Approximations of the allelic frequency spectrum in general supercritical branching populations

Benoît Henry

▶ **To cite this version:**

Benoît Henry. Approximations of the allelic frequency spectrum in general supercritical branching populations. 2018. hal-01445838v2

HAL Id: hal-01445838
https://hal.science/hal-01445838v2

Preprint submitted on 12 Feb 2018

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

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

Approximation of the allelic frequency spectrum in general supercritical branching populations

Benoit Henry[1]

Abstract

We consider a general branching population where the lifetimes of individuals are i.i.d. with arbitrary distribution and where each individual gives birth to new individuals at Poisson times independently from each other. In addition, we suppose that individuals experience mutations at Poissonian rate θ under the infinitely many alleles and neutrality assumptions assuming that types are transmitted from parents to offspring. This mechanism leads to a partition of the population by type, called the allelic partition. The main object of this work is the frequency spectrum $A(k,t)$ which counts the number of families of size k in the population at time t. The process $(A(k,t),\ t \in \mathbb{R}_+)$ is an example of non-Markovian branching process belonging to the class of general branching processes counted by random characteristics. In this work, we propose methods of approximation to replace the frequency spectrum by simpler quantities. Our main goal is study the asymptotic error made during these approximations through central limit theorems. In a last section, we perform several numerical analysis using this model, in particular to analyze the behavior of one of these approximations with respect to Sabeti's Extended Haplotype Homozygosity [20].

MSC 2000 subject classifications: Primary 60J80; secondary 92D10, 60J85, 60G51, 60K15, 60F05.

Key words and phrases. allelic partition – allelic frequency spectrum – branching processes – neutral mutations – splitting tree – Central Limit Theorem.

1 Introduction

In this paper, we consider a general branching population where the lifetimes of the individuals and their reproductions processes are independent and follow the same distribution. Moreover, we assume that their lifetimes are distributed according to an arbitrary probability distribution $\mathbb{P}_V$ and that the births occur, during their lifetime, according to a Poisson process with constant rate b. The tree underlying this dynamics is called a splitting tree. This class of random trees was introduced in [13] by Geiger and Kersting and has been widely studied in the last decade [16, 17, 18].

We suppose, in addition, that neutral mutations occur on individuals and that each new mutation confers to its holder a brand new type (i.e. never seen in the population): this is the *infinitely many*

[1]IECL, UMR CNRS 7502, Université de Lorraine, Site de Nancy, B.P. 70239, F-54506 Vandoeuvre-lès-Nancy Cedex, France
E-mail: benoit.henry@univ-lorraine.fr

alleles assumption. This allows modeling the occurrence of a new type in a population (such as a new species or a new phenotype in a given specie). We also suppose that every individual inherits the type of its parent. This model leads to a partition of the population by types. The frequency spectrum of the population alive at time t is defined as the sequence of number $(A(k,t),\ k \in \mathbb{N}^*)$ ($\mathbb{N}^*$ refers to the set of positive integers) where, for each k, $A(k,t)$ is the number of families (i.e. sets of individuals carrying the same type) of size k in the population. The famous example of Ewens sampling formula gives explicit expression for the law of the frequency spectrum [12] when the genealogy is given by the Kingman's coalescent. However, due to its central role in biology, the frequency spectrum has been studied in many other population models. Among coalescent processes, the frequency spectrum has, for instance, also been studied in the context of Beta [5], Bolthausen-Sznitman [3] or Lambda [4] coalescents. Other works studied similar quantities in the case of Galton-Waston branching processes (see [6] or [14]). In our model, the frequency spectrum has also been widely studied in the past [9, 10, 11, 8].

Another object of interest is the process $(N_t,\ t \in \mathbb{R}_+)$ which counts the number of living individuals in the population at a given time t. This process is known as binary homogeneous Crump-Mode-Jagers process. One of the main result of the theory of such process is the law of large number which gives in our particular case that $e^{-\alpha t}N_t$ converges almost surely to a random variable $\mathcal{E}$ which is exponential conditionally on non-extinction (for some positive constant α).

As for $e^{-\alpha t}N_t$, it is also known that the quantities $e^{-\alpha t}A(k,t)$ converge almost surely to $c_k\mathcal{E}$, where c_k is an explicit deterministic constant. This result can be easily obtained by conjunction of the works of [9] and [19] using the theory of general branching processes counted by random characteristics (a complete statement can be found in [11]). An alternative proof avoiding the use of the general branching processes theory can be found in [8].

It appears that the frequency spectrum $(A(k,t),\ k \in \mathbb{N}^*)$ is a quantity which is hard to manipulate from the probabilistic point of view (see [9, 10, 8]). This implies that such a model is inconvenient for practical applications. In this work we propose to use the laws of large numbers in order to replace $(A(k,t))_{k\geq 1}$ by more manipulable quantities and propose to investigate the error made during this approximation. The first possible approximation is the following.

Approximation 1:

$$(A(k,t),\ k \in \mathbb{N}^*) \approx (c_k)_{k\geq 1}e^{\alpha t}\mathcal{E}.$$

However, this is unsatisfactory for practical applications since the random variable $\mathcal{E}$ is not observable at finite times. Another idea is to exploit the fact that the random variable appearing in the law of large numbers for $A(k,t)$ and for N_t is the same. This leads to the second approximation.

Approximation 2:

$$(A(k,t),\ k \in \mathbb{N}^*) \approx (c_k)_{k\geq 1}N_t.$$

In order to investigate the errors made during this approximation (at least asymptotically), one would like to have central limit theorems associated to the law of large numbers for the frequency spectrum. In a previous work [15], we showed that the error in the convergence of $e^{-\alpha t}N_t$ is of order $e^{\alpha t/2}$ and obtained a central limit theorem for this error. An important aspect of the method introduced in [15] is that it can be used to derive CLTs for other branching processes counted by random characteristics.

In particular, the main goal of this work is to obtain central limit theorems for the convergence of the frequency spectrum. More precisely, we show that the error in both approximations converges (when renormalized) to a Laplace random variable which is obtained through a Gaussian mixing with the limiting random variable of $e^{\alpha t}N_t$ (which also improve the results of [15]).

The original motivation of this study (and of other works on this model [9, 10, 8]) comes from the works of Sabeti and al. [20] where the frequency spectrum is used to detect positive selection of an allele in an increasing population. More specifically, suppose that you want to detect the positive selection of an allele on a given gene. The main idea is that, under neutral evolution, the allele under consideration needs a long time to reach a high frequency in the population. Hence, if the frequency of the allele w.r.t. its age is significantly higher than the expected frequency (w.r.t. its age and under neutral growth), this anomaly would suggest a positive selection of this allele. The main problem is now to be able to estimate how old the allele is. Sabeti and al. remarked that the allelic partition can be used as a clock to estimate the age of an allele. More precisely, the type is defined as a given part of the genome (of the specie) of length x (measured in a meaningful unit such as kilo-bases). Hence, two individuals have different types if this portion of DNA has differences in the sequence of basis. As a consequence, the allelic partition of the subpopulation carrying the allele becomes thinner as x increases (because the higher x is, the higher is the probability that a mutation occurred on a sequence of x bases). Finally, the speed of fragmentation of the allelic partition, when x increases, should give clues on the age of the allele. One of the purposes of this model is to understand how the frequency spectrum evolves under neutral evolution and to provide tools to pursue such analysis. In this work, we discuss some aspects of this idea and give some directions in order to construct rigorous tests for the positive selection (see Section 8).

The paper is organized as follows. Section 2 is devoted to the mathematical description of the model and to preliminary results which are used in the sequel. Section 3 gives the mains theoretical results of this work and, in particular, central limit theorems which allow to study the error in our proposed approximations. Section 4 gives an outline of the strategy of proof. Section 6 and 7 are devoted to the proofs of Theorem 3.4 and 3.1 respectively. Finally, in Section 8 we perform some numerical studies on the model to stress the quality of our approximation. The discussions about the method of Sabeti and al. are given in this last section. An appendix contains some technical proofs.

2 Model and preliminaries

In this work, we consider a branching population with the following dynamic: starting with a single individual (called the *ancestor*) whose lifetime is distributed according to an arbitrary probability distribution $\mathbb{P}_V$ on $(0, \infty]$, this *ancestor* gives birth to new individuals at a Poissonian rate b. Each birth event giving a single new individual. From this point, each child of the ancestor lives and gives birth according to the same mechanism independently from the other individuals in the population. This formal description can be made rigorous through the definition of a probability distribution on the set of chronological trees. For the details of such construction, we refer the reader to [16]. The first quantity of interest when studying such population is the number N_t of alive individuals

in the population at a fixed time t (assuming that the time $t = 0$ is birth-date of the ancestor). The process $(N_t,\ t \in \mathbb{R}_+)$ is known as binary homogeneous Crump-Mode-Jagers process and is a simple example of non-Markovian branching process. In the sequel, we denote by $W(t)$ the expectation of N_t conditionally on the non-extinction at time t. That is

$$W(t) := \mathbb{E}\left[N_t \mid N_t > 0\right].$$

In [16], the author shows that the random variable N_t is geometrically distributed conditionally on $\{N_t > 0\}$ with parameter $\frac{1}{W(t)}$. In addition, the author of [16] showed that the Laplace transform of W can be linked to the Laplace transform of $\mathbb{P}_V$ through the relation

$$\int_{[0,\infty)} W(s)e^{-\lambda s}\, ds = \frac{1}{\psi(\lambda)},\ \forall \lambda > \alpha,$$

where

$$\psi(x) = x - b\int_{(0,\infty]} \left(1 - e^{-rx}\right)\ \mathbb{P}_V(dr), \quad x \in \mathbb{R}_+, \tag{2.1}$$

and α is the largest root of ψ,

$$\alpha = \sup\{x \in \mathbb{R}_+ \mid \psi(x) = 0\}.$$

The function ψ is called the Laplace exponent of the tree and characterizes its law. In particular, the Laplace transform of $\mathbb{P}_V$ can be expressed in terms of ψ,

$$\int_{\mathbb{R}_+} e^{-\lambda v}\ \mathbb{P}_V(dv) = 1 + \frac{\psi(\lambda) - \lambda}{b}, \quad \forall \lambda \in \mathbb{R}_+. \tag{2.2}$$

In this work, we assume that α is a strictly positive real number. This case is called the supercritical case and is equivalent to $b\mathbb{E}[V] > 1$, where V is some random variable with distribution $\mathbb{P}_V$. In the supercritical case, the real number α is called the Malthusian parameter of the population because it corresponds to the mean exponential growth rate of the population. Before going further, let us remark that equation (2.2) leads easily to the following identity:

$$\int_{\mathbb{R}_+} e^{-\alpha v}\ \mathbb{P}_V(dv) = 1 - \frac{\alpha}{b}. \tag{2.3}$$

Many previous works [9, 10, 11] demonstrate that some properties of the splitting tree were easier to study on the tree describing only the genealogical relation between the lineages of the individuals alive at time t. For instance, in the model with mutations, the difference between two individuals in term of type lies only on the time past since their lineages has diverged. Hence, this particular genealogical tree, known as *coalescent point processes* (CPP), contains the essential information to study the allelic partition. In order to derive the law of that genealogical tree, one needs to characterize the joint law of the *times of coalescence* between pairs of individuals in the population, which are the times since their lineages have split.

In [16], the author defines an order on the set of individuals alive at a fixed time t, conditionally on $\{N_t > 0\}$, and considers the sequence of times of coalescence $(H_i)_{0 \le i \le N_t - 1}$ between two consecutive individuals (that is H_i is the time passed since the lineage of individuals i and $i + 1$ have diverged)

with the convention that the older lineage is the first one (i.e. $H_0 = t$). Moreover, in [16], the author shows that the random vector $(H_i)_{0\leq i\leq N_t-1}$ can be produced from a sequence $(H_i)_{i\geq 1}$ of i.i.d. random variable stopped at its first value greater than t and such that

$$\mathbb{P}(H_1 > s) = \frac{1}{W(s)}, \quad s \in \mathbb{R}_+.$$

To summarize, given the population is still alive at time t, one can forget about the details of the splitting tree and code the genealogy by a new object called the *coalescent point process* (CPP). Its law is the law of a sequence $(H_i)_{0\leq i\leq N_t-1}$, where the family $(H_i)_{i\geq 1}$ is i.i.d. with the same law as H, stopped before its first value H_{N_t} greater than t, and H_0 is deterministic equal to t (see Figure 1).

Although we do not use directly the CPP in this work, this object allowed us to obtain [8] formulas for the moments of the frequency spectrum which are widely used in the sequel. For this reason, we recall the properties needed to understand the methods.

Remark 2.1. *Let N be a integer valued random variable. In the sequel we said that a random vector with random size $(X_i)_{1\leq i\leq N}$ form an i.i.d. family of random variables independent of N, if and only if*

$$(X_1, \ldots, X_N) \stackrel{d}{=} \left(\tilde{X}_1, \ldots, \tilde{X}_N\right),$$

where $\left(\tilde{X}_i\right)_{i\geq 1}$ is a sequence of i.i.d. random variables distributed as X_1 independent of N.

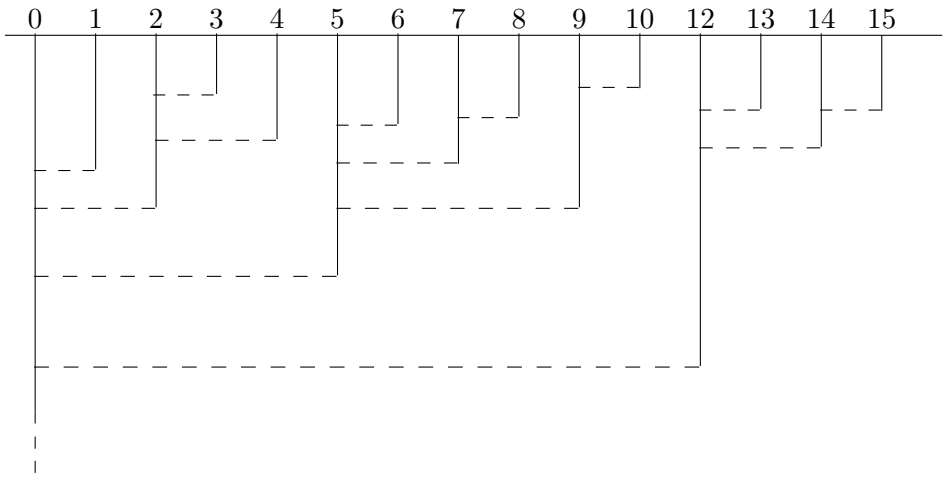

Figure 1: A coalescent point process for 16 individuals, hence 15 branches. (Image by A. Lambert)

Before going further, let us point out that if we define N_t as the first value of the sequence $(H_i)_{i\geq 1}$ greater than t, i.e.

$$N_t = \inf\{i \geq 1 \mid H_i > t\},$$

then N_t is indeed geometric with the expected parameter. More precisely, for a positive integer k,

$$\mathbb{P}(N_t = k \mid N_t > 0) = \frac{1}{W(t)}\left(1 - \frac{1}{W(t)}\right)^{k-1}. \tag{2.4}$$

In particular,

$$\mathbb{E}\left[N_t \mid N_t > 0\right] = W(t). \tag{2.5}$$

Moreover, it can be showed (see [19]), that

$$\mathbb{E}\left[N_t\right] = W(t) - W \star \mathbb{P}_V(t), \tag{2.6}$$

and

$$\mathbb{P}\left(N_t > 0\right) = 1 - \frac{W \star \mathbb{P}_V(t)}{W(t)}, \tag{2.7}$$

where

$$W \star \mathbb{P}_V(t) := \int_{[0,t]} W(t-s)\ \mathbb{P}_V(ds).$$

Now, let us introduce the mathematical formalism for the mutation process used in this work (this formalism comes from [8]). Since only the mutations occurring on the lineages of living individuals at time t can be observed, it follows from standard properties on Poisson point processes, that the mutation process can be defined directly on the CPP. So, let $\mathcal{P}$ be a Poisson random measure on $[0,t] \times \mathbb{N}$ with intensity measure $\theta\lambda \otimes C$, where C is the counting measure on $\mathbb{N}$, then the mutation random measure $\mathcal{N}$ on the CPP is defined by

$$\mathcal{N}\left(da, di\right) = \mathbb{1}_{H_i > t-a} \mathbb{1}_{i < N_t}\ \mathcal{P}\left(da, di\right),$$

where an atom at (a, i) means that the ith branch experiences a mutation at time $t-a$. We suppose that each individual inherits the type of its parent. This rule yields a partition of the population by types. The distribution of the sizes of the families in the population is called the frequency spectrum and is defined as the sequence $(A(k,t))_{k\geq 1}$ where $A(k,t)$ is the number of types carried by exactly k individuals in the alive population at time t, excluding the family holding the ancestral type of the population (i.e. individuals holding the same type as the root at time 0). This last family is called *clonal*, as the ancestral type.

In the study of the frequency spectrum, an important role is played by the law of the clonal family. We denote by $Z_0(t)$ the size of this family at time t.

To study this family, it is easier to consider the clonal splitting tree constructed from the original splitting tree by cutting every branches beyond mutations. This clonal splitting tree is a standard splitting tree without mutations, where individuals are killed as soon as they die or experience a mutation. The new lifespan law is therefore the minimum between an exponential random variable of parameter θ and V. This distribution is denoted $\mathbb{P}_\theta$. It is straightforward by simple manipulations of Laplace transforms that the Laplace exponent of the corresponding tree is

$$\psi_\theta(x) = x - \int_{(0,\infty]} \left(1 - e^{-rx}\right)\ \mathbb{P}_\theta(dr) = \frac{x\psi(x+\theta)}{x+\theta}.$$

We denote by W_θ the corresponding scale function. This leads to,

$$\mathbb{P}\left(Z_0(t) = k \mid Z_0(t) > 0\right) = \frac{1}{W_\theta(t)} \left(1 - \frac{1}{W_\theta(t)}\right)^{k-1}, \quad k \geq 1.$$

When $\alpha > \theta$ (resp. $\alpha = \theta$, $\alpha < \theta$), this new tree is supercritical (resp. critical, sub-critical) and we talk about *clonal supercritical case* (resp. *critical*, *sub-critical* case).

Moreover, the law of Z_0 conditionally on the event $\{N_t > 0\}$ can be obtained, and is given by

$$\mathbb{P}\left(Z_0(t) = k \mid N_t > 0\right) = \frac{e^{-\theta t} W(t)}{W_\theta(t)^2} \left(1 - \frac{1}{W_\theta(t)}\right)^{k-1}, \quad \forall k \geq 1. \tag{2.8}$$

For the rest of this paper, unless otherwise stated, the notation $\mathbb{P}_t$ refers to $\mathbb{P}\left(. \mid N_t > 0\right)$ whereas $\mathbb{P}_\infty$ refers to the probability measure conditioned on the non-extinction event (which has positive probability in the supercritical case).

Finally, we recall the asymptotic behaviors of the scale functions $W(t)$ and $W_\theta(t)$ which are widely used in the sequel.

Lemma 2.2. *([9, Thm. 3.21]) There exist a positive constant γ such that,*

$$e^{-\alpha t}\psi'(\alpha)W(t) - 1 = \mathcal{O}\left(e^{-\gamma t}\right).$$

In the case that $\theta < \alpha$ (clonal supercritical case),

$$W_\theta(t) \underset{t\to\infty}{\sim} \frac{e^{(\alpha-\theta)t}}{\psi_\theta(\alpha-\theta)}.$$

In the case that $\theta > \alpha$ (clonal sub-critical case),

$$W_\theta(t) = \frac{\theta}{\psi(\theta)} + \mathcal{O}\left(e^{-(\theta-\alpha)t}\right).$$

In the case where $\theta = \alpha$ (clonal critical case),

$$W_\theta(t) \underset{t\to\infty}{\sim} \frac{\theta t}{\psi'(\alpha)}.$$

For a purpose, a more precise description of the asymptotic behavior of W is needed. It is given by the following result.

Lemma 2.3. *[15, Prop. 5.1] There exists a positive non-increasing càdlàg function F such that*

$$W(t) = \frac{e^{\alpha t}}{\psi'(\alpha)} - e^{\alpha t}F(t), \quad t \geq 0,$$

and

$$\lim_{t\to\infty} e^{\alpha t}F(t) \xrightarrow[t\to\infty]{} \mu,$$

with

$$\mu := \begin{cases} \frac{1}{b\mathbb{E}[V]-1} & \text{if } \mathbb{E}[V] < \infty, \\ 0 & \text{otherwise,} \end{cases}$$

where V is some random variable with distribution $\mathbb{P}_V$.

From this lemma and (2.7), one can easily deduce that

$$\mathbb{P}\left(N_t > 0\right) \xrightarrow[t\to\infty]{} \mathbb{P}\left(\text{NonEx}\right) = \frac{\alpha}{b}, \tag{2.9}$$

where NonEx refer to the non-extinction event.

In [8], we show that a CPP stopped at time t with scale function W can be constructed by grafting independent CPP stopped at a fixed time $a \leq t$ on a CPP stopped at time $t-a$ with an explicit scale function different of W whose total population is denoted $N_{t-a}^{(t)}$ (see Figure 2). Moreover, we showed

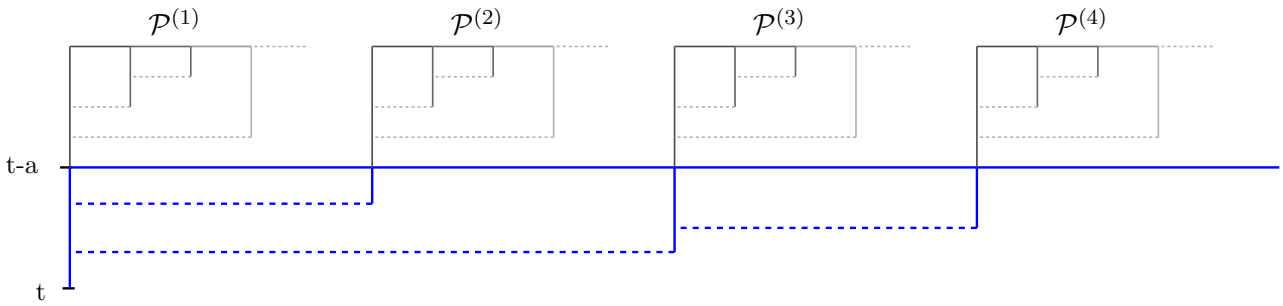

Figure 2: Adjunction of independent CPPs on the blue CPP. $N_{t-a}^{(t)}$ is the number of individual in the blue CPP.

that the frequency spectrum can be expressed as an integral with respect to the random measure $\mathcal{N}$ along the CPP, that is

$$A(k,t) = \int_{[0,t]\times\mathbb{N}} \mathbb{1}_{Z_0^{(u)}(a)=k}\, \mathcal{N}\left(da, du\right), \quad \forall k \in \mathbb{N}^*,$$

where $Z_0^{(u)}$ refers to the clonal family of the uth grafted sub-CPP (see Figure 2). In other words, since the mass of $\mathcal{N}$ is concentrated on mutation points, this boils down to count, for each mutation, if the clonal descent (of the newest type) is represented by k alive individuals at time t. Let us point out that $Z_0^{(u)}(a)$ can be interpreted (without grafting CPPs) as the number of individual at time t which descend from the uth individual (among the individuals alive a time $t-a$) and carrying the same type as this uth individual.

More generally, denoting by $(A^{(u)}(k,a),\ k \in \mathbb{N}^*)$ the frequency spectrum of the uth grafted sub-CPP, we have, for any positive integers $k_1, \cdots, k_l$ with $l \in \mathbb{N}^*$,

$$\prod_{i=1}^{l} A(k_i,t) = \sum_{i=1}^{l} \int_{[0,t]\times\mathbb{N}} \mathbb{1}_{Z_0^{(u)}(a)=k_i} \sum_{u_{1:l-1}=1}^{N_{t-a}^{(t)}} \prod_{\substack{j=1\\ i\neq j}}^{l-1} A^{(u_j)}(k_j,a)\, \mathcal{N}\left(da, du\right), \tag{2.10}$$

where $\sum_{u_{1:l-1}=1}^{N_{t-a}^{(t)}}$ denotes for the multi-sum

$$\sum_{u_1=1}^{N_{t-a}^{(t)}} \cdots \sum_{u_{l-1}=1}^{N_{t-a}^{(t)}}.$$

Moreover, in [8, Thm, 3.1], we show that the expectation of such integral can be computed as if the process $(u,a) \mapsto Z_0^{(u)}(a)$ (or $(u,a) \mapsto A^{(u)}(k,a)$) was independent of the random measure $\mathcal{N}$ (which is clearly not the case). Equation (2.10) is used later to obtain some moments estimates useful to prove our theorems. In particular, this allows to prove that (see [8, Thm, 5.2]) for any positive integer k and l,

$$\mathbb{E}_t\left[A(k,t)\right] = W(t)\int_0^t \frac{\theta e^{-\theta s}}{W_\theta(s)^2}\left(1-\frac{1}{W_\theta(s)}\right)^{k-1} ds, \tag{2.11}$$

and

$$\begin{aligned}
\mathbb{E}_t\left[A(k,t)A(l,t)\right] &= 2W(t)^2\int_0^t \frac{\theta e^{-\theta s}}{W_\theta(s)^2}\left(1-\frac{1}{W_\theta(s)}\right)^{k-1} ds \int_0^t \frac{\theta e^{-\theta s}}{W_\theta(s)^2}\left(1-\frac{1}{W_\theta(s)}\right)^{l-1} ds \\
&- W(t)\int_0^t 2\theta\ \frac{e^{-\theta a}W(a)}{W_\theta(a)^2}\left(1-\frac{1}{W_\theta(a)}\right)^{l-1}\int_0^a \frac{\theta e^{-\theta s}}{W_\theta(s)^2}\left(1-\frac{1}{W_\theta(s)}\right)^{k-1} ds\ da \\
&- W(t)\int_0^t 2\theta\ \frac{e^{-\theta a}W(a)}{W_\theta(a)^2}\left(1-\frac{1}{W_\theta(a)}\right)^{k-1}\int_0^a \frac{\theta e^{-\theta s}}{W_\theta(s)^2}\left(1-\frac{1}{W_\theta(s)}\right)^{l-1} ds\ da \\
&+ W(t)\mathbb{E}\int_0^t \theta W(a)^{-1}\left(\mathbb{E}\left[A(k,a)\mathbb{1}_{Z_0(a)=l}\right] + \mathbb{E}\left[A(l,a)\mathbb{1}_{Z_0(a)=k}\right]\right) da \\
&+ \mathbb{1}_{l=k}W(t)\int_0^t \frac{\theta e^{-\theta a}}{W_\theta(a)^2}\left(1-\frac{1}{W_\theta(a)}\right)^{k-1} da.
\end{aligned} \tag{2.12}$$

These tools also allow, for instance, to prove next two results [16, 11, 8].

Theorem 2.4. *There exists a random variable $\mathcal{E}$, such that*

$$e^{-\alpha t}N_t \xrightarrow[t\to\infty]{} \frac{\mathcal{E}}{\psi'(\alpha)}, \quad a.s.\ and\ in\ L^2.$$

Moreover, under $\mathbb{P}_\infty$, $\mathcal{E}$ is exponentially distributed with parameter 1.

Theorem 2.5. *For any positive integer k,*

$$e^{-\alpha t}A(k,t) \xrightarrow[t\to\infty]{} \frac{c_k\mathcal{E}}{\psi'(\alpha)}, \quad a.s.\ and\ in\ L^2,$$

where $\mathcal{E}$ is the random variable of the Theorem 2.4 and

$$c_k = \int_0^\infty \frac{\theta e^{-\theta a}}{W_\theta(a)}\left(1-\frac{1}{W_\theta(a)}\right)^{k-1} da. \tag{2.13}$$

3 Main results

The almost sure convergences stated in Section 2 suggests studying the second order properties of these convergences to get central limit theorems. Our main result, Theorem 3.1, allows to study the asymptotic error in the second approximation proposed in the introduction of this work. In addition, we prove more standard central limit theorem which are interesting from the theoretical point of view.

Before going further, we recall that the Laplace distribution with mean $\mu \in \mathbb{R}^n$ and covariance matrix K is the probability distribution whose characteristic function is given by

$$\lambda \in \mathbb{R}^n \mapsto \frac{1}{1 + \frac{1}{2}\langle \lambda, K\lambda \rangle - i\langle \mu, \lambda \rangle},$$

where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean scalar product. This law is denoted by $L(\mu, K)$. We also recall that, if G is a Gaussian random vector with mean μ and covariance matrix K and $\mathcal{E}$ is an exponential random variable with parameter 1 independent of G, then $\sqrt{\mathcal{E}}G$ is Laplace $L(\mu, K)$.

3.1 Central limit theorem for the error between $A(k,t)$ and $c_k N_t$

Our first theorem concerns the error between $A(k,t)$ and $c_k N_t$.

Theorem 3.1. *Suppose that $\theta > \alpha$, then*

$$\psi'(\alpha)\left(e^{-\alpha\frac{t}{2}}\left(A(k,t) - c_k N_t\right)\right)_{k\in\mathbb{N}^*} \xrightarrow[t\to\infty]{(d)} \mathcal{L} \text{ w.r.t. } \mathbb{P}_\infty,$$

where $\mathcal{L}$ is $\mathbb{R}^{\mathbb{N}^*}$-valued random variable with distribution $L(0, M)$ where the constants c_k are defined in (2.13) and the covariance matrix M is defined, for any positive integer l and k, by

$$\begin{aligned} M_{l,k} =& 2\psi'(\alpha)\int_0^\infty \frac{\theta e^{-\theta a}}{W_\theta(a)^2}\left(1 - \frac{1}{W_\theta(a)}\right)^{l-1}\left(\mathbb{E}_a\left[A(k,a)\right] - c_k W(a)\right)\, da \\ &+ 2\psi'(\alpha)\int_0^\infty \frac{\theta e^{-\theta a}}{W_\theta(a)^2}\left(1 - \frac{1}{W_\theta(a)}\right)^{k-1}\left(\mathbb{E}_a\left[A(l,a)\right] - c_l W(a)\right) da \\ &- \psi'(\alpha)\int_0^\infty \theta W(a)^{-1}\mathbb{E}_a\left[\left(A(k,a) - c_k N_a\right)\mathbb{1}_{Z_0(a)=l}\right]\, da \\ &- \psi'(\alpha)\int_0^\infty \theta W(a)^{-1}\mathbb{E}_a\left[\left(A(l,a) - c_l N_a\right)\mathbb{1}_{Z_0(a)=l}\right]\, da \\ &+ \mathbb{1}_{l=k}\int_0^\infty \frac{\theta e^{-\theta s}}{W_\theta(s)^2}\left(1 - \frac{1}{W_\theta(s)}\right)^{k-1} ds. \end{aligned} \tag{3.1}$$

The above theorem can be enhanced to obtain more information on the limiting distribution.

Corollary 3.2. *Under the hypothesis of Theorem 3.1. Let G be a centered $\mathbb{R}^{\mathbb{N}^*}$-valued Gaussian random variable, with covariance matrix M independent of the random variable $\mathcal{E}$ of Theorem 2.4, then*

$$\psi'(\alpha)\left(e^{-\alpha\frac{t}{2}}\left(A(k,t) - c_k N_t\right)\right)_{k\in\mathbb{N}^*} \xrightarrow[t\to\infty]{(d)} \sqrt{\mathcal{E}}G \text{ w.r.t. } \mathbb{P}_\infty.$$

In addition, this convergence holds jointly with the weak convergence of $\psi'(\alpha)e^{-\alpha t}N_t$ to $\mathcal{E}$.

Remark 3.3.

- *Let us point out that the above convergences has to be understood as convergence of processes for the product topology of $\mathbb{R}^{\mathbb{N}}$. In particular, it is well-known that for this topology (see [7]) the convergence of finite dimensional distributions is enough to ensure the convergence as processes.*

- *Note that an explicit formula for* $\mathbb{E}_t[A(k,t)]$ *is given by* (2.11). *Proposition 4.5 of [8] gives explicit formulas for* $\mathbb{E}_t\left[A(k,t)\mathbb{1}_{Z_0(t)=l}\right]$. *And a formula for* $\mathbb{E}_t\left[N_t\mathbb{1}_{Z_0(t)=k}\right]$ *can be found in Proposition 4.1 of [9].*

The proof of these results can be found in Section 7.

3.2 Central limit theorem for the convergence of Theorem 2.5

Our second result is a central limit theorem related to the convergence of Theorem 2.5.

Theorem 3.4. *Suppose that* $\theta > \alpha$. *Then, we have, under* $\mathbb{P}_\infty$,

$$\left(e^{-\alpha\frac{t}{2}}\left(\psi'(\alpha)A(k,t) - e^{\alpha t}c_k\mathcal{E}\right)\right)_{k\in\mathbb{N}^*} \xrightarrow[t\to\infty]{(d)} \mathcal{L},$$

where $\mathcal{L}$ *is* $\mathbb{R}^{\mathbb{N}^*}$*-valued random variable with distribution* $L(0,H)$ *where the constants* c_k *are defined in* (2.13) *and the covariance matrix* H *is defined, for any positive integer* k *and* l, *by*

$$H_{k,l} = M_{k,l} - 4\psi'(\alpha)\gamma c_k c_l - 2\mu\psi'(\alpha), \quad i,j \geq 1,$$

where μ *is defined in Lemma 2.3, and*

$$\gamma := \begin{cases} 1 & \text{if } \mathbb{E}[V] < \infty, \\ -\alpha^{-1}\mathbb{P}_V(\{\infty\}) & \text{otherwise,} \end{cases}$$

where V *is some random variable with distribution* $\mathbb{P}_V$.

We have a similar extension for this theorem as for the previous one.

Corollary 3.5. *Under the hypothesis of Theorem 3.1. Let* G *be a centered* $\mathbb{R}^{\mathbb{N}^*}$*-valued Gaussian random variable, with covariance matrix* H *independent of the random variable* $\mathcal{E}$ *of Theorem 2.4, then*

$$\psi'(\alpha)\left(e^{-\alpha\frac{t}{2}}\left(A(k,t) - c_k N_t\right)\right)_{k\in\mathbb{N}^*} \xrightarrow[t\to\infty]{(d)} \sqrt{\mathcal{E}}G \text{ w.r.t. } \mathbb{P}_\infty.$$

In addition, this convergence holds jointly with the weak convergence of $\psi'(\alpha)e^{-\alpha t}N_t$ *to* $\mathcal{E}$.

The proof of Theorem 3.4 can be found in Section 6.

Remark 3.6. *The proofs of the two theorems are very similar. Since the hardest case is the one of Theorem 3.4, we only detail this case in the sequel. The proof of the corollaries are only detailed in the case of Theorem 3.1 in Section 7. However, Remark 7.2 highlights the only difference between the two cases.*

4 Strategy of proof

The proof of this theorem is based on the proof of the central limit theorem for the process $(N_t,\ t \in \mathbb{R}_+)$ given in [15]. The structure of the proof follows the same lines and is detailed in Section 4 of [15]. In a sake of completeness, we recall the ideas. For this reason, the results which are straightforward rewording of the proofs given in [15] are not detailed. However, we think it is necessary to recall some aspects of [15], in particular from [15, Section 4].

The idea of the proof is based on a decomposition of the tree as the one of Figure 3. More precisely, if we fix two times u and t with $u < t$, each individual composing the population at time u induces a subtree of the whole tree made of its residual lifetime and its descent. To formalize this, let us recall that, for any fixed time u, there is a natural order (for instance given by an exploration process [16], see also Figure 3) of the individuals alive at this time. Moreover, we denote, for $1 \leq i \leq N_u$, O_i the residual lifetime of the ith individual alive at time u. The tree of the ith individual is denoted $\mathbb{T}(O_i)$ where O_i refers to the residual lifetime of individual i. Indeed, since the descent of each of these individuals are made of independent random quantities (by construction), it follows that the family $(O_i)_{1\leq i\leq N_u}$ is the only source of dependencies between the family $(\mathbb{T}(O_i))_{1\leq i\leq N_u}$. Roughly speaking, in view of [15], one would like to decompose the error

$$e^{-\frac{\alpha}{2}t}\left(A(k,t) - c_k N_t\right),$$

as the sum of the errors in each subtree,

$$e^{-\frac{\alpha}{2}(t-u)} \sum_{i=1}^{N_u} e^{-\frac{\alpha}{2}u}\left(A(k, t-u, O_i) - N_t(O_i)\right), \tag{4.1}$$

where $A(k, t-u, O_i)$ denotes the number of families of size k in tree $\mathbb{T}(O_i)$ at time $t-u$ (seen as a standalone tree, i.e time 0 for $\mathbb{T}(O_i)$ corresponds to time u in the whole tree). The notation $N_{t-u}(O_i)$ refers to the number of individual in tree $\mathbb{T}(O_i)$ at time $t-u$. Doing so, the error can be re-expressed as a geometric sum (under $\mathbb{P}_u$) of errors with controlled moments, leading to a Laplace distribution when N_u gets big.

Unfortunately, one cannot decompose the error this way. This is due to the fact that subtrees may share individuals with a common type. Hence, among the individuals of a family of size k (which contribute to $A(k,t)$), some might belong to subtree $\mathbb{T}(O_i)$ whereas some others might be in $\mathbb{T}(O_j)$ (for $i \neq j$). However, the decomposition of Equation (4.1) holds true on the event

$$\Gamma_{u,t} = \{\text{"there is no family in the population at time t which is older than } u\text{"}\},$$

which occurs with high probability for $u << t$ and $\theta > \alpha$ (since in the clonal subcritical case, the families tend to extinct).

Another important aspect is to obtain estimates on the moments of the error. To get such estimates, the following estimates which comes from [15, Corollary 6.3] are useful.

$$\frac{1}{\mathbb{P}\left(N_t > 0\right)} = \frac{b}{\alpha} - \frac{b\mu\psi'(\alpha)}{\alpha} e^{-\alpha t} + o(e^{-\alpha t}). \tag{4.2}$$

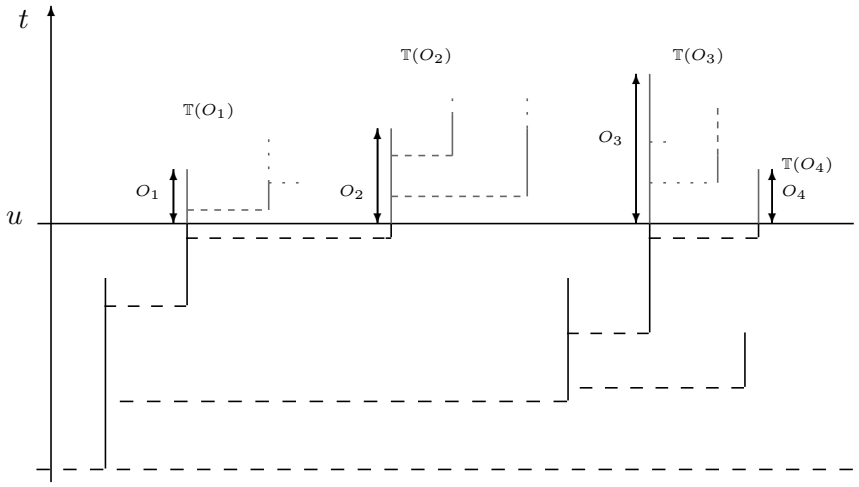

Figure 3: Residual lifetimes with subtrees associated to living individuals at time u.

We also have

$$\mathbb{E}_t\left[N_t\mathcal{E}\right] = \frac{2e^{\alpha t}}{\psi'(\alpha)} - \frac{1}{\psi'(\alpha)} - 3\mu + o(1). \tag{4.3}$$

Finally, let us recall that the law of the vector $(O_2, \ldots, O_{N_u})$ is given by the following lemma which also comes from [15].

Lemma 4.1. *Let u in $\mathbb{R}_+$, we denote by O_i for i an integer between 1 and N_u the residual lifetime of the ith individuals alive at time u. Then under $\mathbb{P}_u$, the family $(O_i,\ i \in \{1, \ldots, N_u\})$ form a family of independent random variables, independent of N_u, and, expect O_1, having the same distribution, given by, for $2 \leq i \leq N_t$,*

$$\mathbb{P}_u(O_i \in dx) = \int_{\mathbb{R}_+} \frac{W(u-y)}{W(u)-1} b\mathbb{P}\left(V - y \in dx\right)\ dy.$$

Moreover, it follows that the family $(N_s(O_i), s \in \mathbb{R}_+)_{1 \leq i \leq N_u}$ is an independent family of process, i.i.d. for $i \geq 2$, and independent of N_u.

To end this reminder, let us recall the decomposition of the limiting random variable $\mathcal{E}$ (defined for instance in Theorem 2.4) at a fixed time u. First, from the construction of the splitting tree (see also [15]), we have, almost surely,

$$N_t(O_i) = \int_{[0,t]} N^{\xi_u}_{t-u} \mathbb{1}_{O_i \geq u}\ \xi(du) + \mathbb{1}_{O_i \geq t}, \quad \forall t \geq 0,$$

where

1. $(N_t,\ t \in \mathbb{R}_+)_{i \geq 1}$ is an i.i.d. family of random processes with the same law as $(N_t,\ t \in \mathbb{R}_+)$,

2. ξ is some Poisson random measure with constant rate b,

3. the objects of item 1 and 2 are independent, and independent of O_i,

4. ξ_a denotes $\xi([0,a])$.

The idea behind these equations is that the tree $\mathbb{T}(O_i)$ is constructed by grafting on a branch with length O_i a random number of trees with the same distribution as the whole splitting tree $\mathbb{T}$. Hence, the number of individual in this tree at time t, is the number individual in each of the grafted trees (taken at a time corresponding to t minus the grafting time), plus 1 if the first branch has length greater than t. In addition, in [15] we showed that $\psi'(\alpha)e^{-\alpha t}N_t(O_i)$ converges in L^2 to a random variable $\mathcal{E}(O_i)$, and that we have the following lemma.

Lemma 4.2. *[15, Lemma 6.8] For any time $u > 0$, we have the following decomposition of $\mathcal{E}$ (see Theorem 2.4 for the definition),*

$$\mathcal{E} = e^{-\alpha u} \sum_{i=1}^{N_u} \mathcal{E}_i\left(O_i\right), \quad a.s.$$

Moreover, under $\mathbb{P}_u$, the random variables $\left(\mathcal{E}_i\left(O_i\right)\right)_{i\geq 1}$ are independent, independent of N_u, and identically distributed for $i \geq 2$.

Before proving Theorem 3.4, we need an important number of estimates on the moments of the error. This is the point of the next section.

5 Preliminary moment estimates

As explained in Section 4, one needs to have estimates on the error in the sum of Equation (4.1). There are two steps to obtain these estimates:

- Get the estimates when the lifetime $V_\emptyset$ of the ancestral individual is distributed according to $\mathbb{P}_V$.
- Deduce the estimates when $V_\emptyset$ follows an arbitrary distribution, and finally take law of O_i for $V_\emptyset$.

In both case, the lifetime distribution of the other individuals is still assumed to be $\mathbb{P}_V$. So, according to plan, we begin with the standard splitting tree case.

5.1 Case $V_\emptyset \overset{\mathcal{L}}{=} \mathbb{P}_V$

One of the main difficulties is to get estimates on moments like

$$\mathbb{E}\left[\left(\psi'(\alpha)A(k,t) - e^{\alpha t}c_k\mathcal{E}\right)^n\right], \text{ for } n = 2 \text{ or } 3.$$

or

$$\mathbb{E}\left[\left(A(k,t) - c_k N_t\right)^n\right], \text{ for } n = 2 \text{ or } 3.$$

We begin with the following lemma.

Lemma 5.1. *Let k and l be two positive integers, then*

$$e^{\alpha t}\mathbb{E}\left[\left(e^{-\alpha t}\psi'(\alpha)A(k,t)-\psi'(\alpha)c_k e^{-\alpha t}N_t\right)\left(e^{-\alpha t}\psi'(\alpha)A(l,t)-c_l e^{-\alpha t}\psi'(\alpha)N_t\right)\right]\xrightarrow[t\to\infty]{} M_{k,l}$$

where M is defined in Equation (3.1).

Proof. We give the details for $l\neq k$, the case $l=k$ is a direct adaptation of what follows with the indicator function $\mathbb{1}_{l=k}$ of (2.12) in mind. We recall that using the calculus made in Remark 5.6 of [8], we have

$$\mathbb{E}_t\left[A(k,t)N_t\right]=2W(t)^2c_k(t)-2W(t)\int_{[0,t]}\theta\mathbb{P}_a\left(Z_0(a)=k\right)da \\ +W(t)\int_{[0,t]}\theta W(a)^{-1}\mathbb{E}_a\left[N_a\mathbb{1}_{Z_0(a)=k}\right]da,$$

with

$$c_k(t):=\int_0^t\frac{\theta e^{-\theta a}}{W_\theta(a)^2}\left(1-\frac{1}{W_\theta(a)}\right)^{k-1}da,\quad\forall k\geq 1,\ t\in\mathbb{R}_+.$$

Moreover, from (2.12) and Lemma 2.3, we have

$$\psi'(\alpha)^2\mathbb{E}_t\left[A(k,t)A(l,t)\right]=2W(t)^2c_k(t)c_l(t)+RW(t)+o(e^{-\alpha t}),$$

with

$$\begin{aligned}R:=&-\psi'(\alpha)\int_0^\infty 2\theta W(a)^{-1}\mathbb{P}_a\left(Z_0(a)=k\right)\mathbb{E}_a\left[A(l,a)\right]da\\&+\psi'(\alpha)\int_0^\infty 2\theta W(a)^{-1}\mathbb{P}_a\left(Z_0(a)=l\right)\mathbb{E}_a\left[A(k,a)\right]da\\&+\psi'(\alpha)\int_0^\infty\theta W(a)^{-1}\left(\mathbb{E}_t\left[A(k,a)\mathbb{1}_{Z_0(a)=l}\right]+\mathbb{E}_t\left[A(l,a)\mathbb{1}_{Z_0(a)=k}\right]\right)da.\end{aligned}$$

These identities allow us to obtain, using also that N_t is geometrically distributed under $\mathbb{P}_t$,

$$\begin{aligned}&\mathbb{E}_t\left[\left(A(k,t)-c_kN_t\right)\left(A(l,t)-c_lN_t\right)\right]=2W(t)^2c_k(t)c_l(t)+e^{-\alpha t}R+o(e^{-\alpha t})\\&-2c_lc_k(t)W(t)^2+2c_lW(t)\int_{[0,t]}\theta\mathbb{P}_a\left(Z_0(a)=k\right)da-c_lW(t)\int_{[0,t]}\theta W(a)^{-1}\mathbb{E}_a\left[N_a\mathbb{1}_{Z_0(a)=k}\right]da\\&-2c_kc_l(t)W(t)^2+2c_lW(t)\int_{[0,t]}\theta\mathbb{P}_a\left(Z_0(a)=l\right)da-c_kW(t)\int_{[0,t]}\theta W(a)^{-1}\mathbb{E}_a\left[N_a\mathbb{1}_{Z_0(a)=l}\right]da\\&+c_kc_lW(t)^2\left(2-\frac{1}{W(t)}\right)\\&=2W(t)^2\left(c_k(t)-c_l\right)\left(c_l(t)-c_k\right)+e^{-\alpha t}\frac{R}{\psi'(\alpha)}+o(e^{-\alpha t}),\\&+2c_lW(t)\int_{[0,t]}\theta\mathbb{P}_a\left(Z_0(a)=k\right)da-c_lW(t)\int_{[0,t]}\theta W(a)^{-1}\mathbb{E}_a\left[N_a\mathbb{1}_{Z_0(a)=k}\right]da\\&+2c_lW(t)\int_{[0,t]}\theta\mathbb{P}_a\left(Z_0(a)=l\right)da-c_kW(t)\int_{[0,t]}\theta W(a)^{-1}\mathbb{E}_a\left[N_a\mathbb{1}_{Z_0(a)=l}\right]da\\&-c_kc_lW(t).\end{aligned}$$

Taking the limit as t goes to infinity leads to

$$\begin{aligned} M_{k,l} := \lim_{t\to\infty} \psi'(\alpha)^2 e^{-\alpha t}\mathbb{E}_t\left[\left(A(k,t)-c_kN_t\right)\left(A(l,t)-c_lN_t\right)\right] = R \\ + 2\psi'(\alpha)c_l\int_{[0,\infty)}\theta\mathbb{P}_a\left(Z_0(a)=k\right)da - \psi'(\alpha)c_l\int_{[0,\infty)}\theta W(a)^{-1}\mathbb{E}_a\left[N_a\mathbb{1}_{Z_0(a)=k}\right]da \\ + 2\psi'(\alpha)c_l\int_{[0,\infty)}\theta\mathbb{P}_a\left(Z_0(a)=l\right)da - \psi'(\alpha)c_k\int_{[0,\infty)}\theta W(a)^{-1}\mathbb{E}_a\left[N_a\mathbb{1}_{Z_0(a)=l}\right]da \\ - \psi'(\alpha)c_kc_l. \end{aligned}$$

□

Our next goal is to get the same type of results for the error in the CLT involving $\mathcal{E}$ of Theorem 3.4. For this, we need the following lemma.

Lemma 5.2. *Consider $\mathcal{E}(O_2)$ as defined in Section 4. Then, we have*

$$\mathbb{E}_t\left[\mathcal{E}(O_2)\right] = \int_{\mathbb{R}_+}\alpha e^{-\alpha u}\mathbb{P}_t(O_2\geq u)\ du = \psi'(\alpha) + \psi'(\alpha)\gamma e^{-\alpha t} + o(e^{-\alpha t})$$

with

$$\gamma := \begin{cases} 1 & \text{if } \mathbb{E}[V]<\infty, \\ -\alpha^{-1}\mathbb{P}_V(\{\infty\}) & \text{otherwise.} \end{cases}$$

In particular, if V is not integrable and $\mathbb{P}(V=\infty)=0$, then $\gamma=0$.

Proof. As recalled in Section 4, we have that there exists

- an i.i.d. family of random processes $(N_t^i,\ t\in\mathbb{R}_+)_{i\geq 1}$ (corresponding to the population counting processes induced by each child of individual 2) with the same distribution as $(N_t,\ t\in\mathbb{R}_+)$,
- a Poisson random measure ξ with constant rate b independent of the above family,

such that all these objects are independent of O_2 (under $\mathbb{P}_t$) and

$$N_t(O_2) = \int_{[0,t]} N_{t-u}^{\xi_u}\mathbb{1}_{O_2\geq u}\ \xi(du) + \mathbb{1}_{O_2\geq t}, \quad \forall t\geq 0,$$

where $\xi_u := \xi([0,u])$. In addition, as shown in Lemma 6.6 of [15], in L^1 ,

$$\psi'(\alpha)e^{-\alpha t}N_t(O_2) \xrightarrow[t\to\infty]{} \mathcal{E}(O_2) := \int_{\mathbb{R}_+} e^{-\alpha u}\mathcal{E}_{\xi_u}\mathbb{1}_{u\leq O_i}\ \xi(du),$$

with

$$\psi'(\alpha)e^{-\alpha t}N_t^i \xrightarrow[t\to\infty]{} \mathcal{E}_i, \quad \text{almost surely and in } L^1.$$

Hence, it follows from Lebesgue's theorem that

$$\psi'(\alpha)e^{-\alpha t}\mathbb{E}_t\left[N_t(O_2)\right] = \psi'(\alpha)\int_{[0,t]} e^{-\alpha u}e^{-\alpha(t-u)}\mathbb{E}\left[N_{t-u}\right]\mathbb{P}_t(O_2\geq u)b\ du + \psi'(\alpha)e^{-\alpha t}\mathbb{P}_t(O_2\geq t),$$

converges, as t goes to infinity, to

$$\int_{\mathbb{R}_+} \alpha e^{-\alpha u} \mathbb{P}_t(O_2 \geq u)\ du,$$

which is equal to $\mathbb{E}_t\left[\mathcal{E}(O_2)\right]$. Now, according to Lemma 4.1, we have that

$$\int_{\mathbb{R}_+} \alpha e^{-\alpha u} \mathbb{P}_t(O_2 \geq u)\ du = \int_{\mathbb{R}_+} \alpha e^{-\alpha u} \int_0^t b \frac{W(t-y)}{W(t)-1} \mathbb{P}(V \geq u+y)\ dy\ du.$$

which gives, thanks to Lemma 2.3,

$$\begin{aligned}
&\int_{\mathbb{R}_+} \alpha e^{-\alpha u} \mathbb{P}_t(O_2 \geq u)\ du \\
&\qquad = b(W(t)-1)^{-1} \int_{\mathbb{R}_+} \alpha e^{-\alpha u} \int_{[0,t]} \left(\frac{e^{\alpha(t-y)}}{\psi'(\alpha)} - e^{\alpha(t-y)} F(t-y) \right) \mathbb{P}(V \geq u+y)\ dy\ du \\
&\qquad = \frac{b e^{\alpha t}}{\psi'(\alpha)(W(t)-1)} \int_{\mathbb{R}_+ \times [0,t]} \alpha e^{-\alpha(u+y)} \mathbb{P}(V \geq u+y)\ dy\ du \qquad (5.1) \\
&\qquad + \frac{b\mu}{W(t)-1} \int_{\mathbb{R}_+ \times [0,t]} \alpha e^{-\alpha u} \mathbb{P}(V \geq u+y)\ dy\ du \\
&\qquad + \frac{b}{W(t)-1} \int_{\mathbb{R}_+ \times [0,t]} \left(\mu - e^{\alpha(t-y)} F(t-y) \right) \alpha e^{-\alpha u} \mathbb{P}(V \geq u+y)\ dy\ du,
\end{aligned}$$

where we recall that μ is defined in Lemma 2.3. In particular, we have

$$\begin{aligned}
\int_{\mathbb{R}_+^2} \alpha e^{-\alpha(u+v)} \mathbb{P}(V \geq u+v)\ du\ dv &= \int_{\mathbb{R}_+} \alpha u e^{-\alpha u} \mathbb{P}(V \geq u)\ du \\
&= \int_{\mathbb{R}_+} \left(\frac{1-e^{-\alpha x}}{\alpha} - x e^{-\alpha x} \right)\ \mathbb{P}_V(dx),
\end{aligned}$$

which gives using (2.3) and (2.1),

$$\int_{\mathbb{R}_+^2} \alpha e^{-\alpha(u+v)} \mathbb{P}(V \geq u+v)\ du\ dv = \frac{\psi'(\alpha)}{b}.$$

In addition,

$$\begin{aligned}
\int_{\mathbb{R}_+ \times (t,\infty)} \alpha e^{-\alpha(u+v)} \mathbb{P}(V \geq u+v)\ du\ dv &= \int_{(t,\infty)} \alpha e^{-\alpha u} \mathbb{P}(V \geq u)(u-t)\ du \\
&= e^{-\alpha t} \int_0^\infty \alpha u e^{-\alpha u} \mathbb{P}(V > t+u)\ du.
\end{aligned}$$

Thanks to Lebesgue's theorem, we hence have

$$\int_0^\infty \alpha u e^{-\alpha u} \mathbb{P}(V > t+u)\ du \xrightarrow[t\to\infty]{} \alpha^{-1} \mathbb{P}_V(\{\infty\}),$$

which is eventually 0 if $\mathbb{P}_V$ does not have mass at infinity. As usual, let V be a random variable with distribution $\mathbb{P}_V$. Then, if $\mathbb{E}[V] < \infty$, similar computations give

$$\int_{\mathbb{R}_+\times\mathbb{R}_+} \alpha e^{-\alpha u}\mathbb{P}(V \geq u+y)\ dy\ du = \mathbb{E}[V] - \frac{1}{b} = \frac{1}{\mu b}.$$

Finally, plugging the above computations in (5.1) gives

$$\int_{\mathbb{R}_+} \alpha e^{-\alpha u}\mathbb{P}_t(O_2 \geq u)\ du = \begin{cases} \psi'(\alpha) + \psi'(\alpha)e^{-\alpha t} + o(e^{-\alpha t}) \text{ if } \mathbb{E}[V] < \infty, \\ \psi'(\alpha) - \psi'(\alpha)e^{-\alpha t}\alpha^{-1}\mathbb{P}_V(\{\infty\}) + o(e^{-\alpha t}) \text{ otherwise.} \end{cases}$$

□

Using the preceding lemma, we can now get the quadratic error in the convergence of the frequency spectrum.

Lemma 5.3 (Quadratic error for the convergence of $A(k,t)$.)**.** *Let k and l two positive integers. Then under the hypothesis of Theorem 3.4, we have*

$$e^{-\alpha t}\mathbb{E}_t\left[\left(\psi'(\alpha)A(k,t) - e^{\alpha t}\mathcal{E}c_k\right)\left(\psi'(\alpha)A(l,t) - e^{\alpha t}\mathcal{E}c_l\right)\right] \xrightarrow[t\to\infty]{} M_{k,l} - 4\psi'(\alpha)\gamma c_k c_l - 2\mu\psi'(\alpha),$$

where the sequence $(c_k)_{k\geq 1}$ is defined by (2.13)*, γ is defined in Lemma 5.2 and μ is defined in Lemma 2.3.*

Proof. The proof of this lemma is based on the decomposition of $\mathcal{E}$ as

$$\mathcal{E} = e^{-\alpha t}\sum_{i=1}^{N_t}\mathcal{E}(O_i).$$

According to Lemma 4.1, we know that the family $(\mathcal{E}(O_i))_{1\leq i\leq N_t}$ is (under $\mathbb{P}_t$) a family of independent random variable (which is i.i.d. for $i \geq 2$), independent of N_t. Hence,

$$\mathbb{E}_t\left[A(k,t)\mathcal{E}\right] = e^{-\alpha t}\mathbb{E}_t\left[A(k,t)(N_t-1)\mathbb{E}_t\left[\mathcal{E}(O_2)\right]\right] + e^{-\alpha t}\mathbb{E}_t\left[A(k,t)\right]\mathbb{E}_t\left[\mathcal{E}(O_1)\right].$$

First of all, we have that the r.h.s. of the above equality is bounded si $e^{-\alpha t}\mathbb{E}_t[A(k,t)]$ converges as t goes to infinity. From this, we get

$$\begin{aligned}
&\mathbb{E}_t\left[\left(\psi'(\alpha)e^{-\alpha t}A(k,t) - \mathcal{E}c_k\right)\left(\psi'(\alpha)e^{-\alpha t}A(l,t) - \mathcal{E}c_l\right)\right] \\
=&\psi'(\alpha)^2 e^{-2\alpha t}\mathbb{E}_t\left[\left(A(k,t) - N_t c_k\right)\left(A(l,t) - N_t c_l\right)\right] \\
&+ \psi'(\alpha)e^{-2\alpha t}c_l\mathbb{E}_t\left[A(k,t)N_t\right]\left(\psi'(\alpha) - \mathbb{E}_t\left[\mathcal{E}(O_2)\right]\right) \\
&+ \psi'(\alpha)e^{-2\alpha t}c_k\mathbb{E}_t\left[A(l,t)N_t\right]\left(\psi'(\alpha) - \mathbb{E}_t\left[\mathcal{E}(O_2)\right]\right) \\
&+ \psi'(\alpha)^2 e^{-2\alpha t}\mathbb{E}_t\left[N_t^2\right] - \mathbb{E}_t\left[\mathcal{E}^2\right] \\
&+ o(e^{-\alpha t})
\end{aligned}$$

Now, according to Lemma 7.1 and Lemma 5.2, we have

$$\psi'(\alpha)^2 e^{-2\alpha t}c_l\mathbb{E}_t\left[A(k,t)N_t\right] \times e^{\alpha t}\left(1 - \mathbb{E}_t\left[\mathcal{E}(O_2)\right]\right) \xrightarrow[t\to\infty]{} -2c_l c_k\gamma,$$

where γ is defined in Lemma 5.2. Now, using (4.2), we get

$$2 - \mathbb{E}_t[\mathcal{E}^2] = 2\mu\psi'(\alpha)e^{-\alpha t} + o(e^{-\alpha t}),$$

which leads, using Lemma 2.3, to

$$e^{\alpha t}\left(\psi'(\alpha)^2 e^{-2\alpha t}\mathbb{E}_t\left[N_t^2\right] - \mathbb{E}_t\left[\mathcal{E}^2\right]\right) \xrightarrow[t\to\infty]{} -2\mu\psi'(\alpha).$$

Finally, using Lemma 5.1, we get

$$\mathbb{E}_t\left[\left(\psi'(\alpha)e^{-\alpha t}A(k,t) - \mathcal{E}c_k\right)\left(\psi'(\alpha)e^{-\alpha t}A(l,t) - \mathcal{E}c_l\right)\right] \xrightarrow[t\to\infty]{} M_{k,l} - 4\gamma c_k c_l - 2\mu\psi'(\alpha).$$

□

Lemma 5.4 (Boundedness of the third moment)**.** *Let k_1, k_2, k_3 three positive integers, then*

$$\mathbb{E}\left[\prod_{i=1}^{3}\left|e^{-\frac{\alpha}{2}t}\left(\psi'(\alpha)A(k_i,t) - e^{\alpha t}\mathcal{E}c_{k_i}\right)\right|\right] = \mathcal{O}(1).$$

Proof. We have,

$$\mathbb{E}\left[\left|\prod_{i=1}^{3}\frac{\left(\psi'(\alpha)A(k_i,t) - e^{\alpha t}\mathcal{E}c_{k_i}\right)}{e^{\frac{\alpha}{2}t}}\right|\right] \leq \prod_{i=1}^{3}\left(\mathbb{E}\left[\left|\frac{\left(\psi'(\alpha)A(k_i,t) - e^{\alpha t}\mathcal{E}c_{k_i}\right)}{e^{\frac{\alpha}{2}t}}\right|^3\right]\right)^{\frac{1}{3}}.$$

Hence, we only have to prove the Lemma for $k_1 = k_2 = k_3 = k$. Hence,

$$\mathbb{E}\left[\left|\frac{\left(\psi'(\alpha)A(k,t) - e^{\alpha t}\mathcal{E}c_k\right)}{e^{\frac{\alpha}{2}t}}\right|^3\right] \leq 8\mathbb{E}\left[\left|\frac{\psi'(\alpha)A(k,t) - c_k N_t}{e^{\frac{\alpha}{2}t}}\right|^3\right] + 8c_k^3\mathbb{E}\left[\left|\frac{N_t - e^{\alpha t}\mathcal{E}}{e^{\frac{\alpha}{2}t}}\right|^3\right].$$

The last term have been treated in the proof of [15, Lemma 6.4], and the boundedness of

$$\mathbb{E}\left[\left|\frac{\psi'(\alpha)A(k,t) - c_k N_t}{e^{\frac{\alpha}{2}t}}\right|^3\right],$$

follows from the following Lemma 5.5 and Hölder's inequality.

□

Lemma 5.5. *For all $k \geq 1$,*

$$\mathbb{E}\left[\left(\frac{A(k,t) - c_k N_t}{e^{\frac{\alpha}{2}t}}\right)^4\right],$$

is bounded.

Due to technicality, the proof of this lemma is postponed to the end in appendix.

5.2 Arbitrary initial distribution case

The following lemmas are the counter part of Lemmas 6.5, 6.6, and 6.7 of [15]. They play the same role in the proof of Theorem 3.4 as in the proof of the central limit theorem given in [15]. In the sequel, we denote by $(A(k,t,\Xi))_{k\geq 1}$, the frequency spectrum of the splitting tree where the lifetime of the ancestral individual is Ξ but where the other individuals have lifetimes distributed according to $\mathbb{P}_V$. This is done in the same spirit as for $N_t(\Xi)$ in [15], which denotes the number of individuals at time t in a splitting tree where the first individual has lifetime Ξ. So, from the construction of the splitting tree, it is easily seen that there exists an i.i.d. family of processes $(N_t^i,\ t\in\mathbb{R}_+)_{i\geq 1}$, and an independent Poisson point measure ξ on $\mathbb{R}$ with intensity b such that

$$N_t(\Xi) = \int_{[0,t])} N_{t-u}^{(i)} \mathbb{1}_{\Xi\geq u}\ \xi(du) + \mathbb{1}_{\Xi\geq t}, \quad \forall t\in\mathbb{R}_+.$$

Now defining

$$\mathcal{E}_i := \lim_{t\to\infty} \psi'(\alpha) e^{-\alpha t} N_t^i, \forall i\geq 1, \quad \text{almost surely,}$$

we can set

$$\mathcal{E}(\Xi) := \int_{[0,\infty)} \mathcal{E}_{(\xi_u)} e^{-\alpha u} \mathbb{1}_{\Xi>u}\ \xi(du).$$

With this in hands, we can study the asymptotic behavior of the moments of

$$\psi'(\alpha)A(k,t,\Xi) - e^{\alpha t}\mathcal{E}(\Xi)c_k.$$

This first lemma gives the asymptotic of the quadratic error.

Lemma 5.6 (L^2 convergence in the general case)**.** *Consider the general frequency spectrum*

$$(A(k,t,\Xi))_{k\geq 1},$$

then, for all k, $\psi'(\alpha)e^{-\alpha t}A(k,t,\Xi)$ converge to $\mathcal{E}(\Xi)$ (see 5.2) in L^2 as t goes to infinity. Moreover,

$$e^{-\alpha t}\mathbb{E}\left[\left(\psi'(\alpha)A(k,t,\Xi) - e^{\alpha t}\mathcal{E}(\Xi)c_k\right)\left(\psi'(\alpha)A(l,t,\Xi) - e^{\alpha t}\mathcal{E}(\Xi)c_l\right)\right]$$

converges, as t goes to infinity, to

$$\frac{\alpha}{b} H_{k,l} \int_{\mathbb{R}_+} e^{-\alpha u}\mathbb{P}(\Xi > u)\, b du,$$

where the convergence is uniform w.r.t. Ξ, and $H_{k,l}$ is defined in Theorem 3.4.

In the case where Ξ is distributed as O_2 for $u=\beta t$ and $0<\beta<\frac{1}{2}$, we get

$$e^{-\alpha t}\mathbb{E}\left[\left(\psi'(\alpha)A(k,t,O_2) - e^{\alpha t}\mathcal{E}(O_2)c_k\right)\left(\psi'(\alpha)A(l,t,O_2) - e^{\alpha t}\mathcal{E}(O_2)c_l\right)\right] \xrightarrow[t\to\infty]{} \psi'(\alpha)K_{k,l}.$$

This next two lemma give bounds on the first and third moment of the error.

Lemma 5.7 (First moment)**.** *The first moments are asymptotically bounded, that is, for all $k\geq 1$,*

$$\mathbb{E}\left(\psi'(\alpha)A(k,t,\Xi) - e^{\alpha t}c_k\mathcal{E}(\Xi)\right) \leq \mathcal{O}(1),$$

uniformly with respect to Ξ.

Lemma 5.8 (Boundedness in the general case.). *Let k_1, k_2, k_3 three positive integers, then*

$$\mathbb{E}\left[\left|\prod_{i=1}^{3}\frac{\left(\psi'(\alpha)A(k_i,t,\Xi)-e^{\alpha t}\mathcal{E}(\Xi)c_{k_i}\right)}{e^{\frac{\alpha}{2}t}}\right|\right]=\mathcal{O}(1),$$

uniformly with respect to Ξ.

We do not detail the proofs of these results since they are direct adaptations of the proofs of Lemmas 6.5, 6.6, and 6.7 of [15].

6 Proof of Theorem 3.4

The following result is based on the fact that, in the clonal sub-critical case, the lifetime of a family is expected to be small. It follows that one can expect that all the family of size k live in different subtrees as soon as $t >> u$. This is the point of the following lemma.

Lemma 6.1. *Suppose that $\alpha < \theta$. If we denote by $\Gamma_{u,t}$ the event,*

$$\Gamma_{u,t}=\{\text{"there is no family in the population at time } t \text{ which is older than } u\text{"}\},$$

then, for all β in $(0, 1-\frac{\alpha}{\theta})$, we have

$$\mathbb{P}_{\beta t}\left(\Gamma_{\beta t,t}\right)\xrightarrow[t\to\infty]{}1.$$

Proof. The proof of this lemma, as the calculation of the moments of $A(k,t)$ relies on the representation of the genealogy of the living population at time t as a coalescent point process [8, Prop. 5.1]. Moreover, we denote by $\widetilde{N}_u^{(t)}$ the number of living individuals at time u who have alive descent at time t. In [8], we showed that, under $\mathbb{P}_t$, $\widetilde{N}_u^{(t)}$ is geometrically distributed with parameter $\frac{W(t-u)}{W(t)}$.

Now, $\mathbb{1}_{\Gamma_{u,t}}$ can be rewritten as

$$\mathbb{1}_{\Gamma_{u,t}}=\prod_{i=1}^{\widetilde{N}_u^{(t)}}\mathbb{1}_{\left\{Z_0^i(t-u)=0\right\}},$$

where $Z_0^i(t-u)$ denotes the number of individuals alive at time t descending from the ith individual alive at time u and carrying its type (in Figure 2, the clonal type of the sub-CPP). Moreover, from Proposition 4.3 of [8], we know that that under $\mathbb{P}_t$, the family $Z_0^{(i)}(t-u)$ is an i.i.d. family of random variables distributed as $Z_0(t-u)$ under $\mathbb{P}_{t-u}$, and $\widetilde{N}_u^{(t)}$ is independent of $Z_0^{(i)}(t-u)$ (still under $\mathbb{P}_t$).

Then,

$$\mathbb{P}_t\left(\Gamma_{u,t}\right)=\mathbb{E}_t\left[\mathbb{P}_{t-u}\left(Z_0(t-u)=0\right)^{\widetilde{N}_u^{(t)}}\right]=\frac{\mathbb{P}_{t-u}\left(Z_0(t-u)=0\right)\frac{W(t-u)}{W(t)}}{1-\mathbb{P}_{t-u}\left(Z_0(t-u)=0\right)\left(1-\frac{W(t-u)}{W(t)}\right)}.$$

Using (2.8), some calculus leads to,

$$\mathbb{P}_t\left(\Gamma_{u,t}\right)=1-\frac{1}{1+\frac{W_\theta(t-u)}{e^{-\theta(t-u)}W(t)}\left(1-\frac{e^{-\theta(t-u)}W(t-u)}{W_\theta(t-u)}\right)}. \tag{6.1}$$

Now, since,

$$\mathbb{P}_t\left(\Gamma_{u,t}\right) = \mathbb{P}_u\left(\Gamma_{u,t}\right)\frac{\mathbb{P}\left(N_u>0\right)}{\mathbb{P}\left(N_t>0\right)} - \frac{\mathbb{P}\left(\Gamma_{u,t}, N_t=0, N_u>0\right)}{\mathbb{P}\left(N_t>0\right)},$$

taking $u=\beta t$, we obtain, using Lemma 2.2 in (6.1), and that

$$\mathbb{P}\left(N_t=0, N_{\beta t}>0\right) = \mathbb{P}\left(N_{\beta t}>0\right) - \mathbb{P}\left(N_t>0\right) \underset{t\to\infty}{\to} 0,$$

the desired result. □

Proof of Theorem 3.4. Fix $0<u<t$. Note that the event $\Gamma_{u,t}$ of Lemma 6.1 can be rewritten as

$$\mathbb{1}_{\Gamma_{u,t}} = \prod_{i=1}^{N_u} \mathbb{1}_{\left\{Z_0^i(t-u,O_i)=0\right\}}, \tag{6.2}$$

where $Z_0^i(t-u,O_i)$ denote the number of individuals alive at time t carrying the same type as the ith alive individual at time u, that is the ancestral family of the tree constructed from the residual lifetime of the ith individual and its descent (see Section 4).

Let N be a positive integer and $K=(k_1,\cdots,k_N)$ be a multi-integer. We denote by $\mathcal{L}^{(K)}$ (resp. $A(K,t)$) the random vector $\left(\mathcal{L}^{k_1},\ldots,\mathcal{L}^{k_N}\right)$ (resp. $(A(k_1,t),\ldots,A(k_N,t))$) with

$$\mathcal{L}_t^k = \frac{\psi'(\alpha)A(k,t) - c_k e^{\alpha t}\mathcal{E}}{e^{\frac{\alpha}{2}t}}. \tag{6.3}$$

On the event $\Gamma_{u,t}$, we have almost surely,

$$A(k_l,t) = \sum_{i=1}^{N_u} A^{(i)}(k_l,t-u,O_i), \quad \forall l=1,\ldots,N,$$

where the family $\left(A^{(i)}\left(k_l,t-u,O_i\right)\right)_{1\le i\le N_u}$ stands for the frequency spectrum the subtrees $(\mathbb{T}(O_i))_{1\le i\le N_u}$. Hence, using Lemma 4.2,

$$\mathcal{L}_t^{k_l} = \sum_{i=1}^{N_u} \frac{\psi'(\alpha)A^{(i)}(k_l,t-u,O_i) - e^{\alpha(t-u)}\mathcal{E}_i(O_i)c_{k_l}}{e^{\frac{\alpha}{2}u}e^{\frac{\alpha}{2}(t-u)}}. \tag{6.4}$$

In the sequel, we denote, for all l and $i\ge 1$,

$$\tilde{A}^{(i)}\left(k_l,t-u,O_i\right) = \frac{\psi'(\alpha)A^{(i)}\left(k_l,t-u,O_i\right) - e^{\alpha(t-u)}\mathcal{E}_i(O_i)c_{k_l}}{e^{\frac{\alpha}{2}(t-u)}}.$$

$\tilde{A}^{(i)}\left(K,t-u,O_i\right)$ denotes the corresponding random vector. In particular, according to Lemma 4.1 and Lemma 4.2, we have that the family $(\tilde{A}^{(i)}\left(k_l,t-u,O_i\right))_{1\le i\le N_u}$ is independent (under $\mathbb{P}_u$).

Now, let

$$\varphi_K(\xi) := \mathbb{E}\left[\exp\left(i<\tilde{A}\left(K,t-u,O_2\right),\xi>\right)\mathbb{1}_{Z_0^2(t-u,O_2)=0}\right],$$
$$\tilde{\varphi}_K(\xi) := \mathbb{E}\left[\exp\left(i<\tilde{A}\left(K,t-u,O_1\right),\xi>\right)\mathbb{1}_{Z_0^1(t-u,O_1)=0}\right],$$

with $\langle \cdot,\ \cdot \rangle$ denotes the standard Euclidean scalar product.

From this point, following closely the proof of Theorem 3.2 of [15]: that is, using Lemmas 4.1 and 4.2, and Equation (6.4), we have that

$$\mathbb{E}_{\beta t}\left[e^{i<\mathcal{L}_t^{(K)},\xi>}\mathbb{1}_{\Gamma_{\beta t,t}}\right] = \tilde{\varphi}_K(e^{-\frac{\alpha\beta}{2}t}\xi)\mathbb{E}_{\beta t}\left[\varphi_K(e^{-\frac{\alpha\beta}{2}t}\xi)^{N_{\beta t}-1}\right],$$

for $u = \beta t$ with $\beta \in \left(0, \frac{1}{2} \wedge (1 - \frac{\alpha}{\theta})\right)$. This gives using that $N_{\beta t}$ has geometric distribution under $\mathbb{P}_{\beta t}$ that

$$\mathbb{E}_{\beta t}\left[e^{i<\mathcal{L}_t^{(K)},\xi>}\mathbb{1}_{\Gamma_{\beta t,t}}\right] = \tilde{\varphi}_K(e^{-\frac{\alpha\beta}{2}t}\xi)\frac{1}{W(\beta t) - (W(\beta t)-1)\varphi_K(e^{-\frac{\alpha\beta}{2}t}\xi)}. \tag{6.5}$$

Now, a Taylor expansion of φ_K gives

$$\varphi(\xi) = 1 + i\sum_{p=1}^{|K|}\xi_p\mathbb{E}\left[\tilde{A}(k_p,(1-\beta)t,O_2)\mathbb{1}_{Z_0^2((1-\beta)t,O_2)=0}\right] - \frac{1}{2}\sum_{p,q=1}^{|K|}\mathcal{M}_{k_p,k_q}(t)\xi_p\xi_q + R_{(1-\beta)t}(\xi),$$

where $|K|$ is the length of the multi-integer K and

$$\begin{aligned}\mathcal{M}_{k_i,k_j}(t) := \mathbb{E}\Bigg[&\left(\frac{\psi'(\alpha)A^{(i)}\left(k_i,(1-\beta)t,O_i\right) - e^{\alpha((1-\beta)t)}\mathcal{E}_i(O_i)c_{k_i}}{e^{\frac{\alpha}{2}((1-\beta)t)}}\right)\\ &\times\left(\frac{\psi'(\alpha)A^{(i)}\left(k_j,(1-\beta)t,O_i\right) - e^{\alpha((1-\beta)t)}\mathcal{E}_i(O_i)c_{k_j}}{e^{\frac{\alpha}{2}((1-\beta)t)}}\right)\mathbb{1}_{Z_0^2((1-\beta)t,O_2)=0}\Bigg].\end{aligned} \tag{6.6}$$

We now need to handle the indicator function $\mathbb{1}_{Z_0(t-u,O_i)=0}$ in the Taylor development of φ_K. We show how it can be done for one of the second order terms, the method is similar for the other terms. It follows from Hölder's inequality that

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{\psi'(\alpha)A^{(i)}\left(k_l,(1-\beta)t,O_i\right) - e^{\alpha((1-\beta)t)}\mathcal{E}_i(O_i)c_{k_l}}{e^{\frac{\alpha}{2}((1-\beta)t)}}\right)^2\mathbb{1}_{Z_0^2((1-\beta)t,O_2)>0}\right]\\ &\qquad\leq \mathbb{E}\left[\left|\frac{\psi'(\alpha)A^{(i)}\left(k_l,(1-\beta)t,O_i\right) - e^{\alpha(1-\beta)t}\mathcal{E}_i(O_i)c_{k_l}}{e^{\frac{\alpha}{2}(1-\beta)t}}\right|^3\right]^{\frac{2}{3}}\mathbb{P}\left(Z_0^2((1-\beta)t,O_2)>0\right)^{\frac{1}{3}},\end{aligned} \tag{6.7}$$

from which it follows, using Lemma 5.8, that the r.h.s. of this last inequality is

$$\mathcal{O}\left(\mathbb{P}\left(Z_0^2((1-\beta)t,O_2)>0\right)^{\frac{1}{3}}\right).$$

Now, using (6.2) and Lemma 6.1, it is easily seen that

$$\mathbb{P}\left(Z_0^2((1-\beta)t,O_2)>0\right)\xrightarrow[t\to\infty]{} 0.$$

Finally, using Cauchy-Schwarz inequality in (6.6) and Lemma 5.3, we get

$$\mathcal{M}_{k_i,k_j}(t)\xrightarrow[t\to\infty]{}\psi'(\alpha)H_{k_i,k_j},$$

where H is defined in Theorem 3.4 (see also Lemma 5.3).

Now, in (6.6), we have, using similar methods and Lemma 5.7, that

$$e^{-\frac{\alpha\beta}{2}t}\mathbb{E}\left[\tilde{A}(k_p,(1-\beta)t,O_2)\mathbb{1}_{Z_0^2((1-\beta)t,O_2)=0}\right]=\mathcal{O}(e^{-\frac{\alpha}{2}t}).$$

Finally, using the above computations, we get with (6.6)

$$\varphi(e^{-\frac{\alpha\beta}{2}t}\xi)=1-e^{-\alpha\beta t}\frac{1}{2}\sum_{p,q=1}^{|K|}\mathcal{M}_{k_p,k_q}(t)\xi_p\xi_q+R_{(1-\beta)t}(e^{-\frac{\alpha\beta}{2}t}\xi)+\mathcal{O}(e^{-\frac{\alpha}{2}t}),\tag{6.8}$$

and a similar treatment for the third order term in the reminder R using Lemma 5.8 gives

$$R_{(1-\beta)t}(e^{-\frac{\alpha\beta}{2}t}\xi)=\mathcal{O}(e^{-\frac{3\alpha\beta}{2}t}).$$

Putting all these together in (6.5) (and taking into account that $\beta<\frac{1}{2}\wedge(1-\frac{\alpha}{\theta})$) entails

$$\mathbb{E}_{\beta t}\left[e^{i<\mathcal{L}_t^{(K)},\xi>}\mathbb{1}_{\Gamma_{\beta t,t}}\right]=\frac{1}{1+e^{-\alpha\beta t}W(\beta t)\frac{1}{2}\sum_{i,j=1}^{|K|}\xi_i\xi_j\mathcal{M}_{k_i,k_j}(t)+o(1)}.\tag{6.9}$$

These allow us to conclude, from (6.9), that

$$\mathbb{E}_{\beta t}\left[e^{i<\mathcal{L}_t^{(K)},\xi>}\mathbb{1}_{\Gamma_{\beta t,t}}\right]\xrightarrow[t\to\infty]{}\frac{1}{1+\frac{1}{2}\sum_{i,j=1}^{N}\mathcal{H}_{k_i,k_j}\ \xi_i\xi_j},$$

where K is the multi-integer $(k_1,\dots,k_N)$. To end the proof, note that,

$$\left|\mathbb{E}_\infty\left[e^{i<\mathcal{L}_t^{(K)},\xi>}\right]-\mathbb{E}_{\beta t}\left[e^{i<\mathcal{L}_t^{(K)},\xi>}\mathbb{1}_{\Gamma_{\beta t,t}}\right]\right|\le\mathbb{E}\left[\left|\frac{\mathbb{1}_{\mathrm{NonEx}}}{\mathbb{P}\left(\mathrm{NonEx}\right)}-\frac{\mathbb{1}_{N_{\beta t}>0}\mathbb{1}_{\Gamma_{\beta t,t}}}{\mathbb{P}\left(N_{\beta t}>0\right)}\right|\right]\underset{t\to\infty}{\to}0,$$

thanks to Lemma 6.1. □

7 Proof of Theorem 3.1

Since all the ideas of the proof of this theorem have been developed the preceding sections, we do not detail all the proof. We only details the steps which need clarification.

7.1 More on moments

Our first step is the computation of the covariance matrix $\mathcal{M}$ of the Laplace limit law. According to the proof of Theorem 3.4, it is given, for two positive integer l and k, by

$$\mathcal{M}_{k,l}:=\lim_{t\to\infty}\frac{W(\beta t)}{e^{\alpha\beta t}}\mathbb{E}\Bigg[\left(\frac{\psi'(\alpha)A^{(i)}\left(k,(1-\beta)t,O_i\right)-\psi'(\alpha)c_kN_{(1-\beta)t}}{e^{\frac{\alpha}{2}((1-\beta)t)}}\right)$$
$$\times\left(\frac{\psi'(\alpha)A^{(i)}\left(l,(1-\beta)t,O_i\right)-c_lN_{(1-\beta)t}}{e^{\frac{\alpha}{2}((1-\beta)t)}}\right)\mathbb{1}_{Z_0^2((1-\beta)t,O_2)>0}\Bigg],$$

which is equal, thanks to (6.7), Lemma 5.1 and an easy adaptation of Lemma 6.6 in [15], to

$$\mathcal{M}_{k,l} = \lim_{t\to\infty} \frac{b\psi'(\alpha)}{\alpha}\frac{W(\beta t)}{e^{\alpha\beta t}} e^{\alpha t}\mathbb{E}\left[\left(e^{-\alpha t}A(k,t) - c_k e^{-\alpha t}N_t\right)\left(e^{-\alpha t}A(l,t) - c_l e^{-\alpha t}N_t\right)\right] = M_{k,l}.$$

Now, in order to obtain the joint convergence, we need some new moment estimates.

Lemma 7.1. *Let k be some positive integer, then*

$$\mathbb{E}\left[e^{-\alpha t}N_t\left(\frac{A(k,t) - c_k N_t}{e^{\frac{\alpha}{2}t}}\right)\right] \xrightarrow[t\to\infty]{} 0$$

Proof. In order to have an explicit expression for

$$\mathbb{E}_t\left[N_t\left(A(k,t) - c_k N_t\right)\right],$$

we use again Remark 5.6 of [8] to get

$$\mathbb{E}_t\left[N_t A(k,t)\right] = 2\int_0^t W(t)^2\left(1 - \frac{W(a)}{W(t)}\right)\frac{\theta e^{-\theta a}}{W_\theta(a)^2}\left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} da + W(t)\int_0^t \theta\frac{\mathbb{E}_a\left[N_a \mathbb{1}_{Z_0(a)=k}\right]}{W(a)}\, da.$$

Now, using that N_t has a geometric distribution under $\mathbb{P}_t$ (with parameter $W(t)^{-1}$), we get

$$c_k\mathbb{E}_t\left[N_t^2\right] = W(t)^2\left(2 - \frac{1}{W(t)}\right)\int_0^\infty \frac{\theta e^{-\theta a}}{W_\theta(a)^2}\left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} da.$$

Combining this two equality leads to

$$\mathbb{E}_t\left[N_t\left(A(k,t) - c_k N_t\right)\right] = 2W(t)^2\int_t^\infty \frac{\theta e^{-\theta a}}{W_\theta(a)^2}\left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} da + \mathcal{O}(tW(t)).$$

Now, since Lemma 2.2 (clonal sub-critical case) and that $W(t) \sim \psi'(\alpha)e^{\alpha t}$ (Lemma 2.2 also) entails

$$e^{-\frac{3\alpha}{2}t}W(t)^2\int_t^\infty \frac{\theta e^{-\theta a}}{W_\theta(a)^2}\left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} da = \mathcal{O}\left(e^{-\frac{3\alpha}{2}t}W(t)^2\int_t^\infty \theta e^{-\theta a}\, da\right),$$

we then get the result using (2.9) and that $\theta > \alpha$. □

Remark 7.2. *In the case of Theorem 3.4, one would need that*

$$\mathbb{E}\left[e^{-\alpha t}N_t\left(\frac{\psi'(\alpha)A(k,t) - c_k e^{\alpha t}\mathcal{E}}{e^{\frac{\alpha}{2}t}}\right)\right] \xrightarrow[t\to\infty]{} 0,$$

to obtain the joint convergence with $e^{-\alpha t}N_t$ as stated in the theorem. In fact, this result easily follows from the above proof and the estimate

$$\mathbb{E}_t\left[N_t\mathcal{E}\right] - \psi'(\alpha)e^{-\alpha t}\mathbb{E}_t\left[N_t^2\right] = \frac{2e^{\alpha t}}{\psi'(\alpha)} - 2\psi'(\alpha)e^{-\alpha t}W(t)^2 + o(e^{\frac{\alpha}{2}t}) = o(e^{\frac{\alpha}{2}t}),$$

which is deduced using (4.3)*, Lemma 2.3 and the geometric law of N_t under $\mathbb{P}_t$.*

7.2 Insights for the joint convergence

In this section, we detail the points which require clarification to obtain the joint convergence as stated in Theorems 3.4 and 3.1. In the case of Theorem 3.1, this means the joint convergence of

$$\left(\psi'(\alpha)e^{-\alpha t}N_t,\ e^{-\frac{\alpha}{2}t}\left(A(k,t)-c_kN_t\right)_{k\geq 1}\right),$$

as t goes to infinity. To this end, rather than considering the characteristic function

$$\mathbb{E}_u\left[e^{i<\mathcal{L}_t^{(K)},\xi>}\mathbb{1}_{\Gamma_{u,t}}\right],$$

as in Section 6 (we also recall that $\mathcal{L}_t^{(K)}$ and $\Gamma_{u,t}$ are respectively defined in Equation (6.3) and Lemma 6.1), we consider

$$\mathbb{E}_u\left[\exp\left(i<\mathcal{L}_t^{(K)},\xi>+i\lambda\psi'(\alpha)e^{-\alpha t}N_t\right)\mathbb{1}_{\Gamma_{u,t}}\right], \tag{7.1}$$

with, this time,

$$\mathcal{L}_t^{k_l}=\sum_{i=1}^{N_u}\frac{\psi'(\alpha)A^{(i)}(k_l,t-u,O_i)-\psi'(\alpha)N_{t-u}(O_i)c_{k_l}}{e^{\frac{\alpha}{2}u}e^{\frac{\alpha}{2}(t-u)}}.$$

and, for all l and $i\geq 1$,

$$\tilde{A}^{(i)}\left(k_l,t-u,O_i\right)=\frac{\psi'(\alpha)A^{(i)}\left(k_l,t-u,O_i\right)-\psi'(\alpha)N_{t-u}(O_i)c_{k_l}}{e^{\frac{\alpha}{2}(t-u)}},$$

where we refers the reader to Section 4 for the definition of $N_{t-u}(O_i)$. Equation (7.1) can be rewritten, following the proof of Theorem 3.4, as

$$\tilde{\varphi}_K(\xi,\lambda)\mathbb{E}_u\left[\varphi_K(e^{-\frac{\alpha}{2}u}\xi,e^{-\frac{\alpha}{2}u}\lambda)^{N_u-1}\right]=\tilde{\varphi}_K(e^{-\frac{\alpha}{2}u}\xi,e^{-\frac{\alpha}{2}u}\lambda)\frac{W(u)^{-1}}{1-(1-W(u)^{-1})\varphi_K(e^{-\frac{\alpha}{2}u}\xi,e^{-\frac{\alpha}{2}u}\lambda)}, \tag{7.2}$$

with, this time,

$$\begin{aligned}\varphi_K\left(\xi,\theta\right)&:=\mathbb{E}\left[\exp\left(i<\tilde{A}\left(K,t-u,O_2\right),\xi>+i\lambda\psi'(\alpha)e^{-\alpha(t-u)}N_{t-u}(O_2)\right)\mathbb{1}_{Z_0^2(t-u,O_2)=0}\right],\\ \tilde{\varphi}_K\left(\xi,\theta\right)&:=\mathbb{E}\left[\exp\left(i<\tilde{A}\left(K,t-u,O_1\right),\xi>+i\lambda\psi'(\alpha)e^{-\alpha(t-u)}N_{t-u}(O_1)\right)\mathbb{1}_{Z_0^1(t-u,O_1)=0}\right].\end{aligned}$$

Now, a simple Taylor expansion gives

$$\begin{aligned}\varphi(\xi,\lambda)=1&+i\sum_{p=1}^{|K|}\xi_p\mathbb{E}\left[\tilde{A}(k_p,t-u,O_2)\right]+\lambda i\mathbb{E}\left[\psi'(\alpha)e^{-\alpha(t-u)}N_{t-u}(O_2)\right]\\&-\frac{1}{2}\sum_{p,q=1}^{|K|}\mathbb{E}\left[\tilde{A}(k_p,t-u,O_2)\tilde{A}(k_q,t-u,O_2)\right]\xi_p\xi_q\\&-\frac{1}{2}\sum_{p=1}^{|K|}e^{-\alpha(t-u)}\mathbb{E}\left[\tilde{A}(k_p,t-u,O_2)\psi'(\alpha)N_{t-u}(O_2)\right]\xi_p\lambda+R_{t-u}(\xi,\lambda).\end{aligned} \tag{7.3}$$

Remark 7.3. *In the above expression, the moments does not involve* $\mathbb{1}_{Z_0^i(t-u,O_1)=0}$ *at will. Since according to the proof of Theorem 3.4, these indicator functions can be neglect, their presence is hidden in the reminder* R_{t-u}.

Hence, plugging (7.3) in Equation (7.2) gives

$$\begin{aligned}
&\tilde{\varphi}_K(e^{-\alpha u}\xi, e^{-\alpha u}\lambda)\mathbb{E}_u\left[\varphi_K(e^{-\alpha u}\xi, e^{-\alpha u}\lambda)^{N_u-1}\right] \\
&= \frac{1}{W(u)}\Bigg[1-(1-W(u)^{-1})\Big(1+e^{-\alpha(t-u)}\lambda i\mathbb{E}\left[\psi'(\alpha)N_{t-u}(O_2)\right] \\
&\qquad\qquad -\frac{1}{2}\sum_{p,q=1}^{|K|} e^{-\alpha u}\mathbb{E}\left[\tilde{A}(k_p,t-u,O_2)\tilde{A}(k_q,t-u,O_2)\right]\xi_p\xi_q+\tilde{R}(t,u)\Big)\Bigg]^{-1},
\end{aligned}$$

for some $\tilde{R}$ satisfying $\tilde{R}(\beta t,t)=o(e^{-\alpha\beta t})$(with the same choice of β as in the proof of Theorem 3.4). Now, setting $u=\beta t$ as in the proof of Theorem 3.4, and taking the limit as t goes to infinity shows the convergence of the above quantity to

$$\frac{1}{1+\lambda i+\frac{1}{2}\sum_{p,q=1}^{|K|}\mathcal{M}_{i,j}\xi_p\xi_q}.$$

Indeed, the one main difference with the proof of Theorem 3.4 lies on the moment

$$\mathbb{E}\left[\tilde{A}(k_p,t-u,O_2)N_{t-u}(O_2)\right]$$

in the Taylor development of φ_K, which can be shown to go to 0 using Lemma 7.1 and an adaptation of Lemma 6.6 in [15].

To end the proof let us remark that if G is some Gaussian random variable with null mean and covariance matrix K independent of a random variable $\mathcal{E}$ (with exponential distribution and mean 1), then the characteristic function of the couple $(\sqrt{\mathcal{E}}G,\mathcal{E})$ is given by

$$(\xi,\lambda)\in\mathbb{R}^d\times\mathbb{R}\mapsto\frac{1}{1+\lambda i+\frac{1}{2}\langle\xi,K\xi\rangle},$$

where $\langle\cdot,\ \cdot\rangle$ denotes the standard Euclidean scalar product.

8 Numerical studies

The purpose of this section is to analyze our approximation method and the estimation of the error by virtue of numerical experiments. There are several practical difficulties appearing when one tries to perform such study.

The first problem, which involves no conceptual difficulties, lies only on the implementation of the formulas appearing in Theorems 3.4 and 3.1. In particular, the computation of the moments of type $\mathbb{E}[A(k,t)\mathbb{1}_{Z_0(t)=l}]$ are particularly complicated (see Proposition 5.4 in [8]).

Another difficulty is to obtain numerical approximations of the scale functions W and W_θ. For instance, these functions appear in the computation of the covariance matrix of Theorems 3.4 and 3.1 or when one wants to simulate the *coalescent point process*. To obtain such approximations, we need to apply numerically the Laplace inversion operator to the functions $\frac{1}{\psi}$ and $\frac{1}{\psi_\theta}$.

Unfortunately, the Laplace numerical inversion is a rather difficult problem (see for instance [1] or [2]) which is often computationally expensive. As a consequence, the computational cost of performing multiple numerical integration involving W or W_θ can be important when done with a crude method. Moreover, these methods presents rough numerical instabilities when the original function is exponentially increasing (inverting $\lambda \to \frac{1}{1-\lambda}$, whose inverse is $x \to e^x$, is already a tough numerical problem).

For all these reasons, we provide with this work a Matlab toolbox which handle all these difficulties and allows users to perform numerical experiments without having to take care of these issues. This toolbox is available on the author personal homepage.

In this whole section, we are interested in the approximation of the frequency spectrum at a fixed time t by the sequence $N_t(c_k)_{k\geq 1}$ (we recall that c_k is defined in equation (2.13)). As a consequence, the errors in this approximation are computed thanks to Theorem 3.1. The parameters of the model are set as follows:

- $\mathbb{P}_V$ is a Rice distribution with shape parameter 1 and scale parameter 1.
- $b = 1$.
- $\theta = 1$.

For such parameters α approximately equals to 0.5. Figure 4 shows the evolution of the frequency spectrum (for k between 1 and 10) through time. The different quantities seem to growth exponentially with rate α with a time-shift which depend on k. An interesting open question would be to understand the behavior of these shifts. In order to stress our methods of approximation, the first idea is to look to the renormalized frequency spectrum $(\frac{A(k,t)}{c_k})_{k\geq 1}$ which is expected to look like $(N_t,\ t \in \mathbb{R}_+)$. As showed in Figure 5, the approximation seems to be quite accurate for $k = 1, 2$. However, a more quantitative analysis is required. Figure 6 shows the absolute error in the approximation of $A(1,t)$ by $c_1 N_t$. This error is a little disappointing since it since to diverge when t goes to infinity. However, even if, according to Figure 6, the absolute error at time 20 is of order 10^3, the relative error shows that this error is quite small with respect to the value of $A(1,20)$. Another question is about the speed of convergence in the central limit theorem stated in Theorem 3.1. The red curve of Figure 7 shows the density of the Laplace distribution given in Theorem 3.1 in the case of $A(1,t)$ whereas the blue histogram shows the distribution of $\psi'(\alpha)(e^{-\alpha\frac{t}{2}}(A(1,t) - c_k N_t))$ for $t = 10$ ($\alpha t \sim 5$ and $\mathbb{E}_t[N_t] \sim 300$) from 10000 simulations. This figure highlights the fact that even if the taken time t is quite small the distribution $\psi'(\alpha)(e^{-\alpha\frac{t}{2}}(A(1,t) - c_k N_t))$ seems already close to the limiting distribution. Figure 8 shows the same kind of behavior in the multidimensional case. To be more quantitative, Figure 9 shows the evolution in time of the distance between the density of limit distribution given in Theorem 3.1 and a kernel estimation of the distribution of $\psi'(\alpha)(e^{-\alpha\frac{t}{2}}(A(1,t) - c_k N_t))$ (the estimation is made from 10000 simulations at each time). This

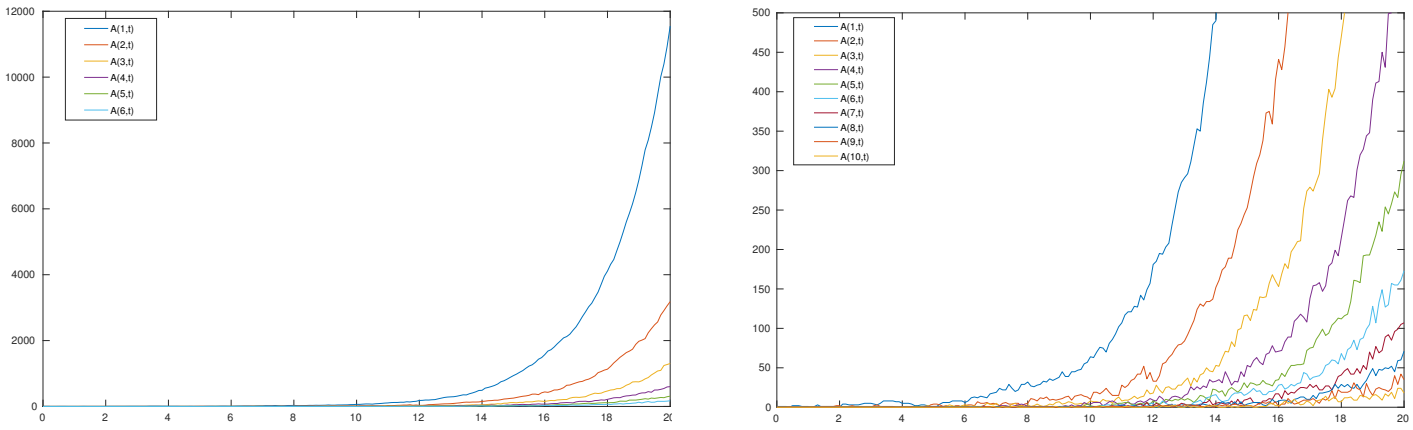

Figure 4: A simulation of the evolution of the frequency spectrum under the given model.

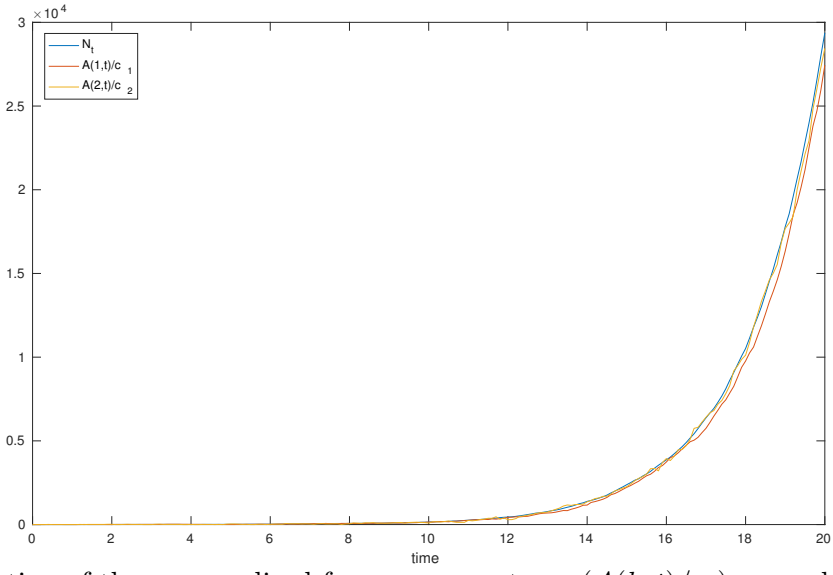

Figure 5: Evolution of the renormalized frequency spectrum $(A(k,t)/c_k)_{k\geq 1}$ under the given model.

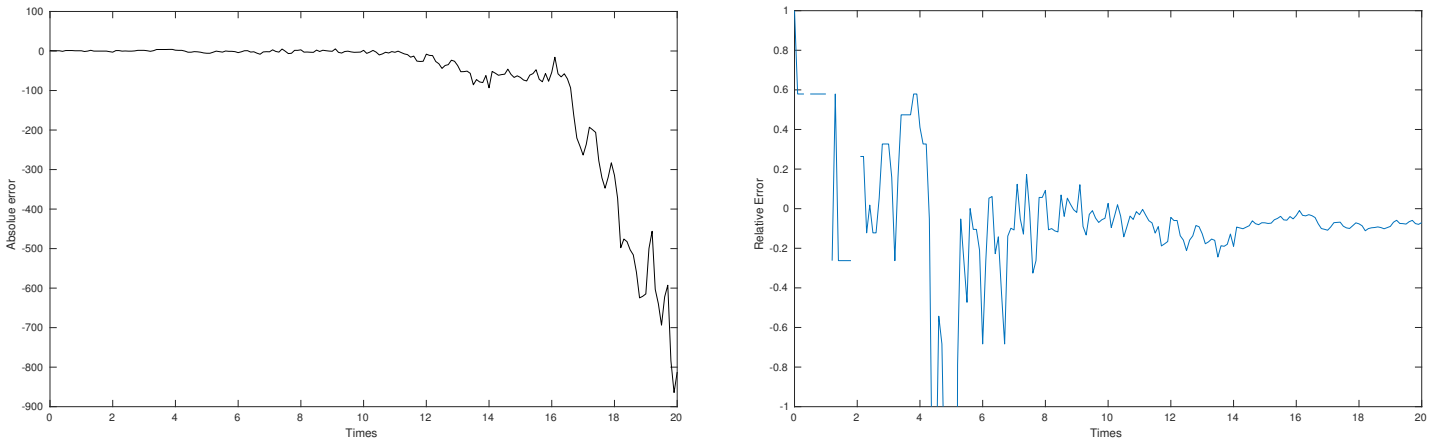

Figure 6: Absolute error (left picture) and relative (right picture) in the approximation of $A(1,t)$ by $c_1 N_t$.

suggest an exponential rate of convergence in Theorem 3.1. In the view of Figure 9, one may think that Berry-Essen type results for Theorem 3.1 would be quite interesting, in particular to understand how the speed of convergence is related to choice of the parameters. Another interesting question which could be partially probed by simulation is the study of the behavior of the error in the clonal supercritical case. Figure 10 shows a kernel estimation (from 10000 simulations) of the density of $\psi'(\alpha)(e^{-\alpha\frac{t}{2}}(A(1,t) - c_k N_t))$ in the clonal supercritical case ($\theta = 0.2 < \alpha$). Figure 10 suggest a totally different behavior with a limit distribution which is asymmetric with respect to 0. In particular, in the view of the shape of the distribution, one could conjecture that the limit is a skew stable distribution.

To end this section, let us goes back to one of the motivation of this work. The following discussion do not claim to be rigorous and is essentially formal. We recall that the Extended Haplotype Homozygosity (EHH) can be used to detect positive selection in a population [20]. In particular, the behavior of the frequency spectrum we are interested in the behavior of the frequency spectrum as the mutation rate increases. In order to have a rigorous model to describe this type phenomenon, we need to introduce a new mutation measure which is different from the one given in Section 2. We define it directly on the CPP but this could be equivalently defined on the splitting tree. So let $\mathcal{P}$ be a Poisson random measure on $[0,t] \times \mathbb{N} \times \mathbb{R}_+$ with intensity measure $\lambda \otimes C \otimes \lambda$, where C is the counting measure on $\mathbb{N}$. Then, for any mutation rate θ in $\mathbb{R}_+$, we define the θ-mutation random measure $\mathcal{N}_\theta$ by

$$\mathcal{N}_\theta(A \times B) = \int_{A \times B \times [0,\theta]} \mathbb{1}_{H_i > t-a} \mathbb{1}_{i < \mathcal{N}_t} \mathcal{P}(di, da, dx),$$

where, as before, an atom at (a,i) means that the ith branch experiences a mutation at time $t-a$. This construction allows to increase the mutation rate in consistent manner. This allows to model the type of an individual at a distance x (such that the mutation rate is a function of x) from the core region of DNA (we refer the reader to [20] for more details). Now, following [8], we can define

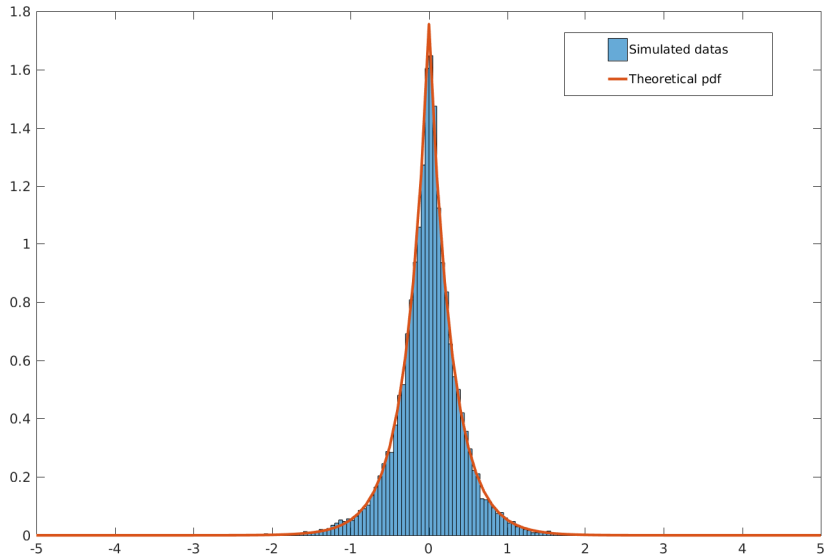

Figure 7: Distribution of the renormalized error and expected limit distribution given by our CLT.

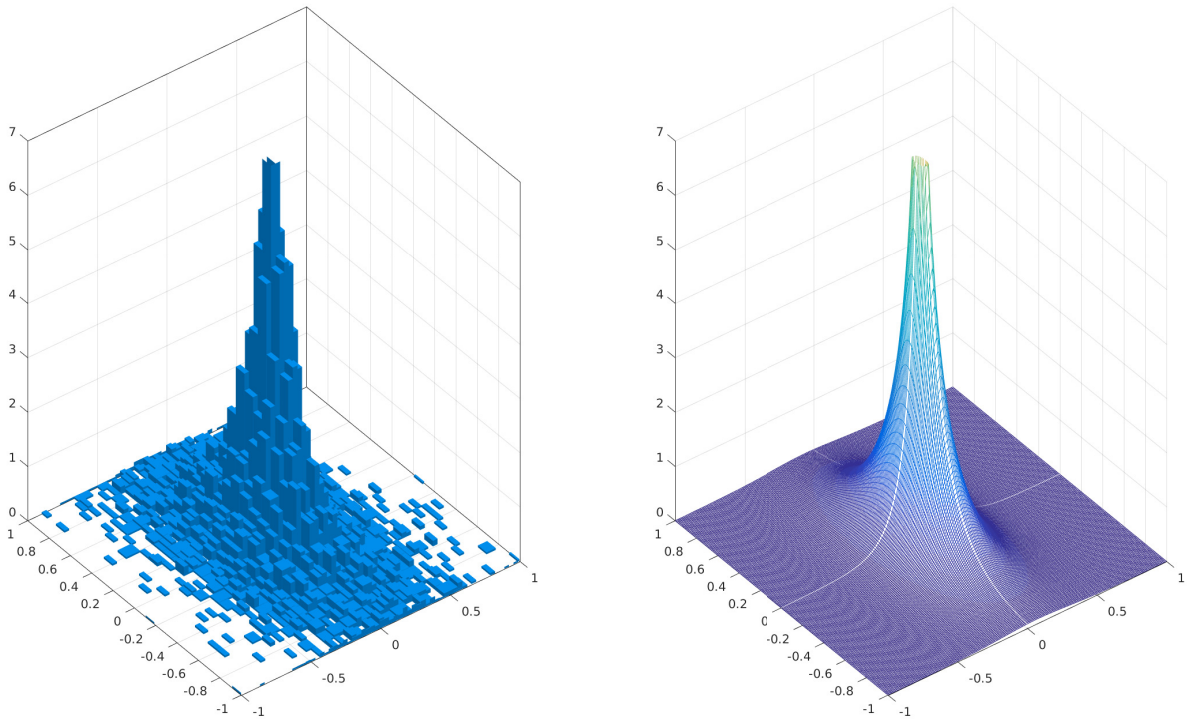

Figure 8: Joint distribution of the renormalized error (left figure) and expected limit distribution (right figure) given by our CLT.

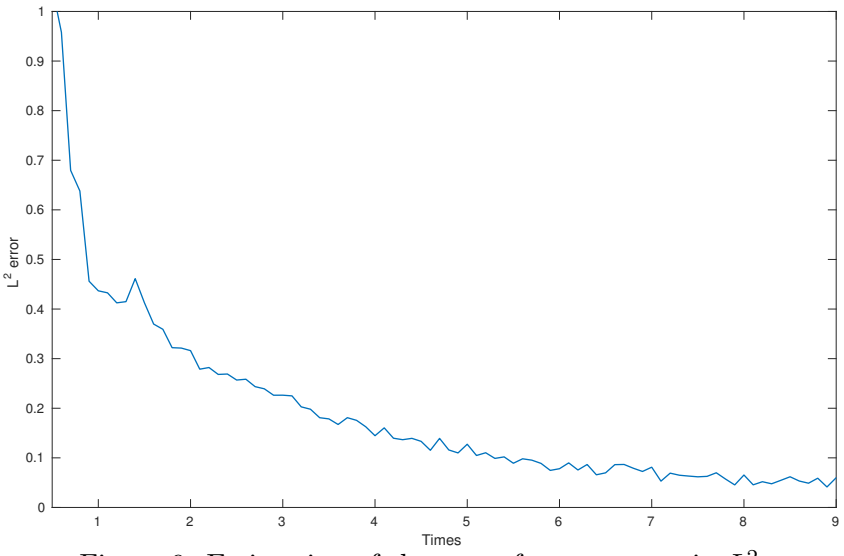

Figure 9: Estimation of the rate of convergence in L^2 norm.

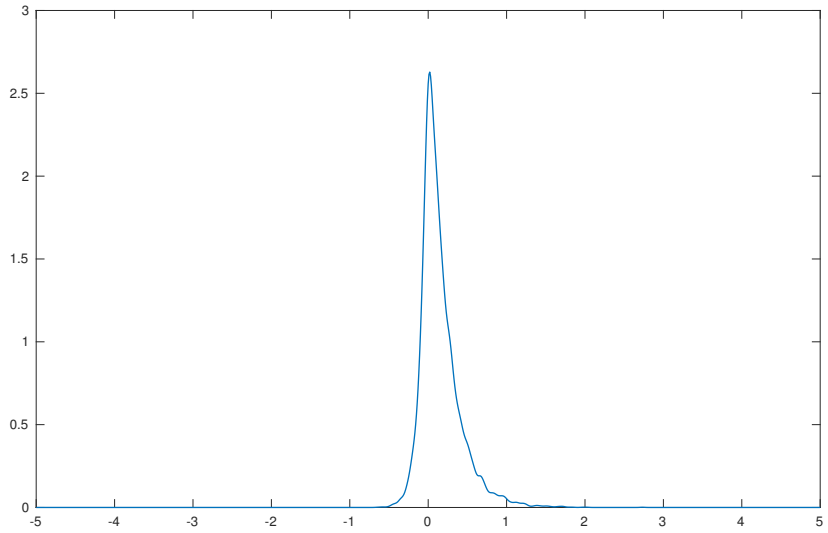

Figure 10: Kernel estimate of the probability density function of the limit distribution in the clonal supercritical case.

the frequency spectrum at mutation rate θ by

$$A^\theta(k,t) = \int_{[0,t]\times\mathbb{N}} \mathbb{1}_{Z_0(i,a)=k} \mathcal{N}_\theta(di, da),$$

where $Z_0(i,a)$ is the number of individual at time t carrying the type of the ith individual at time $t-a$ (see [8] for more details). Let us also define $Z_0^\theta(t)$ the number of individuals carrying the type of the first individual at time 0 when the mutation measure is given by $\mathcal{N}_\theta$. As expected, the allelic partition of the population becomes thinner as θ growth.

Now, the definition of the EHH $G_\theta(t)$ is the probability that two uniformly sampled individuals in the population have the same type, that is

$$G_t(\theta) = \frac{Z_0^\theta(t)(Z_0^\theta(t)-1) + \sum_{k\geq 1} k(k-1)A^\theta(k,t)}{N_t(N_t-1)}.$$

Using that

$$N_t = Z_0^\theta(t) + \sum_{k\geq 1} kA^\theta(k,t),$$

this rewrite

$$G_t(\theta) = \frac{(N_t - \sum_{k\geq 1} kA^\theta(k,t))(N_t - \sum_{k\geq 1} kA^\theta(k,t)-1) + \sum_{k\geq 1} k(k-1)A^\theta(k,t)}{N_t(N_t-1)}.$$

Finally, using the approximation

$$(A(k,t))_{k\geq 1} \approx (c_k)_{k\geq 1} N_t$$

proposed in this work, one could expect that

$$G_t(\theta) \approx \frac{\sum_{k\geq 1} k(k-1)c_k}{N_t} = \frac{\int_0^\infty 2\theta e^{-\theta x}(W_\theta(x)-1)dx}{N_t}.$$

We stress the fact that the above expression makes sens only in the clonal subcritical case (in the other cases the integral in not finite). Now, we can look at the accuracy of this approximation in view of numerical simulation. Figure 11 shows the value of the EHH (when θ increase) from a simulation of the model (blue curve) and the one obtained using our approximation (red curve). In view of Figure 11, the approximation seems pretty accurate. In order to be more quantitative, Figure 12 shows the relative error between the EHH and its approximation for one simulation. This shows that the error, as least for sufficiently large θ, remains under 8%. To end, let us highlight that Theorem 3.1 can be used to give confidence intervals for fixed θ but in order the construct tests of selection from curves like these of Figure 11 one would need to have functional CLT in long time for the process $((A^\theta(k,t) - c_k^\theta N_t)_{k\geq 1},\ \theta \in \mathbb{R}_+)$.

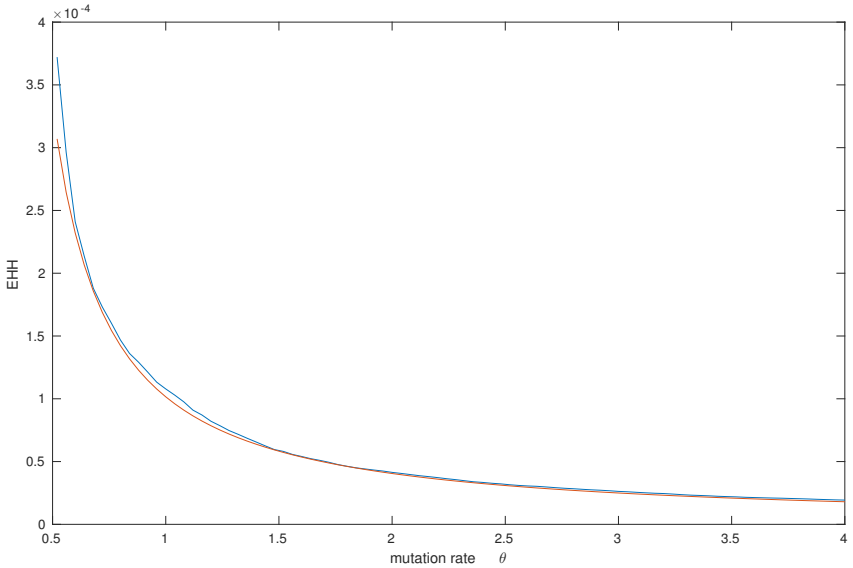

Figure 11: Extended Haplotype Homozygosity (EHH) with the given approximation (red curve) and from simulated data (blue curve)

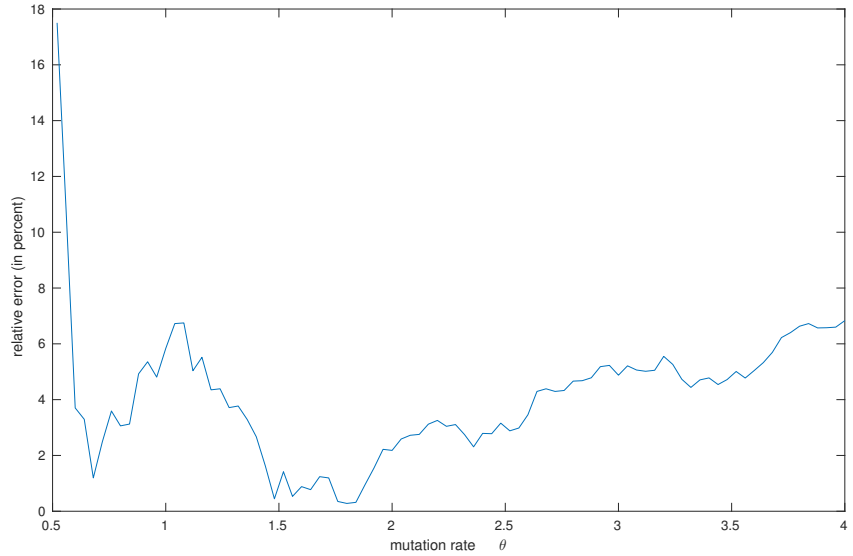

Figure 12: Relative error in the approximation of the EHH

A Formula for the fourth moment of the error

Lemma A.1.

$$\begin{aligned}
\mathbb{E}_t\left[(A(k,t)-c_kN_t)^4\right] &= 4\int_{[0,t]}\theta\frac{W(t)}{W(a)}\mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k}\left(A(k,a)-c_kN_a\right)^3\right]da \\
&+ 48\int_{[0,t]}\theta\frac{W(t)^2}{W(a)^2}\left(1-\frac{W(a)}{W(t)}\right)\mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k}N_aA(k,a)\right]\mathbb{E}_a\left[(c_kN_a-A(k,a))\right]da \\
&+ 24\int_{[0,t]}\theta\frac{W(t)^2}{W(a)^2}\left(1-\frac{W(a)}{W(t)}\right)\mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k}N_a^2\right]\mathbb{E}_a\left[(A(k,a)-c_kN_a)\right]da \\
&+ 24\int_{[0,t]}\theta\frac{W(t)^2}{W(a)^2}\left(1-\frac{W(a)}{W(t)}\right)\mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k}A(k,a)^2\right]\mathbb{E}_a\left[(A(k,a)-c_kN_a)\right]da \\
&+ 8\int_{[0,t]}\theta\frac{W(t)^2}{W(a)^2}\left(1-\frac{W(a)}{W(t)}\right)\mathbb{P}_a\left(Z_0(a)=k\right)\mathbb{E}_a\left[\left(A(k,a)-c_kN_a\right)^3\right]da \\
&+ 48\int_{[0,t]}\theta\frac{W(t)^2}{W(a)^2}\left(1-\frac{W(a)}{W(t)}\right)\mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k}A(k,a)\right]\mathbb{E}_a\left[\left(A(k,a)-c_kN_a\right)^2\right]da \\
&+ 72\int_{[0,t]}\theta\frac{W(t)^3}{W(a)^3}\left(1-\frac{W(a)}{W(t)}\right)^2\mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k}\left(A(k,a)-c_kN_a\right)\right]\mathbb{E}_a\left[(A(k,a)-c_kN_a)\right]^2da \\
&+ 72\int_{[0,t]}\theta\frac{W(t)^3}{W(a)^3}\left(1-\frac{W(a)}{W(t)}\right)^2\mathbb{P}_a\left(Z_0(a)=k\right)\mathbb{E}_a\left[\left(A(k,a)-c_kN_a\right)^2\right]\mathbb{E}_a\left[A(k,a)-N_ac_k\right]da \\
&+ 96\int_{[0,t]}\theta\frac{W(t)^4}{W(a)^4}\left(1-\frac{W(a)}{W(t)}\right)^3\mathbb{P}_a\left(Z_0(a)=k\right)\mathbb{E}_a\left[(A(k,a)-c_kN_a)\right]^3da + c_k^4\mathbb{E}_tN_t^4
\end{aligned}$$

Proof. The proof of this Lemma lies on the calculation of the expectation of each term in the development of

$$\left(A(k,t)-c_kN_t\right)^4.$$

To make this, we intensively use the relation (2.10) and the method developed in [8]. We begin by computing

$$\mathbb{E}_t\left[A(k,t)^4\right].$$

Formula (2.10) gives us,

$$
\begin{aligned}
A(k,t)^4 =& 4\int_{[0,t]\times\mathbb{N}} \mathbb{1}_{Z_0^i(a)=k_i} \sum_{u_{1:3}=1}^{N_{t-a}^{(t)}} \prod_{\substack{j=1\\ i\neq j}}^{3} A^{(u_j)}(k,a)\mathcal{N}\,(da,di) \\
=& 4\int_{[0,t]\times\mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} A^i(k,a)A^i(k,a)A^i(k,a)\mathcal{N}(da,di) \\
&+ 4\int_{[0,t]\times\mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} \sum_{\substack{j_1,j_2,j_3=1\\ j_1\neq j_2\neq j_3\neq i}}^{N_{t-a}^{(t)}} A^{j_1}(k,a)A^{j_2}(k,a)A^{j_3}(k,a)\mathcal{N}(da,di) \\
&+ 12\int_{[0,t]\times\mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} A^i(k,a)A^i(k,a) \sum_{j=1,j\neq i}^{N_{t-a}^{(t)}} A^j(k,a)\mathcal{N}(da,di) \\
&+ 4\int_{[0,t]\times\mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} \sum_{j=1,j\neq i}^{N_{t-a}^{(t)}} A^j(k,a)^3\mathcal{N}(da,di) \\
&+ 12\int_{[0,t]\times\mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} A^i(k,a) \sum_{j_1,j_2=1,j_1\neq j_2\neq i}^{N_{t-a}^{(t)}} A^{j_1}(k,a)A^{j_2}(k,a)\mathcal{N}(da,di) \\
&+ 24\int_{[0,t]\times\mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} A^i(k,a) \sum_{j_1=1,j_1\neq i}^{N_{t-a}^{(t)}} A^{j_1}(k,a)A^{j_1}(k,a)\mathcal{N}(da,di) \\
&+ 12\int_{[0,t]\times\mathbb{N}} \mathbb{1}_{Z_0^i(a)=k} \sum_{j_1,j_2=1,j_1\neq j_2\neq i}^{N_{t-a}^{(t)}} A^{j_1}(k,a)^2A^{j_2}(k,a)\mathcal{N}(da,di). \qquad \text{(A.1)}
\end{aligned}
$$

The decomposition of the sum in form

$$\sum_{u_{1:3}=1}^{N_{t-a}^{(t)}},$$

has then been made to distinguish independence properties in our calculation. Actually, as soon as, $i \neq j$, $A^i(k,a)$ is independent from $A^i(k,a)$ (see [8] for details). It is essential to note that the expectation of these integrals with respect to the random measure $\mathcal{N}$ are all calculated thanks to

Theorem 3.1 of [8]. So, taking the expectation now leads to,

$$\begin{aligned}
\mathbb{E}_t\left[A(k,t)^4\right] =& 4\int_{[0,t]} \theta \mathbb{E}_a\left[N_{t-a}^{(t)}\right] \mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k} A(k,a)^3\right] \theta da \\
&+ 4\int_{[0,t]} \theta \mathbb{P}_a\left(Z_0(a)=k\right) \mathbb{E}_a\left[\left(N_{t-a}^{(t)}\right)_{(4)}\right] \mathbb{E}_a\left[A(k,a)\right]^3 da \\
&+ 12\int_{[0,t]} \theta \mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k} A(k,a)^2\right] \mathbb{E}_a\left[\left(N_{t-a}^{(t)}\right)_{(2)}\right] \mathbb{E}_a\left[A(k,a)\right] da \\
&+ 4\int_{[0,t]} \theta \mathbb{P}_a\left(Z_0(a)=k\right) \mathbb{E}_a\left[\left(N_{t-a}^{(t)}\right)_{(2)}\right] \mathbb{E}_a\left[A(k,a)^3\right] da \\
&+ 12\int_{[0,t]} \theta \mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k} A(k,a)\right] \mathbb{E}_a\left[\left(N_{t-a}^{(t)}\right)_{(3)}\right] \mathbb{E}_a\left[A(k,a)\right]^2 da \\
&+ 24\int_{[0,t]} \theta \mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k} A(k,a)\right] \mathbb{E}_a\left[\left(N_{t-a}^{(t)}\right)_{(2)}\right] \mathbb{E}_a\left[A(k,a)^2\right] da \\
&+ 12\int_{[0,t]} \theta \mathbb{P}_a\left(Z_0(a)=k\right) \mathbb{E}_a\left[\left(N_{t-a}^{(t)}\right)_{(3)}\right] \mathbb{E}_a\left[A(k,a)^2\right] \mathbb{E}_a\left[A(k,a)\right] da.
\end{aligned}$$

Using the same method for all the other terms and that, for any positive real number a lower than t,

$$N_t = \sum_{i=1}^{N_{t-a}^{(t)}} N_a^{(i)},$$

we get Lemma A.1 by reassembling similar terms together. The last term is obtained using the geometric distribution of N_t under $\mathbb{P}_t$. □

B Boundedness of the fourth moment

Lemma B.1. *We begin the proof of the boundedness of the fourth moment by some estimates.*

$$\mathbb{E}_t\left[\left(A(k,t) - c_k N_t\right)\right] = \mathcal{O}\left(e^{-(\theta-\alpha)t}\right), \tag{i}$$

$$\mathbb{E}_t\left[\left(A(k,t) - c_k N_t\right)^3\right] = \mathcal{O}\left(W(t)^2\right), \tag{ii}$$

$$\mathbb{E}_t\left[\left(A(k,t) - c_k N_t\right)^2\right] = \mathcal{O}\left(W(t)\right), \tag{iii}$$

$$\mathbb{E}_t N_t^n = \mathcal{O}(e^{n\alpha t}), \quad n \in \mathbb{N}^*, \tag{iv}$$

$$\mathbb{P}_t\left(Z_0(t)=k\right) = \mathcal{O}(e^{(\alpha-\theta)t}). \tag{v}$$

Proof. Relation (i) is easily obtained using the expectation of N_t and $A(k,t)$ using (2.11), (2.13) and the behaviour of W provided by Proposition 2.3. The relation (iii) has been obtained in the proof of Theorem 6.1 in [8]. The two last relations are easily obtained from (2.4), (2.8) and Lemma 2.2. The relation (ii) is obtained using the following estimation,

$$\left|\mathbb{E}_t\left[(A(k,t)-c_kN_t)^3\right]\right| \leq \mathbb{E}_t\left[N_t\left(A(k,t)-c_kN_t\right)^2\right].$$

We begin the proof by computing the r.h.s. of the previous inequality using the same techniques as in Appendix A.

$$\begin{aligned}
\mathbb{E}\left[A(k,t)^2N_t\right] &= 2\int_0^t \theta\frac{W(t)}{W(a)}\mathbb{E}\left[N_aA(k,a)\mathbb{1}_{Z_0(a)=k}\right]da \\
&+4\int_0^t \theta\frac{W(t)^2}{W(a)^2}\left(1-\frac{W(a)}{W(t)}\right)\mathbb{E}\left[N_a\mathbb{1}_{Z_0(a)=k}\right]\mathbb{E}\left[A(k,a)\right]da \\
&+4\int_0^t \theta\frac{W(t)^2}{W(a)^2}\left(1-\frac{W(a)}{W(t)}\right)\mathbb{E}\left[A(k,a)\mathbb{1}_{Z_0(a)=k}\right]\mathbb{E}\left[N_a\right]da \\
&+4\int_0^t \theta\frac{W(t)^2}{W(a)^2}\left(1-\frac{W(a)}{W(t)}\right)\mathbb{P}_a\left(Z_0(a)=k\right)\mathbb{E}\left[A(k,a)N_a\right]da \\
&+12\int_0^t \theta\frac{W(t)^3}{W(a)^3}\left(1-\frac{W(a)}{W(t)}\right)^2\mathbb{P}_a\left(Z_0(a)=k\right)\mathbb{E}\left[A(k,a)\right]\mathbb{E}\left[N_a\right]da.
\end{aligned}$$

$$\begin{aligned}
2\mathbb{E}\left[A(k,t)N_t^2\right] &= 2\int_0^t \theta\frac{W(t)}{W(a)}\mathbb{E}\left[N_a^2\mathbb{1}_{Z_0(a)=k}\right]da \\
&+8\int_0^t \theta\frac{W(t)^2}{W(a)^2}\left(1-\frac{W(a)}{W(t)}\right)\mathbb{E}\left[N_a\mathbb{1}_{Z_0(a)=k}\right]\mathbb{E}\left[N_a\right]da \\
&+4\int_0^t \theta\frac{W(t)^2}{W(a)^2}\left(1-\frac{W(a)}{W(t)}\right)\mathbb{P}_a\left(Z_0(a)=k\right)\mathbb{E}\left[N_a^2\right]da \\
&+12\int_0^t \theta\frac{W(t)^3}{W(a)^3}\left(1-\frac{W(a)}{W(t)}\right)^2\mathbb{P}_a\left(Z_0(a)=k\right)\mathbb{E}\left[N_a\right]^2da.
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}\left[N_t\left(A(k,t)-c_kN_t\right)^2\right] &= 2\int_0^t \theta\frac{W(t)}{W(a)}\mathbb{E}\left[N_a\left(A(k,a)-c_kN_a\right)\mathbb{1}_{Z_0(a)=k}\right]da \\
&+4\int_0^t \theta\frac{W(t)^2}{W(a)^2}\left(1-\frac{W(a)}{W(t)}\right)\mathbb{E}\left[N_a\mathbb{1}_{Z_0(a)=k}\right]\mathbb{E}\left[A(k,a)-c_kN_a\right]da \\
&+4\int_0^t \theta\frac{W(t)^2}{W(a)^2}\left(1-\frac{W(a)}{W(t)}\right)\mathbb{E}\left[\left(A(k,a)-c_kN_a\right)\mathbb{1}_{Z_0(a)=k}\right]\mathbb{E}\left[N_a\right]da \\
&+4\int_0^t \theta\frac{W(t)^2}{W(a)^2}\left(1-\frac{W(a)}{W(t)}\right)\mathbb{P}_a\left(Z_0(a)=k\right)\mathbb{E}\left[N_a\left(A(k,a)-c_kN_a\right)\right]da \\
&+12\int_0^t \theta\frac{W(t)^3}{W(a)^3}\left(1-\frac{W(a)}{W(t)}\right)^2\mathbb{P}_a\left(Z_0(a)=k\right)\mathbb{E}\left[N_a\right]\mathbb{E}\left[A(k,a)-c_kN_a\right]da \\
&+c_k^2\mathbb{E}_tN_t^3.
\end{aligned}$$

Now, an analysis similar to the one of Lemma 5.5 leads to the result. □

Proof of Lemma 5.5. The ideas of the proof, is to analyses one to one every terms of the expression of

$$\mathbb{E}_t\left[(A(k,t)-c_kN_t)^4\right],$$

given by Lemma A.1 using Lemma B.1 to show that they behave as $\mathcal{O}\left(W(t)^2\right)$. Since the ideas are the same for every terms, we just give a few examples.

First of all, we consider

$$\int_{[0,t]}\frac{W(t)}{W(a)}\mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k}\left(A(k,a)-c_kN_a\right)^3\right]da.$$

Using Lemma B.1 (ii), we have

$$\int_{[0,t]}\frac{W(t)}{W(a)}\mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k}\left(A(k,a)-c_kN_a\right)^3\right]da=\mathcal{O}\left(W(t)^2\right).$$

Now take the term

$$\int_{[0,t]}\frac{W(t)^2}{W(a)^2}\mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k}N_a^2\right]\mathbb{E}_a\left[\left(A(k,a)-c_kN_a\right)\right]da,$$

we have from Lemma B.1 (i) and (iv),

$$\int_{[0,t]}\frac{W(t)^2}{W(a)^2}\mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k}N_a^2\right]\mathbb{E}_a\left[\left(A(k,a)-c_kN_a\right)\right]da\le\int_{[0,t]}\frac{W(t)^2}{W(a)^2}\mathbb{E}_a\left[N_a^2\right]e^{-(\theta-\alpha)a}da=\mathcal{O}\left(W(t)^2\right).$$

Every term in $W(t)$ or $W(t)^2$ are treated this way. Now, we consider the term in $W(t)^4$ which is

$$I:=96\int_{[0,t]}\frac{W(t)^4}{W(a)^4}\mathbb{P}_a\left(Z_0(a)=k\right)\mathbb{E}_a\left[\left(A(k,a)-c_kN_a\right)\right]^3da+24W(t)^4c_k^4,$$

since N_t is geometrically distributed under $\mathbb{P}_t$, and that

$$\mathbb{E}_tN_t^4=24W(t)^4-36W(t)^3+\mathcal{O}(W(t)^2). \tag{B.1}$$

On the other hand, using the law of $Z_0(t)$ given by (2.8) and the expectation of $A(k,t)$ given by (2.11) (under $\mathbb{P}_t$), we have,

$$\begin{aligned}
&96\int_{[0,t]}\frac{W(t)^4}{W(a)^4}\mathbb{P}_a\left(Z_0(a)=k\right)\mathbb{E}_a\left[\left(A(k,a)-c_kN_a\right)\right]^3da\\
&\quad=-96W(t)^4\int_0^t\frac{\theta e^{-\theta a}}{W_\theta(a)^2}\left(1-\frac{1}{W_\theta(a)}\right)^{k-1}\left(\int_0^a\frac{\theta e^{-\theta s}}{W_\theta(s)^2}\left(1-\frac{1}{W_\theta(s)}\right)^{k-1}ds\right)^3da\\
&\quad=-24W(t)^4\left(\int_0^t\frac{\theta e^{-\theta a}}{W_\theta(a)^2}\left(1-\frac{1}{W_\theta(a)}\right)^{k-1}da\right)^4.
\end{aligned}$$

Finally,

$$I = 24W(t)^4 \left(\int_t^\infty \frac{\theta e^{-\theta a}}{W_\theta(a)^2} \left(1 - \frac{1}{W_\theta(a)}\right)^{k-1} da \right)^4 = \mathcal{O}\left(W(t)^4 e^{-4\theta t}\right) = o(1).$$

The last example is the most technical and relies with the term in $W(t)^3$, which is, using (B.1) and Lemma A.1,

$$\begin{aligned} J :=& 72 \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{E}_a \left[\mathbb{1}_{Z_0(a)=k} \left(A(k,a) - c_k N_a\right) \right] \mathbb{E}_a \left[\left(A(k,a) - c_k N_a\right) \right]^2 da \\ &+ 72 \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{P}_a \left(Z_0(a) = k\right) \mathbb{E}_a \left[\left(A(k,a) - c_k N_a\right)^2 \right] \mathbb{E}_a \left[A(k,a) - N_a c_k \right] da \\ &- 288 \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{P}_a \left(Z_0(a) = k\right) \mathbb{E}_a \left[\left(A(k,a) - c_k N_a\right) \right]^3 da - 36 c_k^4 W(t)^3. \end{aligned}$$

On the other hand, using the calculus made in the proof of Theorem 6.3 of [8], we have

$$\begin{aligned} &\mathbb{E}_a \left[\left(A(k,a) - c_k N_a\right)^2 \right] \\ &= 4 \int_{[0,a]} \frac{W(a)^2}{W(s)^2} \left(1 - \frac{W(s)}{W(a)}\right) \mathbb{P}_s \left(Z_0(s) = k\right) \mathbb{E}_s \left(A(k,s) - c_k N_s\right) ds \\ &\quad + 2 \int_{[0,a]} \frac{W(s)}{W(a)} \mathbb{E}_s \left[\mathbb{1}_{Z_0(s)=k} \left(A(k,s) - c_k N_s\right) \right] ds + c_k^2 W(a)^2 \left(2 - \frac{1}{W(a)}\right). \end{aligned}$$

Substituting this last expression in J leads to

$$\begin{aligned} J =& -144 \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{E}_a \left[\mathbb{1}_{Z_0(a)=k} \left(A(k,a) - c_k N_a\right) \right] \int_{[a,\infty]} \frac{\mathbb{P}\left(Z_0(a) = k\right)}{W(s)^2} \mathbb{E}_s \left[\left(A(k,s) - c_k N_s\right) \right] ds da \\ &+ 144 W(t)^3 \int_{[0,t]} \frac{1}{W(a)} \mathbb{E}_a \left[\mathbb{1}_{Z_0(a)=k} \left(A(k,a) - c_k N_a\right) \right] \int_{[a,t]} \frac{1}{W(s)^2} \mathbb{P}_s \left(Z_0(s) = k\right) \mathbb{E}_s \left[A(k,s) - N_s c_k \right] da \\ &- 144 c_k^2 \int_{[0,t]} \frac{W(t)^3}{W(a)} \mathbb{P}_a \left(Z_0(a) = k\right) \mathbb{E}_a \left[A(k,a) - N_a c_k \right] da \\ &+ 144 \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{P}\left(Z_0(a) = k\right) \mathbb{E}_a \left[A(k,a) - N_a c_k \right]^3 da \\ &- 288 \int_{[0,t]} \frac{W(t)^3}{W(a)^2} \mathbb{P}_a \left(Z_0(a) = k\right) \int_{[0,a]} \frac{1}{W(s)} \mathbb{P}_s \left(Z_0(s) = k\right) \mathbb{E}_s \left(A(k,s) - c_k N_s\right) ds \mathbb{E}_a \left[A(k,a) - N_a c_k \right] da \\ &+ 72 \int_{[0,t]} \frac{W(t)^3}{W(a)} \mathbb{P}_a \left(Z_0(a) = k\right) c_k^2 \left(2 - \frac{1}{W(a)}\right) \mathbb{E}_a \left[A(k,a) - N_a c_k \right] da \\ &- 288 \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{P}_a \left(Z_0(a) = k\right) \mathbb{E}_a \left[\left(A(k,a) - c_k N_a\right) \right]^3 da - 36 c_k^4 W(t)^3. \end{aligned}$$

Using many times that,

$$\begin{aligned}
&\int_{[0,t]} \frac{\theta \mathbb{P}\left(Z_0(a)=k\right)}{W(s)^2} \mathbb{E}_s\left[\left(A(k,s)-c_k N_s\right)\right] ds \\
&\quad = -\int_{[0,t]} \frac{\theta e^{-\theta s}}{W_\theta(s)^2}\left(1-\frac{1}{W_\theta(s)}\right)^{k-1} \int_{[s,\infty]} \frac{\theta e^{-\theta u}}{W_\theta(u)^2}\left(1-\frac{1}{W_\theta(u)}\right)^{k-1} du ds \\
&\quad = \frac{c_k^2}{2} - \frac{1}{2}\left(\int_{[t,\infty]} \frac{\theta e^{-\theta s}}{W_\theta(s)^2}\left(1-\frac{1}{W_\theta(s)}\right)^{k-1} ds\right)^2,
\end{aligned}$$

thanks to (2.8), (2.11), and (2.6), we finally get

$$\begin{aligned}
J = &-144\left(c_k^2 - c_k(t)^2\right) \int_{[0,t]} \frac{W(t)^3}{W(a)^3} \mathbb{E}_a\left[\mathbb{1}_{Z_0(a)=k}\left(A(k,a)-c_k N_a\right)\right] da \\
&+ 36 W(t)^3 \left(c_k^2 \left(\int_{[t,\infty]} \frac{W(t)^3}{W(a)^3} \mathbb{E}_a\left[A(k,a)-N_a c_k\right]^3 da\right)^2 - \left(\int_{[t,\infty]} \frac{W(t)^3}{W(a)^3} \mathbb{E}_a\left[A(k,a)-N_a c_k\right]^3 da\right)^4\right) \\
&+ 144\left(c_k - c_k(t)\right)^2 \int_{[0,t]} \frac{W(t)^3}{W(a)} \mathbb{E}_a\left[A(k,a)-N_a c_k\right] da \\
&+ 36 W(t)^3 \left(c_k - c_k(t)\right)^4.
\end{aligned}$$

This shows that J is $\mathcal{O}\left(W(t)^2\right)$. □

References

[1] Joseph Abate, Gagan L Choudhury, and Ward Whitt. An introduction to numerical transform inversion and its application to probability models. In *Computational probability*, pages 257–323. Springer, 2000.

[2] Joseph Abate and Ward Whitt. A unified framework for numerically inverting laplace transforms. *INFORMS Journal on Computing*, 18(4):408–421, 2006.

[3] Anne-Laure Basdevant, Christina Goldschmidt, et al. Asymptotics of the allele frequency spectrum associated with the bolthausen-sznitman coalescent. *Electronic Journal of Probability*, 13:486–512, 2008.

[4] Julien Berestycki, Nathanaël Berestycki, Vlada Limic, et al. Asymptotic sampling formulae for Λ-coalescents. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 50, pages 715–731. Institut Henri Poincaré, 2014.

[5] Julien Berestycki, Nathanaël Berestycki, and Jason Schweinsberg. Beta-coalescents and continuous stable random trees. *The Annals of Probability*, pages 1835–1887, 2007.

[6] Jean Bertoin. The structure of the allelic partition of the total population for Galton-Watson processes with neutral mutations. *Ann. Probab.*, 37(4):1502–1523, 2009.

[7] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[8] Nicolas Champagnat and Henry Benoit. Moments of the frequency spectrum of a splitting tree with neutral poissonian mutations. *Electron. J. Probab.*, 21:34 pp., 2016.

[9] Nicolas Champagnat and Amaury Lambert. Splitting trees with neutral Poissonian mutations I: Small families. *Stochastic Process. Appl.*, 122(3):1003–1033, 2012.

[10] Nicolas Champagnat and Amaury Lambert. Splitting trees with neutral Poissonian mutations II: Largest and oldest families. *Stochastic Process. Appl.*, 123(4):1368–1414, 2013.

[11] Nicolas Champagnat, Amaury Lambert, and Mathieu Richard. Birth and death processes with neutral mutations. *Int. J. Stoch. Anal.*, pages Art. ID 569081, 20, 2012.

[12] Warren J. Ewens. *Mathematical population genetics. I*, volume 27 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, second edition, 2004. Theoretical introduction.

[13] J. Geiger and G. Kersting. Depth-first search of random trees, and Poisson point processes. In *Classical and modern branching processes (Minneapolis, MN, 1994)*, volume 84 of *IMA Vol. Math. Appl.*, pages 111–126. Springer, New York, 1997.

[14] R. C. Griffiths and Anthony G. Pakes. An infinite-alleles version of the simple branching process. *Adv. in Appl. Probab.*, 20(3):489–524, 1988.

[15] Benoît Henry. Central limit theorem for supercritical binary homogeneous crump-mode-jagers processes. *ESAIM: Probability and Statistics*, 21:113–137, 2017.

[16] Amaury Lambert. The contour of splitting trees is a Lévy process. *Ann. Probab.*, 38(1):348–395, 2010.

[17] Amaury Lambert, Lea Popovic, et al. The coalescent point process of branching trees. *The Annals of Applied Probability*, 23(1):99–144, 2013.

[18] Amaury Lambert and Pieter Trapman. Splitting trees stopped when the first clock rings and Vervaat's transformation. *Journal of Applied Probability*, 50(01):208–227, 2013.

[19] Mathieu Richard. *Arbres, Processus de branchement non Markoviens et processus de Lévy*. Thèse de doctorat, Université Pierre et Marie Curie, Paris 6.

[20] Pardis C Sabeti, David E Reich, John M Higgins, Haninah ZP Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, Jill V Platko, Nick J Patterson, Gavin J McDonald, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, 2002.