# Estimation du noyau de division dans une population structrurée par la taille

Van Ha Hoang [1]

[1] *Laboratoire Paul Painlevé, Université de Lille 1*
*Cité Scientifique, 59655 Villeneuve D'Ascq, France*
*van-ha.hoang@ed.univ-lille1.fr*

**Résumé.** Dans ce travail, nous considérons une population de cellules structurée par la taille. La taille des cellules croît de façon déterministe et les cellules se divisent à des temps exponentiels. La population est décrite par une mesure empirique et nous observons les divisions sur l'intervalle de temps continu $[0, T]$. Nous nous intéressons ici au problème d'estimation du noyau de division $h(\cdot)$ (ou noyau de fragmentation) dans le cas de données complètes. Nous construisons un estimateur adaptatif à noyau $K$ fondé sur un choix de fenêtre inspiré par la méthode de Goldenschluger et Lepski. Nous obtenons une inégalité oracle et une vitesse de convergence exponentielle.

**Mots-clés.** Population structurée par la taille, noyau de division, estimation non-paramétrique.

**Abstract.** We consider a size-structured population which represents the cell division. The size of the cells grows in a deterministic way and the cells divide with exponential times. The cell population is described by an empirical measure and we observe the divisions in the continuous time interval $[0, T]$. We address here the problem of estimating the division kernel $h(\cdot)$ (or fragmentation kernel) in the case of complete data. Then, we construct an adaptive estimator of $h$ based on a kernel function $K$ with a fully data-driven bandwidth selection method based on Goldenschluger et Lepski. Finally, we obtain an oracle inequality and an exponential convergence rate.

**Keywords.** size-structured population, division kernel, nonparametric estimation.

# 1 Introduction

Dans ce papier, nous nous intéressons à un modèle stochastique individu-centré en temps continu de population structurée par la taille, dont les individus sont des cellules se divisant de façon binaire (les individus sont caractérisés par des variables qui croissent de manière déterministe avec le temps tels que le volume, la longueur, les niveaux des protéines, le contenu de l'ADN, *etc.*...). De telles classes de modèles ont été étudiées par Athreya and Ney [1], Harris [7], Jagers [9] et font l'objet d'une abondante littérature (*e.g.* Bansaye *et al.* [2], Cloez [3], Tran [13]). Nous avons ici à l'esprit que chaque cellule contient une certaine toxicité qui joue le rôle de la taille dans l'étude de Stewart *et. al.* [12].

Dans notre modèle, une cellule qui contient une toxicité $x \in \mathbb{R}_+$ se divise en deux filles avec une vitesse $R > 0$. La toxicité croît dans la cellule avec un taux de croissance $\alpha > 0$. En général, la vitesse de division et le taux de croissance peuvent être non constants mais nous considérons ici le cas constant. Lorsqu'une cellule se divise, une fraction aléatoire $\Gamma \in [0, 1]$ de la toxicité va à la première fille et une fraction $1 - \Gamma$ à la seconde. Nous supposons que $\Gamma$ a une distribution symétrique par rapport à $\frac{1}{2}$ sur $[0, 1]$ avec une densité $h$ par rapport à la mesure de Lebesgue.

Le modèle individu-centré fournit un cadre naturel à l'estimation statistique et à l' estimation de la vitesse de division $R$ qui constituent les sujet abordés dans les travaux de Doumic *et al.* [4, 5], Hoffmann et Olivier [8]. Dans ce travail, la densité $h$ est le noyau de division que nous souhaitons estimer. L'intérêt de l'estimation de $h$ est de détecter les phénomènes du vieillissement mis en lumière par Stewart *et al.* [12]. Si les divisions sont inégales, on peut considérer qu'une fille est plus âgée. Par conséquent, si les cellules sont suivies dans le temps et si on connaît la densité $h$, ainsi que le taux de croissance des toxicités qu'elles contiennent, on peut déterminer l'effet du vieillissement.

# 2 Modèle microscopique

Dans cette section, nous introduisons une équation différentielle stochastique (EDS) gouvernée par une mesure ponctuelle de Poisson pour décrire la population cellulaire. Nous rappelons la notation de Ulam-Harris-Neveu utilisée pour décrire l'arbre généalogique. La première cellule est marquée par $\emptyset$ et lorsque la cellule $i$ se divise, les deux descendants sont marqués par $i0$ and $i1$. L'ensemble des labels est

$$J = \{\emptyset\} \cup \bigcup_{m=1}^{\infty} \{0, 1\}^m . \tag{1}$$

On note $V_t$ l'ensemble des cellules vivantes au temps $t$, et $V_t \subset J$.

La toxicité $(X_t, t \geq 0)$ satisfait l'équation :

$$dX_t = \alpha dt. \tag{2}$$

À chaque individu $i$, on peut associer une toxicité $X_t^i$ qui suit (2) et correspond à la quantité de toxicité contenue dans la cellule $i$. Au moment de la division au temps $t$, $X_t^{i0} = X_{t-}^i \times \Gamma_i$ et $X_t^{i1} = X_{t-}^i \times (1 - \Gamma_i)$ où $\Gamma_i$ sont des variables aléatoires indépendantes de densité $h$.

Soit $\mathcal{M}_F(\mathbb{R}_+)$ l'espace des mesures finies sur $\mathbb{R}_+$ muni de la topologie de la convergence étroite. Nous décrivons la population cellulaire au temps $t$ par une mesure ponctuelle aléatoire sur $\mathcal{M}_F(\mathbb{R}_+)$ :

$$Z_t(dx) = \sum_{i \in V_t} \delta_{X_t^i}(dx), \quad \text{et} \quad N_t = \langle Z_t, 1 \rangle = \int_{\mathbb{R}_+} Z_t(dx) \tag{3}$$

est le nombre d'individus vivants au temps $t$. Pour une mesure $\mu \in \mathcal{M}_F(\mathbb{R}_+)$ et une fonction positive $f$, on utilise la notation $\langle \mu, f \rangle = \int_{\mathbb{R}_+} f d\mu$.

Soit $Z_0 \in \mathcal{M}_F(\mathbb{R}_+)$ une condition initiale telle que

$$\mathbb{E}(\langle Z_0, 1 \rangle) < +\infty, \tag{4}$$

et soit $Q(ds, di, d\gamma)$ une mesure ponctuelle de Poisson sur $\mathbb{R}_+ \times \mathcal{E} := \mathbb{R}_+ \times J \times [0,1]$ d'intensité $q(ds, di, d\gamma) = R\, ds\, n(di) H(d\gamma)$. $n(di)$ est la mesure de comptage sur $J$ et $ds$ est la mesure de Lebesgue sur $\mathbb{R}_+$. Nous notons $\{\mathcal{F}\}_{t\geq 0}$ la filtration canonique engendrée par $Z_0$ et $Q$. Le processus stochastique $(Z_t)_{t\geq 0}$ est décrit par l' EDS suivante :

**Définition 1.** *Pour toute fonction $f_t(x) = f(x,t) \in \mathcal{C}_b^{1,1}(\mathbb{R}_+ \times \mathbb{R}_+, \mathbb{R})$, nous avons*

$$\begin{aligned}\langle Z_t, f_t \rangle = \langle Z_0, f_0 \rangle &+ \int_0^t \int_{\mathbb{R}_+} \big(\partial_s f_s(x) + \alpha \partial_x f_s(x)\big) Z_s(dx) ds \\ &+ \int_0^t \int_{\mathcal{E}} \mathbb{1}_{\{i \leq N_{s-}\}} \Big[ f_s\left(\gamma X_{s-}^i\right) + f_s\left((1-\gamma) X_{s-}^i\right) - f_s\left(X_{s-}^i\right) \Big] Q(ds, di, d\gamma). \end{aligned} \tag{5}$$

Le deuxième terme du membre de droite de (5) correspond à la croissance des toxicités dans les cellules et le troisième terme donne une description de la division cellulaire où le partage de toxicité en deux filles dépend de la fraction aléatoire $\Gamma$.

# 3 Estimation du noyau de division

### Données et construction de l'estimateur

Supposons que l'on observe complètement l'évolution de la population cellulaire sur $[0,T]$. Au $i^{\text{ème}}$ temps de la division $t_i$, on note $j_i$ l'individu qui se divise en deux filles $X_{t_i}^{j_i 0}$ et $X_{t_i}^{j_i 1}$ et on définit :

$$\Gamma_i^0 = \frac{X_{t_i}^{j_i 0}}{X_{t_i-}^{j_i}} \quad \text{and} \quad \Gamma_i^1 = \frac{X_{t_i}^{j_i 1}}{X_{t_i-}^{j_i}},$$

les fractions aléatoires qui vont dans les filles, avec la convention $\frac{0}{0} = 0$.

$\Gamma_i^0$ et $\Gamma_i^1$ sont échangeables avec $\Gamma_i^0 + \Gamma_i^1 = 1$, $\Gamma_i^0$ et $\Gamma_i^1$ ne sont donc pas indépendantes mais les couples $(\Gamma_i^0, \Gamma_i^1)_{i \in \mathbb{N}^*}$ sont indépendants identiquement distribués selon $(\Gamma^0, \Gamma^1)$ où $\Gamma^0 \sim H(d\gamma)$ et $\Gamma^1 = 1 - \Gamma^0$.

Il est naturel d'utiliser une méthode à noyau pour estimer la densité $h$. Nous définissons un estimateur $\hat{h}_\ell$ de $h$ basé sur les observations $(\Gamma_i^0, \Gamma_i^1)_{i \in \mathbb{N}^*}$.

**Définition 2.** *Soit $N_T$ le nombre aléatoire de divisions sur intervalle du temps $[0,T]$. Pour tout $\gamma \in (0,1)$, définissons*

$$\hat{h}_\ell(\gamma) = \frac{1}{N_T} \sum_{i=1}^{N_T} K_\ell(\gamma - \Gamma_i^1), \tag{6}$$

*où $K_\ell = \frac{1}{\ell} K(\cdot/\ell)$, $K$ est un noyau, $\ell > 0$ est la fenêtre à choisir.*

Il est à noter que $N_T \geq 1$ pour tout $T \in \mathbb{R}_+$ car la population cellulaire commence avec une cellule mère au début.

Dans (6), $\hat{h}_\ell$ depend également de $T$. Cependant, nous omettons $T$ pour simplifier les notations. Les fractions aléatoires $(\Gamma_i^1)_{i\in\mathbb{N}^*}$ et $N_T$ sont indépendantes sous l'hypothèse que le taux de division $R$ est une constante qui ne dépend pas de la toxicité de la cellule qui se divise.

**Estimation adaptative de $h$ par la méthode de Goldenshluger et Lepski**

Nous appliquons la méthode de Goldenschluger et Lepski [6] (notée GL) afin de choisir une fenêtre pour le noyau $\hat{h}_\ell$. Nous étudions d'abord le risque $L_2$ de $\hat{h}_\ell$ conditionnellement à $N_T$.

**Proposition 1.** *Nous avons que*

$$\mathbb{E}\Big[\|\hat{h}_\ell - h\|_2 \big| N_T\Big] \leq \|h - K_\ell \star h\|_2 + \frac{\|K\|_2}{\sqrt{N_T \ell}}. \tag{7}$$

La fenêtre oracle $\bar{\ell}$ est telle que

$$\bar{\ell} := \underset{\ell\in\mathcal{H}}{\operatorname{argmin}} \Big\{ \|h - K_\ell \star h\|_2 + \frac{\|K\|_2}{\sqrt{N_T \ell}} \Big\}. \tag{8}$$

Elle est obtenue en minimisant la décomposition biais-variance dans le membre de droite de (7) mais elle ne peut pas être utilisée en pratique car elle dépend encore de $h$ qui est inconnue. Pour appliquer la méthode GL, nous définissons pour tout $\ell, \ell' \in \mathcal{H}$ :

$$\hat{h}_{\ell,\ell'}(\gamma) := \frac{1}{N_T} \sum_{i=1}^{N_T} \big(K_\ell \star K_{\ell'}\big)(\gamma - \Gamma_i^1) = \Big(K_\ell \star \hat{h}_{\ell'}\Big)(\gamma).$$

Alors, la fenêtre adaptative et l'estimateur de $h$ sont choisis comme suit :

**Définition 3.** *Soit $\epsilon > 0$ et $\chi := (1+\epsilon)(1 + \|K\|_1)$, nous définissons*

$$\hat{\ell} := \underset{\ell\in\mathcal{H}}{\operatorname{argmin}} \Big\{ A(\ell) + \frac{\chi\|K\|_2}{\sqrt{N_T \ell}} \Big\}, \tag{9}$$

*où, pour tout $\ell \in \mathcal{H}$,*

$$A(\ell) := \sup_{\ell'\in\mathcal{H}} \Big\{ \|\hat{h}_{\ell,\ell'} - \hat{h}_{\ell'}\|_2 - \frac{\chi\|K\|_2}{\sqrt{N_T \ell'}} \Big\}_+, \tag{10}$$

*Alors, l'estimateur $\hat{h}$ est donné par*

$$\hat{h} := \hat{h}_{\hat{\ell}}. \tag{11}$$

Le théorème suivant nous donne une inégalité oracle pour le risque $L_2$ de notre procédure adaptative.

**Théorème 1.** *Soit $T > 0$ et $\epsilon > 0$. Supposons que les observations sont enregistrées sur $[0, T]$. Soit $N_0$ le nombre de cellules initiales au temps $t = 0$ et $N_T$ est le nombre aléatoire de divisions sur $[0, T]$. Considérons $\mathcal{H}$ un sous-ensemble dénombrable $\{D^{-1} : D = 1, \ldots, D_{\max}\}$ dans lequel nous choisissons les fenêtres et $D_{\max} = \lfloor \delta N_T \rfloor$ pour $\delta > 0$. Supposons $h \in L^\infty$ et nous considérons la fonction noyau bornée $K$. Soit $\hat{h}$ un estimateur à noyau défini par le noyau $K_{\hat{\ell}}$ où $\hat{\ell}$ est choisi par la méthode GL. Alors*

$$\mathbb{E}\Big[\|\hat{h} - h\|_2^2\Big] \leq C_1 \inf_{\ell \in \mathcal{H}} \left\{ \|K_\ell \star h - h\|_2^2 + \frac{\|K\|_2^2}{\ell} e^{-\rho T} \right\} + C_2 e^{-\rho T}, \tag{12}$$

*où $\rho = \frac{N_0 R}{N_0 + 1}$, $C_1$ est une constante dépendante de $\|K\|_1$ et de $\epsilon$ et $C_2$ est une constante dépendante de $\delta$, $\epsilon$, $\|K\|_1$, $\|K\|_2$, $\|h\|_\infty$.*

Pour obtenir une vitesse de convergence de $\hat{h}$, nous supposons que la densité $h$ et la fonction noyau satisfont les hypothèses suivantes :

**Hypothèse 1.** *Soit $\beta > 0$, $L > 0$ et soit $k = \lfloor \beta \rfloor$. On suppose que la densité $h$ appartient à la classe de Hölder $\mathcal{H}(\beta, L)$ définie par :*

$$h \in \mathcal{H}(\beta, L) \Longleftrightarrow h \in C^k \quad et \quad \big|h^{(k)}(y) - h^{(k)}(x)\big| \leq L\big|x - y\big|^{\beta - k}, \forall x, y.$$

**Hypothèse 2.** *Il existe un entier positif $k = \lfloor \beta \rfloor$ tel que pour $j = 1, 2, \ldots, k$,*

$$\int x^j K(x) dx = 0 \quad et \quad \int |x|^\beta |K(x)| dx < +\infty.$$

Alors, la vitesse de convergence est donnée par la proposition suivante.

**Corollaire 1.** *On se place sous les Hypothèses 1, 2 et les hypothèses du Théorème 1. Alors, pour tout $T > 0$, $\beta > 0$ et $L > 0$, l'estimateur $\hat{h}$ satisfait*

$$\sup_{h \in \mathcal{H}(\beta, L)} \mathbb{E}\|\hat{h} - h\|_2^2 \leq C_3 \exp\left( -\frac{2\beta}{2\beta + 1} \rho T \right), \tag{13}$$

*où $C_3$ est une constante dépendante de $\beta$, $L$ et le noyau $K$.*

**Remarque 1.** *Avec l'estimateur $\hat{h}$, nous obtenons la vitesse de convergence $\exp\left(-\frac{2\beta}{2\beta+1}\rho T\right)$ dans (13). Comme notre modèle repose sur une division binaire des cellules, le nombre de cellules augmente de façon exponentielle. Ainsi, l'estimateur $\hat{h}$ est construit sur la taille d'un échantillon qui augmente géométriquement avec $T$, expliquant la vitesse de convergence exponentielle dans (13). Ce résultat est à comparer aux vitesses de convergence polynomiales non-paramétriques habituelles, $n^{\frac{-2\beta}{2\beta+1}}$ pour $n$ observations (voir par exemple [14]).*

# Bibliographie


[1] Athreya, K.B. and Ney, P.E (1970). Branching Processes. *Springer edition*.

[2] Bansaye, V. and Tran, V.C. (2011). Branching Feller diffusion for cell division with parasite infection. *ALEA, Lat. Am. J. Probab. Math. Stat.*,8:95-127.

[3] Cloez, B. (2011). Limit theorems for some branching measure-valued processes. hal-00598030.

[4] Doumic, M. and Hoffmann, M. and Krell, N. and Robert, L. (2012). Statistical estimation of a growth-fragmentation model observed on a genealogical tree. arXiv:1210.3240.

[5] Doumic, M. and Hoffmann, M. and Reynaud-Bouret, P. and Rivoirard, V. (2012). Nonparametric estimation of the division rate of a size-structured population. *SIAM Journal on Numerical Analysis*, 50(2):925-950.

[6] Goldenschluger, A. and Lepski, O. (2011). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608-1632.

[7] Harris, T.E. (1963). The Theory of Branching Processes. *Springer, Berlin*.

[8] Hoffmann, M. and Olivier, A. (2014). Nonparametric estimation of the division rate of an age dependent branching process. arXiv:1412.5936.

[9] Jagers, P. (1969). A general stochastic model for population development. *Scandinavian Actuarial Journal*, (1-2):84-103.

[10] Massart, P. (2007). Concentration Inequalities and Model Selection. *Springer, Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, 6-23, 2003*.

[11] Reynaud-Bouret, P. and Rivoirard, V. and Grammont, F. and Tuleau-Malot, C. (2014). Goodness-of-Fit Tests and Nonparametric Adaptive Estimation for Spike Train Analysis. *Journal of Mathematical Neuroscience*, doi:10.1186/2190-8567-4-3.

[12] Stewart, E.J. and Madden, R. and Paul, G. and Taddei, F. (2005). Aging and Death in an Organism That Reproduces by Morphologically Symmetric Division. *PLOS Biology*, 3(2):10.1371/journal.pbio.0030045.

[13] Tran, V.C. (2008). Large population limit and time behaviour of a stochastic particle model describing an age-structured population. ESAIM: Probability and Statistics, 12:345-386.

[14] Tsybakov, A. B (2004).Introduction to Nonparametric Estimation. *Springer series in Statistics*.