Statistical Estimation of the Division Kernel of a Size-structured Population

Van Ha Hoang
Laboratoire Paul Painlevé, Université Lille 1

Colloque Jeunes Probabilistes et Statisticiens 2016
(April 18 - 22, Les Houches)

Outline

Microscopic model

- We study a stochastic individual-based model of size-structured population in continuous time.
- Individuals are cells undergoing binary divisions.
- Genealogical tree:
  - A cell divides at a rate $R > 0$ and the toxicity grows with rate $\alpha > 0$.
  - When a cell divides, a random fraction Γ of the toxicity goes in the first daughter cell and a fraction $1 - \Gamma$ in the second one. We assume that Γ has a symmetric distribution on $[0, 1]$ with a density h.
- Along branches: the toxicity $(X_t, t \geq 0)$ satisfies

$$\mathrm{d}X_t = \alpha \, \mathrm{d}t.$$

- Goal: estimate the density h (called the division kernel). The interest of estimating h is to detect aging phenomena.

Empirical measure

Let $\mathcal{M}_F(\mathbb{R}_+)$ be the space of finite measures on $\mathbb{R}_+$ embedded with the topology of weak convergence, we describe the population of cells at time t by a random point measure in $\mathcal{M}_F(\mathbb{R}_+)$:

$$Z_t(\mathrm{d}x) = \sum_{i=1}^{N_t} \delta_{X_t^i}(\mathrm{d}x),$$

where

$$N_t = \langle Z_t, 1 \rangle = \int_{\mathbb{R}_+} Z_t(\mathrm{d}x),$$

is the number of individuals living at time t. For a measure $\mu \in \mathcal{M}_F(\mathbb{R}_+)$ and a positive function f, we use the notation $\langle \mu, f \rangle = \int_{\mathbb{R}_+} f \, \mathrm{d}\mu$.

Stochastic differential equation

- Z_t is described by a SDE driven by a Poisson point measure.
- If $Z_0 \in \mathcal{M}_F(\mathbb{R}_+)$ is such that $\mathbb{E}\left(\langle Z_0, 1\rangle\right) < +\infty$, then for all test function $f_t(x) = f(x,t) \in \mathcal{C}_b^{1,1}(\mathbb{R}_+ \times \mathbb{R}_+, \mathbb{R})$, the population of cells is described by:

$$\langle Z_t, f_t\rangle = \langle Z_0, f_0\rangle + \int_0^t \int_{\mathbb{R}_+} \left(\partial_s f_s(x) + \alpha \partial_x f_s(x)\right) Z_s(dx)ds$$
$$+ \int_0^t \int_0^1 \left[f_s(\gamma x) + f_s\left((1-\gamma)x\right) - f_s(x)\right] Rh(\gamma)\, \mathrm{d}\gamma\, \mathrm{d}s + \mathfrak{M}_t^f,$$

where $\mathfrak{M}_t^f$ is a square integrable martingale.

Influence of the distribution on the mean age

Definition

The mean age of the cell population up to time $t \in \mathbb{R}_+$ is defined by:

$$\bar{X}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} X_t^i = \frac{\langle Z_t, f \rangle}{N_t},$$

where $f(x) = x$.

Following the work of Bansaye *et al.* (2011), we note that the long time behavior of the mean age is related to the law of an auxiliary process Y started at $Y_0 = \frac{X_0}{N_0}$ with infinitesimal generator characterized by $\forall f \in \mathcal{C}_b^{1,1}(\mathbb{R}_+, \mathbb{R})$,

$$Af(x) = \alpha f'(x) + 2R \int_0^1 \left(f(\gamma x) - f(x) \right) h(\gamma) d\gamma.$$

Influence of the distribution on the mean age

The auxiliary process Y satisfies ergodic properties (see Bansaye *et al.*, 2011) which entail the following theorem.

Theorem

For $t \in \mathbb{R}_+$,

$$\lim_{t\to+\infty} \bar{X}_t = \lim_{t\to+\infty} \mathbb{E}(Y_t) = \frac{\alpha}{R}.$$

The theorem shows that the average of the mean age tends to the constant α/R when the time t is large.

Influence of the distribution on the mean age

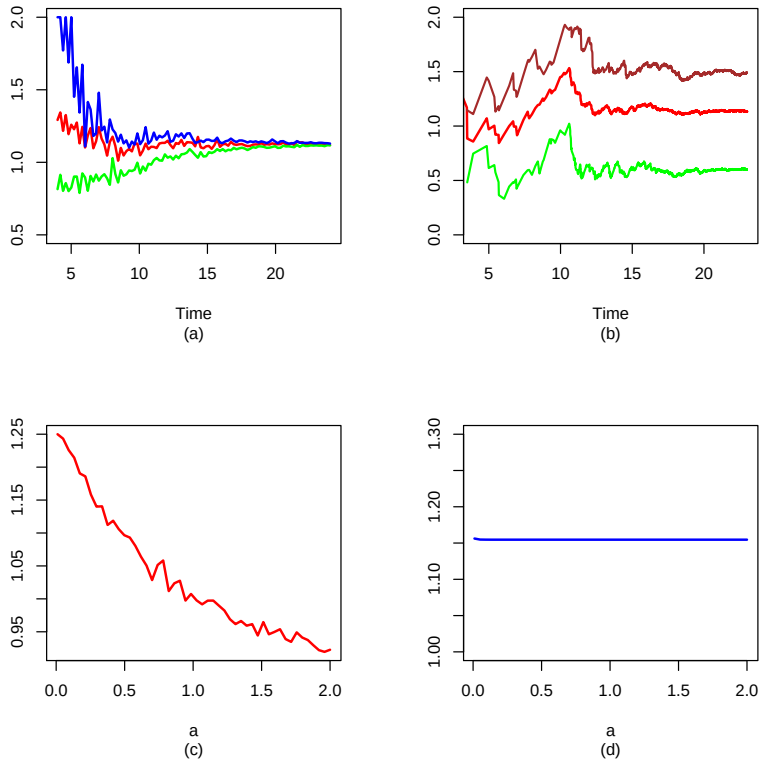


Figure : *(a) Average mean,* 1^{st} *and* 3^{rd} *quartiles for the sample of means for* 50 *trees. (b) Average mean,* 1^{st} *and* 3^{rd} *quartiles for one tree. (c) Average of* $Q_{75} - Q_{25}$ *with* $a \in [0, 2]$ *at* $t = 12$*. (d) Mean age with* $a \in [0, 2]$ *at* $t = 12$*.*

Data and the estimator

- Suppose that we observe the evolution of the cell population in $[0, T]$, $T > 0$.
- At the i^{th} division time t_i, denote j_i the individual who splits into two daughters $X_{t_i}^{j_i 0}$ and $X_{t_i}^{j_i 1}$. We define

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

  the random fractions that go into the daughter cells, with the convention $\frac{0}{0} = 0$.
- The couples $(\Gamma_i^0, \Gamma_i^1)_{i \in \mathbb{N}^*}$ are independent with distribution (Γ^0, Γ^1) where Γ^1 has a symmetric distribution with density h and $\Gamma^0 = 1 - \Gamma^1$.
- We construct an estimator of h based on $(\Gamma_i^0, \Gamma_i^1)_{i \in \mathbb{N}^*}$.

Data and the estimator

Definition

Let M_T be the random number of divisions in $[0, T]$. For all $\gamma \in (0, 1)$, define

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

where $K_\ell = \frac{1}{\ell} K(\cdot/\ell)$ is a bounded kernel function, $\ell > 0$ is the bandwidth to be chosen.

Remark

M_T and Γ_i^1 are independent.

Difficulty: The random number of divisions is random.

Adaptive estimation of h by GL method

- Let $\hat{h}_\ell$ be the kernel estimator, our objective is to find a bandwidth that minimizes L_2-risk.
- Since M_T is random, we study the mean integrated squared error (MISE) conditionally to M_T.
- The L_2-risk of $\hat{h}_\ell$ given M_T satisfies:

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

- From a finite family of bandwidths H, we propose a bandwidth $\bar{\ell}$ where

$$\bar{\ell} := \underset{\ell \in H}{\operatorname{argmin}} \left\{ \|h - K_\ell \star h\|_2 + \frac{\|K\|_2}{\sqrt{M_T \ell}} \right\}.$$

$\bar{\ell}$: oracle bandwidth.

Adaptive estimation of h by GL method

- Goldenschluger and Lepski developed a fully data-driven bandwidth selection method (GL method) corresponding to oracle inequality.
- To apply GL method, we set for any ℓ, $\ell' \in H$:

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

Definition

Given $\epsilon > 0$ and setting $\chi := (1+\epsilon)(1+\|K\|_1)$, we define

$$\hat{\ell} := \underset{\ell \in H}{\operatorname{argmin}} \left\{ A(\ell) + \frac{\chi \|K\|_2}{\sqrt{M_T \ell}} \right\} \quad \text{and} \quad \hat{h} := \hat{h}_{\hat{\ell}},$$

where, for any $\ell \in H$,

$$A(\ell) := \sup_{\ell' \in H} \left\{ \|\hat{h}_{\ell,\ell'} - \hat{h}_{\ell'}\|_2 - \frac{\chi \|K\|_2}{\sqrt{M_T \ell'}} \right\}_+,$$

Oracle inequality

Theorem

Let N_0 be the number of mother cells at the beginning of divisions and M_T is the random number of divisions in $[0, T]$. Define

$$\varrho(T)^{-1} = \begin{cases} \dfrac{e^{-RT+\ln(RT)}}{1-e^{-RT}}, & \text{if } N_0 = 1, \\ e^{-RT}, & \text{if } N_0 > 1. \end{cases}$$

Assume $h \in L^\infty([0,1])$ and let $\hat{h}$ be a kernel estimator defined as above. For some $\delta > 0$, consider $H \subset \{D^{-1} : D = 1, \ldots, \lfloor \delta M_T \rfloor\}$, then

$$\mathbb{E}\left[\|\hat{h} - h\|_2^2\right] \leq C_1 \inf_{\ell \in H} \left\{ \|K_\ell \star h - h\|_2^2 + \frac{\|K\|_2^2}{\ell} \varrho(T)^{-1} \right\} + C_2 \varrho(T)^{-1}$$

where C_1 is a constant depending on N_0, $\|K\|_1$ and ϵ and C_2 is a constants depending on N_0, δ, ϵ, $\|K\|_1$, $\|K\|_2$, $\|h\|_\infty$.

Rate of convergence

Definition

Let $\beta^* > 0$. A function $K : \mathbb{R} \to \mathbb{R}$ is a kernel of order β^* if

- $\int K(x)dx = 1$,
- $\int |x|^{\beta^*}|K(x)|dx < \infty$,
- For $k = \lfloor \beta^* \rfloor$, $\forall 1 \leq j \leq k$, $\int x^j K(x)dx = 0$.

Theorem

Let $\beta^* > 0$ and K be a kernel of order β^*. Let $\beta \in (0, \beta^*)$. Assume that $h \in \mathcal{H}(\beta, L)$ ($h \in \mathcal{C}^{\lfloor \beta \rfloor}$ and $h^{(\beta)}$ is $\beta - \lfloor \beta \rfloor$ continuous). Then, for any $T > 0$, the kernel estimator $\hat{h}$ satisfies

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

where C_3 is a constant depending on N_0, δ, ϵ, $\|K\|_1$, $\|K\|_2$, $\|h\|_\infty$, β and L.

Lower bound on L_2 risk

Theorem

For any $T > 0$, $\beta > 0$ and $L > 0$. Assume that $h \in \mathcal{H}(\beta, L)$, then there exists a constant $C_4 > 0$ such that for any estimator $\hat{h}_T$ of h

$$\sup_{h \in \mathcal{H}(\beta, L)} \mathbb{E}\|\hat{h}_T - h\|_2^2 \geq C_4 e^{-\frac{2\beta}{2\beta+1} RT}.$$

Numerical computation of $\hat{h}$

- Simulations with two distributions as division kernels:
  - Beta$(2,2)$: symmetric divisions.
  - Beta mixture $\frac{1}{2}\,\text{Beta}(2,6) + \frac{1}{2}\,\text{Beta}(6,2)$: asymmetric divisions.
- Using the classical Gaussian kernel $K(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, we estimate $\hat{h}$ with *GL* method. We have $\|K\|_1 = 1$, $\|K\|_2 = 2^{-1/2}\pi^{-1/4}$ and $K_\ell \star K_{\ell'} = K_{\sqrt{\ell+\ell'}}$, hence it is not difficult to calculate in practice $\hat{h}_{\ell,\ell'}$ as well as $\hat{h}_{\ell'}$.
- Compare $\hat{h}$ estimated by the GL bandwidth with those by the cross-validation (CV) bandwidth, the rule-of-thumb (RoT) bandwidth and the oracle bandwidth.

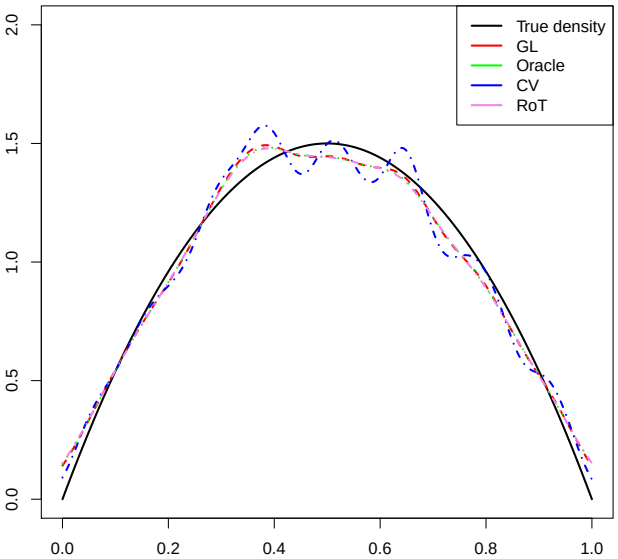
Monte-Carlo simulation

- To estimate the MISE, we implement Monte-Carlo simulations with respect to $T = 13, 17$ and 20. Each sample is simulated with division rate $R = 0.5$. The number of repetitions for each simulation is $m = 100$.
- We compute the mean of relative error $\bar{e} = (1/m)\sum_{i=1}^{m} e_i$ and the standard deviation $\sigma_e = \sqrt{(1/m)\sum_{i=1}^{m}(e_i - \bar{e})^2}$ where

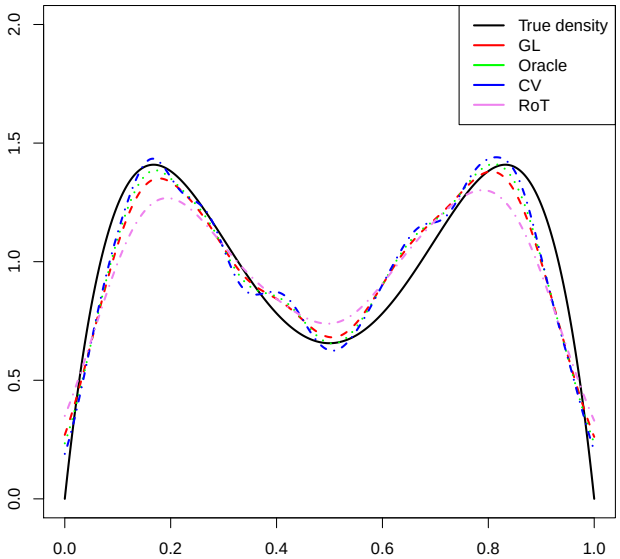
$$e_i = \frac{\|\hat{h} - h\|_2}{\|h\|_2}, \quad i = 1, \ldots, m.$$

- For a further comparison, we compute the relative error in a parametric setting by comparing the true density h with the density of Beta$(\hat{a}, \hat{a})$ where $\hat{a}$ is the Maximum Likelihood (ML) estimator of a.

Numerical Illustrations



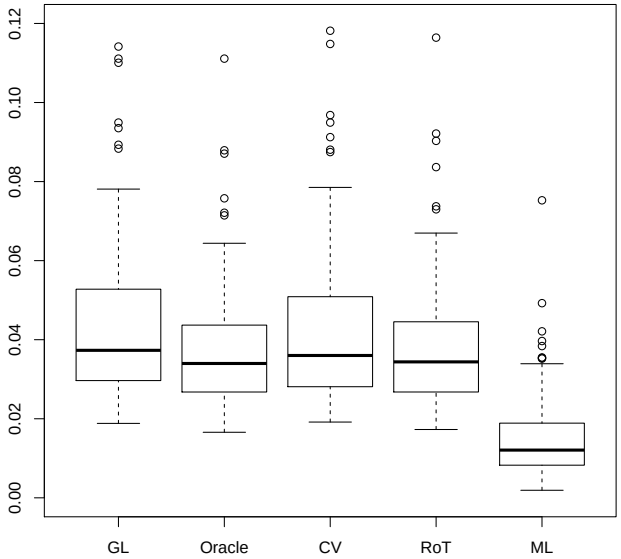
(a) Beta$(2, 2)$



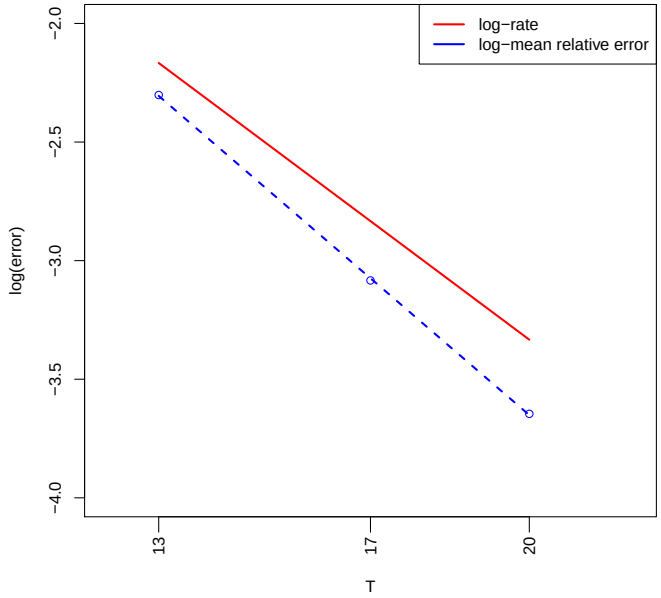
(b) $\frac{1}{2}$ Beta$(2, 6) + \frac{1}{2}$ Beta$(6, 2)$

Figure : Reconstruction of $\hat{h}$ with $T = 13$.

Numerical Illustrations



(a)



(b)

Figure : (a): Errors of estimated densities of Beta$(2, 2)$ when $T = 17$. (b): The log-mean relative error for the reconstruction of Beta$(2, 2)$ compared to the log-rate (solid line) computed with $\beta = 1$.

Thank you for your attention

References


V. Bansaye, J.-F. Delmas, L. Marsalle, and V. C. Tran. Limit theorems for markov processes indexed by continuous time galton-watson trees. *The Annals of Applied Probability*, 21, 2011.


M. Doumic, M. Hoffmann, P. Reynaud-Bouret, and V. Rivoirard. Nonparametric estimation of the division rate of a size-structured population. *SIAM Journal on Numerical Analysis*, 50, 2012.


A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: orcale inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011.


V. H. Hoang. Estimating the division kernel of a size-structured population. arXiv:1509.02872v3, 2015.


M. Hoffmann and A. Olivier. Nonparametric estimation of the division rate of an age dependent branching process. *arXiv:1412.5936*, 2014.


E. J. Stewart, R. Madden, G. Paul, and F. Taddei. Aging and Death in an Organism That Reproduces by Morphologically Symmetric Division. *PLOS Biology*, 3, 2005.


V. C. Tran. Large population limit and time behaviour of a stochastic particle model describing an age-structured population. *ESAIM: Probability and Statistics*, 12, 2008.