Estimating the Lambda measure in multiple-merger coalescents

Verónica Miró Pina[1,2,3], Émilien Joly[4], and Arno Siri-Jégousse[3*]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain.
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[3]Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas, Universidad Nacional Autónoma de México, CDMX, México.
[3]Centro de Investigación en Matemáticas, AC (CIMAT), Guanajuato, México
[*]e-mail: arno@sigma.iimas.unam.mx

Abstract

Multiple-merger coalescents, also known as Λ-colescents, have been used to describe the genealogy of populations that have a skewed offspring distribution or that undergo strong selection. Inferring the characteristic measure Λ, which describes the rates of the multiple-merger events, is key to understand these processes. So far, most inference methods only work for some particular families of Λ-coalescents that are described by only one parameter, but not for more general models. This article is devoted to the construction of a non-parametric estimator of Λ that is based on the observation at a single time of the so-called Site Frequency Spectrum (SFS), which describes the allelic frequencies in a present population sample. First, we produce estimates of the multiple-merger rates by solving a linear system, whose coefficients are obtained by appropriately subsampling the SFS. Then, we use a technique that aggregates the information extracted from the previous step through a kernel type of re-construction to give a non-parametric estimation of the measure Λ. We give a consistency result of this estimator under mild conditions on the behavior of Λ around 0. We also show some numerical examples of how our method performs.

1 Introduction

Inferring the evolutionary history of a sample of a present-time population is a very challenging task for both biological and mathematical communities. If no information is available on the past of the population, it is natural to consider Markov processes to model its genealogical tree. The family of Λ-coalescents, also called multiple-merger coalescents [Donnelly and Kurtz, 1999, Pitman, 1999, Sagitov, 1999], provides a rich class of processes to this purpose. They appear naturally as limit genealogies in population models, with possibly highly variable reproductive success [Sagitov, 1999]. Their flexibility (and richness) comes from the fact that their dynamics are fully described by a measure Λ on the interval $[0,1]$. The intrinsic exchangeability of the branches at a given time in Λ-coalescents makes them particularly suited for modeling genealogies in a very general framework. The family of Λ-coalescents contains

the Kingman coalescent [Kingman, 1982b,a], which is considered as the neutral model for random genealogies. We refer to Berestycki [2009], Gnedin et al. [2014] for a mathematical survey that contains most of the justifications of the classical formulas that we use in this paper.

Some subclasses of Λ-coalescent, such as Beta-coalescents [Schweinsberg, 2003] or Dirac-coalescents [Eldon and Wakeley, 2006], are intensively used in practice since they are quite easily interpretable and are characterized by only one parameter. The inference task is then reduced to the estimation of this parameter, that represents the level of reproductive skewness in the population. These one-parameter families are now well admitted to fit to populations with highly skewed offspring distribution, e.g. marine species [Sigurgíslason and Árnason, 2003, Niwa et al., 2016], tuberculosis cells [Menardo et al., 2020] or cancer cells [Kato et al., 2017]. Additionally, the class of Beta-coalescents contains the Bolthausen-Sznitman coalescent that provides a genealogical process for modeling populations under strong selection effects [Neher and Hallatschek, 2013, Desai et al., 2013, Schweinsberg, 2017, Cortines and Mallein, 2017, Schertzer and Wences, 2023].

Most of the previous inference work on Λ-coalescents has focused on estimating the characterizing parameter of the one parameter families cited above [Birkner et al., 2013, Eldon et al., 2015, Koskela, 2018, Hobolth et al., 2019]. But the inference of more general Λ measures is still a challenging issue. To our knowledge, it was only treated in Koskela et al. [2018], with Bayesian non-parametric methods. The authors consider a forward in time model of type frequencies, more adapted to time series observations, that are difficult to obtain from natural populations. In this paper we present an estimator that uses only contemporaneous observations.

More general models for populations with varying size [Freund, 2020], under bottlenecks [González Casanova et al., 2021], with diploid reproduction mechanism [Birkner et al., 2018], or recombination, are part of a larger class of coalescents, offering more possibilities for inference and model selection (see for example [Eldon et al., 2015, Koskela, 2018, Koskela and Wilke Berenguer, 2019, Freund and Siri-Jégousse, 2021] although they are not the purpose of our present study.

In this work, we use approximations via Bernstein polynomials to establish a non-parametric estimator of the measure Λ, which is built from estimates of the transition rates of the Λ-coalescent process. These estimates are the solutions of a linear system where the coefficients are the expected branch lengths (or SFS) starting from different initial configurations of the sample. This technique relies on generalizations of classical recursions for functionals of the coalescent (see Proposition 2.1).

The type of observations that we use are SNP[1] matrices that summarize mutation data: at every locus (column) where a mutation is observed, the individuals (lines) are assigned a 1 if they carry the mutation and a 0 otherwise (see Table 1 as an example). From the SNP matrix we can compute the SFS, which is the vector such that its i-th component contains the number of observed mutations shared by exactly i individuals in the sample. This is a statistic widely used in population genetics, both for model selection [Eldon et al., 2015, Freund and Siri-Jégousse, 2021] and for parameter inference [Fu, 1995, Kersting et al., 2021]. In this paper, we use an extension of the SFS, the so-called *weighted SFS*, which can also be read from the SNP matrix by subsampling the individuals (lines), as explained at the end of the Results section.

In addition, we assume given a collection of SNP matrices $S_1, \ldots, S_n$ governed by the same subjacent Λ-coalescent. We also assume that those matrices are independent (in the probabilistic sense) so that no coalescence in one genealogical branch of one matrix S_i have an influence on the other matrices. For

[1] Single Nucleotide Polymorphisms (SNP) are mutations that affect one single base position in the DNA. For the purpose of this paper, SNPs can be substitutions or indels.

example, one can think of SNP matrices obtained from different replicates in an evolutionary experiment, that have evolved independently, under similar conditions (see for example Lenski et al. [1991] for *E. coli*, Johnson et al. [2021] for yeast or Teotónio et al. [2017] for *C. elegans*). For applications to real populations, it is not possible to obtain independent SNP matrices from the same population. However, ignoring dependencies between genomic regions that are at sufficient recombination distance is a natural approximation and it has been used for example in McVean and Cardin [2005] to define the sequentially Markov coalescent. In our case, one can make a similar approximation by considering that SNP matrices obtained from different chromosomes are *almost* independent. By computing the weighted SFS from this collection of almost independent SNP matrices, we can finely estimate the transition rates of the Λ-coalescent.

Summary of the method

In order to obtain a non-parametic estimator of the characteristic measure Λ, one has to follow these steps

1. For each SNP matrix, compute the weighted SFS using the technique provided in (9)
2. By averaging over the weighted SFS obtained from the different SNP matrices, obtain an estimator of the branch lengths of the tree, as given in (10).
3. Solving the linear equations in Proposition 2.1, obtain estimates of the transition rates of the Λ-coalescent.
4. Use equation (6) to get an estimator of the coalescent measure. Consistency of this estimator is proven in Theorem 2.1.

2 Results

2.1 Definition of the estimator

The Λ-coalescents are exchangeable continuous-time Markov chains taking their values in the partitions of the integers. The chain jumps from one state to another by coagulating some blocks of the partition to form one new block. The exchangeability implies that the jump rates only depend on the number of blocks of the partitions. Thus, the dynamics of this chain are totally described by the array $(\lambda_{k,j})_{k\geq j\geq 2}$ where $\lambda_{k,j}$ stands for the rate at which a given j-tuple, among k present blocks, coalesces into one new block. Moreover, the class of exchangeable coalescents is consistent. This means that a coalescent starting from a partition π made of n blocks, and restricted to the set of partitions of $[m]$, with $m \leq n$, has the same law as a coalescent starting from the partition π restricted to $[m]$. This consistency property implies that, for $k \geq j \geq 2$,

$$\lambda_{k,j} = \lambda_{k+1,j} + \lambda_{k+1,j+1} \tag{1}$$

which is equivalent to

$$\lambda_{k,j} = \int_0^1 x^{j-2}(1-x)^{k-j}\Lambda(dx) = \int_0^1 x^{j}(1-x)^{k-j}\nu(dx) \tag{2}$$

where ν is a measure on $[0,1]$ satisfying $\int_0^1 x^2\nu(dx) < \infty$, and $\Lambda(dx) = x^2\nu(dx)$. Equation (2) is the fundamental link between the unknown measure Λ and the coalescent rates $(\lambda_{k,j})_{k\geq j\geq 2}$. One has to note that both measures Λ and ν characterize the coalescent, nevertheless the measure ν is easier to estimate than the measure Λ. Besides this fact, the estimation of ν can only insure a good estimation outside of an open interval containing 0 due to the presence of x^{-2} in its definition. In this work, we suppose that the measure ν is absolutely continuous with respect to the Lebesgue measure over the interval $(0,1]$, that is $\nu(dx) = \nu(x)dx$.

The rate at which the number of blocks of the coalescent jumps from k to j is

$$r_{k,j} = \binom{k}{j}\lambda_{k,j}. \tag{3}$$

These new rates are then referred to as the combinatorial coalescent rates and the total transition rate from a state with k blocks is given by

$$d_k = \sum_{j=2}^{k} r_{k,j}.$$

This section is focused on the practical reconstruction of the measure ν when estimations $(\hat{r}_{m,j})_{2\leq j\leq m}$ of the $(r_{m,j})_{2\leq j\leq m}$ are given. Note that the number m is fixed so that the only information needed is a good estimation of the coalescent rates of a group of m individuals. This reconstruction can be seen as a map $(\hat{r}_{m,j})_j \mapsto \hat{\nu}_{m,h}$ taking values in the continuous functions on $(0,1]$, where h holds for a window hyperparameter in a kernel estimation technique that we explain in details below.

The particular form of the coefficient $\lambda_{m,j}$ obtained from (2) allows us to use Bernstein polynomials as a technical tool for the approximation step. The Bernstein polynomial basis is defined as a double index sequence given by

$$B_j^m(x) = \binom{m}{j} x^j (1-x)^{m-j}$$

for all $0 \leq j \leq m$. In the following, we call Bernstein polynomial of order m a polynomial that belongs to the vector space spanned by the $(B_j^m)_{0\leq j\leq m}$ where m is fixed. The ideas behind the construction of the estimator $\hat{\nu}_{m,h}$ of ν are the following.

1. Choose a non-negative even function ϕ that is continuous, has a compact support containing 0 and such that $\int_{\mathbb{R}} \phi(x)dx = 1$. Such a function is called a regularizing function. Without loss of generality, we assume that the support of ϕ is included in a segment of the form $[-\tau,\tau]$. For $h > 0$, define $K_h(x) = h^{-1}\phi(h^{-1}x)$. The function K_h is called a kernel of window h. Classical functional theorems show that a continuous function f is well approximated by its smoothed version given by the convolution $f \star K_h$ defined, for any fixed x, by

   $$f \star K_h(x) = \int_{\mathbb{R}} K_h(x-y)f(y)dy \underset{h\to 0}{\longrightarrow} f(x).$$

   The convolution will apply on the function ν that has support included in $(0,1]$ so the integration of the function $K_h^x : y \to K_h(x-y)$ is only on the segment $(0,1]$.

2. Since the function K_h^x is smooth on its compact support, it can be approximated by a sequence of Bernstein polynomials $P_{m,h}^x$ so that

$$\|K_h^x - P_{m,h}^x\|_\infty \underset{m\to\infty}{\longrightarrow} 0.$$

Those polynomials have the explicit definition

$$P_{m,h}^x(y) = \sum_{j=0}^{m} \binom{m}{j} K_h^x(\frac{j}{m}) y^j (1-y)^{m-j}.$$

3. According to points 1. and 2., the quantity

$$\int_0^1 P_{m,h}^x(y)\nu(y)dy \tag{4}$$

is a good candidate for the estimator of $\nu(x)$. From the definition in (4), it is not clear to see that it is completely defined by the values of the $r_{m,j}$. But it can be rewritten in the following way

$$\int_0^1 P_{m,h}^x(y)\nu(y)dy = \sum_{j=0}^{m} \binom{m}{j} K_h^x\left(\frac{j}{m}\right) \int_0^1 y^j(1-y)^{m-j}\nu(y)dy. \tag{5}$$

In the sum of the right hand side, the terms for $j=0$ or $j=1$ won't lead to quantities that only depend on $r_m = (r_{m,j})_{2\le j\le m}$. This is not a problem if one takes x to be outside of the support of K_h for the terms $j=0,1$. This leads to the restriction $x - m^{-1} \ge h\tau$. Then, we define

$$\nu_{m,h}(x) = \sum_{j=2}^{m} K_h^x\left(\frac{j}{m}\right) r_{m,j},$$

if $x \in [m^{-1}+\tau h, 1]$ and 0 elsewhere.

4. One has to note that $\nu_{m,h}(x)$ is still not an estimator of $\nu(x)$ since it is not computable only with respect to the estimated values $\hat{r}_m := (\hat{r}_{m,j})_{2\le j\le m}$. From (5) and (4), for any $x \in [m^{-1}+\tau h, 1]$, we define the final estimator $\hat{\nu}_{m,h}$ by a plug in of the estimated values $(\hat{r}_{m,j})_{2\le j\le m}$,

$$\hat{\nu}_{m,h}(x) = \sum_{j=2}^{m} K_h^x\left(\frac{j}{m}\right) \hat{r}_{m,j} \tag{6}$$

and we define $\hat{\nu}_{m,h}(x) = 0$ for $x < m^{-1}+\tau h$. In particular, we see that the interval $[m^{-1}+\tau h, 1]$ tends to $(0,1]$ when we let $m\to\infty$ and $h\to 0$.

We establish in Theorem 2.1 that this estimator is consistent and we provide some control on the error. For that purpose we need to define the modulus of continuity. For a function $f:[0,1]\to\mathbb{R}$ we define, for any $\delta > 0$,

$$\omega_f(x,\delta) = \sup_{y:|x-y|\le\delta} |f(x)-f(y)|.$$

Theorem 2.1. *Let $\tau > 0$, $h > 0$ and $m \geq 2$. Let ϕ be a non-negative α-Hölder function of support $[-\tau, \tau]$ and such that $\phi(x) \leq \mathbb{1}_{\{[-\tau,\tau]\}}(x)$. Define $K_h(x) = h^{-1}\phi(h^{-1}x)$ and for $x \in [m^{-1} + \tau h, 1]$, let $\hat{\nu}_{m,h}(x)$ as in* (6)*. Then,*

$$\begin{aligned} \|\hat{\nu}_{m,h} - \nu\|_{1,[m^{-1}+h\tau,1]} \leq & \sup_{x\in[m^{-1}+\tau h,1]} \omega_\nu(x,\tau h) + C\Lambda[0,1/m] \\ + & \ 1.5c\frac{\nu[1/m,1]}{h^{1+\alpha}m^{\alpha/2}} + \|\hat{r}_m - r_m\|_1 \end{aligned}$$

where $\|\cdot\|_{1,[m^{-1}+h\tau,1]}$ holds for the L_1 norm on the interval $[m^{-1} + h\tau, 1]$, and C, c are some positive constants.

The proof of this Theorem can be found in Appendix A.

In words, there are four different sources of error in the reconstruction of ν. The first one comes from the modulus of continuity of ν. Since the measure ν is continuous, this quantifies how much it varies in a small interval of size τh. This is the most general way of stating the theorem. In most of practical cases, the function ν is sufficiently regular to weaken this assumption. In particular, if ν is Lipschitz of constant k, we can replace the term $\omega_\nu(x,\delta)$ by $k\delta$ when x and δ are such that $x - \delta \geq m^{-1} + \tau h$ to insure to avoid any problem of discontinuity. Taking $\delta = \tau h$, the window parameter h can be seen as a smoothing effect parameter. As every smoothing parameter it will have to be balanced in such a way to avoid over-estimation or under-estimation. This first term in the bound drives us in taking h the smallest possible whereas the third term in the bound is governing h the other way around. The second source of error comes from the mass of Λ close to 0. The third one to the mass of ν in $[1/m, 1]$. Observe that, since $\int_0^1 x^2\nu(dx) < \infty$, we have that $\nu[1/m,1] = o(m^2)$. This implies that we can always find a Hölder exponent α such that the term $\nu[1/m,1]/(h^{1+\alpha}m^{\alpha/2})$ converges to 0. This also reflects the trade-off between the regularity of ϕ and the behavior of ν near 0. The fourth source of error is error in the estimation of the $r_{m,j}$'s. The next section is devoted to providing a method to estimate these coalescent rates.

2.2 Techniques for the estimation of the coalescent rates

Let us turn to the problem of estimating the $m-1$ combinatorial coalescent rates $(r_{m,j})_{2\leq j\leq m}$. The estimator $\hat{r}_m = (\hat{r}_{m,j})_{2\leq j\leq m}$ is defined as the solution of a linear system of $m-1$ equations. Set $e_j = (0,\ldots,0,1,0,\ldots,0)$ where the value 1 is at j-th coordinate.

To do so, we consider the set of partitions of the integer m, $\mathscr{C}_m = \{\bar{c} = (c_1,\ldots,c_m), \sum_j jc_j = m\}$. We denote the number of blocks of $\bar{c} \in \mathscr{C}_m$ by $|\bar{c}| = \sum_j c_j$. Then, we define the multi-dimensional block counting process of the coalescent with initial configuration $\bar{c}$ as $(\Pi(t,\bar{c}))_{t\geq 0} = (N_1(t,\bar{c}),\ldots,N_m(t,\bar{c}))_{t\geq 0}$, where $N_j(t,\bar{c})$ holds for the number of blocks of size j in the coalescent at time t. The total number of blocks in the coalescent process at time t is then given by $|\Pi(t,\bar{c})| = \sum_j N_j(t,\bar{c})$.

The partial length of order j ($1 \leq j \leq m-1$) of the coalescent starting from $\bar{c}$ is defined by

$$L_j(\bar{c}) = \int_0^{D(\bar{c})} N_j(t,\bar{c})dt$$

where $D(\bar{c})$ is the depth of the tree, also called time to the most recent common ancestor (TMRCA), $D(\bar{c}) = \inf\{t \geq 0, \Pi(t,\bar{c}) = e_m\} = \inf\{t \geq 0, |\Pi(t,\bar{c})| = 1\}$.

Proposition 2.1. *Let $\bar{c}_0 = (c_1,\ldots,c_m) = m.e_1 = (m,0,0,\ldots)$ and $\bar{b}_j = (m-j).e_1 + e_j, j \in \{2,\ldots,m\}$. Then we have. for $i \in \{2,\ldots,m\}$,*

$$d_m \mathbb{E}[L_i(\bar{c}_0)] = c_i + \sum_{j=2}^{m-1} r_{m,j} \mathbb{E}[L_i(\bar{b}_j)], \quad i \in \{1, \dots, m-1\}. \tag{7}$$

Equations (7) provide a linear system of $m-1$ equations for $m-1$ unknown values $(r_{m,j})_{2 \le j \le m}$ which can be rewritten as

$$AR = B,$$

where A is a $(m-1) \times (m-1)$ matrix such that the last column contains the expected branch lengths, $(\mathbb{E}[L_i(\bar{c}_0)])^T_{1 \le i \le m-1}$ and for $j \in \{1, m-2\}$, the j-th column contains the difference between the expected branch lengths and partial lengths starting from b_j, i.e. $(\mathbb{E}[L_i(\bar{c}_0)] - \mathbb{E}[L_i(\bar{b}_j)])^T_{1 \le i \le m-1}$; B is the vector $(m, 0, \dots, 0)$ and $R = (r_{m,j})_{2 \le j \le m}$.

The precision in the estimation of the vector r_m will thus depend on the precision of the estimation of the expectations in (7). The proof of Proposition 2.1 is a consequence of the Markov property applied at the first jump time (see Appendix B).

Remark 2.1. *There is a classical relation between the rates $(r_{m,j})_{2 \le j \le m}$ and the expectation of the total branch lengths of the associated coalescents starting from $j \in \{2, \dots, m\}$ individuals, denoted by $B(j)$. It is also obtained thanks to the Markov property applied at the first jump time:*

$$d_m \mathbb{E}[B(m)] = m + \sum_{j=2}^{m-1} r_{m,m-j+1} \mathbb{E}[B(j)]. \tag{8}$$

A first approach to estimate the coalescence rates would consist in considering this equation and add a number of recursive equations (1) *to obtain a linear system of $m-1$ equations. However this very rigid strategy, based only on the total number of blocks of the coalescent, leads to uncontrolled error as the error resulting from the estimate of $r_{m,2}$ will spread additively in the rest of the estimates. The strategy based on recursive equations for the partial lengths of the coalescent, as proposed in Proposition 2.1, leads to a more stable system.*

2.3 Obtaining the partial lengths from the SNP matrix

In this section we provide a method to estimate the expected values in (7). We assume that mutations happen at rate θ, as a Poisson process along the branches of the genealogical tree resulting from the coalescent process considered in the previous section. This mutation rate can also be seen as a scaling parameter since the law of the mutations is not changed if one multiplies this parameter θ and the transition rates are all multiplied by a common constant. Since the data is only considering the mutations affecting individuals in the sample, this scaling effect is untraceable, resulting in the model being not identifiable. To avoid dealing with those issues, we fix θ to be 1 in the sequel, which corresponds to measuring time in units of mutation rate. In the infinite sites model, we suppose that each mutation affects a distinct site so that they can all be observed in present-time data.

We assume provided a sample of m individuals, whose genotypes are encoded in vectors $X_1, \dots, X_m$. Formally, the vector X_j describing the genotype of individual j is a vector $X_j = (X_{j,k})_k \in \{0,1\}^p, p \in \mathbb{N}$ of binary values where $X_{j,k} = 1$ if and only if the individual j carries the mutation at site k (also called the SNP k). All this information is stored in a SNP matrix, where the vectors X_j are written as rows.

See Table 1 for an example. Observe that the size of the vectors is not relevant here as the matrix can be adjusted by withdrawing columns of 0's, which are the sites where no genetic variation is detected.

We also define the site frequency spectrum $(M_1(\bar{c}_0), M_2(\bar{c}_0), \ldots, M_{m-1}(\bar{c}_0))$ as the vector such that $M_i(\bar{c}_0)$ counts the number of mutations carried by exactly i individuals, that is $M_i(\bar{c}_0) = \text{Card}\{k, \sum_{j=1}^m X_{j,k} = i\}$. Also, the total number of mutations observed in the sample is $M(\bar{c}_0) = \text{Card}\{k, \sum_{j=1}^m X_{j,k} \geq 1\}$. Notice that these quantities depend on the initial configuration $(\bar{c}_0)$. In fact, considering a sample of m different individuals corresponds to starting the coalescent process from m singletons, i.e. from $\bar{c}_0 = m.e_1 = (m, 0, \ldots, 0)$. Our assumption on $\theta = 1$ leads to

$$\mathbb{E}[M_i(\bar{c}_0)] = \mathbb{E}[L_i(\bar{c}_0)].$$

As a result, any estimator of the expected value $\mathbb{E}[M_i(\bar{c}_0)]$ is also an estimator of the expected value of the partial length of order i.

Ind.	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
X_1	1	0	0	0	0	0	1
X_2	1	0	0	0	0	1	1
X_3	1	0	0	0	0	1	1
X_4	0	0	1	1	0	0	0
X_5	0	0	1	1	1	0	0
X_6	0	1	0	0	0	0	0

Table 1: Example of a SNP matrix, where 0 denotes the ancestral type and 1 denotes the derived type. Here the sample size is $m = 6$ and the total number of mutations is $M(\bar{c}_0) = 7$.

In order to make use of the linear system (7), one has to be able to estimate the partial length $\mathbb{E}[L_i(\bar{b}_j)]$ of a coalescent tree started at any partition $\bar{b}_j$ of the integer m of the form $\bar{b}_j = (m-j)e_1 + e_j, \quad j \in \{2, \ldots, m\}$. This can be done by a trick on the site frequency spectrum based on the consistency and the Markov properties of the coalescent. Indeed, the law of the coalescent tree starting from a partition $\bar{c}$ is the same than that of the coalescent tree starting from $|\bar{c}|$ individuals where individuals are associated to weights given by the partition $\bar{c}$. We will therefore define a *weighted* SFS, by subsampling m' individuals with replacement. The weight of each individual is given by the number of times it appears in the subsample. A mutation carried by an individual with weight j will count as a mutation carried by j individuals. The weighted SFS corresponding to a subsample with weights $\bar{c}$ is denoted by $(M_1(\bar{c}), \ldots, M_{m'-1}(\bar{c}))$.

Let us give an example using the SNP matrix of Table 1. In this matrix $m = 6$, $M(\bar{c}_0) = 7$ and the observed SFS is $(2, 3, 2, 0, 0)$. Now, think of $\bar{c} = \bar{b}_3 = 3e_1 + e_3$, i.e. a sample where three individuals have weight 1 and one individual has weight 3. We chose at random one individual to which we associate the block of size 3, e.g. X_1 and three individuals to which we associate weight 1 (X_2, X_3 and X_4). Then the weighted version of X_1 is $(3, 0, 0, 0, 0, 0, 3)$. The corresponding weighted SFS is $(2, 1, 0, 0, 2)$. Recall that, due to the exchangeability of the coalescent process, the law of the weighted SFS does not depend on the labels of the individuals which are assigned to each weight.

Formally, for a general partition $\bar{c}$ of any integer $m' \geq m$ such that $|\bar{c}| \leq m$, let $(h_1, \ldots, h_m)$ be any partition vector associated to $\bar{c}$, such that $\sum h_j = m'$, $\text{Card}\{j, h_j = i\} = c_i$ and $\text{Card}\{j, h_j = 0\} = m - |\bar{c}|$. Then,

$$M_i(\bar{c}) = \text{Card}\{k, \sum_{j=1}^{m} h_j X_{j,k} = i\}. \tag{9}$$

Observe that $m' = m$ in the example above.

Now, suppose that we can observe n i.i.d. SNP matrices of m individuals and that we can obtain n realizations $(M_1^k(\bar{c}), \ldots, M_{m'-1}^k(\bar{c}))$ for $k \in \{1, \ldots, n\}$. The estimator of $\mathbb{E}[L_i(\bar{c})]$ is given by

$$\hat{L}_i(\bar{c}) = \frac{1}{n} \sum_{k=1}^{n} M_i^k(\bar{c}). \tag{10}$$

3 A numerical example

We apply our method to simulated data, where the coagulation measure Λ is known, so we can compare the estimated measure $\hat{\nu}_{m,h}$ to the true measure ν. The goal of this section is not to quantify the error in the reconstruction, which is already bounded in Theorem 2.1, but rather to illustrate our method.

We used the population genetics simulator `msprime` [Baumdicker et al., 2022, Kelleher et al., 2016]. There are two Λ-coalescent models available in `msprime`: the Beta-coalescent [Schweinsberg, 2003] and the Dirac-coalescents [Eldon and Wakeley, 2006]. Since we have assumed that the Λ measure is continuous with respect to the Lebesgue measure, we chose the Beta-coalescent model. It is a one parameter family in which the characterizing parameter α denotes the degree of skewness of the offspring distribution of the population. Smaller values of α correspond to a more skewed offspring distribution, while $\alpha \to 2$ corresponds to the Kingman coalescent. For this example we fixed $\alpha = 1.2$, but we obtain qualitatively similar results for other values such that $1 < \alpha < 2$ (which is the range of values allowed by `msprime`). Observe that in this model the characteristic measure $\nu(x)$ behaves like $x^{-1-\alpha}$ near 0, which implies it has no first moment, providing a "worst-case scenario" for the non-parametric estimation.

We simulated the SNP matrix of a population of size m. We estimated the partial lengths by computing the SFS and the weighted SFSs, as explained in the previous section. We averaged over n independent replicates. Then we solved the linear system (7) in Proposition 2.1. Since some of the coalescent rates (the $r_{m,j}$'s) are small, sometimes solving the linear system $AR = B$ with the estimated values of the partial lengths $\hat{L}_i(\bar{c})$'s yielded negative values of the $r_{m,j}$'s, so, instead we used `scipy.optimize`, to minimize $AR - B$ with the constraint that the $r_{m,j}$'s must be positive.

To reconstruct the measure ν, we followed our method using a Gaussian kernel for K_h. As we can observe from Figure 1, even when the true values of the $r_{m,j}$'s were used for the reconstruction, there is a rather large error close to 0. This is because the family of Λ measures that we were able to simulate using `msprime` have a lot of mass close to 0. We show results for two different values of h. We observe that the error in the reconstruction depends on h, and that for $m = 50$, $h = 0.02$ is a good parameter choice.

In a second step, we applied the whole method to our simulated data, i.e. we first estimated the $r_{m,j}$'s from our simulated SNP matrices and then we reconstructed the measure ν.

We simulated different values of the sample size m and (population-size rescaled) mutation rate θ. Our method does not allow to estimate the mutation rate θ, but as discussed in the previous section, changing θ corresponds to changing the time-scale, i.e. multiplying the $r_{m,j}$'s by some constant. Since the aim of this project is not to get this constant, in order to visualize the results, we normalized the measures by the integral of Λ in $[m^{-1} + h\tau, 1]$, which is an approximation for the total coalescence rate, i.e. the constant we need.

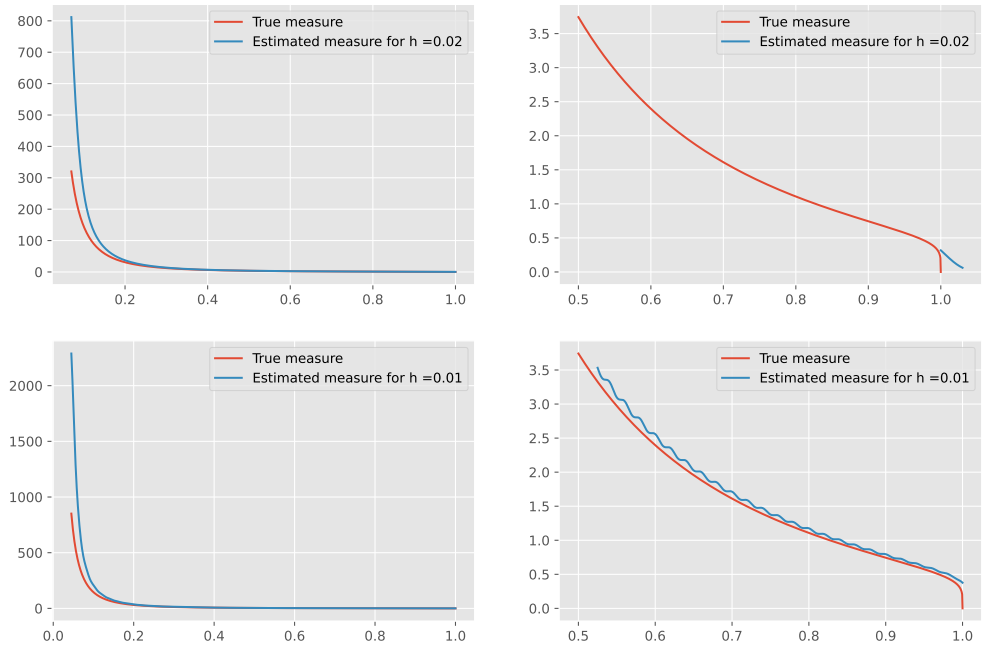

Figure 1: **Estimation of ν from the coalescent rates**. Here the true value of the $r_{m,j}$'s for a Beta-coalescent with $\alpha = 1.2$ and $m = 50$ is used to reconstruct the measure ν, using two different values of h. Panels on the right are zoomed versions of the panels on the left (notice the difference in the x-axis).

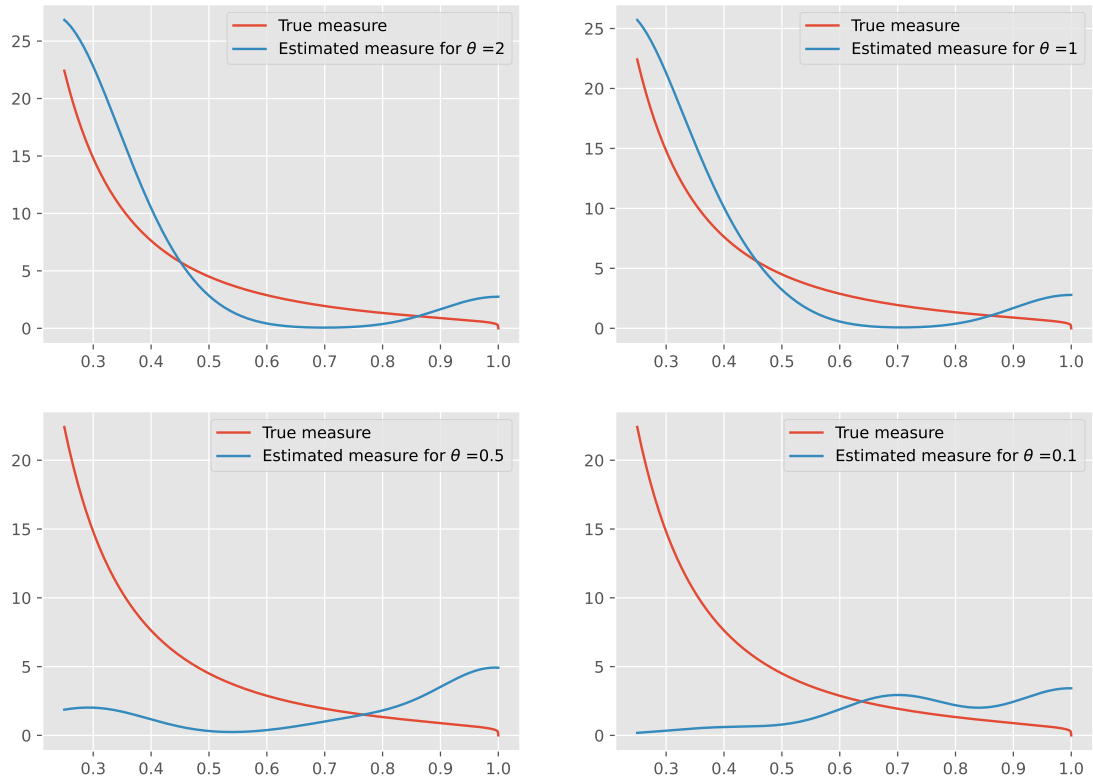

Figure 2: **Estimation of ν from simulated SNP matrices**. Here $n = 20$, $m = 10$, $h = 0.1$, $\tau = 2$.

We fixed the number of independent replicates to $n = 20$. This would correspond, for a species with 20 chromosomes, to consider that each chromosome is an almost independent replicate. Figure 2 shows the obtained results, which strongly depend on the mutation rate. For θ larger than 1 (which corresponds more or less to 500 mutations per sample in our model), the estimation of ν is quite good, in the sense that it has a similar behavior to the true ν. However, for lower mutation rates, the estimated measure does not match with the real measure.

If we increase the number of independent samples, the quality of our estimates improves (Figure 3). However, obtaining such a large number of almost independent samples, with enough mutation on each requires either a high-mutation high-recombination regime that is difficult to observe from real data. This parameter regime seems more likely to be achieved if the independent samples are replicates from an evolutionary experiment.

Finally, increasing m does not necessarily provide better estimates. This is because the effect of increasing the sample size is counteracted by an increase in the size of the linear system given by (7) and

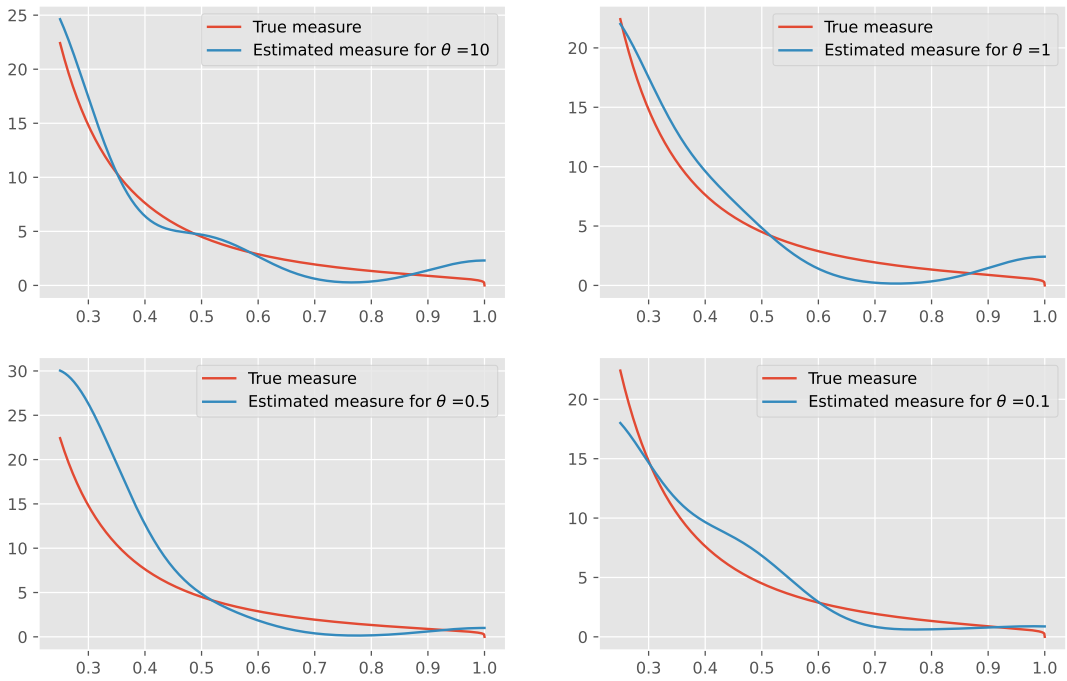

Figure 3: **Estimation of ν from simulated SNP matrices**. Here $n = 100$, $m = 10$, $h = 0.1$, $\tau = 2$.

the errors in the estimation of the partial lengths propagate through the system.

4 Discussion

We have developed the first non-parametric estimator of the measure Λ of Λ-coalescents from observations taken at a single time point. Thanks to this method we don't have to rely on a one parameter family, such as Beta-coalescents, so we can cover many more models and scenarios of evolution. We are also able to quantify the consistency of this estimator.

The method works in three steps. First, we estimate the branch lengths of the coalescent from SNP matrices thanks to a new observable that we called the weighted SFS. Second, we obtain the coalescent rates from the branch lengths by inverting the linear system (7). Third, we estimate the coalescent measure from the coalescent rates thanks to a non-parametric method based on Bernstein polynomials, as explained in the first paragraph of the Results section. Our (new) mathematical results focus only on step 3. Notice that steps 1 and 2 may be useful for other statistical techniques, in particular if SFS based statistics are needed, or if coalescent rates are needed.

The error generated at step 1 does only come from the central limit theorem so it is minimized if we can observe data from a large number of independent replicates or if the mutation rate is high. It can increase a lot if those assumptions do not hold. To obtain independent or almost independent samples we can use: different replicates of an evolutionary experiment or mutations in different chromosomes or different segments of chromosome assuming that recombination is large enough so that different segments are almost independent (as in SMC, see McVean and Cardin [2005]). We tried the weighted SFS method by considering distinct samples from a same population as independent samples. To do so, we first simulated the ancestry of a very large population using `mrprime` and then subsampled individuals with or without replacement. This yielded large errors in the estimation of the expected branch lengths. This is because of the shared evolutionary history between individuals of a same population, that makes the ancestry trees of the different samples highly correlated (especially for the deeper nodes).

The error produced at step 2 is due to the propagation of the error in step 1 at the moment of inverting a linear system. This part might be improved by using some more efficient numerical tools, or by adding constraints or equations in the system, e.g. recursive equations derived from (1). A more promising track to follow is that we may need not all the rates $r_{m,j}$ since they get too small when j goes close to m and do not bring substantial information for the estimation of Λ.

Finally, we managed to bound the error generated at step 3 in Theorem 2.1 thanks to the theory of non-parametric statistics. The error naturally explodes close to 0 when the measure Λ accumulates too much mass at 0. This is typically the case in Beta-coalescents with parameter α between 1 and 2. Thus, obtaining a non-explosive measure close to 0 could also mean that the evolution of the population does not fall into the domain of attraction of classical coalescents.

In this work we did not focus on the estimation of the mutation rate θ. Recall that it is not possible to discriminate between the couple (θ, Λ) and $(c\theta, c\Lambda)$ for some constant c from one date set taken at a single time. We refer to Eldon et al. [2015], Freund and Siri-Jégousse [2021] for some further techniques involving estimation of the mutation rate and model selection.

5 Code availability

Code for the numerical example is available at : `https://github.com/vmiropina/bootstrap_coalescent/`.

6 Funding

Most of this work was conducted when VMP was a postdoc at UNAM, funded by the DGAPA-UNAM postdoctoral program. VMP also acknowledges support of the Spanish Ministry of Science and Innovation to the EMBL partnership, the Centro de Excelencia Severo Ochoa and the CERCA Programme / Generalitat de Catalunya. The research of EJ was partially funded by CONACYT, Ciencia de Frontera project 1043167. The research of ASJ was partially funded by UNAM-DGAPA PAPIIT grant IN104722.

References

F. Baumdicker, G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, G. Tsambos, S. Zhu, B. Eldon, E. C. Ellerman, J. G. Galloway, A. L. Gladstein, G. Gorjanc, B. Guo, B. Jeffery, W. W. Kretzschumar, K. Lohse, M. Matschiner, D. Nelson, N. S. Pope, C. D. Quinto-Cortés, M. F. Rodrigues, K. Saunack, T. Sellinger, K. Thornton, H. van Kemenade, A. W. Wohns, Y. Wong, S. Gravel, A. D. Kern, J. Koskela, P. L. Ralph, and J. Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3):iyab229, 2022.

N. Berestycki. Recent progress in coalescent theory. *Ensaios Matematicos*, 16(1):1–193, 2009.

M. Birkner, J. Blath, and B. Eldon. Statistical properties of the site-frequency spectrum associated with lambda-coalescents. *Genetics*, 195(3):1037–1053, 2013. ISSN 0016-6731. doi: 10.1534/genetics.113.156612. URL `https://www.genetics.org/content/195/3/1037`.

M. Birkner, H. Liu, and A. Sturm. Coalescent results for diploid exchangeable population models. *Electron. J. Probab.*, 23(49):44, 2018.

A. Cortines and B. Mallein. A N-branching random walk with random selection. *ALEA, Lat. Am. J. Probab. Math. Stat.*, 14:17–137, 2017.

M. M. Desai, A. M. Walczak, and D. S. Fisher. Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics*, 193(2):565–585, 2013.

P. Donnelly and T. G. Kurtz. Particle representations for measure-valued population models. *Ann. Probab.*, 27(1):166–205, 1999.

B. Eldon and J. Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172(4):2621–2633, 2006. doi: 10.1534/genetics.105.052175.

B. Eldon, M. Birkner, J. Blath, and F. Freund. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics*, 199(3):841–856, 2015.

F. Freund. Cannings models, population size changes and multiple-merger coalescents. *J. Math. Biol.*, 80(5):1497–1521, 2020.

F. Freund and A. Siri-Jégousse. The impact of genetic diversity statistics on model selection between coalescents. *Comput. Stat. Data Anal.*, 156:107055, 2021.

Y.-X. Fu. Statistical properties of segregating sites. *Theor. Pop. Biol.*, 48(2):172–197, 1995.

A. Gnedin, A. Iksanov, and A. Marynych. Λ-coalescents: a survey. *J. Appl. Probab.*, 51A(Celebrating 50 Years of The Applied Probability Trust):23–40, 2014. ISSN 0021-9002. doi: 10.1239/jap/1417528464. URL `http://dx.doi.org/10.1239/jap/1417528464`.

A. González Casanova, V. Miró Pina, and A. Siri-Jégousse. The symmetric coalescent and Wright-Fisher models with bottlenecks. *Ann. Appl. Probab.*, 32(2):235–268, 2021.

A. Hobolth, A. Siri-Jégousse, and M. Bladt. Phase-type distributions in population genetics. *Theor. Pop. Biol.*, 127:16–32, 2019.

M. S. Johnson, S. Gopalakrishnan, J. Goyal, M. E. Dillingham, C. W. Bakerlee, P. T. Humphrey, T. Jagdish, E. R. Jerison, K. Kosheleva, K. R. Lawrence, J. Min, A. Moulana, A. M. Phillips, J. C. Piper, R. Purkanti, A. Rego-Costa, M. J. McDonald, A. N. Nguyen Ba, and M. M. Desai. Phenotypic and molecular evolution across 10,000 generations in laboratory budding yeast populations. *eLife*, 10:e63910, jan 2021. ISSN 2050-084X. doi: 10.7554/eLife.63910. URL `https://doi.org/10.7554/eLife.63910`.

M. Kato, D. A. Vasco, R. Sugino, D. Narushima, and A. Krasnitz. Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer. *Royal Sic. Open Sci.*, 4(9), 2017. doi: 10.1098/rsos.171060. URL `http://rsos.royalsocietypublishing.org/content/4/9/171060`.

J. Kelleher, A. M. Etheridge, and G. McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.*, 12(5):e1004842, 2016.

G. Kersting, A. Siri-Jégousse, and A. H. Wences. Site frequency spectrum of the bolthausen-sznitman coalescent. *ALEA. Lat. Am. J. Prob. Math. Stat.*, 18:1483–1505, 2021.

J. F. Kingman. On the genealogy of large populations. *J. Appl. Probab.*, pages 27–43, 1982a.

J. F. Kingman. The coalescent. *Stochastic Process. Appl.*, 13(3):235–248, 1982b.

J. Koskela. Multi-locus data distinguishes between population growth and multiple merger coalescents. *Stat. Appl. Genet. Mol. Biol.*, 17(3), 2018.

J. Koskela and M. Wilke Berenguer. Robust model selection between population growth and multiple merger coalescents. *Math. Biosci.*, 311:1–12, 2019.

J. Koskela, P. A. Jenkins, and D. Spanò. Bayesian non-parametric inference for *Lambda*-coalescents: Posterior consistency and a parametric method. *Bernoulli*, 24(3):2122–2153, 2018.

R. Lenski, M. Rose, S. Simpson, and S. Tadler. Long-term experimental evolution in escherichia coli. i. adaptation and divergence during 2,000 generations. *Am. Nat.*, 138(6):1315–41, 1991.

G. McVean and N. Cardin. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 1459(360):1387–93, 2005.

F. Menardo, S. Gagneux, and F. Freund. Multiple merger genealogies in outbreaks of mycobacterium tuberculosis. *Mol. Biol. Evol.*, 38(1):290–306, 2020.

R. A. Neher and O. Hallatschek. Genealogies of rapidly adapting populations. *Proc. Natl. Acad. Sci. U.S.A.*, 110(2):437–442, 2013.

H. S. Niwa, K. Nashida, and T. Yanagimoto. Reproductive skew in japanese sardine inferred from dna sequences. *ICES J. Mar. Sci.*, 73(9):2181–2189, 2016.

J. Pitman. Coalescents with multiple collisions. *Ann. Probab.*, 27(4):1870–1902, 1999. ISSN 0091-1798. doi: 10.1214/aop/1022677552. URL `http://dx.doi.org/10.1214/aop/1022677552`.

H. Queffélec and C. Zuily. *Analyse pour l'agrégation-Agrégation/Master Mathématiques*. Dunod, 2020.

S. Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.*, 36(4): 1116–1125, 1999.

E. Schertzer and A. H. Wences. Relative vs absolute fitness in a population genetics model. how stronger selection may promote genetic diversity. *arXiv preprint arXiv:2301.07762*, 2023.

J. Schweinsberg. Coalescent processes obtained from supercritical galton–watson processes. *Stochastic Process. Appl.*, 106(1):107–139, 2003.

J. Schweinsberg. Rigorous results for a population model with selection ii: genealogy of the population. *Electron. J. Probab.*, 22, 2017.

H. Sigurgíslason and E. Árnason. Extent of mitochondrial dna sequence variation in atlantic cod from the faroe islands: a resolution of gene genealogy. *Heredity*, 91(6):557–564, 2003. doi: 10.1038/sj.hdy.6800361. URL `https://doi.org/10.1038/sj.hdy.6800361`.

H. Teotónio, S. Estes, P. Phillips, and C. Baer. Experimental evolution with caenorhabditis nematodes. *Genetics*, 206(2):697–716, 2017.

7 Appendix A: Proof of Theorem 2.1

In this section we prove the consistency of the proposed estimator. One source of error comes from the uncertainty on the terms $\hat{r}_{m,j}$. There are also two other sources of approximation error given by the fact that we approximate ν by $\nu \star K_h$ and that we approximate K_h^x by $P_{m,h}^x$. To prove the theorem, we need to tackle these three sources of error. The first results that we present deals with the smooth approximation of ν by $\nu \star K_h$.

Proposition 7.1. *Let ϕ be such that for all $x \in \mathbb{R}$, $\phi(x) \leq \mathbb{1}_{\{[-\tau,\tau]\}}(x)$, for some $\tau > 0$. Then, for any $\delta > 0$ and any $x \in (0,1]$ such that $x - \delta \geq m^{-1} + \tau h$,*

$$|\nu(x) - \nu \star K_h(x)| \leq \omega_\nu(x,\delta) + 4 \sup_{x \in [m^{-1}+\tau h, 1]} |\nu(x)|. \left((\tau - \frac{\delta}{h}) \vee 0 \right).$$

Proof. Using the fact that K_h is of integral 1, we see that for any $\delta > 0$ and any x,

$$\begin{aligned} |\nu(x) - \nu \star K_h(x)| &\leq \int_{\mathbb{R}} |\nu(x) - \nu(x+y)| K_h(y) dy \\ &= \int_{-\delta}^{\delta} |\nu(x) - \nu(x+y)| K_h(y) dy + \int_{\mathbb{R} \setminus [-\delta,\delta]} |\nu(x) - \nu(x+y)| K_h(y) dy. \end{aligned}$$

Observe that the last term is well defined since the support of K_h is included in $[-\tau h, \tau h]$. Then,

$$\begin{aligned} |\nu(x) - \nu \star K_h(x)| &\leq \omega_\nu(x,\delta) + 2 \sup_{x \in [m^{-1}+\tau h, 1]} |\nu(x)| \int_{\mathbb{R} \setminus [-\delta,\delta]} K_h(y) dy \\ &= \omega_\nu(x,\delta) + 4 \sup_{x \in [m^{-1}+\tau h, 1]} |\nu(x)| \int_{\delta}^{+\infty} K_h(y) dy. \end{aligned}$$

This last integral is obviously 0 if the support of K_h and $[\delta, +\infty)$ are disjoint i.e., if $\tau h < \delta$. Otherwise,

$$\int_{\delta}^{+\infty} K_h(y) dy \leq \int_{\delta}^{h\tau} \frac{1}{h} \phi(\frac{y}{h}) dy \leq \int_{\delta}^{h\tau} \frac{1}{h} dy = \tau - \frac{\delta}{h}.$$

□

Next, we deal with the approximation of K_h^x by the Bernstein polynomials $P_{m,h}^x$.

Proposition 7.2. *Let f be a continuous function on $[0,1]$. Then*

$$\|f - P_m(f)\|_\infty \leq 1.5 \|\omega_f(\cdot, m^{-1/2})\|_\infty$$

where $P_m(f) = \sum_{j=0}^{m} \binom{m}{j} f(\frac{j}{m}) y^j (1-y)^{m-j}$.

The proof can be found in Theorem II, p. 518 of Queffélec and Zuily [2020] where the constant can be taken equal to 1.5 from the proof page 519. Proposition 7.2 will be applied on the kernel function K_h. It is obvious to see that for any $\delta > 0$,

$$\omega_{K_h^x}(y,\delta) = \omega_{K_h}(y,\delta) = h^{-1} \omega_\phi(y, h^{-1}\delta).$$

Thus the regularity of K_h^x is directly inherited from the regularity of ϕ. For example, if ϕ is α-Hölder,

$$\|\omega_{K_h^x}(\cdot, m^{-1/2})\|_\infty \leq \frac{c}{h^{1+\alpha} m^{\alpha/2}} \tag{11}$$

for some constant $c > 0$.

The following simple result is the verification that the plug in strategy of (6) is valid when one controls the L_1 error of estimation on the vector $r_m = (r_{m,j})_{2 \leq j \leq m}$.

Proposition 7.3. *For any $h > 0$ and any $m \geq 2$ one has that*

$$\|\hat{\nu}_{m,h} - \nu_{m,h}\|_1 \leq \|\hat{r}_m - r_m\|_1.$$

Proof. The proof of the proposition is obtained by noticing that the terms $K_h(x - j/m)$ have integral equal to 1. More formally, notice that the support of $\hat{\nu}_{m,h}$ or $\nu_{m,h}$ is included in $[m^{-1} + \tau h, 1]$ and so

$$\begin{aligned}
\|\hat{\nu}_{m,h} - \nu_{m,h}\|_1 &\leq \int_{m^{-1}+\tau h}^1 \left| \sum_{j=2}^m K_h^x\left(\frac{j}{m}\right) r_{m,j} - \sum_{j=2}^m K_h^x\left(\frac{j}{m}\right) \hat{r}_{m,j} \right| dx \\
&\leq \int_{m^{-1}+\tau h}^1 \sum_{j=2}^m K_h^x\left(\frac{j}{m}\right) |r_{m,j} - \hat{r}_{m,j}| dx \\
&\leq \sum_{j=2}^m |r_{m,j} - \hat{r}_{m,j}| \int_{\mathbb{R}} K_h^x\left(\frac{j}{m}\right) dx = \|\hat{r}_m - r_m\|_1,
\end{aligned}$$

which concludes the proof. □

When one sums up the previous results we end with the following result of consistency.

Proof of Theorem 2.1. Noticing that for any x,

$$|\hat{\nu}_{m,h}(x) - \nu(x)| \leq |\hat{\nu}_{m,h}(x) - \nu_{m,h}(x)| + |\nu_{m,h}(x) - \nu \star K_h(x)| + |\nu \star K_h(x) - \nu(x)|$$

and using the previous results, we see that it is enough to show that

$$|\nu_{m,h}(x) - \nu \star K_h(x)| \leq C\Lambda[0, 1/m] + 1.5c \frac{\nu[1/m, 1]}{h^{1+\alpha} m^{\alpha/2}}.$$

But since x belongs to $[m^{-1} + h\tau, 1]$, we have that $\nu_{m,h}(x) = \int_0^1 P_{m,h}^x(y)\nu(y)dy$ and so

$$\begin{aligned}
|\nu_{m,h}(x) - \nu \star K_h(x)| &\leq \int_0^1 |P_{m,h}^x(y) - K_h^x(y)|\nu(y)dy \\
&= \int_0^{\frac{1}{m}} \frac{|P_{m,h}^x(y)|}{y^2} y^2 \nu(y)dy + \int_{\frac{1}{m}}^1 |P_{m,h}^x(y) - K_h^x(y)|\nu(y)dy.
\end{aligned}$$

On the one hand, the polynomial $|P_{m,h}^x(y)|/y^2$ is uniformly bounded so there exists a constant C such that

$$\int_0^{\frac{1}{m}} \frac{|P_{m,h}^x(y)|}{y^2} y^2 \nu(y)dy \leq C\Lambda[0, 1/m].$$

On the other hand, by Proposition 7.2 and (11), we have that

$$\int_{\frac{1}{m}}^{1} |P^x_{m,h}(y) - K^x_h(y)|\nu(y)dy \leq \|K^x_h - P^x_{m,h}\|_\infty \int_{\frac{1}{m}}^{1} \nu(y)dy \leq 1.5c\frac{\nu[1/m,1]}{h^{1+\alpha}m^{\alpha/2}}$$

□

8 Appendix B: Proof of Proposition 2.1

In this section we prove Proposition 2.1 by establishing a more general recursion result. Consider any initial configuration $\bar{c} \in \mathscr{C}_m$. By the strong Markov property (applied at the time of the first jump of the process), we obtain that, for $i \in \{1,\dots,m-1\}$,

$$d_{|\bar{c}|}\mathbb{E}[L_i(\bar{c})] = c_i + \sum_{\bar{b}:\bar{b}\preccurlyeq\bar{c}} S_{\bar{c},\bar{b}}\mathbb{E}[L_i(\bar{b})] \tag{12}$$

where $S_{\bar{c},\bar{b}}$ is the rate at which the multi-dimensional block counting process jumps from $\bar{c}$ to $\bar{b}$, i.e.

$$S_{\bar{c},\bar{b}} = \lambda_{|\bar{c}|,\sum(c_j-b_j)\mathbf{1}_{b_j<c_j}} \prod_{j,b_j<c_j} \binom{c_j}{c_j-b_j},$$

already defined in Eq. (20) of Hobolth et al. [2019]. In the summation in (12), the notation $\bar{b} \preccurlyeq \bar{c}$ holds for the vectors $\bar{b}$ obtained by a single step merging of a multiple number of lineage into one. This imposes, in particular, that there exists a unique index $i_0 \leq m$ such that $b_{i_0} > c_{i_0}$ and in this case, $b_{i_0} = c_{i_0} + 1$. This last fact is actually a characterization of the partial order $\preccurlyeq$. An additional fact is that

$$d_{|\bar{c}|} = \sum_{\bar{b}:\bar{b}\preccurlyeq\bar{c}} S_{\bar{c},\bar{b}}.$$

Proposition 2.1 is obtained in the particular example where the initial configuration only contains singletons, that is $\bar{c}_0 = m.e_1 = (m,0,0,\dots)$. In this case we recover the classical functional leading to the site frequency spectrum. After the first merge, the $m-2$ attainable partitions in (12) are of the form

$$\bar{b}_j = (m-j)e_1 + e_j, \quad j \in \{2,\dots,m\}.$$

An important observation is that, in this case, $S_{\bar{c}_0,\bar{b}_j} = r_{m,j}$.