(*A01-public*)

Résumé : On étudie un modèle issu de la biologie qui décrit par exemple l'évolution d'une colonie de bactéries, et qui combine croissance cellulaire et mécanisme de division.

Mots clefs : Loi exponentielle, fonction caractéristique, convergences.

► *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. La présentation, bien que totalement libre, doit être organisée et le jury apprécie qu'un plan soit annoncé en préliminaire. L'exposé doit être construit en évitant la paraphrase et mettant en lumière les connaissances, à partir des éléments du texte. Il doit contenir des illustrations informatiques réalisées sur ordinateur, ou, à défaut, des propositions de telles illustrations. Des pistes de réflexion, indicatives et largement indépendantes les unes des autres, vous sont proposées en fin de texte.*

1. La modélisation

On s'intéresse au destin d'une cellule mère. La masse de la cellule augmente linéairement au cours du temps, de manière déterministe. Au bout d'un temps aléatoire, la cellule se divise en deux parties égales (mitose). Ces deux cellules filles subissent ensuite le même mécanisme, de manière indépendante, etc. Cela crée un arbre binaire de descendance de la cellule d'origine. Les branches de cet arbre sont de même loi. Intéressons-nous à l'évolution de la masse cellulaire le long d'une branche quelconque de l'arbre partant de la cellule mère. Si X_t représente la masse à l'instant $t \geq 0$ alors le processus $(X_t)_{t\geq 0}$ est à valeurs dans $\mathbb{R}_+$, ses trajectoires sont presque sûrement linéaires de pente 1 par morceaux, continues à droite avec limite à gauche, et chaque saut correspond à une division par 2, cf. figure 1.

Posons $T_0 = 0$. Si $0 \leq T_1 \leq T_2 \leq \cdots$ sont les temps (aléatoires) de division alors pour tout entier $n \in \mathbb{N}$ et tout réel $t \in [T_n, T_{n+1}[$:

$$X_t = X_{T_n} + t - T_n \quad \text{et} \quad X_{T_{n+1}} = \frac{1}{2} X_{T_{n+1}^-} = \frac{1}{2}(X_{T_n} + T_{n+1} - T_n). \tag{1}$$

On suppose que la masse de la cellule mère X_0 est une variable aléatoire de carré intégrable, indépendante de la suite des temps de saut $(T_n)_{n\geq 1}$. On suppose également, pour simplifier, que les temps $(T_n)_{n\geq 0}$ sont espacés par des durées $E_1, E_2, \dots$ qui sont des variables aléatoires i.i.d. de loi exponentielle de paramètre $\lambda > 0$. On a donc $T_1 = E_1$, $T_2 = E_1 + E_2$, etc. Les variables $(T_n)_{n\geq 1}$ constituent les instants de division tandis que les variables $(E_n)_{n\geq 1}$ constituent les durées entre les divisions.

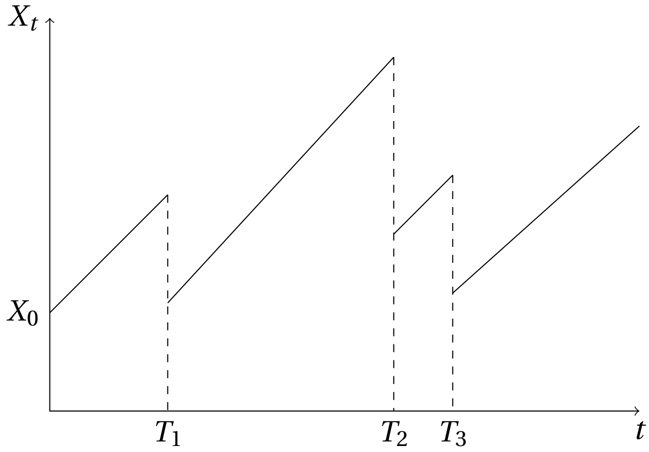

FIGURE 1. Allure d'une trajectoire du processus $(X_t)_{t\geq 0}$.

Rappel statistique

L'estimation du paramètre λ à partir de l'observation des temps $(T_n)_{n\geq 1}$ peut être menée grâce aux convergences classiques :

Théorème 1 (Convergences classiques)**.**

$$\frac{1}{n}T_n \xrightarrow[n\to\infty]{p.s.} \frac{1}{\lambda} \quad et \quad \sqrt{n}\left(\frac{1}{n}T_n - \frac{1}{\lambda}\right) \xrightarrow[n\to\infty]{loi} \mathcal{N}\left(0, \frac{1}{\lambda^2}\right). \tag{2}$$

2. Étude de la masse le long d'une branche de l'arbre

On étudie ici le processus $(X_t)_{t\geq 0}$. Pour simplifier et tirer profit de la construction récursive, on considère plutôt dans un premier temps le processus pris aux instants de saut, en posant $\widetilde{X}_n = X_{T_n}$ pour tout $n \geq 0$.

2.1. *Masses aux temps de saut*

Théorème 2 (Convergence en loi de $(\widetilde{X}_n)_{n\geq 0}$)**.** *La suite des lois des variables $(\widetilde{X}_n)_{n\geq 0}$ converge vers une mesure de probabilité $\widetilde{\mu}$ sur $\mathbb{R}_+$ dont la fonction caractéristique est donnée pour tout $u \geq 0$ par*

$$\int_0^\infty e^{iux}\,\widetilde{\mu}(dx) = \prod_{k=1}^{\infty} \frac{\lambda}{\lambda - 2^{-k}iu}. \tag{3}$$

Cette loi $\widetilde{\mu}$ ne dépend pas de $\widetilde{X}_0$. L'espérance de $\widetilde{\mu}$ vaut $1/\lambda$ *et sa variance* $1/(3\lambda^2)$.

Démonstration. Par récurrence, on obtient, pour tout $n \geq 1$,

$$\widetilde{X}_n = 2^{-n}X_0 + \sum_{k=1}^{n} 2^{-k}E_{n-k+1}. \tag{4}$$

On observe que $\widetilde{X}_n$ a même loi que $\widetilde{Z}_n = 2^{-n}X_0 + \sum_{k=1}^{n} 2^{-k}E_k$. La suite $(2^{-n}X_0)_{n\geq 1}$ converge p.s. vers 0 quand $n \to \infty$. D'autre part, le théorème de convergence monotone montre que la variable aléatoire $\widetilde{Z} = \sum_{k=1}^{\infty} 2^{-k}E_k$ est intégrable et donc p.s. finie. Par conséquent $\widetilde{Z}_n$ converge p.s. et donc en loi vers $\widetilde{Z}$, dont on notera $\widetilde{\mu}$ la loi sur $\mathbb{R}_+$. Comme $\widetilde{X}_n$ et $\widetilde{Z}_n$ ont même loi, $\widetilde{X}_n$ converge également en loi vers $\widetilde{Z}$.

La formule pour la fonction caractéristique de $\widetilde{\mu}$ découle du calcul de la fonction caractéristique de $\sum_{k=1}^{n} 2^{-k}E_k$. Enfin, les deux premiers moments de $\widetilde{\mu}$ s'obtiennent par exemple en utilisant les formules

$$\mathbb{E}(\widetilde{X}_n) = 2^{-n}\mathbb{E}(X_0) + \lambda^{-1}\sum_{k=1}^{n} 2^{-k} \quad \text{et} \quad \mathbb{V}\mathrm{ar}(\widetilde{X}_n) = 2^{-2n}\mathbb{V}\mathrm{ar}(X_0) + \lambda^{-2}\sum_{k=1}^{n} 2^{-2k}. \tag{5}$$

□

2.2. *Retour au processus en temps continu*

Note : Cette partie peut être sautée en première lecture.

Concernant le processus $(X_t)_{t\geq 0}$ lui-même, des arguments plus élaborés permettent d'établir l'existence d'une mesure de probabilité μ sur $\mathbb{R}_+$ et d'une variable aléatoire Z de loi μ telles que X_t converge en loi vers Z en $t \to \infty$, pour toute condition initiale X_0. On pourra admettre ce résultat. On souhaite à présent décrire la vitesse de cette convergence. Soit $\mathscr{P}$ l'ensemble des mesures de probabilités de moment d'ordre 1 fini. Pour tous α et β dans $\mathscr{P}$, on définit la *distance de couplage* $D(\alpha, \beta)$ entre α et β par

$$D(\alpha;\beta) = \inf_{\Pi} \int |x-y|\,\Pi(dx,dy) \tag{6}$$

où l'infimum porte sur l'ensemble des lois de probabilité Π sur $\mathbb{R}_+ \times \mathbb{R}_+$ dont les lois marginales sont α et β. Cet ensemble n'est pas vide car il contient la loi produit $\alpha \otimes \beta$. On admettra que l'application $D : \mathscr{P} \times \mathscr{P} \to \mathbb{R}_+$ est bien une distance sur $\mathscr{P}$, et que pour toute suite $(U_n)_{n\geq 0}$ de variables aléatoires réelles intégrables et toute variable aléatoire réelle intégrable U, on a, en notant $\mathscr{L}(U_n)$ (resp. $\mathscr{L}(U)$) la loi de U_n (resp. U)

$$\lim_{n\to\infty} D(\mathscr{L}(U_n), \mathscr{L}(U)) = 0 \quad \text{si et seulement si} \quad U_n \xrightarrow[n\to\infty]{\text{loi}} U \quad \text{et} \quad \mathbb{E}(U_n) \xrightarrow[n\to\infty]{} \mathbb{E}(U). \tag{7}$$

L'identité (4) suggère que $(X_t)_{t\geq 0}$ *oublie* exponentiellement vite sa condition initiale. Cette intuition est confirmée par le résultat suivant en terme de distance de couplage D.

Théorème 3 (Oubli exponentiellement rapide de la condition initiale)**.** *Soient deux copies $(X_t)_{t\geq 0}$ et $(Y_t)_{t\geq 0}$ du processus associées aux conditions initiales respectives $X_0 = x$ p.s. et $Y_0 = y$ p.s. Alors pour tout $t \geq 0$:*

$$D(\mathscr{L}(X_t);\mathscr{L}(Y_t)) \leq |x-y| \exp\left(-\frac{1}{2}\lambda t\right). \tag{8}$$

Démonstration. Il suffit de proposer une construction de $(Y_t)_{t\geq 0}$ de sorte que pour tout $t \geq 0$:

$$\mathbb{E}(|X_t - Y_t|) \leq |x-y| \exp\left(-\frac{1}{2}\lambda t\right). \tag{9}$$

Si $(T_n)_{n\geq 0}$ est la suite des temps de saut de $(X_t)_{t\geq 0}$, nous construisons $(Y_t)_{t\geq 0}$ avec la même suite $(T_n)_{n\geq 1}$ mais avec cette fois-ci $Y_0 = y$. Soit $(C_t)_{t\geq 0}$ le processus de comptage de la suite de *tops* $(T_n)_{n\geq 1}$, défini pour tout $t \geq 0$ par

$$C_t = \sum_{n=1}^{\infty} \mathrm{I}_{\{T_n \leq t\}}. \tag{10}$$

L'événement $\{C_t = k\}$ correspond à dire qu'il y a exactement k tops dans l'intervalle $[0, t]$. Comme les temps de sauts de $(X_t)_{t\geq 0}$ et $(Y_t)_{t\geq 0}$ sont identiques, on a $|X_t - Y_t| = |x - y|2^{-k}$ sur l'événement $\{C_t = k\}$, pour tout entier $k \geq 0$, et donc

$$\mathbb{E}(|X_t - Y_t|) = \sum_{k=0}^{\infty} \mathbb{E}\left(|X_t - Y_t|\mathrm{I}_{\{C_t = k\}}\right) = |x - y| \sum_{k=0}^{\infty} \frac{1}{2^k}\mathbb{P}(C_t = k). \tag{11}$$

Or C_t suit la loi de Poisson de moyenne λt et donc $\mathbb{E}(|X_t - Y_t|) = |x - y| \exp\left(-\frac{1}{2}\lambda t\right)$. □

3. Étude de la taille de la population

Reconsidérons l'arbre de descendance $\mathscr{A}$ de la cellule mère de l'introduction. Il s'agit d'un arbre binaire infini. Sur chaque nœud de l'arbre, nous avons disposé une variable aléatoire exponentielle de paramètre λ, que nous appelons à présent *marque*, qui modélise la durée écoulée avant la division. Ces variables aléatoires sont i.i.d.

Le *score* d'un nœud de $\mathscr{A}$ est la somme des marques de tous les nœuds de ses ancêtres (lui compris) jusqu'au nœud racine. La *profondeur* d'un nœud de $\mathscr{A}$ est égale au nombre de ses ancêtres (lui compris), c'est-à-dire au nombre d'éléments de la somme dans le calcul de son score. Le nœud racine est de profondeur 1, et son score suit une loi exponentielle de paramètre λ. Pour tout entier $k \geq 1$, l'arbre $\mathscr{A}$ possède exactement 2^{k-1} nœuds de profondeur k, et le score d'un nœud de profondeur k suit la loi de la somme de k variables aléatoires exponentielles indépendantes de paramètre λ.

Pour tout temps $t \geq 0$, on construit le sous–arbre $\mathscr{A}_t$ en supprimant tous les nœuds de $\mathscr{A}$ dont le score dépasse t. Ainsi, $\mathscr{A}_t$ est un arbre binaire fini, qui représente l'histoire de la population de cellules jusqu'à l'instant t. Les nœuds terminaux (c'est-à-dire sans nœuds enfants) de $\mathscr{A}_t$ représentent les divisions qui ont eu lieu avant l'instant t. On note N_t leur nombre. Il y a alors $N_t + 1$ cellules à l'instant t. Pour $t = 0$, on a $N_0 = 0$ (la cellule mère ne s'est pas encore divisée). On souhaite décrire la loi du processus de comptage $(N_t)_{t\geq 0}$.

Étant donnée une suite de variables aléatoires $(R_n)_{n\geq 1}$, on rappelle qu'on définit le processus de comptage $(C_t)_{t\geq 0}$ des tops donnés par $(R_n)_{n\geq 1}$ de la manière suivante : pour tout $t \geq 0$,

$$C_t = \sum_{n=1}^{\infty} \mathrm{I}_{\{R_n \leq t\}}. \tag{12}$$

On a alors le résultat suivant.

Théorème 4 (Loi de la taille de la population)**.** *Le processus $(N_t)_{t\geq 0}$ suit la même loi que le processus de comptage de tops espacés par des durées aléatoires indépendantes $S_1, S_2, S_3 \ldots$ de loi exponentielle de paramètre $\lambda, 2\lambda, 3\lambda, \ldots$.*

Par conséquent, pour tout $t \geq 0$, la variable N_t suit la loi géométrique sur $\mathbb{N}^*$ de paramètre $e^{-\lambda t}$, ce qui signifie que pour tout $n \geq 0$:

$$\mathbb{P}(N_t \geq n) = (1 - e^{-\lambda t})^n. \tag{13}$$

En particulier, on a pour tout $t \geq 0$:

$$\mathbb{E}(N_t) = e^{\lambda t} - 1 \quad \text{et} \quad \mathbb{V}\mathrm{ar}(N_t) = (1 - e^{-\lambda t})e^{2\lambda t}. \tag{14}$$

Démonstration. On a $N_0 = 0$ et les trajectoires du processus $(N_t)_{t \geq 0}$ sont constantes par morceaux, avec des sauts d'amplitude 1. L'événement $\{N_t = 0\}$ signifie « le score du nœud racine est plus grand que t ». L'événement $\{N_t = 1\}$ signifie « le score du nœud racine est plus petit que t et les scores des deux nœuds enfants du nœud racine sont plus grands que t ». Ceci précise les lois de S_1 et S_2. Plus généralement, les lemmes 5 et 6 ci-dessous permettent d'établir que $(N_t)_{t \geq 0}$ a la même loi que le processus de comptage de tops espacés par des durées aléatoires indépendantes $S_1, S_2, S_3 \ldots$ de loi exponentielle de paramètre $\lambda, 2\lambda, 3\lambda, \ldots$. Cela conclut la première partie du théorème.

Pour la seconde partie, on note qu'on a par conséquent

$$\mathbb{P}(N_t \geq n) = \mathbb{P}(S_1 + \cdots + S_n \leq t) \tag{15}$$

pour tout $t \geq 0$ et tout $n \geq 0$. Le lemme 7 permet de conclure. □

La preuve des deux premiers lemmes ci-dessous est omise car classique.

Lemme 5 (Minimum)**.** *Si $E_1, \ldots, E_n$ sont des variables aléatoires indépendantes de lois exponentielles de paramètre $\lambda_1, \ldots, \lambda_n$ alors $\min(E_1, \ldots, E_n)$ suit une loi exponentielle de paramètre $\lambda_1 + \cdots + \lambda_n$.*

Lemme 6 (Amnésie)**.** *Si E est une variable aléatoire de loi exponentielle de paramètre λ alors pour tout $t \geq 0$, la loi conditionnelle $\mathscr{L}(E - t \,|\, E > t)$ est aussi exponentielle de paramètre λ.*

Lemme 7 (Maximum)**.** *Soit $S_1, S_2, S_3, \ldots$ des variables aléatoires exponentielles indépendantes de paramètres respectifs $\lambda, 2\lambda, 3\lambda, \ldots$ pour un réel $\lambda > 0$ fixé. Alors pour tout $n \geq 1$, la variable aléatoire $S_1 + \cdots + S_n$ suit la même loi que $\max(E_1, \ldots, E_n)$ où $E_1, \ldots, E_n$ sont des variables aléatoires exponentielles i.i.d. de paramètre λ.*

Démonstration. On procède par récurrence sur n. Le résultat est trivialement vrai pour $n = 1$. Montrons que si le résultat est vrai pour n alors il est vrai pour $n + 1$. On a

$$\mathbb{P}(\max(E_1, \ldots, E_{n+1}) \leq t) = \mathbb{P}(E_1 \leq t) \cdots \mathbb{P}(E_{n+1} \leq t) = (1 - e^{-\lambda t})^{n+1}. \tag{16}$$

Par ailleurs, on a en utilisant l'hypothèse de récurrence :

$$\begin{aligned}
\mathbb{P}(S_1 + \cdots + S_{n+1} \leq t) &= \int_0^t \mathbb{P}(S_1 + \cdots + S_n + s \leq t \,|\, S_{n+1} = s)\,(n+1)\lambda e^{-(n+1)\lambda s}\,ds \\
&= (n+1)\lambda \int_0^t \mathbb{P}(S_1 + \cdots + S_n \leq t - s)\, e^{-(n+1)\lambda s}\,ds \\
&= (n+1)\lambda \int_0^t (1 - e^{-\lambda(t-s)})^n\, e^{-(n+1)\lambda s}\,ds.
\end{aligned} \tag{17}$$

Le développement de $(1 - e^{-\lambda(t-s)})^n$ grâce à la formule du binôme permet de conclure. □

Suggestions et pistes de réflexion

► *Les pistes de réflexion suivantes ne sont qu'indicatives et il n'est pas obligatoire de les suivre. Vous pouvez choisir d'étudier, ou non, certains des points proposés, de façon plus ou moins approfondie, mais aussi toute autre question à votre initiative. Vos investigations comporteront une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats. À défaut, si vos illustrations informatiques n'ont pas abouti, il est conseillé d'expliquer ce que vous auriez souhaité mettre en œuvre.*

- *Aspects mathématiques*
  - Comment construire un intervalle de confiance asymptotique pour λ à partir de l'observation de T_n ? Quelle est la loi de T_n ? Peut-on déterminer un intervalle de confiance non–asymptotique pour λ ?
  - La suite $(\sum_{k=1}^{n} 2^{-k} E_{n-k+1})_{n\geq 1}$ converge-t-elle presque sûrement lorsque $n \to \infty$?
  - Comment obtenir les moments de $\widetilde{\mu}$ à partir de sa fonction caractéristique ?
  - Les scores des nœuds de même profondeur sont-ils indépendants ?
- *Programmes et simulations*
  - On pourra simuler des trajectoires des processus $(X_t)_{t\geq 0}$ et $(N_t)_{t\geq 0}$
  - On pourra donner un histogramme approché de $\widetilde{\mu}$ et μ pour différents λ
  - On pourra illustrer l'estimation de λ et construire un intervalle de confiance.
  - On pourra vérifier par une simulation que $e^{-\lambda t} N_t$ converge en loi quand $t \to \infty$.
- *Modélisation*
  - On pourra considérer la masse totale de la population au temps t
  - On pourra discuter l'hypothèse faite sur $(T_n)_{n\geq 1}$
  - On pourra raffiner la modélisation (prise en compte d'une valeur seuil pour la masse cellulaire ; dépendance de λ en la masse ou le temps ; division en plusieurs fragments de tailles inégales)