adaptée à ce type de bruit.

**Abstract** – DNA replication is a highly regulated process. Its dynamics can be analyzed by single molecule experiments. Replication speed i.e. the local synthesis rate, a key parameter of this dynamics, is then extracted from the detection of signal patterns along marked nucleotide incorporation profiles. Here, we explore sparse coding methods for this estimation: LASSO, one of the most used formulation, and two methods aiming at increasing the precision for non-Gaussian noise. We propose an hybrid discrete-continuous method having better performances than a full discrete approach based on the Kullback-Leibler divergence although specifically adapted to this type of noise.

## 1 Réplication de l'ADN en molécule unique et codage parcimonieux

La duplication des chromosomes est un mécanisme fondamental permettant, lors d'une division cellulaire, que chaque cellule fille reçoive une copie complète du patrimoine génétique. Elle repose sur la réplication de l'ADN qui démarre de multiples emplacements le long des chromosomes ou origines de réplication. De chaque origine émergent deux machineries de synthèse ou fourches de réplication se déplaçant en sens opposé. La copie d'une molécule d'ADN est achevée lorsque toutes les paires de fourches de réplication convergentes, issues de deux origines successives, se sont rejointes.

Un des objectifs des études sur la réplication de l'ADN est de caractériser la dynamique de ce processus. Dans ce cadre, l'analyse de molécules uniques permet de caractériser le taux de synthèse de l'ADN d'une seule fourche, à savoir sa vitesse de progression. Pour cela, les cellules sont mises en présence de BrdU (un nucléotide modifié qui est incorporé à la place de la thymidine) pendant un temps court ( $\sim 2$  min) suivi de son élimination (i.e. *chasse*) [11]. On mesure alors par séquençage nanopore le profil d'incorporation de BrdU le long des fibres d'ADN extraites des cellules. Ces profils prennent la forme de pulses asymétriques localisés le long des régions traversées par les fourches pendant la présence de BrdU (Figure 1 (haut)). La montée concave rapide correspond au temps de présence du BrdU alors que la redescente convexe plus lente est caractéristique de la chasse, ce qui permet de déterminer le sens de propagation des fourches. L'étendue du pulse est proportionnelle à la vitesse de la fourche pendant la présence de BrdU. Plus une fourche est rapide plus la longueur d'ADN recopiée pendant le pulse de BrdU est grande.

**Contributions et plan** – Nous nous intéressons à la détection des fourches de réplication et de leurs caractéristiques le long des profils d'incorporation de BrdU (cf. Figure 1 (haut)) par des outils de codage parcimonieux dont les principaux éléments d'état-de-l'art seront rappelés dans la section 2. La forme asymétrique Y des pulses de BrdU ayant été déterminée expérimentalement [11], il s'agit d'utiliser un dictionnaire espace-échelle d'atomes présentés en section 3. Chaque atome encode alors les propriétés d'une fourche de réplication : la localisation le long du chromosome (position  $x$ ) ainsi que le sens et la vitesse de progression de la fourche ( $\text{sign}(s)$ ,  $|s|$ ). Cette méthode vise à décomposer le signal d'intérêt comme une somme de motifs décalés dans le temps et chercher ses paramètres. Ces trois informations réunies nous permettent une caractérisation fine de la dynamique de réplication de l'ADN le long des chromosomes [11]. Néanmoins, une contrainte forte est que le bruit sous-jacent ne suit pas une statistique gaussienne, qui est une hypothèse des principales techniques de codage parcimonieux. Notre objectif est, d'une part, d'identifier dans quelle mesure la prise en compte de la statistique du bruit améliore la précision de l'estimation de la position et de la vitesse des fourches de réplication (cf. section 4). D'autre part, nous évaluons également l'impact d'une approche hybride continue-discrète que nous proposons en section 5, permettant de gérer efficacement des dictionnaires espace-échelle, et que nous comparons à une formulation discrète standard. La section 6 compare les différentes stratégies algorithmiques dans le contexte de la réplication.

## 2 Eléments de l'état-de-l'art

**Codage parcimonieux et facteurs limitants** – Rappelons, qu'une représentation parcimonieuse est une représentation

compacte définie par la combinaison linéaire d'atomes dans un dictionnaire  $D$ . Formellement, à partir d'un signal  $z \in \mathbb{R}^n$  bruité par un bruit gaussien, le problème de codage parcimonieux s'écrit :

$$\min_{\alpha \in \mathbb{R}^l} \|D\alpha - z\|_2^2 + \lambda \|\alpha\|_*, \quad (1)$$

où  $\|\cdot\|_*$  impose la parcimonie de  $\alpha$ . La formulation LASSO se réduit au cas où la pénalisation est une norme  $\ell_1$ . Dans notre étude, la classe de dictionnaire qui nous intéresse particulièrement est celle des dictionnaires convolutifs composée de  $K$  filtres  $d_k \in \mathbb{R}^m$  avec  $m \ll n$ . Cela conduit à un problème d'apprentissage parcimonieux convolutif de la forme :

$$\min_{(a_1, \dots, a_K) \in \mathbb{R}^n \times \dots \times \mathbb{R}^n} \left\| \sum_{k=1}^K d_k * a_k - z \right\|_2^2 + \lambda \sum_{k=1}^K \|a_k\|_1. \quad (2)$$

La formulation sous forme de convolution est avantageuse tant en terme de représentation compacte que d'implémentation efficace par FFT.

Plusieurs facteurs peuvent affecter la précision d'estimation par les approches mentionnées précédemment. Parmi eux, le niveau et la nature du bruit, qui peut corrompre le signal rendant difficile ou impossible la récupération de sa position correcte. La seconde limitation provient de la discrétisation sur une grille des valeurs de position ne permettant d'atteindre qu'un nombre fini  $n$  de positions pour chaque filtre.

**Au-delà du bruit gaussien** – Dans le cas d'un bruit gaussien, sous des hypothèses spécifiques de parcimonie et régularité, le LASSO est capable de retrouver le support de la solution ([6, 13]). Dans de nombreux domaines applicatifs, d'autres types de bruit peuvent apparaître, comme par exemple le bruit de type Poisson ou le bruit de Speckle. Des méthodes spécialisées pour ces types de bruits ont été proposées ([7, 8]) et modifient essentiellement le terme d'attache aux données.

**Approches continues** – En ce que concerne les limites de la discrétisation, les principales méthodes de codage parcimonieux discrètes, comme les méthodes Matching Pursuit et LASSO, ont été étendues à des variables à temps continu, donnant lieu aux méthodes M.P. continu et BLASSO ([4, 12]). Diverses études montrent la capacité de ces formulations à augmenter la précision de l'estimation par rapport à une approche discrète ([5, 1, 9, 3]). La différence principale d'une approche à temps continu repose sur l'estimation de paramètres continus modélisés par un espace de paramètres  $\Theta$ . Dans ce contexte, le dictionnaire  $D$  sous forme matricielle ou convolutive est remplacé par sa forme fonctionnelle :  $\mathcal{D} = \{\varphi(\cdot; \theta) \in \mathcal{L}^2(\Theta) : \theta \in \Theta\}$ , où  $\Theta \subset \mathcal{R}^l$  avec  $l \in \mathbb{N}$ . Si dans une formulation discrète de type LASSO, l'objectif est d'estimer un vecteur  $a$ , la formulation continue raisonne sur des sommes finies de mesures de Dirac appartenant à l'espace des mesures noté  $\mathcal{M}(\Theta)$ . Formellement, soit  $m \in \mathcal{M}(\Theta)$  tel que  $m = \sum_{i=1}^L c_i \delta_{\theta_i}$ , où  $\theta_i \in \Theta$  et  $c_i \in \mathbb{R}$  et  $L \in \mathbb{N}$ , on définit l'opérateur  $\Phi$  comme :

$$\Phi(m) = \int_{\Theta} \varphi(\theta) dm(\theta) = \sum_{i=1}^L c_i \varphi(\cdot; \theta_i). \quad (3)$$

et le BLASSO est alors décrit par le problème d'optimisation suivant :

$$\min_{m \in \mathcal{M}(\Theta)} \left\| z - \Phi(m) \right\|_2^2 + \gamma |m|(\Theta), \quad (4)$$

où  $|m|(\Theta) = \|c\|_1$  est la variation totale de la mesure  $m$ .

### 3 Dictionnaire espace-échelle

Notre travail se concentre sur des ensembles de paramètres avec la forme  $\Theta = \mathcal{X} \times \mathcal{S}$ . Le dictionnaire  $\mathcal{D}$  est défini par :

$$\mathcal{D} = \left\{ \{\varphi(\cdot; (x, s))\}_{(x,s) \in \Theta} : \varphi(\cdot; (x, s)) = \zeta_s \psi\left(\frac{\cdot - x}{s}\right), \zeta_s \in \mathbb{R} \right\} \quad (5)$$

où  $\psi \in C(\mathcal{X})$  est un atome de référence et  $\zeta_s$  la normalisation par rapport à l'échelle  $s$ . Le signe de  $s$  donne l'orientation de la fourche, comme nous observons dans la Figure 1.

Le dictionnaire correspondant dans le cas convolutionnel ; discret (2) repose sur une discrétisation de l'espace d'échelle  $\tilde{S} = \{s_1, \dots, s_K\}$  pour obtenir des atomes du type :

$$\tilde{\mathcal{D}} = \left\{ d_k \in \mathbb{R}^m : d_k = \zeta_{s_k} \psi\left(\frac{\cdot}{s_k}\right), \zeta_{s_k} \in \mathbb{R}, s_k \in \tilde{S} \right\}. \quad (6)$$

Le  $(\cdot)$  représente la variable qui sera discrétisée avec la même dimension que le signal  $z$ .

### 4 Approche discrète adaptée au cas non-gaussien

La formulation LASSO peut être obtenue à partir d'une formulation bayésienne reposant sur une hypothèse de bruit gaussien. Par conséquent dans le cas d'une statistique de bruit qui n'est pas gaussienne, la formulation variationnelle prend la forme

$$\text{forme fonctionnelle} - \log p(z|D\alpha) + \lambda \|\alpha\|_1 \quad (7)$$

où  $p(z|D\alpha)$  modélise la vraisemblance. Dans le cas spécifique d'un bruit Poisson,  $-\log p(z|D\alpha) \propto \text{KL}(z|D\alpha)$  où KL est la divergence de Kullback-Leibler :

$$\text{KL}(z|w) = \sum_n z_n \log \left( \frac{z_n}{w_n} \right) + w_n - z_n. \quad (8)$$

Par conséquent, dans un cadre variationnel général, la version convolutionnelle de (7) est donnée par :

$$\min_{(a_1, \dots, a_K) \in \mathbb{R}^n \times \dots \times \mathbb{R}^n} - \log p\left(z \mid \sum_{k=1}^K d_k * a_k\right) + \lambda \sum_k \|a_k\|_1. \quad (9)$$

où les atomes  $d_k$  sont définis en (6).

Cette formulation explicite rend possible l'utilisation de certains algorithmes proximaux, comme celui de Douglas-Rachford [10], qui traitent le cas de la somme de fonctions convexes non nécessairement différentiables. Nous ré-écrivons (9) comme :

$$\min_{a \in \mathbb{R}^{n \times K}} - \log p(z \mid u) + \lambda \|a\|_1 + i_{\mathcal{C}}(a, u) \quad (10)$$

où  $\mathcal{C} = \{(a, u) : \sum d_k * a_k = u\}$  et  $a = (a_1, \dots, a_K)$ . En suivant [2], l'algorithme permettant de minimiser (9) est présenté par l'Algorithme 1 où la suite  $(a^{[\ell]})_{\ell}$  converge vers un minimiseur de (9).

### 5 Approche hybride discrète-continue

Dans un second temps, nous proposons de résoudre le problème (4) dans le contexte d'un dictionnaire espace-échelle comme celui défini dans la section 3. Cependant, la méthode de Frank-Wolfe sous-jacente à la gestion de dictionnaires continus, repose

---

**Algorithme 1** : Codage parcimonieux convolutionnel en présence de bruit non-gaussien

---

**Entrées** :  $z \in \mathbb{R}^n$ ,  $w_a^{[0]} \in \mathbb{R}^{n \times K}$  et  $w_u^{[0]} = \sum_k d_k * w_a^{[0]}$ .

2 **pour**  $\ell = 0, \dots, N$  **faire**

3      $(a^{[\ell]}, u^{[\ell]}) = \text{proj}_C(w_a^{[\ell]}, w_u^{[\ell]})$

4      $w_a^{[\ell+1]} = w_a^{[\ell]} - a^{[\ell]} + \text{prox}_{\lambda \|\cdot\|_1}(2a^{[\ell]} - w_a^{[\ell]})$

5      $w_u^{[\ell+1]} = w_u^{[\ell]} - u^{[\ell]} + \text{prox}_{-\log p}(2u^{[\ell]} - w_u^{[\ell]})$

6 **fin**

---

sur une fonction difficilement tractable numériquement. Nous proposons donc une approche hybride qui discrétise la variable d'échelle :  $\tilde{S} = \{s_1, \dots, s_K\}$ , et estime la position sous une forme continue. L'objectif de cette approche est d'augmenter la précision de l'estimation de la position  $x$  tout en gardant une bonne estimation de l'échelle  $s$ . Dans cette section, la statistique du bruit n'est pas traitée. Seul l'impact d'un choix d'un dictionnaire continu est considéré.

Nous considérons  $m_k \in \mathcal{M}(\mathcal{X})$ . Le problème hybride décrit :

$$\min_{m \in \mathcal{M}(\mathcal{X})^K} \left\| \sum_{k=1}^K \Phi_k(m_k) - z \right\|_2^2 + \gamma \sum_{k=1}^K |m_k|(\mathcal{X}). \quad (11)$$

où  $m = (m_1, \dots, m_K)$ ,  $m_k = \sum_{j=1}^{N_k} c_{k,j} \delta_{x_{k,j}}$ , où  $x_{k,j} \in \mathcal{X}$ . Les opérateurs  $\Phi_k$  pour  $k \in \{1, \dots, K\}$  sont définis comme :

$$\Phi_k(m_k) = \zeta_{s_k} \sum_{j=1}^{N_k} c_{k,j} \psi\left(\frac{\cdot - x_{k,j}}{s_k}\right). \quad (12)$$

Nous proposons l'algorithme 2 adapté à ce formalisme hybride discret-continu qui converge vers la solution du problème (11).

L'algorithme hybride est composé de trois étapes : (i) choisir l'échelle  $s_k$  et faire une première estimation de la composante de translation  $x_{k,j}$ ; (ii) Ajuster les amplitudes  $c_{k,j}$  basée sur l'estimation précédente de  $x_{k,j}$  pour tout  $i$  et  $k$  et (iii) Résoudre le problème non-convexe (11) pour raffiner les paramètres  $\{\alpha_{k,j}, x_{k,j}\}$  en utilisant comme point initial l'estimation donnée par les étapes 1 et 2. Notons que les étapes 3 et 4 peuvent être simplifiées dans la cas où il n'y a pas de superposition d'atomes. Dans ce cas, ces étapes peuvent être réalisées en considérant seulement l'échelle  $s^{[\ell]}$ , choisie dans l'étape 1.

## 6 Détection de fourches

**NanoForkSpeed (NFS).** La méthode actuelle pour la localisation, le sens et la vitesse de progression de la fourche est appelée NFS [11]. Cette méthode est basée sur une approximation linéaire par morceaux du signal lissé, qui n'est pas sensible à la forme exacte du pulse et par conséquent des artefacts de mesures sont confondus avec des pulses. L'approche par codage parcimonieux a comme objectif de raffiner cette approche avec la possibilité de filtrer les faux positifs.

Les atomes de  $\mathcal{D}$  sont décrits par la formule (5). La fonction  $\psi$  est proposée à partir du cadre expérimental. La formule exacte de  $\psi$  peut être retrouvée dans [11, Section Methods] et est composée de deux parties exponentielles aux paramètres différents. Nous pouvons observer son comportement sur la Figure 1. Les translations suivant  $x$  nous permettent de fournir

---

**Algorithme 2** : Codage parcimonieux hybride discret-continu

---

**Entrées** : Soit  $\gamma > 0$ ,  $N_k^{[0]} = 0$ , et  $\sum_1^0 \cdot = 0$

1 **pour**  $\ell = 0, \dots, N$  **faire**

2     **Etape 1** : Estimer un spike additionnel :

$$(\bar{x}, s_{\bar{k}}) = \arg \max_{x \in \mathcal{X}} \eta(x, s)$$

4     avec  $s \in \{s_1, s_2, \dots, s_K\}$

$$\eta(x, s) = \left| \frac{1}{\gamma} \left\langle z - \sum_{k=1}^K \sum_{j=1}^{N_k^{[\ell]}} c_{k,j}^{[\ell]} \zeta_{s_k} \psi\left(\frac{\cdot - x_{k,j}^{[\ell]}}{s_k}\right), \varphi_s(\cdot, x) \right\rangle \right|$$

6     et définir  $\begin{cases} x_{\bar{k}}^{[\ell+\frac{1}{2}]} = (x_{\bar{k}}^{[\ell]}, \bar{x}) \\ x_k^{[\ell+\frac{1}{2}]} = x_k^{[\ell]} \text{ pour tout } k \neq \bar{k}. \end{cases}$

7     **Test d'arrêt** :  $\eta(\bar{x}, s_{\bar{k}}) \leq 1$

8     **Etape 2** : Estimer les amplitudes :

$$c^{[\ell+\frac{1}{2}]} = \arg \min_c \left\| z - \sum_{k=1}^K \sum_{j=1}^{N_k^{[\ell+\frac{1}{2}]}} c_{k,j} \zeta_{s_k} \psi\left(\frac{\cdot - x_{k,j}^{[\ell+\frac{1}{2}]}}{s_k}\right) \right\|_2^2 + \gamma \|c\|_1$$

10    **Etape 3** : Déterminer une solution  $m^*$  du problème

non-convexe (11) avec  $\sum_k N_k^{[\ell+\frac{1}{2}]}$  spikes en initialisant par  $\bar{m}_k = \sum_j c_{k,j}^{[\ell+1/2]} \delta_{x_{k,j}^{[\ell+1/2]}}$ .

11    **Etape 4** : Mettre à jour pour tout  $k$

$$N_k^{[\ell+1]} = |m_k^*|, \quad x_k^{[\ell+1]} = x_k^* \text{ et } c_k^{[\ell+1]} = c_k^*.$$

13 **fin**

---

la position de la concentration maximale de BrdU, tandis que l'échelle  $s$  représente la vitesse et l'orientation de la fourche.

Un exemple de données  $z \in \mathbb{R}^n$  est affiché en noir sur la Figure 1. Il correspond au résultat de la moyenne de la concentration de BrdU dans une fenêtre glissante de 100 paires de base (bp) (cf. [11] pour plus de détails).

**Cadre expérimental.** Les tests sont faits pour  $P = 200$  signaux bruités  $z$  de taille  $n = 3000$ , pouvant avoir une ou deux fourches. Les positions et directions des fourches sont choisies aléatoirement en interdisant leur chevauchement et la vitesse de chaque fourche est tirée uniformément dans l'intervalle  $[500, 3000]$ . Pour les méthodes discrète et hybride, la vitesse est discrétisée en  $K = 10$  échelles dans l'intervalle  $[500, 3000]$  bp ainsi que leur opposées. Le choix de  $K$  a été fait pour équilibrer précision et temps de calcul. Nous proposons de considérer un bruit poisson de facteur  $\beta = 100, 140$  ou  $180$ . Le bruit est généré en chaque point à partir du signal non-bruité  $\hat{z}$  et :  $z_i = \text{Poisson}(\hat{z}_i \beta) / \beta$ .

L'objectif de cette section expérimentale est d'évaluer d'une part la prise en compte d'une statistique poissonnienne adaptée au bruit sous-jacent dans une approche de type discrétisée LASSO et d'autre part le gain d'une approche hybride par rapport à LASSO dans le cas d'un dictionnaire hybride discret-continu temps-échelle. Par conséquent, les résultats numériques obtenus avec les trois méthodes : LASSO (2), KL-L1 (9), et Hybride (11) sont comparées dans le contexte de détection de position et vitesse des fourches.

Le choix du paramètre  $\lambda$  dans les cas de LASSO et KL-L1 est

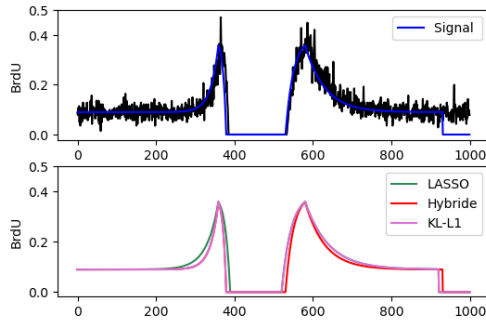


FIGURE 1 : (haut) Exemple de signal simulé contenant deux fourches, avant (bleu) et après (noir) ajout du bruit Poisson ( $\beta = 140$ ). Le pulse de gauche (resp. de droite) correspond à une fourche de réplication se déplaçant vers la gauche (resp. droite) à une vitesse  $s = 500$  bp/min (resp.  $-1000$  bp/min). (bas) Atomes obtenus par les méthodes indiquées. Dans notre implémentation le support des atomes est tronqué à une longueur fixe de 200 points (20 kb).

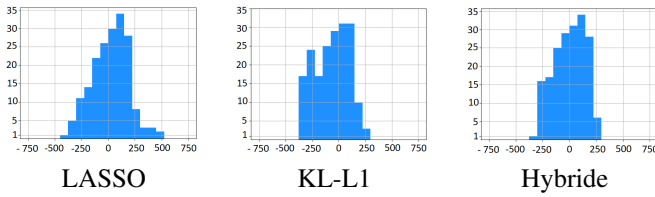


FIGURE 2 : Histogrammes de l'erreur d'estimation de la vitesse (en bp/min) pour  $\beta = 140$ .

fait pour minimiser l'erreur moyenne par rapport à la position et vitesse. Dans le cas de la méthode hybride, le paramètre est choisi pour détecter le moment où l'étape 1 de l'Algorithme 2, détecte une corrélation avec seulement du bruit. Dans nos simulations,  $\lambda = 0.5$ , pour la méthode LASSO,  $\lambda = 2$ , pour la méthode KL et  $\gamma = 0.002$ , pour la méthode hybride.

**Résultats.** Les résultats sont présentés sur les Figure 1 et 2 et le Tableau 1. Dans le tableau, nous observons que l'approche KL-L1 a principalement un impact positif pour l'estimation de la vitesse là où la détection de la position conduit à des performances très proches de LASSO. Nous observons aussi que KL-L1 diffère plus fortement de LASSO quand le bruit est plus élevé (i.e.  $\beta$  plus petit).

La comparaison entre KL-L1 et LASSO par rapport à la vitesse peut être raffinée en regardant les histogrammes de la Figure 2. Nous observons que la distribution de l'erreur en KL-L1 est concentrée dans une intervalle plus petit, entre  $-400$  et  $300$  (en bp/min). Néanmoins, la distribution de LASSO est plus symétrique. La méthode hybride apparaît comme la plus avantageuse concernant la détection de vitesse et position. L'amélioration de l'estimation de la position est estimée à 32% de l'erreur en moyenne et à 20% par rapport à la vitesse.

Cette progression de la méthode hybride doit être relativisée par rapport au temps d'exécution relativement élevé par rapport à LASSO et KL-L1. Pour un signal avec une fourche, LASSO prends 24s, KL-L1 60s, et la méthode hybride 252s.

## 7 Conclusion

Dans le contexte de l'analyse de la réplication de l'ADN en molécule unique, nous avons proposé de traiter le problème de détection de fourches du point de vue du codage parcimonieux. Deux méthodes ont été proposées et comparées avec

Méthode	$\beta$	Erreur absolue en position (bp)	Erreur absolue en vitesse (bp/min)
		Moyenne $\pm$ SEM	Moyenne $\pm$ SEM
LASSO	100	232 $\pm$ 13	139 $\pm$ 7
KL-L1	100	224 $\pm$ 10	127 $\pm$ 5
Hybride	100	128 $\pm$ 5	120 $\pm$ 5
LASSO	140	208 $\pm$ 11	139 $\pm$ 7
KL-L1	140	216 $\pm$ 10	124 $\pm$ 7
Hybride	140	131 $\pm$ 6	114 $\pm$ 6
LASSO	180	199 $\pm$ 10	129 $\pm$ 7
KL-L1	180	200 $\pm$ 12	127 $\pm$ 8
Hybride	180	102 $\pm$ 5	100 $\pm$ 6

TABLE 1 : Erreur moyenne absolue de la position et vitesse de 200 fourches  $\pm$  l'erreur type de la moyenne

la méthode standard LASSO : une méthode utilisant la divergence de KL comme attache aux données, et l'autre, appelée hybride, inspirée par la formulation BLASSO. Nous concluons que la méthode qui utilise la divergence de KL n'a pas augmenté la précision d'estimation de manière significative mais que la méthode hybride a amélioré notablement la précision de la détection de la position et de la vitesse des fourches.

Ce travail a été soutenu par l'Agence Nationale de la Recherche (ANR-18-CE45-0002 et ANR-19-CE12-0028).

## Références

- [1] J.-M. AZAÏS, Y. DE CASTRO et F. GAMBOA : Spike detection from inaccurate samplings. *Applied and Computational Harmonic Analysis*, 38(2):177–195, 2015.
- [2] L. M. BRICENO-ARIAS, P. L. COMBETTES, J.-C. PESQUET et N. PUSTELNIK : Proximal algorithms for multicomponent image recovery problems. *Journal of Mathematical Imaging and Vision*, 41(1-2):3–22, 2011.
- [3] E. CANDÈS et C. FERNANDEZ-GRANDA : Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19:1229–1254, 2013.
- [4] E. CLÉMENT, R. GRIBONVAL, C. SOUSSEN et C. HERZET : OMP and continuous dictionaries : Is k-step recovery possible ? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5546–5550, 2019.
- [5] Q. DENOYELLE, V. DUVAL, G. PEYRÉ et E. SOUBIES : The sliding frank-wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001, 2019.
- [6] D.L. DONOHO, M. ELAD et V.N. TEMLYAKOV : Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006.
- [7] D. GUO, X. QU, X. DU, K. WU et X. CHEN : Salt and pepper noise removal with noise detection and a patch-based sparse representation. *Advances in Multimedia*, 2014:1–14, 03 2014.
- [8] KUSTIYO, ROKHMATULOH, A H SAPUTRO et D KUSHARDONO : Speckle noise reduction of sentinel-1 sar data using fast fourier transform temporal filtering to monitor paddy field area. *IOP Conference Series : Earth and Environmental Science*, 739(1):012086, apr 2021.
- [9] B. LAVILLE, L. BLANC-FÉRAUD et G. AUBERT : Off-the-grid variational sparse spike recovery : Methods and algorithms. *Journal of Imaging*, 7(12), 2021.
- [10] P. L. LIONS et B. MERCIER : Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.
- [11] B. THEULOT *et al.* : Genome-wide mapping of individual replication fork velocities using nanopore sequencing. *Nature Communications*, 13:3295, 2022.
- [12] D. VINCENT et G. PEYRÉ : Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.
- [13] M. J. WAINWRIGHT : Sharp thresholds for high-dimensional and noisy sparsity recovery using  $\ell_1$ -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.