LOCAL BANDWIDTH SELECTION FOR KERNEL DENSITY ESTIMATION IN BIFURCATING MARKOV CHAIN MODEL

Siméon Valère Bitseki Penda, Angelina Roche

LOCAL BANDWIDTH SELECTION FOR KERNEL DENSITY ESTIMATION IN BIFURCATING MARKOV CHAIN MODEL

S. VALÈRE BITSEKI PENDA AND ANGELINA ROCHE

Abstract. We propose an adaptive estimator for the stationary distribution of a bifurcating Markov Chain on $\mathbb{R}^d$. Bifurcating Markov chains (BMC for short) are a class of stochastic processes indexed by regular binary trees. A kernel estimator is proposed whose bandwidth is selected by a method inspired by the works of Goldenshluger and Lepski [18]. Drawing inspiration from dimension jump methods for model selection, we also provide an algorithm to select the best constant in the penalty.

Keywords: Bandwidth selection ; Bifurcating autoregressive process ; Bifurcating Markov chains ; Binary trees ; Nonparametric kernel estimation.

Mathematics Subject Classification (2010): 62G05, 62G10, 62G20, 60J80, 60F05, 60F10, 60J20, 92D25.

1. Introduction

First introduced by Basawa and Zhou (2004) [2], Bifurcating Markov chains models (BMCM for short) has recently received particular attention for its application to cell lineage study. Guyon (2007) [21], have proposed such a model to detect cellular aging in *Escherichia Coli* and proved laws of large numbers and central limit theorem for this class of stochastic process. Bitseki-Penda *et al.* (2014) [7] have then completed these asymptotic results and proved concentration inequalities.

To the best of our knowledge, kernel density estimation for the BMCM were considered first by Doumic & *al.* [17], where they estimate the division rate of population of cells reproducing by symmetric division, *i.e.* cell reproduction where each fission produces two equal daughter cells. After this work, Bitseki & *al.* [8], have used the wavelets methodology to study the nonparametric estimation of the density of BMCM. They propose an adaptive estimator in dimension 1. Recently, Bitseki and Olivier [9] have studied the Nadaraya-Watson type estimators of a BMCM that they called nonlinear bifurcating autoregressive process. The latter model can be seen as an adaptation of nonlinear autoregressive process on binary regular tree. We mention that, except in [8], all the estimations done in the previous works are non adaptive. In particular, the question of data-driven bandwidth selection was not addressed in [17] and [9]. The main objective of this work is then to propose a data-driven method for choosing the bandwidth for the kernel estimator of the invariant measure in the multi-dimensional BMCM, following ideas from the works of Goldenshluger and Lepski (2011) [18].

The idea of the method is to select the bandwidth minimizing an empirical criterion imitating the bias-variance decomposition of the risk of the kernel estimator. More precisely, let $\mathcal{H}$ be a collection of bandwidths and let $(\widehat{\nu}_h)_{h\in\mathcal{H}}$ be a family of kernel estimators of an unknown density ν. Then, we select the bandwidth $\widehat{h}$ as

$$\widehat{h} = \arg\min_{h\in\mathcal{H}}\{\widehat{A}(h) + bV(h)\}$$

with $A(h)$ an empirical version of the bias of the estimator $\widehat{\nu}_h$ and $V(h)$ a penalty term with the same order than the variance. The now so-called *Goldenshuger-Lepski* methodology ([18], but also [19, 26, 27, 24, 22]), initially developed for density estimation, has been applied in many contexts such as deconvolution problems [14], conditional cumulative distribution function estimation [11, 10], regression problems [10, 12], conditional density estimation [10, 5], hazard rate estimation [10], white noise model [23], kernel empirical risk minimization (including robust regression) [13], Lévy processes [3], Cox model [20], stochastic differential model [16]. Under suitable assumptions on the kernel, it is shown in [18] that this selection rule leads to a minimax adaptive estimator on a general class of regular functions, for a general class of $\mathbb{L}_s$-risks. Pointwise versions of the Goldenshluger-Lepski selection rule have been less considered (Comte et Lacour (2013) [14], Rebelles (2015) [27] and Chagny and Roche (2016) [12]). The interest of such an approach is that the bandwidth is selected with a local criterion which realizes the best bias-variance compromise at the point where the estimator is calculated. On the contrary, integrated versions of the Goldenshluger-Lepski selection rule select the same bandwidth at all points. Let us mention that, in all the articles cited above, the theoretical results rely on concentration inequalities for sums of i.i.d. random variables such as Bernstein Inequality or Talagrand Inequality. In our context, such results are not applicable. Hence, we prove a Bernstein-type Inequality for functionals of BMC, where the functions are kernels and convolution of kernels. Compare to those obtained in [8], our inequalities are more complete in the sense that the deviation parameter can take all the positive values. More precisely, its values do not depend on the size of the samples, with is essential for our theoretical results.

Ideally, the penalty term V, called "the majorant" by Goldenshluger and Lepski, depends entirely on the kernel and the observations (it does not depend on the density ν). The selected estimator is then $\widehat{\nu}_{\widehat{h}}$. However, for the BMCM, the variance term V in the previous selection rule contains a term which may depend on the unknown density ν. Moreover, this term is generally not estimable from observations. To resolve this problem, we propose a modification of Goldenshluger-Lepski rule's selection, inspired by the works of Lacour and Massart (2016) [22]. As suggested in [22], the constant term in $bV(x,h)$ is then selected automatically from the data with an algorithm inspired by the works of Arlot and Massart (2009) [1].

The paper is organized as follows. The model is defined in Section 2. Section 3 is devoted to the definition of the estimator. In Section 4, we provide a numerical study of our estimator. The proofs are given in Section 5.

2. Definitions

We are now going to give a precise definition of a BMC. First we note that this class of stochastic processes has been introduced by Guyon [21] in order to understand the mechanisms of cell division. Indeed, these stochastic processes are well adapted to study a population (or more generally, any dynamic system) where each individual (or more generally, each particle) in one generation gives birth to two individuals in the next one. In the sequel, we will then use the language of the population dynamic to define the sets of interest.

Let $(\Omega, \mathcal{F}, (\mathcal{F}_m, m \in \mathbb{N}), \mathbb{P})$ be a filtered probability space. Let $(X_u, u \in \mathbb{T})$ be a sequence of random variables defined on $(\Omega, \mathbb{P})$, taking values in $\mathbb{R}^d$, where $d \geq 1$, and indexed by the infinite binary tree $\mathbb{T} = \bigcup_{m=0}^{\infty} \{0,1\}^m$, with the convention that $\{0,1\}^0 = \emptyset$. We equip $\mathbb{R}^d$ with its usual Borel σ-field. Now, we will see $\mathbb{T}$ as a given population. Then each individual u of this population is represented by a sequence of 0's and 1's, and has two descendants, $u0$ and $u1$. The initial individual of the population is $\emptyset$. For all $m \in \mathbb{N}$, let $\mathbb{G}_m$ be the set of individuals belonging to the m-th generation, and $\mathbb{T}_m$ the set of individuals belonging to the m first generations. We have:

$$\mathbb{G}_m = \{0,1\}^m,\ \mathbb{T}_m = \bigcup_{q=0}^{m} \mathbb{G}_q \ \text{ and } \ \mathbb{T} = \bigcup_{m\geq 0} \mathbb{G}_m.$$

For an individual $u \in \mathbb{G}_m$, we set $|u| := m$ its length (i.e. the generation to which it belongs).

2.1. **Bifurcating Markov chain** [21]**.**

Definition 1 ($\mathbb{T}$-transition probability)**.** *Let $\mathcal{P} : \mathbb{R}^d \times \mathcal{B}((\mathbb{R}^d)^2) \to [0,1]$ (with $\mathcal{B}(\mathbb{R}^d)^2 = \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^d)$ the usual product σ-field on $(\mathbb{R}^d)^2$). Then $\mathcal{P}$ is a* $\mathbb{T}$-transition probability *if*

- *$x \to \mathcal{P}(x, A)$ is measurable for all $A \in \mathcal{B}(\mathbb{R}^d)^2$.*
- *$A \to \mathcal{P}(x, A)$ is a probability measure on $((\mathbb{R}^d)^2, \mathcal{B}(\mathbb{R}^d)^2)$ for all $x \in \mathbb{R}^d$.*

For a $\mathcal{B}(\mathbb{R}^d)^3$-measurable function $f : (\mathbb{R}^d)^3 \to \mathbb{R}$, we denote (when it is defined) by $\mathcal{P}f$ the $\mathcal{B}(\mathbb{R}^d)$-measurable function

$$x \in \mathbb{R}^d \mapsto \int_{(\mathbb{R}^d)^2} f(x,y,z)\mathcal{P}(x,dy,dz).$$

Definition 2 (Bifurcating Markov chain)**.** *Let μ be a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $\mathcal{P}$ a $\mathbb{T}$-transition probability. We say that $(X_u, u \in \mathbb{T})$ is a $(\mathcal{F}_m)$-bifurcating Markov chain with initial distribution μ and $\mathbb{T}$-transition probability $\mathcal{P}$ (denoted $\mathbb{T}$-BMC in the sequel) if*

- *X_u is $\mathcal{F}_m$ measurable for all $u \in \mathbb{G}_m$.*
- *$X_\emptyset$ has distribution μ.*
- *For all $m \in \mathbb{N}$, and for all family $(f_u, u \in \mathbb{G}_m)$ of $\mathcal{B}(\mathbb{R}^d)$-measurable functions from $(\mathbb{R}^d)^3$ to $\mathbb{R}$,*

$$\mathbb{E}\left[\prod_{u\in\mathbb{G}_m} f_u(X_u, X_{u0}, X_{u1})|\mathcal{F}_m\right] = \prod_{u\in\mathbb{G}_m} \mathcal{P}f_u(X_u),$$

where $u0 := (u,0) \in \mathbb{G}_{m+1}$ and $u1 := (u,1) \in \mathbb{G}_{m+1}$.

2.2. **Tagged-branched chain** [21, 9]**.** Let $(X_u, u \in \mathbb{T})$ be a $\mathbb{T}$-BMC with initial distribution μ and $\mathbb{T}$-transition probability $\mathcal{P}$. We denote by $\mathcal{P}_0$ and $\mathcal{P}_1$ respectively the first and second marginals of $\mathcal{P}$. More precisely

$$\mathcal{P}_0(x,B) = \mathcal{P}(x, B\times\mathbb{R}^d) \ \text{ and } \ \mathcal{P}_1(x,B) = \mathcal{P}(x, \mathbb{R}^d \times B),$$

for all $x \in \mathbb{R}^d$ and all $B \in \mathcal{B}(\mathbb{R}^d)$. Let $\mathcal{Q}$ be the mixture of $\mathcal{P}_0$ and $\mathcal{P}_1$ with equal weights

$$\mathcal{Q} = \frac{1}{2}\mathcal{P}_0 + \frac{1}{2}\mathcal{P}_1.$$

The Markov chain $Y := (Y_m)_{m\in\mathbb{N}}$ on $\mathbb{R}^d$ with initial value $Y_0 := X_\emptyset$ and transition probability $\mathcal{Q}$ is called the *tagged-branch chain*.

In all the paper, we will denote by $\mathcal{Q}^m$ the mth iterated of $\mathcal{Q}$ recursively defined by the formulas

$$\mathcal{Q}^0(x,\cdot) = \delta_x \quad \text{and} \quad \mathcal{Q}^{m+1}(x,B) = \int_S \mathcal{Q}(x,dy)\mathcal{Q}^m(y,B) \quad \forall B \in \mathcal{B}(\mathbb{R}^d).$$

It is well known that $\mathcal{Q}^m$ is a transition probability in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. In particular, we have $\mathbb{E}[f(Y_m)] = \mu\mathcal{Q}^m f$ for all measurable function $f : \mathbb{R}^d \to \mathbb{R}$.

In the sequel, we will assume that the Markov chain Y is ergodic- that is say, that there exists a unique distribution ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)$ such that, for all measurable function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\lim_{m\to\infty}\mathbb{E}[f(Y_m)]=\int_{\mathbb{R}^d} f d\nu.$$

We will also assume that the distribution ν has a density, that we also denote by ν, with respect to the Lebesgue measure.

As previous works have shown (see for example [21, 7]), the analysis of a BMC $(X_u, u\in\mathbb{T})$ is strongly related to the asymptotic behavior of the tagged-branched chain $(Y_m, m\in\mathbb{N})$, and therefore to the knowledge of the invariant distribution ν. We stress that this distribution is unknown and it is not directly observable, in such a way that its estimation from the data is of great interest. The aim is to estimate ν from the observation of a subpopulation $(X_u, u\in\mathbb{T}_n)$.

3. Estimation of the stationary distribution ν

3.1. **Definition of the estimator.** We suppose that we observe the process $(X_u, u\in\mathbb{T})$ up to the n-th generation. We denote by $|\mathbb{T}_n|=2^{n+1}-1$ the cardinality of $\mathbb{T}_n$. Based on the observation of $(X_u, u\in\mathbb{T}_n)$, we propose the following estimator of ν

$$\widehat{\nu}_h(x)=\frac{1}{|\mathbb{T}_n|}\sum_{u\in\mathbb{T}_n}K_h(x-X_u)\quad\forall x\in\mathbb{R}^d, \tag{1}$$

where for all $x=(x_1,...,x_d)^t\in\mathbb{R}^d$ and for all $h=(h_1,...,h_d)^t\in]0,+\infty[^d$

$$K_h(x)=K(x_1/h_1,...,x_d/h_d)/\prod_{j=1}^d h_j$$

and K is a kernel, that is to say a function $K:\mathbb{R}^d\to\mathbb{R}$ which verifies $\int_{\mathbb{R}^d}K(u)du=1$.

The vector h is usually called a bandwidth. In the sequel, $|h|$ will denote the product $h_1\times\cdots\times h_d$. It is well known in the kernel estimation theory that the choice of h is of great interest. Indeed the amount of smoothing is controlled by a judicious choice of the bandwidth. Now, in order to tackle the issue of the choice of this bandwidth, we will need the following assumptions.

Assumption 3 (Assumptions on the kernel)**.**

$$\|K\|_1=\int_{\mathbb{R}^d}|K(t)|dt<+\infty,\quad \|K\|_2^2=\int_{\mathbb{R}^d}|K(t)|^2dt<+\infty\quad \textit{and}\quad \|K\|_\infty=\sup_{t\in\mathbb{R}^d}|K(t)|<+\infty.$$

Assumption 4 (Uniform geometric ergodicity condition)**.** *There exists two constants $\rho\in(0,1/2)$ and $M>0$ such that for all bounded ν-integrable function g, for all $x\in\mathbb{R}^d$ and $m\geq 0$*

$$|\mathcal{Q}^m g(x)-\int_{\mathbb{R}^d}g d\nu|\leq M\|g\|_\infty\rho^m.$$

This assumption is verified for severals models of BMC. We refer for example to [9] where Bitseki and Olivier have shown it for NBAR processes (see [9] Lemma 20). For a precise definition of NBAR processes, we refer to section 4.

Assumption 5 (Assumption on $\mathcal{P}$, $\mathcal{P}_0$, $\mathcal{P}_1$, $\mathcal{Q}$ and ν)**.** *We assume that the transitions $\mathcal{P}$, $\mathcal{P}_0$, $\mathcal{P}_1$ and $\mathcal{Q}$ admit densities with respect to the Lebesgue measure that we denote with the same notations. Moreover, we assume that*

$$\|\mathcal{P}\|_\infty<+\infty,\ \|\mathcal{P}_0\|_\infty<+\infty,\ \|\mathcal{P}_1\|_\infty<+\infty,\ \|\nu\|_\infty<+\infty,\ \textit{and}\ \|\mathcal{Q}\|_\infty<+\infty.$$

3.2. **Bias-variance decomposition.** We consider a pointwise quadratic risk

$$\mathbb{E}[(\widehat{\nu}_h(x) - \nu(x))^2],$$

where $x \in \mathbb{R}^d$ is a given point.

The results of [9] (see the proof of Proposition 21) allows to obtain the following upper-bound on the risk, where we have make explicit the constants that appear.

Proposition 6. *Under Assumption 3 to Assumption 5, we have*

$$\mathbb{E}[(\widehat{\nu}_h(x) - \nu(x))^2] \leq 2(K_h * \nu(x) - \nu(x))^2 + 2\frac{C(P,\nu)}{|\mathbb{T}_n||h|}, \tag{2}$$

where $*$ denotes the convolution product $f*g(x) = \int_{\mathbb{R}^d} f(x-t)g(t)dt$ for all functions f, g integrable over $\mathbb{R}^d$ and

$$C(P,\nu) = \frac{C_I}{(\sqrt{2}-1)^2}$$

with

$$C_I = (1 + \frac{1}{1-2\rho^2})(\|\mathcal{Q}\|_\infty + \|\nu\|_\infty)^2 + M^2 + C_P$$

and

$$C_P = 2\|K\|_2^2(\|\mathcal{Q}\|_\infty + \|\nu\|_\infty) + \|\mathcal{P}\|_\infty + \|\nu\|_\infty(\|K\|_2^2 + \|\mathcal{P}_0\|_\infty + \|\mathcal{P}_1\|_\infty)$$

Inequality (2) can be seen as a bias-variance decomposition in the sense that

$$\mathbb{E}_\nu[\widehat{\nu}_h(x)] = \frac{1}{|\mathbb{T}_n|}\sum_{u\in\mathbb{T}_n}\mathbb{E}_\nu[K_h(x-X_u)] = \int_{\mathbb{R}^d} K_h(x-t)d\nu(t) = K_h * \nu(x),$$

where $\mathbb{E}_\nu$ is the expectation with respect to the measure ν and $C(P,\mu)/|\mathbb{T}_n||h|$ is an upper-bound on the variance term $\mathbb{E}\left[(\widehat{\nu}_h(x) - K_h * \nu(x))^2\right]$.

3.3. **Selection rule.** Given a family of bandwidths $\mathcal{H}_n \subset [0, +\infty[^d$, we have a family of estimators $(\widehat{\nu}_h)_{h\in\mathcal{H}_n}$ and the aim is to select an estimator in this family with risk close to the unknown oracle risk

$$\mathbb{E}\left[(\widehat{\nu}_{h^*}(x) - \nu(x))^2\right] \text{ with } h^* = \arg\min_{h\in\mathcal{H}_n}\mathbb{E}\left[(\widehat{\nu}_h(x) - \nu(x))^2\right].$$

Imitating the decomposition given in Equation (2) we could consider the following selection rule inspired by the work of [18]

$$\widehat{h} = \arg\min_{h\in\mathcal{H}_n}\{A(x,h) + bV(x,h)\} \tag{3}$$

where

- $\mathcal{H}_n \subset [0, +\infty[^d$ is a finite collection of bandwidths;
- $A(x,h) = \max_{h'\in\mathcal{H}_n}\left((\widehat{\nu}_{h'}(x) - K_h * \widehat{\nu}_{h'}(x))^2 - aV(x,h')\right)_+$ with $b \geq a \geq 1$;
- $V(x,h) = C(P,\mu)\log(|\mathbb{T}_n|)/|\mathbb{T}_n||h|$.

The constants a and b can be different, as suggested in Section 5 of [22]. In Section 3.4, we provide a method to choose both a and b, as well as the quantity $C(P,\mu)$ appearing in the variance term $V(x,h)$. The lower bound $a = 1$ is a critical minimal value, in the sense that the procedure fails for lower values of a (see [22]).

We prove the following oracle-inequality on the selected estimator $\widehat{\nu}_{\widehat{h}}$.

Theorem 7. *Under Assumption 4, if* $\min_{h\in\mathcal{H}_n}|h| \geq \log(|\mathbb{T}_n|)/|\mathbb{T}_n|$, *then there exists a minimal value* $a_{\min} > 0$ *independent of* n, *such that, for all* $a > a_{\min}$,

$$\mathbb{E}\left[(\widehat{\nu}_{\widehat{h}}(x) - \nu(x))^2\right] \leq C_1 \min_{h\in\mathcal{H}_n}\left\{\mathcal{B}_h(x) + \frac{\log(|\mathbb{T}_n|)}{|\mathbb{T}_n||h|}\right\} + \frac{C_2}{|\mathbb{T}_n|}, \tag{4}$$

where $C_1, C_2 > 0$ *do not depend on* n *nor* x *and*

$$\mathcal{B}_h(x) = \max_{h'\in\mathcal{H}_n}(K_{h'} * \nu(x) - K_h * K_{h'} * \nu(x))^2.$$

Remark 8. *The form of the bias term* $\mathcal{B}_h(x)$ *in Inequality (4) is very similar to the one obtained for pointwise adaptive kernel density estimation in* [27, Theorem 1]. *It can be replaced by an upper-bound, e.g.* $\|K\|_1\|\nu - K_h * \nu\|_\infty$, *coming from the Young's inequality, as in* [14, Theorem 2].

Once the previous theorem is proved, an immediate corollary follows

Corollary 9. *Suppose that the assumptions of Theorem 7 are verified and that* $\nu \in \Sigma(\beta, L, D)$ *where* $\beta = (\beta_1, ..., \beta_d) \in]0, +\infty[^d$ *and* $\Sigma(\beta, L, D)$ *is the set of all* β*-Hölder densities on the open set* $D \subseteq \mathbb{R}^d$ *i.e. the set of all functions* $f : D \to \mathbb{R}$ *which admits, for all* $j = 1, ..., d$, *partial derivatives with respect to* x_j *up to the order* $\lfloor\beta_j\rfloor$ *and verifies, for all* $x = (x_1, ..., x_d) \in D$, $x'_j \in \mathbb{R}$ *such that* $(x_1, ..., x_{j-1}, x'_j, x_{j+1}, ..., x_d) \in D$,

$$\left|\frac{\partial^{\lfloor\beta_j\rfloor} f}{\partial x_j^{\lfloor\beta_j\rfloor}}(x_1, , ..., x_{j-1}, x'_j, x_{j+1}, ..., x_d) - \frac{\partial^{\lfloor\beta_j\rfloor} f}{\partial x_j^{\lfloor\beta_j\rfloor}}(x)\right| \leq L|x_j - x'_j|^{\beta_j - \lfloor\beta_j\rfloor}.$$

Moreover, suppose that K *is a kernel of order* $\ell \in \mathbb{N}$ *(with* $\ell \geq \max_{j=1,...,d}\{\beta_j\}$*) that is to say,* $\int_{\mathbb{R}} K(t)dt = 1$, $\int_{\mathbb{R}} x^j K(x) = 0$ *for all* $j = 1, ..., \ell$ *and that*

$$\int_{\mathbb{R}} |x|^\ell |K(u)| < +\infty.$$

Suppose also that, for all n, *there exists* $h^* = (h_1^*, ..., h_d^*) \in \mathcal{H}_n$ *such that* $h_j^* \asymp \left(\frac{\log(|\mathbb{T}_n|)}{|\mathbb{T}_n|}\right)^{\bar{\beta}/(\beta_j(2\bar{\beta}+d))}$ *where* $\bar{\beta} = d/(1/\beta_1 + ... + 1/\beta_d)$ *is the harmonic mean of* β. *Then, there exists a constant* $C > 0$ *such that*

$$\sup_{x\in D}\mathbb{E}\left[(\widehat{\nu}_{\widehat{h}}(x) - \nu(x))^2\right] \leq C\left(\frac{|\mathbb{T}_n|}{\log(|\mathbb{T}_n|)}\right)^{-2\bar{\beta}/(2\bar{\beta}+1)}.$$

This corollary is a direct consequence of Theorem 7 and [28] Proposition 2.1.

Remark that the assumption on the bandwidth collection is verified e.g. by

$$\mathcal{H}_n = \left\{h_{\max}k^{-\alpha}, k = 1, ..., (|\mathbb{T}_n|h_{\max}/\log(|\mathbb{T}_n|))^{1/\alpha}\right\}^d$$

for all constant $h_{\max} > 0$ and all $\alpha > 1$.

3.4. **Estimation of the constant appearing in the variance.** Due to the term $C(P, \mu)$ which is hardly estimable, the variance term $V(x, h)$ is not calculable in practice. We propose an algorithm to estimate it, based on slope estimation, as developed in [1, 22].

As suggested in [22], we take $b = 2a$, hence in order to calculate the estimator, it is sufficient to have a *good* value of $\kappa = aC(P, \mu)$. The following algorithm is inspired by the procedure described in [1], for model selection purposes. However, note that, in our case, both terms $A(x, h)$ and $V(x, h)$ depends on the constant κ, which is not the case in model selection contexts where only the penalty term depends on the constant. The selection of the grid of κ is then different here.

Algorithm 10.

1. *Initialization : set*

$$\kappa_m = \frac{|\mathbb{T}_n|}{\log(|\mathbb{T}_n|)} \max_{h,h' \in \mathcal{H}_n} |h'|(\widehat{\nu}_{h'}(x) - K_h * \widehat{\nu}_{h'}(x))^2 \text{ and } \kappa_1 = 0. \tag{5}$$

2. *While* $s \leq s_{\max}$
   (i) *Generate a sequence* $(\kappa_j)_{1 \leq j \leq m}$ *such that and* $\kappa_j = \kappa_1 + \frac{j-1}{m-1}(\kappa_m - \kappa_1)$, *for all* $j = 1, ..., m$.
   (ii) *Calculate* $\widehat{h}_j := \widehat{h}(\kappa_j)$ *as the minimizer of the criterion* (3) *with* $a = \kappa / C(P, \mu)$ *and* $b = 2a$.
   (iii) *Set*

$$j_{jump} = \arg\max_{j=1,...,m-1} \left| \frac{1}{|\widehat{h}_j|} - \frac{1}{|\widehat{h}_{j+1}|} \right|,$$

   $\kappa_1 = \kappa_{j_{jump}}$, $\kappa_1 = \kappa_{j_{jump}+1}$ *and* $s = s + 1$.
3. *Return* $\widehat{h}_{j_{jump}+1}$.

The algorithm search the value of κ for which the variance of the estimator increases significantly and select a slightly larger value. This allows to select an estimator with a reasonable variance. The same reasoning has given rise to the so-called *dimension jump* method for model selection purposes (see [1]). The chosen value for the initialization of κ_m comes from the following reasoning. Setting $\kappa \geq \kappa_m$ as suggested: by definition, for all $h, h' \in \mathcal{H}_n$,

$$\frac{|\mathbb{T}_n|h'}{\log(|\mathbb{T}_n|)}(\widehat{\nu}_{h'}(x) - K_h * \widehat{\nu}_{h'}(x))^2 \leq \kappa$$

which implies that $A(x, h) = 0$ for all $h \in \mathcal{H}_n$ and that the criterion (3) will select the smaller bandwidth in $\mathcal{H}_n$. On the contrary, if $\kappa > \kappa_m$, let $h, h' \in \mathcal{H}_n$ for which the minimum is attained in (5), we have

$$\frac{|\mathbb{T}_n|}{\log(|\mathbb{T}_n|h')}(\widehat{\nu}_{h'}(x) - K_h * \widehat{\nu}_{h'}(x))^2 > \kappa$$

which implies that $A(x, h) > 0$ and the criterion may select a bandwidth which is not the smaller one. Hence, the values of κ for which the criterion (3) may select suitable values of the bandwidth can not be greater than κ_m. That is the reason why we consider an initial grid in the interval $[0, \kappa_m]$. However, the initial value κ_m may be very large compared to the optimal value of κ. The loop 2. allows to search among small values of κ while avoiding the choice of a too large m which could be very expensive in terms of computation time. In practice $s_{\max} = 2$ and $m = 20$ seems to be a reasonable choice.

Remark 11. *The choice* $b = 2a$ *is arbitrary since, in theory, any value* b *such that* $b \geq a$ *might work as soon as a sufficiently large. Other choices may be made such as* $b = a$ *which is the usual choice of Lepski's method,* $b = a + \varepsilon$, *with* $\varepsilon > 0$ *or* $b = (1 + \varepsilon)a$.

4. Simulations

We shall now illustrate the previous results on simulated data coming from various models of bifurcating Markov chains.

4.1. **Bifurcating autoregressive processes.**

Bifurcating autoregressive processes (BAR, for short) were first introduced by Cowan and Staudte [15] in order to study the data from cell division, where each individual in one generation gives birth to two children in the next generation. This model has been widely studied over the last thirty years (see for e.g. [9] and references therein). Recently, Bitseki and Olivier in [9] have proposed an extension of BAR process initially introduced by Cowan and Staudte. Their model is defined as follows.

Let $X_u \in \mathbb{R}$ be a quantitative data associated to the cell $u \in \mathbb{T}$, for example the growth rate of *E. Coli*. Then the quantities X_{u0} and X_{u1} associated to $u0$ and $u1$ the two children of u are linked to X_u through the following autoregressive equations

$$\mathcal{L}(X_\emptyset) = \mu, \quad \text{and for } u \in \mathbb{T}, \quad \begin{cases} X_{u0} = f_0(X_u) + \varepsilon_{u0}, \\ \\ X_{u1} = f_1(X_u) + \varepsilon_{u1}, \end{cases} \tag{6}$$

where μ is a distribution probability on $\mathbb{R}$ and $f_0, f_1 : \mathbb{R} \rightsquigarrow \mathbb{R}$. The noise $\big((\varepsilon_{u0}, \varepsilon_{u1}), u \in \mathbb{T}\big)$ forms a sequence of independent and identically distributed bivariate centered random variables with common density g_ε on $\mathbb{R} \times \mathbb{R}$. The process $(X_u, u \in \mathbb{T})$ defined by (6) is a bifurcating Markov chain with $\mathbb{T}$-transition probability

$$\boldsymbol{\mathcal{P}}(x, dy, dz) = g_\varepsilon(y - f_0(x), z - f_1(x))dydz.$$

Under some assumptions on μ, f_0, f_1 and g_ε, it has been shown in [9] that the process X satisfies all the good properties needed for our theoretical results (we refer to [9] for more details). We note that the previous model can be seen as an adaptation of nonlinear autoregressive model when the data have a binary tree structure. Furthermore, the original BAR process in [15] is defined for linear link functions f_0 and f_1 with $f_0 = f_1$.

Now, for our numerical illustrations, we build a BAR process living in $S := [0, 1]$ as follows. First, we choose $X_\emptyset$ such that $\mathcal{L}(X_\emptyset) = Beta(2, 2)$, where $Beta(2, 2)$ is the standard Beta distribution with shape parameters $(2, 2)$. Then for $u \in \mathbb{T}$ and conditionally on $X_u = x$, we construct X_{u0} and X_{u1} independently in such a way that $\mathbb{P}(X_{u0} \in dy, X_{u1} \in dz) = \boldsymbol{\mathcal{P}}(x, y, z)dydz$, where $\boldsymbol{\mathcal{P}}(x, \cdot, \cdot) := \mathcal{P}(x, \cdot) \otimes \mathcal{P}(x, \cdot)$ and

$$\mathcal{P}(x, y) := (1 - x)\,\frac{y(1 - y)^2}{B(2, 3)} + x\,\frac{y^2(1 - y)}{B(3, 2)}, \quad x, y \in [0, 1]$$

with $B(\alpha, \beta)$ the normalizing constant of a standard Beta distribution with shape parameters α and β. Now, one can prove that this process is stationary, it has an explicit invariant density, which is crucial to evaluate the quality of estimation of our method: this is a standard Beta distribution with shape parameters $(2, 2)$. One can also prove that

$$\mathbb{E}\left[X_{u0}|X_u\right] = \mathbb{E}\left[X_{u1}|X_u\right] = 1/5X_u + 2/5,$$

in such a way that the equations (6) are satisfied with $f_0(x) = f_1(x) = 1/5x + 2/5$ (for more details, we refer for e.g. to [25]). Now, it is no hard to verify that this process satisfies our required assumptions.

We simulate the n first generations of the process X, with $n = 10$ (hence the size of $\mathbb{T}_n$ is $|\mathbb{T}_n| = 2^{10} = 1024$). We consider the Gaussian kernel $K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$. The results are given in Figure 1.

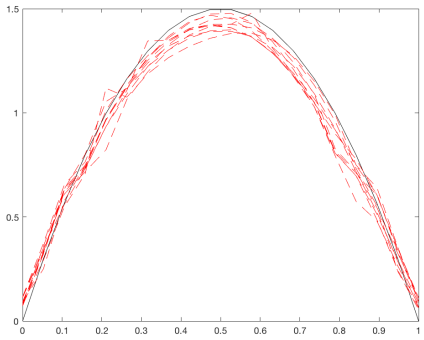

FIGURE 1. Plot of 10 estimators (red dashed lines) obtained from independent copies of X. The black solid line represents the function ν to estimate.

4.2. **Estimation of splitting rate in a growth-fragmentation model.**

We are now interested in growth-fragmentation models. Theses models describes (for e.g.) the evolution of cells which grow and divide randomly over time (see for e.g. [17] and references therein). The model we are going to study is a simplification of the one studied in [17]; it is defined as follows. Let S be a subset of $[0,\infty)$ and let $B : S \mapsto [0,\infty)$ be a continuous function (the splitting rate). Each cell $u \in \mathbb{T}$ grows exponentially with a common rate $\tau > 0$ and when it reaches a certain size x, it splits at rate $B(x)$, and gives birth to two offspring ($u0$ and $u1$) of size $x/2$. Next, this two offspring, $u0$ and $u1$, start a new life independently of each other. Clearly, the process $(X_u, u \in \mathbb{T})$, where X_u is the size of the cell u at birth, is a BMC. It is proved in [17] that the $\mathbb{T}$-transition probability of this BMC is $\boldsymbol{\mathcal{P}}(x,\cdot,\cdot) := \mathcal{P}(x,\cdot) \otimes \mathcal{P}(x,\cdot)$, where the density of $\mathcal{P}(x,\cdot)$ is given by

$$\mathcal{P}(x,y) := \frac{B(2y)}{\tau y} \exp\left(- \int_{x/2}^{y} \frac{B(2z)}{\tau z} dz \right) \mathbb{1}_{\{y \geq x/2\}},$$

for $x \in S$ and $y \in S/2$. It can also be seen that the probability transition of the tagged-branch chain is $\mathcal{Q} = \mathcal{P}$.

Doumic *et al.* [17] have proved that $\mathcal{P}$ admits an invariant probability measure ν having a density, that we still denote by $\nu(\cdot)$, with respect to the Lebesgue measure. It is also known from [17] that the rate function $B(\cdot)$ and the invariant density $\nu(\cdot)$ verify

$$B(x) = \frac{\tau x}{2} \frac{\nu_B(x/2)}{\int_{x/2}^{x} \nu_B(z) dz}$$

in such a way that a natural estimator for $B(x)$, based on the observation of $(X_u, u \in \mathbb{T}_n)$ is

$$\widehat{B}_n(x) = \frac{\tau x}{2} \frac{\widehat{\nu}_h(x/2)}{\left(\frac{1}{|\mathbb{T}_n|} \sum_{u \in \mathbb{T}_n} \mathbf{1}_{\{x/2 \leq X_u < x\}} \right) \vee \varpi_n},$$

where $\widehat{\nu}_h$ is the kernel estimator defined in (1) and ϖ_n is a threshold which ideally go to 0. Moreover, Bitseki *et al.* [8] have proved that under suitable assumptions on the splitting rate $B(\cdot)$, the process X satisfies all the good properties needed for our theoretical results. For all the previous assertions, we refer to [17, 8] for more details. The strategy is then to use our results for the estimation of $\nu_B(x/2)$ for all $x \in S$.

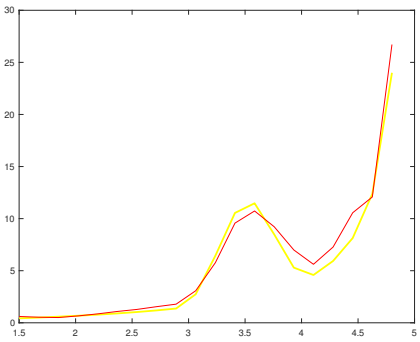

FIGURE 2. Plot of 1 estimator (red solid line) obtained from a copy of X. The yellow solid line represents the function B to estimate.

Now, for our numerical illustrations, we will work with the splitting rate using in [8]. We choose $\tau = 2$, $S = (0, 5)$ and for all $x \in S$, B has the form

$$B(x) = \frac{x}{x-5} + 3T\big(2(x - \tfrac{7}{2})\big)$$

where $T(x) = (1+x)\mathbf{1}_{\{-1\le x<0\}} + (1-x)\mathbf{1}_{\{0\le x\le 1\}}$ is a tent shaped function. With this choice of B, the required assumptions for our theoretical results are satisfied (see [8]).

We simulate the n first generations of the process X, with $n = 15$. The results are given in Figure 2.

5. PROOFS

The proof rely on the lemma below, which is a Bernstein-type inequality.

Lemma 12. *Let $(X_u, u \in \mathbb{T}_n)$ be a bifurcating Markov chain on $\mathbb{R}^d$ with initial distribution μ and $\mathbb{T}$-transition probability P. Under the assumption of uniform geometric ergodicity, we have for all $\delta > 0$*

$$\mathbb{P}\left(\frac{1}{|\mathbb{T}_n|}\left|\sum_{u\in\mathbb{T}_n}\left(K_h * K_{h'}\left(x - X_u\right) - \mathbb{E}_\nu\left[K_h * K_{h'}\left(x - X_u\right)\right]\right)\right| > \delta\right) \tag{7}$$
$$\le 2\exp\left(\frac{\delta c_{K,Q,\nu,M} c'_\rho}{\frac{4c_{\rho,M}\|K\|_1\|K\|_\infty\delta}{3} + c_{K,Q,\nu,M}c'_\rho}\right)\exp\left(-\frac{\delta^2|\mathbb{T}_n||h'|}{2\left(c_{K,Q,\nu,M}c'_\rho + \frac{4c_{\rho,M}\|K\|_1\|K\|_\infty\delta}{3}\right)}\right),$$

where

$$c_{\rho,M} = \frac{M(1+\rho)}{1-2\rho},\ c'_\rho = 3 + \frac{2}{1-2\rho},$$
$$c_{K,Q,\nu,M} = 8\max\{2\|K\|_1^2\|K\|_2^2(\|Q\|_{\infty,\infty} + \|\nu\|_\infty); \max\{\|Q\|_{\infty,\infty} + \|\nu\|_\infty; M\|K\|_1\|K\|_\infty\}^2\}$$

We also have for all $\delta > 0$,

$$(8)\quad \mathbb{P}\left(\frac{1}{|\mathbb{T}_n|}\left|\sum_{u\in\mathbb{T}_n}\left(K_h\left(x-X_u\right)-\mathbb{E}_\nu\left[K_h\left(x-X_u\right)\right]\right)\right|>\delta\right)$$
$$\leq 2\exp\left(\frac{\delta c'_{K,Q,\nu,M}c'_\rho}{\frac{4c_{\rho,M}\|K\|_\infty\delta}{3}+c'_{K,Q,\nu,M}c'_\rho}\right)\exp\left(-\frac{\delta^2|\mathbb{T}_n||h|}{2\left(c'_{K,Q,\nu,M}c'_\rho+\frac{4c_{\rho,M}\|K\|_\infty\delta}{3}\right)}\right)$$

where $c'_{K,Q,\nu,M}=8\max\{M\|K\|_\infty;(\|\mathcal{Q}\|_{\infty,\infty}+\|\nu\|_\infty)\|K\|_1;(\|\mathcal{Q}\|_{\infty,\infty}+\|\nu\|_\infty)\|K\|_2^2\}$.

Remark 13. *As mentioned above, these inequalities are more complete than those obtained in* [8]*, since the deviation parameter* δ *does not depend on the size of the data. We stress that this fact is essential for our theoretical results.*

Proof. We will do the proof of (7). The proof of (8) follows the same lines.

Let $\lambda>0$ and $\delta>0$. By Chernoff inequality, we have

$$\mathbb{P}\left(\frac{1}{|\mathbb{T}_n|}\sum_{u\in\mathbb{T}_n}\left(K_h*K_{h'}\left(x-X_u\right)-\mathbb{E}_\nu\left[K_h*K_{h'}\left(x-X_u\right)\right]\right)>\delta\right)$$
$$\leq\exp\left(-\delta\lambda|\mathbb{T}_n|\right)\mathbb{E}\left[\exp\left(\lambda\sum_{u\in\mathbb{T}_n}g(X_u)\right)\right],$$

where the function g is defined by

$$g(y)=K_h*K_{h'}(x-y)-\mathbb{E}_\nu\left[K_h*K_{h'}(x-X_\emptyset)\right].$$

For all $u\in\mathbb{G}_{n-1}$, we have, on the one hand

$$|g(X_{u0})+g(X_{u1})-2\mathcal{Q}g(X_u)|\leq 2M(1+\rho)\|K_h*K_{h'}\|_\infty.$$

Using the Young's inequality, we have

$$\|K_h*K_{h'}\|_\infty\leq\|K\|_1\|K\|_\infty/|h'|$$

and therefore

$$|g(X_{u0})+g(X_{u1})-2\mathcal{Q}g(X_u))|\leq 2c_{\rho,M}\|K\|_1\|K\|_\infty/|h'|.$$

On the other hand, we have

$$\mathbb{E}\left[\left(g(X_{u0})+g(X_{u1})-2\mathcal{Q}g(X_u)\right)^2|\mathcal{F}_{n-1}\right]\leq 4\|Q\|_{\infty,\infty}\|K_h*K_{h'}\|_2^2\leq c_{K,Q,\nu,M}/|h'|).$$

Now for all $\lambda\in(0,3/(2c_{\rho,M}\|K\|_1\|K\|_\infty(|h'|)^{-1}))$, we have from Bennett inequality

$$\mathbb{E}\left[\exp\left(\lambda\left(g\left(X_{u0}\right)+g\left(X_{u1}\right)-2\mathcal{Q}g\left(X_u\right)\right)\right)|\mathcal{F}_{n-1}\right]\quad\leq\quad\exp\left(\frac{c_{K,\mathcal{Q},\nu,M}(|h'|)^{-1}\lambda^2}{2\left(1-\frac{2c_{\rho,M}\|K\|_1\|K\|_\infty(|h'|)^{-1}\lambda}{3}\right)}\right)$$

and using the Markov property, this leads us to

$$\mathbb{E}\left[\exp\left(\lambda\sum_{u\in\mathbb{T}_n} g(X_u)\right)\right] \leq \exp\left(\frac{c_{K,\mathcal{Q},\nu,M}(|h'|)^{-1}\lambda^2|\mathbb{G}_{n-1}|}{2\left(1-\frac{2c_{\rho,M}\|K\|_1\|K\|_\infty(|h'|)^{-1}\lambda}{3}\right)}\right)$$
$$\times\mathbb{E}\left[\exp\left(\lambda\sum_{u\in\mathbb{T}_{n-3}} g(X_u)\right)\times\exp\left(\lambda\sum_{u\in\mathbb{G}_{n-2}} I^{(2)}(X_u)\right)\right.$$
$$\left.\times\prod_{u\in\mathbb{G}_{n-2}}\mathbb{E}\left[\exp\left(J^{(2)}(X_u,X_{u0},X_{u1})\right)\Big|\mathcal{F}_{n-2}\right]\right]$$

where

$$I^{(2)}(X_u) = (g+2\mathcal{Q}g+2^2\mathcal{Q}^2g)(X_u)$$
$$J^{(2)}(X_u,X_{u0},X_{u1}) = (g+2\mathcal{Q}g)(X_{u0})+(g+2\mathcal{Q}g)(X_{u1})-2(\mathcal{Q}g+2^2\mathcal{Q}^2g)(X_u).$$

Now for the second step, we will do the same thing with $J^{(2)}(X_u,X_{u0},X_{u1})$ instead of $g(X_{u0})+g(X_{u1})-2\mathcal{Q}g(X_u)$. For all $u\in\mathbb{G}_{n-2}$ we have

$$|J^{(2)}(X_u,X_{u0},X_{u1})| \leq 2c_{\rho,M}\|K\|_1\|K\|_\infty(|h'|)^{-1}.$$

We also have

$$\mathbb{E}[(J^{(2)}(X_u,X_{u0},X_{u1}))^2|\mathcal{F}_{n-2}] \leq 4\mathcal{Q}((g+2\mathcal{Q}g)^2)(X_u).$$

Now we will control the right hand side of the previous inequality. On the one hand, we have for all $y\in\mathbb{R}^d$,

$$\mathcal{Q}g(y) = \int\int K_h(t)\,K_{h'}(x-z-t)\,\mathcal{Q}(y,z)dtdz - \mathbb{E}_\nu[K_h * K_{h'}(x-X_\emptyset].$$

Using the change of variables

$$t = hu \quad \text{and} \quad z = x-hu-h'v,$$

where for $x=(x_1,...,x_d)^t, y=(y_1,...,y_d)^t\in\mathbb{R}^d$, xy denotes the vector $(x_1y_1,...,x_dy_d)^t$, we obtain

$$\int\int K_h(t)\,K_{h'}(x-z-t)\,\mathcal{Q}(y,z)dtdz| \leq \|Q\|_{\infty,\infty}.$$

In the same way we prove that

$$|\mathbb{E}_\nu[K_h * K_{h'}(x-X_\emptyset]| \leq \|\nu\|_\infty.$$

This leads us to

$$|\mathcal{Q}g(y)| \leq \|Q\|_{\infty,\infty}+\|\nu\|_\infty$$

and therefore, for all $m\geq 1$

$$|\mathcal{Q}^m g(y)| \leq \|Q\|_{\infty,\infty}+\|\nu\|_\infty.$$

On the other hand, uniform geometric ergodicity assumption implies that for all $m \geq 1$ and for all $y \in \mathbb{R}^d$,

$$\mathcal{Q}^m g(y) \leq M\rho^m \|K_h * K_{h'}\| \leq M\|K\|_1\|K\|_\infty \rho^m (|h'|)^{-1},$$

and therefore, for all $m \geq 1$ and for all $y \in \mathbb{R}^d$, we have

$$\mathcal{Q}^m g(y) \leq \max\{\|Q\|_{\infty,\infty} + \|\nu\|_\infty, M\|K\|_1\|K\|_\infty\} \times \inf\{1, \rho^m(|h'|)^{-1}\}.$$

For all $y \in \mathbb{R}^d$, we also have

$$\begin{aligned}\mathcal{Q}g^2(y) &\leq 2\mathcal{Q}(K_h * K_{h'}) + 2(\mathbb{E}[K_h * K_{h'}(x - X_\emptyset)])^2 \\ &\leq 2\|K\|_1^2\|K\|_2^2(\|Q\|_{\infty,\infty} + \|\nu\|_\infty),\end{aligned}$$

in such a way that for all $u \in \mathbb{G}_{n-2}$,

$$\begin{aligned}4\mathcal{Q}((g + 2\mathcal{Q}g)^2)(X_u) &\leq 8\mathcal{Q}g^2(X_u) + 8\mathcal{Q}((2\mathcal{Q}g)^2)(X_u) \\ &\leq c_{K,\mathcal{Q},\nu,M}((|h'|)^{-1} + 2\inf\{1, \rho(|h'|)^{-1}\}^2)\end{aligned}$$

and thus, we obtain

$$\mathbb{E}[J^{(2)}(X_u, X_{u0}, X_{u1})|\mathcal{F}_{n-2}] \leq c_{K,\mathcal{Q},\nu,M}((|h'|)^{-1} + 2\inf\{1, \rho(|h'|)^{-1}\}^2).$$

Once again, for all $\lambda \in (0, 3/(2c_{\rho,M}\|K\|_1\|K\|_\infty(|h'|)^{-1}))$ and for all $u \in \mathbb{G}_{n-2}$, we have from Bennett's inequality

$$\mathbb{E}[J^{(2)}(X_u, X_{u0}, X_{u1})|\mathcal{F}_{n-2}] \leq \exp\left(\frac{c_{K,\mathcal{Q},\nu,M}((|h'|)^{-1} + 2\inf\{1, \rho(|h'|)^{-1}\}^2)\lambda^2}{2\left(1 - \frac{2c_{\rho,M}\|K\|_1\|K\|_\infty(|h'|)^{-1}\lambda}{3}\right)}\right).$$

It then follows that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda \sum_{u\in\mathbb{T}_n} g(X_u)\right)\right] &\leq \exp\left(\frac{c_{K,\mathcal{Q},\nu,M}(|h'|)^{-1}\lambda^2|\mathbb{G}_{n-1}|}{2\left(1 - \frac{2c_{\rho,M}\|K\|_1\|K\|_\infty(|h'|)^{-1}\lambda}{3}\right)}\right) \\ &\times \exp\left(\frac{c_{K,\mathcal{Q},\nu,M}((|h'|)^{-1} + 2\inf\{1, \rho(|h'|)^{-1}\}^2)\lambda^2|\mathbb{G}_{n-2}|}{2\left(1 - \frac{2c_{\rho,M}\|K\|_1\|K\|_\infty(|h'|)^{-1}\lambda}{3}\right)}\right) \\ &\times \mathbb{E}\left[\exp\left(\lambda \sum_{u\in\mathbb{T}_{n-4}} g(X_u)\right) \times \exp\left(\lambda \sum_{u\in\mathbb{G}_{n-3}} I^{(3)}(X_u)\right)\right. \\ &\left. \times \prod_{u\in\mathbb{G}_{n-3}} \mathbb{E}\left[\exp\left(J^{(3)}(X_u, X_{u0}, X_{u1})\right)\Big|\mathcal{F}_{n-2}\right]\right]\end{aligned}$$

where

$$\begin{aligned}I^{(3)}(X_u) &= (g + 2\mathcal{Q}g + 2^2\mathcal{Q}^2g + 2^3\mathcal{Q}^3g)(X_u); \\ J^{(3)}(X_u, X_{u0}, X_{u1}) &= (g + 2\mathcal{Q}g + 2^2\mathcal{Q}^2g)(X_{u0}) + (g + 2\mathcal{Q}g + 2^2\mathcal{Q}^2g)(X_{u1}) - 2(\mathcal{Q}g + 2\mathcal{Q}^2g + 2^2\mathcal{Q}^3g)(X_u).\end{aligned}$$

Now, iterating this method, we are led to

$$\mathbb{E}\left[\exp\left(\lambda\sum_{u\in\mathbb{T}_n} g(X_u)\right)\right] \leq \exp\left(\frac{\sum_{m=1}^{n} c_{K,\mathcal{Q},\nu,M}\left((|h'|)^{-1}+\left(\sum_{l=1}^{m-1}2^l((|h'|)^{-1}\rho^l\wedge 1)\right)^2\right)|\mathbb{G}_{n-m|}\lambda^2}{2\left(1-\frac{2c_{\rho,M}\|K\|_1\|K\|_\infty(|h'|)^{-1}\lambda}{3}\right)}\right) \times \exp\left(c_{K,\mathcal{Q},\nu,M}\left((|h'|)^{-1}+\sum_{m=1}^{n}2^m\left(1\wedge\rho^m(|h'|)^{-1}\right)\right)\right).$$

Set $m^*=\lfloor \log|h'|/\log\rho\rfloor$. Then we have

$$\sum_{m=1}^{n}\left((|h'|)^{-1}+\left(\sum_{l=1}^{m-1}2^l((|h'|)^{-1}\rho^l\wedge 1)\right)^2\right)|\mathbb{G}_{n-m|}$$
$$=\sum_{m=m^*+1}^{n}\left((|h'|)^{-1}+\left(\sum_{l=1}^{m^*}2^l+\sum_{l=m^*}^{m-1}2^l h^{-1}\rho^l\right)^2\right)|\mathbb{G}_{n-m|}+\sum_{m=1}^{m^*}\left((|h'|)^{-1}+\left(\sum_{l=1}^{m-1}2^l\right)^2\right)|\mathbb{G}_{n-m|}$$
$$\leq\left(6+(1+\frac{1}{1-2\rho})^2\right)(|h'|)^{-1}|\mathbb{T}_n|.$$

We also have

$$(|h'|)^{-1}+\sum_{m=1}^{n}2^m\left(1\wedge\rho^m(|h'|)^{-1}\right)\leq c'_\rho(|h'|)^{-1}.$$

In view of the above, for all $\lambda\in(0,3/(2c_{\rho,M}\|K\|_1\|K\|_\infty(|h'|)^{-1}))$ we have

$$\mathbb{P}\left(\frac{1}{|\mathbb{T}_n|}\sum_{u\in\mathbb{T}_n}\left(K_h*K_{h'}(x-X_u)-\mathbb{E}_\nu\left[K_h*K_{h'}(x-X_u)\right]\right)>\delta\right) \leq\exp\left(-\lambda\delta|\mathbb{T}_n|+\frac{c_{K,\mathcal{Q},\nu,M}c'_\rho(|h'|)^{-1}|\mathbb{T}_n|\lambda^2}{2\left(1-\frac{2c_{\rho,M}\|K\|_1\|K\|_\infty(|h'|)^{-1}\lambda}{3}\right)}\right)\times\exp\left(\lambda c_{K,\mathcal{Q},\nu,M}c'_\rho(|h'|)^{-1}\right).$$

Taking $\lambda=(\delta(|h'|))/(c_{K,\mathcal{Q},\nu,M}c'_\rho+(4c_{\rho,M}\|K\|_1\|K\|_\infty\delta)/3)$, we obtain

$$\mathbb{P}\left(\frac{1}{|\mathbb{T}_n|}\sum_{u\in\mathbb{T}_n}\left(K_h*K_{h'}(x-X_u)-\mathbb{E}_\nu\left[K_h*K_{h'}(x-X_u)\right]\right)>\delta\right) \leq\exp\left(\frac{\delta c_{K,\mathcal{Q},\nu,M}c'_\rho}{\frac{4c_{\rho,M}\|K\|_1\|K\|_\infty\delta}{3}+c_{K,\mathcal{Q},\nu,M}c'_\rho}\right)\exp\left(-\frac{\delta^2|\mathbb{T}_n|(|h'|)}{2\left(c_{K,\mathcal{Q},\nu,M}c'_\rho+\frac{4c_{\rho,M}\|K\|_1\|K\|_\infty\delta}{3}\right)}\right).$$

The result follows since we can do the same thing for $-g$ instead of g. Now, the proof of (8) follows the same lines and this ends the proof. □

Proof of Theorem 7. We start from the following decomposition, true for all $h \in \mathcal{H}_n$

$$(\widehat{\nu}_{\widehat{h}}(x) - \nu(x))^2 \leq 3(\widehat{\nu}_{\widehat{h}}(x) - K_h * \widehat{\nu}_{\widehat{h}}(x))^2 + 3(K_h * \widehat{\nu}_{\widehat{h}}(x) - \widehat{\nu}_h(x))^2 + 3(\widehat{\nu}_h(x) - \nu(x))^2.$$

Hence, since $K_h * \widehat{\nu}_{\widehat{h}}(x) = K_{\widehat{h}} * \widehat{\nu}_h(x)$, by definition of $A(x,h)$ and then by definition of $\widehat{h}$ and the fact that $a \leq b$

$$\begin{aligned}
(\widehat{\nu}_{\widehat{h}}(x) - \nu(x))^2 &\leq 3((\widehat{\nu}_{\widehat{h}}(x) - K_h * \widehat{\nu}_{\widehat{h}}(x))^2 - aV(x,\widehat{h}) + aV(x,\widehat{h})) \\
&\quad +3((K_{\widehat{h}} * \widehat{\nu}_h(x) - \widehat{\nu}_h(x))^2 - aV(x,h) + aV(x,h)) + 3(\widehat{\nu}_h(x) - \nu(x))^2 \\
&\leq 3(A(x,h) + bV(x,\widehat{h})) \\
&\quad +3(A(x,\widehat{h}) + bV(x,h)) + 3(\widehat{\nu}_h(x) - \nu(x))^2 \\
&\leq 6(A(x,h) + bV(x,h)) + 3(\widehat{\nu}_h(x) - \nu(x))^2. \qquad (9)
\end{aligned}$$

Now it remains to upper-bound $\mathbb{E}[A(x,h)]$. We have

$$\begin{aligned}
(\widehat{\nu}_{h'}(x) - K_h * \widehat{\nu}_{h'}(x))^2 &\leq 3(\widehat{\nu}_{h'}(x) - K_{h'} * \nu(x))^2 + 3(K_h * \widehat{\nu}_{h'}(x) - K_h * K_{h'} * \nu(x))^2 \\
&\quad +3(K_{h'} * \nu(x) - K_h * K_{h'} * \nu(x))^2.
\end{aligned}$$

With a rough upper-bound of the $\max_{h \in \mathcal{H}_n}$ by the $\sum_{h \in \mathcal{H}_n}$ we get

$$\begin{aligned}
\mathbb{E}[A(x,h)] &\leq 3\,\mathbb{E}\left[\max_{h' \in \mathcal{H}_n}\left((\widehat{\nu}_{h'}(x) - K_{h'} * \nu(x))^2 - a\frac{V(x,h')}{6}\right)_+\right] \\
&\quad + 3\,\mathbb{E}\left[\max_{h' \in \mathcal{H}_n}\left(K_h * \widehat{\nu}_{h'}(x) - K_h * K_{h'} * \nu(x))^2 - a\frac{V(x,h')}{6}\right)_+\right] \\
&\quad + 3\max_{h' \in \mathcal{H}_n}(K_{h'} * \nu(x) - K_h * K_{h'} * \nu(x))^2 \\
&\leq 3\sum_{h' \in \mathcal{H}_n}\mathbb{E}\left[\left(\left(\frac{1}{|\mathbb{T}_n|}\sum_{u \in \mathbb{T}_n} K_{h'}(x - X_u) - \mathbb{E}_\nu[K_{h'}(x - X_u)]\right)^2 - a\frac{V(x,h')}{6}\right)_+\right] \\
&+3\sum_{h' \in \mathcal{H}_n}\mathbb{E}\left[\left(\left(\frac{1}{|\mathbb{T}_n|}\sum_{u \in \mathbb{T}_n} K_h * K_{h'}(x - X_u) - \mathbb{E}_\nu[K_h * K_{h'}(x - X_u)]\right)^2 - a\frac{V(x,h')}{6}\right)_+\right] \\
&\quad +3\max_{h' \in \mathcal{H}_n}(K_{h'} * \nu(x) - K_h * K_{h'} * \nu(x))^2 \\
&\leq \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{B}_h(x).
\end{aligned}$$

We first give an upper-bound for $\mathcal{T}_1$. Let $h' \in \mathcal{H}_n$ fixed, now remark that, by Lemma 12,

$$\begin{aligned}
\mathcal{T}_1 &= \int_0^{+\infty} \mathbb{P}\left(\left(\left(\frac{1}{|\mathbb{T}_n|}\sum_{u\in\mathbb{T}_n} K_{h'}(x-X_u) - \mathbb{E}_\nu[K_{h'}(x-X_u)]\right)^2 - a\frac{V(x,h')}{6}\right)_+ \geq t\right) dt \\
&\leq \int_0^{+\infty} \mathbb{P}\left(\left|\frac{1}{|\mathbb{T}_n|}\sum_{u\in\mathbb{T}_n} K_{h'}(x-X_u) - \mathbb{E}_\nu[K_{h'}(x-X_u)]\right| \geq \sqrt{t + a\frac{V(x,h')}{6}}\right) dt \\
&\leq \int_{aV(x,h')/6}^{+\infty} \mathbb{P}\left(\left|\frac{1}{|\mathbb{T}_n|}\sum_{u\in\mathbb{T}_n} K_{h'}(x-X_u) - \mathbb{E}_\nu[K_{h'}(x-X_u)]\right| \geq \sqrt{u}\right) du \\
&\leq \int_{aV(x,h')/6}^{+\infty} \exp\left(\frac{\sqrt{u}c_K c'_\rho}{\frac{4c_\rho\|K\|_\infty\sqrt{u}}{3} + c_K c'_\rho}\right)\exp\left(-\frac{u|\mathbb{T}_n||h|}{2(c_K c'_\rho + (4/3)c_\rho\|K\|_\infty\sqrt{u})}\right) du \\
&\leq I_1 + I_2,
\end{aligned}$$

where

$$\begin{aligned}
I_1 &= \int_{aV(x,h')/6}^{aC(P,\mu)/6} \exp\left(\frac{\sqrt{u}c_K c'_\rho}{\frac{4c_\rho\|K\|_\infty\sqrt{u}}{3} + c_K c'_\rho}\right)\exp\left(-\frac{u|\mathbb{T}_n||h|}{2(c_K c'_\rho + (4/3)c_\rho\|K\|_\infty\sqrt{u})}\right) du \\
I_2 &= \int_{aC(P,\mu)/6}^{+\infty} \exp\left(\frac{\sqrt{u}c_K c'_\rho}{\frac{4c_\rho\|K\|_\infty\sqrt{u}}{3} + c_K c'_\rho}\right)\exp\left(-\frac{u|\mathbb{T}_n||h|}{2(c_K c'_\rho + (4/3)c_\rho\|K\|_\infty\sqrt{u})}\right) du,
\end{aligned}$$

where we recall that, since for all $h' \in \mathcal{H}_n$, $|h'| \geq \frac{\log(|\mathbb{T}_n|}{|\mathbb{T}_n|}$, we have $V(x,h') \leq C(P,\mu)$.

We first upper-bound I_1,

$$\begin{aligned}
I_1 &\leq C_1 \int_{aV(x,h')/6}^{aC(P,\mu)/6} \exp\left(-\frac{u|\mathbb{T}_n||h'|}{2(c_K c'_\rho + (2\sqrt{2}/3\sqrt{3})c_\rho\|K\|_\infty\sqrt{aC(P,\mu)})}\right) du \\
&\leq \frac{C'_1}{|\mathbb{T}_n||h'|}\exp\left(-\sqrt{a}c_1^*|\mathbb{T}_n||h'|\right) \leq C'_1\log(|\mathbb{T}_n|)\exp\left(-\sqrt{a}c_1^*\log(|\mathbb{T}_n|)\right) = C'_1|\mathbb{T}_n|^{-\sqrt{a}c_1^*}, \qquad (10)
\end{aligned}$$

with $C_1 = \exp(\sqrt{aC(P,\mu)}/6)$, $C'_1 = C_1 2(c_K c'_\rho + (2\sqrt{2}/3\sqrt{3})c_\rho\|K\|_\infty\sqrt{aC(P,\mu)})$, $c_1^* = C(P,\mu)/(12(c_K c'_\rho + (2\sqrt{2}/3\sqrt{3})c_\rho\|K\|_\infty\sqrt{C(P,\mu)}))$, using the fact that $a \geq 1$ (remark that c_1^* does not depend on a).

We turn now to I_2, remark that the function $u \mapsto \sqrt{u}c_K c'_\rho/(\frac{4c_\rho\|K\|_\infty\sqrt{u}}{3} + c_K c'_\rho)$ is non decreasing and converges to $3c_K c'_\rho/(4c_\rho\|K\|_\infty)$ when $u \to \infty$, hence it is bounded by this quantity. We have then, using again $a \geq 1$,

$$\begin{aligned}
I_2 &\leq C_2 \int_{aC(P,\mu)/6}^{+\infty} \exp\left(-\frac{\sqrt{u}|\mathbb{T}_n||h'|}{2\left(\frac{\sqrt{6}c_K c'_\rho}{\sqrt{C(P,\mu)}} + (4/3)c_\rho\|K\|_\infty\right)}\right) du \\
&\leq \left(\frac{C'_2}{|\mathbb{T}_n||h'|} + \frac{C''_2}{(|\mathbb{T}_n||h'|)^2}\right)\exp(-\sqrt{a}c_2^*|\mathbb{T}_n||h'|) \\
&\leq C'_2\log(|\mathbb{T}_n|)|\mathbb{T}_n|^{-\sqrt{a}c_2^*} + C''_2\log^2(|\mathbb{T}_n|)|\mathbb{T}_n|^{-\sqrt{a}c_2^*}, \qquad (11)
\end{aligned}$$

with

$$C_2 = \exp(3c_K c'_\rho/(4c_\rho\|K\|_\infty), \quad C'_2 = C_2\sqrt{a}(4c_K c'_\rho + \frac{16}{3\sqrt{6}}\sqrt{C(P,\mu)}c_\rho\|K\|_\infty),$$

$$C''_2 = 8C_2\left(\sqrt{6}c_K c'_\rho/\sqrt{C(P,\mu)} + \frac{4}{3}c_\rho\|K\|_\infty\right)^2 \quad \text{and} \quad c_2^* = \frac{\sqrt{C(P,\mu)}}{2\sqrt{6}\left(\sqrt{6}c_K c'_\rho/\sqrt{C(P,\mu)} + \frac{4}{3}c_\rho\|K\|_\infty\right)}.$$

Hence, gathering (10) and (11), there exists $C' > 0$ depending only on $C(P,\mu)$, K, c_K and ρ, such that

$$\mathcal{T}_1 \quad \leq \quad C\,\mathrm{card}(\mathcal{H}_n)\log^2(|\mathbb{T}_n|)|\mathbb{T}_n|^{-\sqrt{a}c^*} \leq C'|\mathbb{T}_n|^{-1}$$

with $C = 3\max\{C'_1, C'_2, C''_2\}$, $c^* = \min\{c_1^*, c_2^*\}$ as soon as $a > 4/(c^*)^2$. □

6. Acknowledgement

The authors want to thank Claire Lacour for her helpful advices on constant calibration.

References

[1] S. Arlot, and P. Massart. *Data-driven calibration of penalties for least-squares regression.* Journal of Machine Learning Research, 10 (2009), 245–279.

[2] I. V. Basawa and J. Zhou. *Non-Gaussian bifurcating models and quasi-likelihood estimation.* Journal of Applied Probability, 41 (2004), 55-64.

[3] M. Bec and C. Lacour. *Adaptive pointwise estimation for pure jump Lévy processes.* Statistical Inference and Stochastic Processes, 18 (2015), 229–256.

[4] I. Benjamini and Y. Peres. *Markov chains indexed by trees.* The Annals of Probability, 22 (1994), 219–243.

[5] K. Bertin, C. Lacour and V. Rivoirard. *Adaptive pointwise estimation of conditional density function.* Annales de l'Institut Henri Poincaré Probabilités et Statistiques, 52 (2016), 939–980.

[6] L. Birgé and P. Massart. *Minimum contrast estimators on sieves: exponential bounds and rates of convergence.* Bernoulli, 4 (1998), 329–375.

[7] S. V. Bitseki Penda, H. Djellout and A. Guillin. *Deviation inequalities, moderate deviations and some limit theorems for bifurcating Markov chains with application.* The Annals of Applied Probability, 24 (2014), 235–291.

[8] S. V. Bitseki Penda, M. Hoffmann and A. Olivier. *Adaptive estimation for bifurcating Markov chains.* To appear in Bernoulli (2016).

[9] S. V. Bitseki Penda., and A. Olivier. *Autoregressive Functions Estimation in Nonlinear Bifurcating Autoregressive Models.* Statistical Inference for Stochastic Processes, 20 (2017), 179–210.

[10] G. Chagny. *Adaptive warped kernel estimators.* Scandinavian Journal of Statistics, 42 (2015), 336–360.

[11] G. Chagny and A. Roche. *Adaptive and minimax estimation of the cumulative distribution function given a functional covariate.* Electronic Journal of Statistics, 8 (2014), 2352–2404.

[12] G. Chagny and A. Roche. *Adaptive estimation in the functional nonparametric regression model.* Journal of Multivariate Analysis, 146 (2016), 105–118.

[13] M. Chichignoud and S. Loustau. *Bandwidth selection in kernel empirical risk minimization via the gradient.* The Annals of Statistics, 43 (2015), 1617–1646.

[14] F. Comte, and C. Lacour. *Anisotropic adaptive kernel deconvolution.* Annales de l'Institut Henri Poincaré Probabilités et Statistiques, 49 (2013), 569–609.

[15] R. Cowan and R. G. Staudte. *The bifurcating autoregressive model in cell lineage studies.* Biometrics, 42 (1986), 769–783.

[16] C. Dion and V. Genon-Catalot. *Bidimensional random effect estimation in mixed stochastic differential model.* Statistical Inference for Stochastic Processes, 19 (2016), 131–158.

[17] M. Doumic, M. Hoffmann, N. Krell and L. Robert. *Statistical estimation of a growth-fragmentation model observed on a genealogical tree.* Bernoulli, 21 (2015), 1760–1799.

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

[19] A. Goldenshluger and O. Lepski. *On adaptive minimax density estimation on R^d.* Probability Theory and Related Fields, 159 (2014), 479–543.

[20] A. Guilloux, S. Lemler and M.-L. Taupin. *Adaptive kernel estimation of the baseline function in the Cox model with high-dimensional covariates*, Journal of Multivariate Analysis, 148 (2016), 141–159.
[21] J. Guyon. *Limit theorems for bifurcating Markov chains. Application to the detection of cellular aging*. The Annals of Applied Probability, 17 (2007), 1538–1569.
[22] C. Lacour, and P. Massart. *Minimal penalty for the Goldenshluger-Lepski method*. Stochastic Processes and their Applications, 126 (2016), 3774–3789.
[23] O. Lepski. *Adaptive estimation over anisotropic functional classes via oracle approach*. The Annals of Statistics, 43 (2015), 1178–1242.
[24] T. Patschkowski and A. Rohde. *Adaptation to lowest density regions with application to support recovery*. The Annals of Statistics, 44 (2016), 255–287.
[25] M. K. Pitt, C. Chatfield and S. G. Walker *Constructing First Order Stationary Autoregressive Models via Latent Processes*. Scandinavian Journal of Statistics, 29, 4 (2002), 657–663
[26] G. Rebelles. $\mathbb{L}_p$ *adaptive estimation of an anisotropic density under independence hypothesis*. Electronic Journal of Statistics, 9 (2015), 106–134.
[27] G. Rebelles *Pointwise adaptive estimation of a multivariate density under independence hypothesis*. Bernoulli, 20 (2015), 1984–2023.
[28] A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics, Springer-Verlag, New-York, 2009.

S. Valère Bitseki Penda, IMB, CNRS-UMR 5584, Université Bourgogne Franche-Comté, 9 avenue Alain Savary, 21078 Dijon Cedex, France.
E-mail address: simeon-valere.bitseki-penda@u-bourgogne.fr

Angelina Roche, CEREMADE, CNRS-UMR 7534, Université Paris-Dauphine, Place du maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France.
E-mail address: roche@ceremade.dauphine.Fr