J. Appl. Probab. 1–35 (2024)
*doi:*10.1017/jpr.2024.56

MAXIMUM LIKELIHOOD ESTIMATION FOR SPINAL-STRUCTURED TREES

ROMAIN AZAÏS,[*] *Inria team MOSAIC*
BENOÎT HENRY,[**] *IMT Nord Europe*

Abstract

We investigate some aspects of the problem of the estimation of birth distributions (BDs) in multi-type Galton–Watson trees (MGWs) with unobserved types. More precisely, we consider two-type MGWs called spinal-structured trees. This kind of tree is characterized by a spine of special individuals whose BD ν is different from the other individuals in the tree (called normal, and whose BD is denoted by μ). In this work, we show that even in such a very structured two-type population, our ability to distinguish the two types and estimate μ and ν is constrained by a trade-off between the growth-rate of the population and the similarity of μ and ν. Indeed, if the growth-rate is too large, large deviation events are likely to be observed in the sampling of the normal individuals, preventing us from distinguishing them from special ones. Roughly speaking, our approach succeeds if $r < \mathfrak{D}(\mu, \nu)$, where r is the exponential growth-rate of the population and $\mathfrak{D}$ is a divergence measuring the dissimilarity between μ and ν.

Keywords: Multi-type Galton–Watson tree; branching process; parametric inference; latent variable

2020 Mathematics Subject Classification: Primary 60J80; 62M05
Secondary 62F12

1. Introduction

1.1. Problem formulation

A Galton–Watson tree with birth distribution μ is a random tree obtained recursively as follows: starting from the root, the number of children of any node is generated independently according to μ. In light of [4], the log-likelihood of such a tree T observed until generation h is given by

$$\mathcal{L}_h^{\mathrm{GW}}(\mu) = \sum_{v\in T,\ \mathcal{D}(v)<h} \log\ \mu(\#C(v)),$$

where $\mathcal{D}(v)$ denotes the depth of node v, i.e. the length of the path from the root to v, and $C(v)$ stands for the set of children of v. When the mean value $m(\mu)$ of the birth distribution μ is smaller than 1, the model is said to be subcritical and the number of vertices of T as well as its expectation are finite, that is, the genealogy associated to T becomes extinct. If a

Received 10 December 2022; accepted 18 July 2024.

[*] Postal address: ÉNS de Lyon, 46 Allée d'Italie, 69 007 Lyon, France. Email address: romain.azais@inria.fr
[**] Postal address: Institut Mines-Télécom, Univ. Lille, 59 000 Lille, France. Email address: benoit.henry@imt-nordeurope.fr

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

subcritical Galton–Watson model is conditioned to survive until (at least) generation h, the structure of the induced trees is changed according to Kesten's theorem [1, 10]. Indeed, the conditional distribution converges, when h tends to infinity, towards the distribution of Kesten's tree defined as follows. Kesten's tree is a two-type Galton–Watson tree in which nodes can be normal or special such that the following conditions hold.

- The birth distribution of normal nodes is μ, while that of special nodes ν is biased:

$$\nu(k) = \frac{k\mu(k)}{m(\mu)} \quad \text{for all } k \geq 0. \tag{1.1}$$

  As for Galton–Watson trees, the numbers of children are generated independently.

- All the children of a normal node are normal. Exactly one of the children of a special node (picked at random) is special.

It should be noted that the set of children of a special node is non-empty since $\nu(0) = 0$. Consequently, Kesten's tree consists of an infinite spine composed of special nodes, to which subcritical Galton–Watson trees of normal nodes (with birth distribution μ) are attached. Following the reasoning presented in [4] together with the form of the special birth distribution (1.1), the log-likelihood of Kesten's tree is given by

$$\begin{aligned}\mathcal{L}_h^{\mathrm{K}}(\mu) &= \sum_{v \notin \mathcal{S},\, \mathcal{D}(v)<h} \log\, \mu(\#\mathcal{C}(v)) + \sum_{v \in \mathcal{S},\, \mathcal{D}(v)<h} \log\, \nu(\#\mathcal{C}(v)) \\ &= \sum_{\mathcal{D}(v)<h} \log\, \mu(\#\mathcal{C}(v)) + \sum_{v \in \mathcal{S},\, \mathcal{D}(v)<h} \log\, \#\mathcal{C}(v) - h \log\, m(\mu),\end{aligned}$$

where $\mathcal{S}$ denotes the spine of T, i.e. the set of special nodes. Interestingly, maximizing the log-likelihood (with respect to μ) does not require us to observe the types of the nodes. Indeed, the term that involves the spine does not depend on the parameter of the model.

In this paper we investigate spinal-structured trees, which can be seen as a generalization of Kesten's tree. A spinal-structured tree is a two-type Galton–Watson tree, made of normal and special nodes, parametrized by a distribution μ and a non-trivial function $f\colon \mathbb{N} \to \mathbb{R}_+$, such that the following conditions hold.

- The birth distribution of normal nodes is μ, while that of special nodes ν is biased:

$$\nu(k) = \frac{f(k)\mu(k)}{\sum_{l \geq 0} f(l)\mu(l)} \quad \text{for all } k \in \mathbb{N}, \tag{1.2}$$

  assuming that the denominator is positive.

- As for Kesten's tree, a normal node gives birth to normal nodes, whereas if the set of children of a special node is non-empty, then exactly one of them (picked at random) is special.

Whenever $f(0) = 0$, a spinal-structured tree admits an infinite spine made of special nodes, which gives its name to the model. It should be remarked that the model fails to be identifiable because the line spanned by f defines the same probability measure ν. As a consequence, without loss of generality, we assume

$$\sum_{l \geq 0} f(l)\mu(l) = 1.$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

Taking this into account, the log-likelihood of spinal-structured trees is given by

$$\mathcal{L}_h^{\text{SST}}(\mu, f) = \sum_{\mathcal{D}(v)<h} \log\ \mu(\#C(v)) + \sum_{v\in\mathcal{S},\ \mathcal{D}(v)<h} \log f(\#C(v)).$$

For any birth distribution μ, any biased distribution ν can be written as (1.2) with a suitable choice of f (except of course distributions ν such that, for some k, $\nu(k) > 0$ and $\mu(k) = 0$). The parametrization (μ, f) instead of (μ, ν) makes it clearer that spinal-structured trees form a generalization of Kesten's tree, which is obtained if and only if f is linear, considering that μ is subcritical. In addition, Galton–Watson trees can be seen as spinal-structured trees assuming that f is constant. Our goal in this work is to investigate the problem of estimating μ and f through the maximization of $\mathcal{L}_h^{\text{SST}}$ without knowledge of the types of the individuals. The main advantage of the parametrization (μ, f) is that, just as for Kesten's tree, it allows us to maximize the log-likelihood with respect to μ without observing the types of the nodes. However, maximizing it with respect to f entails observation of the types of the nodes.

1.2. Motivation

The motivation for this paper is twofold: first, it provides a step forward in the theoretical understanding of type identification in multi-type Galton–Watson trees (MGWs); second, it offers preliminary theoretical foundations for statistically testing whether or not population data have been conditioned to survive. These two points are detailed below.

Spinal-structured trees can be seen as particular instances of two more general models. If the special individuals are interpreted as immigrants, the underlying population process is a Galton–Watson model with immigration given by ν. And more generally, like every Galton–Watson process with immigration, it can also be seen as a particular case of MGWs. The problem of the estimation of birth distributions in MGWs has been heavily studied, for example in [5], [6], [11], and [13], and references therein, but in all these works the types of the individuals are assumed to be known. A small number of works, e.g. [9], [15], and [16], have investigated this problem with unobserved types, but none of these provide theoretical results: they only investigate numerical aspects. Using the special case of spinal-structured trees, this paper aims to demonstrate the theoretical difficulties involved in type estimation and propose a statistical strategy for dealing with them. In particular, we shall show that we are able to estimate the underlying parameters when population growth is not too large compared with the dissimilarity of the two birth distributions. This phenomenon is likely to hold true for more complicated problems.

When estimating the parameters of an observed population using a stochastic model, the latter must first be accurately chosen. To the best of our knowledge, even in the simple framework of Galton–Watson models, there is no method in the literature for rigorously determining from the data whether or not the population has been conditioned to survive. However, as mentioned above, estimating the parameters under the wrong model introduces significant biases that can lead to wrong conclusions about the population. Spinal-structured trees generalize both Galton–Watson trees (f is constant, not investigated hereafter) and Galton–Watson trees conditioned to survive ($f \propto \text{Id}$). By estimating f, and even better by testing the shape of f, we can conclude which model to apply. The results of this paper will allow us to make progress in this direction (see in particular Section 7.2).

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

1.3. Aim of the paper

The present paper is dedicated to the development and study of an estimation method for the unknown parameters μ and f, as well as the unknown type of the nodes, from the observation T_h of one spinal-structured tree until generation h. The estimation algorithm that we derive below is based on the maximization of $\mathcal{L}_h^{\mathrm{SST}}$ with the major difficulty that types are unobserved. Once the calculations are done, it can be succinctly described as follows.

(i) Naive estimation of μ:

$$\widehat{\mu}_h(i) = \frac{1}{\#T_h} \sum_{v \in T_h} \mathbb{1}_{C(v)=i}.$$

(ii) Estimation of the spine, i.e. the set of special nodes:

$$\widehat{\mathcal{S}}_h = \arg\max_{\mathbf{s}\in\mathfrak{S}_h} d_{KL}(\overline{\mathbf{s}}, \mathcal{B}\widehat{\mu}_h),$$

where $\mathfrak{S}_h$ denotes the set of spine candidates (branches still alive at generation h), $\overline{\mathbf{s}}$ is the empirical distribution of the number of children along the spine candidate $\mathbf{s}$, $\mathcal{B}\widehat{\mu}_h(i) \propto i\,\widehat{\mu}_h(i)$, and $d_{KL}(p, q)$ denotes the Kullback–Leibler divergence between distributions p and q.

(iii) Unbiased estimation of μ (without estimated special nodes $\widehat{\mathcal{S}}_h$):

$$\widehat{\mu}_h^\star(i) = \frac{1}{\#T_h - h} \sum_{v \in T_h \setminus \widehat{\mathcal{S}}_h} \mathbb{1}_{C(v)=i}.$$

(iv) Estimation of f:

$$\widehat{f}_h(i) = \frac{1}{\widehat{\mu}_h^\star(i) h} \sum_{v \in \widehat{\mathcal{S}}_h} \mathbb{1}_{C(v)=i}.$$

Even in such a structured instance of MGWs, the convergence of these estimates is far from easy to establish. In Theorem 3.2 we state that if the distribution of surviving normal nodes is not too close to the special birth distribution ν compared to the exponential growth-rate of the tree, then $\widehat{\mu}_h^\star$ and $\widehat{f}_h$ almost surely converge towards μ and f. In addition, the recovered part of the spine is almost surely of order h when h goes to infinity. We insist on the fact that these two results are true for any growth regime of the tree (subcritical $m(\mu) < 1$, critical $m(\mu) = 1$, or supercritical $m(\mu) > 1$). Nevertheless, the reason behind these convergence results is not the same in the subcritical regime (where almost all of the spine can be recovered in an algorithmic fashion) and in the critical and supercritical regimes (where the number of spine candidates explodes). The theoretical convergence properties related to the asymptotics of the estimators $\widehat{\mu}_h^\star$, $\widehat{f}_h$, and $\widehat{\mathcal{S}}_h$ are shown under the following main conditions, which are essential to the proofs of convergence.

- The maximum number of children in the tree is $N \geq 1$, i.e. $\mu \in \mathcal{M}$, where $\mathcal{M}$ denotes the set of probability distributions on $\{0, \dots, N\}$. By construction (1.2), ν also belongs to $\mathcal{M}$.
- $f(0) = 0$ so that $\nu(0) = 0$, that is, the tree admits an infinite spine of special nodes.

For the sake of readability and conciseness of the proofs, we also assume the following conditions.

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

- The support of μ is $\{0, \ldots, N\}$.
- $f(k) > 0$ for any $k > 0$, which implies that the support of ν is $\{1, \ldots, N\}$.

The article is organized as follows. Section 2 describes how some parts of the spine can be algorithmically recovered in a deterministic fashion. Section 3 is devoted to our estimation procedure and theoretical results:

- Section 3.1 for the preliminary estimation of μ;
- Section 3.2 for the identification of a candidate for the spine, named the Ugly Duckling;
- Section 3.3 for the final estimation of μ and the estimation of f;
- Section 3.4 for the statement of our main result, Theorem 3.2.

The proof of Theorem 3.2 in the subcritical case can be found in Section 4. The proof in the supercritical case involves large deviation-type estimates, for which we need information on the rate function. The rate function is studied in Section 5 and the information needed is stated in Theorem 5.1. We finally consider the proof in the critical and supercritical cases in Section 6. The final Section 7 is devoted to numerical illustrations of the results (Section 7.1) and an application to asymptotic tests for populations conditioned on surviving (Section 7.2). Appendix A concerns the proof of some intermediate lemmas.

2. Algorithmic identification of the spine

Here we propose an algorithm to (at least partially) identify the spine of a spinal-structured tree T observed until generation h. A node v of T is called observed if $\mathcal{D}(v) < h$. It means that the number of children of v can be considered as part of the data available to reconstruct the spine of T (even if the depth of these children is h). The tree restricted to the observed nodes is denoted by T_h.

We will also need the notion of observed height of a subtree $T[v]$ of T. If v is a node of T, $T[v]$ denotes the tree rooted at v and composed of v and all its descendants in T. In the literature on trees, the height $\mathcal{H}(T[v])$ of a subtree $T[v]$ is the length of the longest path from its root v to its leaves. In the context of this work, T is only observed until generation h, and thus the height of $T[v]$ can be unknown since the leaves of $T[v]$ can be inaccessible. For this reason, we define the observed height of $T[v]$ as

$$\mathcal{H}_o(T[v]) = \min\left(\mathcal{H}(T[v]), \quad l - \mathcal{D}(v)\right),$$

where l is the length of the minimal path from v to unobserved nodes. It should be noted that $\mathcal{H}_o$ implicitly depends on h. Either $l = +\infty$, if v has no unobserved descendant, or $l = h - \mathcal{D}(v)$. In addition, v has no unobserved descendant if and only if $\mathcal{H}(T[v]) + \mathcal{D}(v) < h$.

The following result makes it possible to partially identify the spine.

Proposition 2.1. *Let T be a spinal-structured tree observed until generation h and let v be an observed node of T.*

- *If $\mathcal{H}_o(T[v]) + \mathcal{D}(v) < h$, then v is normal.*
- *If v is special, the children of v are observed, and*

$$\exists!\, c \in \mathcal{C}(v) \text{ such that } \mathcal{H}_o(T[c]) + \mathcal{D}(c) \geq h,$$

then c is special.

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

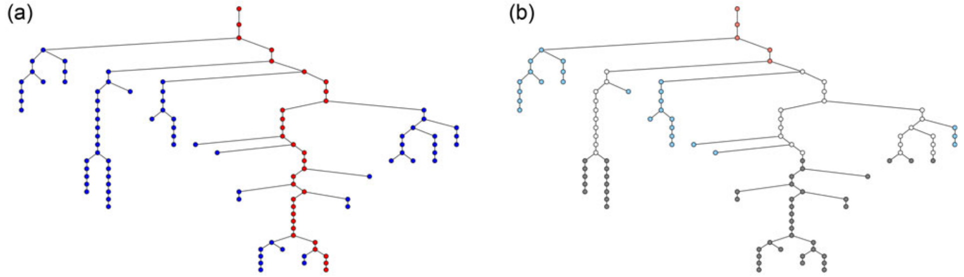


FIGURE 1. (a) A spinal-structured tree simulated until generation 30 with normal nodes in blue and special nodes in red. We assume that it is observed until generation $h = 15$ and in (b) we identify the type of the nodes using Proposition 2.1 with the following color code: light blue for identified normal nodes, light red for identified special nodes, gray for unobserved nodes, and white for unidentified types.

Proof. The proof relies on the fact that special nodes have an infinite number of descendants. First, if v is such that $\mathcal{H}_o(T[v]) + \mathcal{D}(v) < h$, it means that its subtree has become extinct before generation h and thus v is normal. Second, if v is special, exactly one of its children is special. All the subtrees rooted at the children of v that become extinct are composed of normal nodes. Consequently, if only one subtree among its children has not become extinct before generation h, it is necessarily special. □

It should be noticed that if an observed node is not covered by the two previous conditions, it can be either special or normal. Indeed, if the c_k are the children of a special node v that do not become extinct before generation h, only one of them is special, while the others are normal. In fact, only the distribution of the subtrees rooted at the c_k can be used to differentiate them. An application of Proposition 2.1 is presented in Figure 1.

If a node v at depth $\mathcal{D}(v) = h - 1$ has been identified as special, i.e. if v is the only node with children at depth $h - 1$, it means that the spine has been algorithmically reconstructed, and is formed by v and all its ancestors. Otherwise, if the type of two or more nodes at depth $h - 1$ has not been identified, each of them is part of a possible spine. More formally, the set of possible spines $\mathfrak{S}_h$ is made of all the branches from the root to v whenever $\mathcal{D}(v) = h - 1$ and the type of v has not been identified as normal. With this notation, if $\#\mathfrak{S}_h = 1$, then the spine has been fully reconstructed. In all cases, $\bigcap_{\mathbf{s}\in\mathfrak{S}_h} \mathbf{s}$ is exactly the set of nodes identified as special, while the complement $\bigcup_{\mathbf{s}\in\mathfrak{S}_h} \mathbf{s} \setminus \bigcap_{\mathbf{s}\in\mathfrak{S}_h} \mathbf{s}$ is composed of all the nodes that cannot be identified in an algorithmic way.

Spine candidates can be indexed by their first unobserved node. Given a node v in T, the sequence of ancestors of v is denoted by $\mathcal{A}(v)$,

$$\mathcal{A}(v) = \left(\mathcal{P}^h(v), \mathcal{P}^{h-1}(v), \ldots, \mathcal{P}(v)\right), \tag{2.1}$$

where $\mathcal{P}(v)$ is the parent of v in T and recursively $\mathcal{P}^h(v) = \mathcal{P}(\mathcal{P}^{h-1}(v))$. If $\mathcal{D}(v) = h$, then $\mathcal{A}(v)$ is an element of $\mathfrak{S}_h$. Throughout the paper, when there is no ambiguity, we identify $\mathcal{A}(v)$ with the sequence of numbers of children along $\mathcal{A}(v)$, i.e. $(\#\mathcal{C}(u) \colon u \in \mathcal{A}(v))$.

3. Ugly Duckling

In this section we aim to develop an estimation method for the unknown parameters μ and f as well as the spine $\mathcal{S}$ of a spinal-structured tree observed until generation h.

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

The algorithm presented below takes advantage of the specific behavior of spinal-structured trees. We also present our main result of convergence that holds for any growth regime of the normal population, i.e. whatever the value of $m(\mu)$.

3.1. Estimation of μ

As remarked in the Introduction, maximizing $\mathcal{L}_h^{\text{SST}}$ with respect to μ does not require us to observe the type of the nodes. Consequently, as f is unknown, we can still construct a first estimate of μ as

$$\widehat{\mu}_h = \arg\max_{\mu \in \mathcal{M}} \mathcal{L}_h^{\text{SST}}(\mu, f).$$

Standard calculus shows that

$$\widehat{\mu}_h = \left(\frac{1}{\#T_h} \sum_{v \in T_h} \mathbb{1}_{C(v)=i} \right)_{i \in \{0,\dots,N\}}. \tag{3.1}$$

We can notice that the optimum in f of $\mathcal{L}_h^{\text{SST}}$ depends on the unknown spine $\mathcal{S}$, and is thus of no use at this stage.

3.2. Selection of the spine

In $\mathfrak{S}_h$, the spine $\mathcal{S}$ is the unique element whose component-wise distribution is ν defined from (1.2). In that sense, finding $\mathcal{S}$ is a sample selection problem, where, however, the distribution at stake ν is unknown. Our approach consists in estimating the spine by the sample that differs the most from the expected behavior of a sample made of normal nodes.

However, it should be observed that the $\mathfrak{S}_h$ consist of surviving lineages. Thus μ *is not* the component-wise distribution of the samples of normal nodes in $\mathfrak{S}_h$, and, as a consequence, is not the right distribution for comparison. Identifying the law of $\mathbf{s} \in \mathfrak{S}_h$ can be done thanks to the so-called many-to-one formula presented in the following theorem (see [12]).

Theorem 3.1. *Let G be a Galton–Watson tree with birth distribution μ and let h be an integer. Then, for any bounded measurable function $\varphi\colon \mathbb{R}^h \to \mathbb{R}$, we have*

$$\mathbb{E}\left[\sum_{\{u \in G\colon \mathcal{D}(u)=h\}} \varphi(\mathcal{A}(u)) \right] = m(\mu)^h \, \mathbb{E}[\varphi(X_0, \dots, X_{h-1})], \tag{3.2}$$

where $\mathcal{A}(u)$ is defined in (2.1) and $X_1, \dots, X_{h-1}$ is an independent and identically distributed (i.i.d.) family of random variables with common distribution $\mathcal{B}\mu$, where the operator $\mathcal{B}$ is defined, for any $p \in \mathcal{M}$, by

$$\mathcal{B}p(i) = \frac{ip_i}{m(p)} \quad \textit{for all } i \in \{1, \dots, N\}. \tag{3.3}$$

With this new information in hand, we can now define the estimate of the spine as the *Ugly Duckling* in $\mathfrak{S}_h$,

$$\widehat{\mathcal{S}}_h = \arg\max_{\mathbf{s} \in \mathfrak{S}_h} d_{KL}(\overline{\mathbf{s}}, \mathcal{B}\widehat{\mu}_h), \tag{3.4}$$

where $\overline{X}$ denotes the empirical measure associated to the vector X and $d_{KL}(p, q)$ denotes the Kullback–Leibler divergence between distributions p and q. In this formula, we compare $\overline{\mathbf{s}}$ to $\mathcal{B}\widehat{\mu}_h$, with $\widehat{\mu}_h$ given by (3.1), because the true distribution μ is obviously unknown.

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

Remark 3.1. Another approach would have consisted in selecting the spine as the most likely sample under ν, which is unknown but can be estimated from an estimate of μ (e.g. $\widehat{\mu}_h$ defined in (3.1)) and an estimate of f. However, as explained in Section 3.1, the optimum of $\mathcal{L}_h^{\text{SST}}$ in f depends on the spine. As a consequence, this approach would have resulted in an iterative algorithm where f is estimated from the spine, and conversely the spine from f, likely highly dependent on the initial value.

3.3. Correction of μ and estimation of f

We can remark that the estimate (3.1) of μ is the empirical distribution of the numbers of children in the tree. However, the tree is made of h special nodes that do not follow μ, which biases the estimation. Now we know how to estimate the spine, i.e. the set of special nodes in the tree, we can take this into account and correct the estimator of μ as

$$\widehat{\mu}_h^{\star}(i) = \frac{1}{\#T_h - h} \sum_{v \in T_h \setminus \widehat{\mathcal{S}}_h} \mathbb{1}_{C(v)=i} \quad \text{for all } i \in \{0, \ldots, N\}. \tag{3.5}$$

Then we can estimate f by maximizing (under the constraint $\sum_{i=0}^{N} f(i)\widehat{\mu}_h^{\star}(i) = 1$) $\mathcal{L}_h^{\text{SST}}$, where the unknown spine $\mathcal{S}$ has been replaced by $\widehat{\mathcal{S}}_h$, which results in

$$\widehat{f}_h(i) = \frac{1}{\widehat{\mu}_h^{\star}(i) h} \sum_{v \in \widehat{\mathcal{S}}_h} \mathbb{1}_{C(v)=i} \quad \text{for all } i \in \{0, \ldots, N\}. \tag{3.6}$$

It should be noted that, by construction, no node of the spine estimate has no child, which implies $\widehat{f}_h(0) = 0$.

3.4. Theoretical results

The purpose of this section is to study the behavior of the Ugly Duckling method for large observation windows, i.e. $h \to \infty$. The main difficulty arising in our problem is to recover a substantial part of the spine. Depending on the growth-rate of the population, this question takes different forms. Indeed, the number of spine candidates $\#\mathfrak{S}_h$ is highly dependent on the growth-rate $m(\mu)$ of the normal population in the tree.

First, in the subcritical case $m(\mu) < 1$, the trees of normal individuals grafted onto the spine tends to become extinct. In other words, the set of spines $\mathfrak{S}_h$ is essentially reduced to $\mathcal{S}$ or at least to small perturbations of $\mathcal{S}$. Thus a macroscopic part of the spine can be directly identified without further difficulty following the algorithm of Section 2. The only point that needs clarification is that if the unidentified part of the spine is not large enough to perturb the estimation, then we would not be able to guarantee that our estimators are convergent.

In the critical and supercritical cases, identifying the spine becomes substantially harder as the set $\mathfrak{S}_h$ may have a large size and contain potentially long lineages of non-special individuals. In particular, if the number of possible spines is large, one may observe that the empirical distribution of the number of children along some lineages $\mathbf{s} \in \mathfrak{S}_h$ may experience large deviations from its distribution, so that

$$d_{KL}(\overline{\mathbf{s}}, \mathcal{B}\mu) \gg d_{KL}(\overline{\mathcal{S}}, \mathcal{B}\mu).$$

In such a situation, one would not be able to distinguish which of $\mathbf{s}$ or $\mathcal{S}$ is the spine.

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

It follows that our ability to identify the spine relies on a dissimilarity/population-growth trade-off.

- On the one hand, if the growth-rate of the population is small, the number of possible spines is small and none of the normal spines greatly deviate from its expected distribution. Thus we can identify the sample with law ν even if laws $\mathcal{B}\mu$ and ν are similar (but without being too close).
- On the other hand, if the growth-rate is large (i.e. $m(\mu) \gg 1$), then one may expect large deviation samples. In such a situation, we would not be able to recover the spine unless distributions $\mathcal{B}\mu$ and ν are very different.

One good way to measure the dissimilarity between two distributions p and q in our context is given by the following divergence:

$$\mathfrak{D}(p, q) = \inf_{\substack{x,y,z)\in\mathcal{M}^3 \\ \delta\geq 0}} \Bigg\{ d_{KL}(x, p) + d_{KL}(y, q) + \delta d_{KL}(z, q) \,\Bigg|\, d_{KL}\left(\frac{\delta z + x}{\delta + 1}, p\right) - d_{KL}\left(\frac{\delta z + y}{\delta + 1}, p\right) \geq 0 \Bigg\}. \tag{3.7}$$

This idea is summarized in the following theorem, where our convergence criterion relies on a comparison between $\log(m(\mu))$ and $\mathfrak{D}(\mathcal{B}\mu, \nu)$.

Theorem 3.2. *If* $\log(m(\mu)) - \mathfrak{D}(\mathcal{B}\mu, \nu) < 0$*, then the following convergences hold almost surely:*

$$\widehat{\mu}_h^\star \xrightarrow[h\to\infty]{} \mu$$

and

$$\widehat{f}_h \xrightarrow[h\to\infty]{} f.$$

In addition, an order h of the spine is recovered, that is,

$$\frac{\#\mathcal{S} \cap \widehat{\mathcal{S}}_h}{h} \xrightarrow[h\to\infty]{} 1$$

almost surely.

4. Proof of Theorem 3.2 in the subcritical case

In subcritical cases, note that the criteria of Theorem 3.2 are always satisfied. In addition, it is important to note that the first step of our estimation procedure is in this case a dummy step, as it has (essentially) no use in the following steps. If $m(\mu) < 1$, our estimation works, as for large h we can automatically identify an order h of the special individuals as the lineages of the normal ones tend to become extinct. Thus the main point to check in the proof of the subcritical case is that enough spine is directly identifiable. We directly give the proof of Theorem 3.2 in this case.

Proof of Theorem 3.2, subcritical case $m(\mu) < 1$. The key point is that normal Galton–Watson trees induced by special individuals are very unlikely to reach a large height. As their

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

number is finite at each generation, very few of them reach height h. In particular, they would be rather recent subtrees.

Let K_h denote the length of spine that can be algorithmically identified (using the procedure presented in Proposition 2.1) when the spinal-structured tree is observed up to height h. Now, recalling that the spinal-structured tree T can be constructed by grafting an i.i.d. family of Galton–Watson trees $(G_{i,j})_{i,j\geq 1}$ onto the spine, K_h is given by

$$K_h = \sup\{1 \leq n \leq h \mid \mathcal{H}(G_{i,j}) < h - i \;\forall\, i \in \{1, \ldots, n\},\ \ \forall\, j \in \{1, \ldots, S_i - 1\}\},$$

where $S_1, \ldots, S_h$ denote the numbers of special children of the individuals of the spine. Thus

$$\mathbb{P}(K_h \geq n) = \mathbb{P}\left(\bigcap_{i=1}^{n} \bigcap_{j=1}^{S_i-1} \{\mathcal{H}(G_{i,j}) < h - i\}\right) = \prod_{i=1}^{n} \mathbb{E}\left[p_{h-i}^{S_i-1}\right],$$

where p_l denotes the probability that a tree of type $G_{i,j}$ becomes extinct before reaching height l. We then have

$$\mathbb{P}(K_h \geq n) \geq \prod_{i=1}^{n} p_{h-i}^{\mathbb{E}[S_i-1]} = \left(\prod_{i=1}^{n} p_{h-i}\right)^{m(\nu)-1},$$

by Jensen's inequality. Fixing some $\varepsilon > 0$, we thus have

$$\mathbb{P}\left(1 - \frac{K_h}{h} > \varepsilon\right) = 1 - \mathbb{P}(K_h \geq \lfloor (1-\varepsilon)h \rfloor) \leq 1 - \left(\prod_{i=1}^{\lfloor (1-\varepsilon)h \rfloor} p_{h-i}\right)^{m(\nu)-1}.$$

In the subcritical case, it is known [2] that

$$p_l \geq 1 - \gamma^l$$

for some real number $\gamma \in (0, 1)$. Hence

$$\mathbb{P}\left(1 - \frac{K_h}{h} > \varepsilon\right) \leq 1 - \left(\prod_{i=1}^{\lfloor (1-\varepsilon)h \rfloor} (1 - \gamma^{h-i})\right)^{m(\nu)-1} \leq 1 - (1 - \gamma^{\varepsilon h})^{(m(\nu)-1)\lfloor (1-\varepsilon)h \rfloor}.$$

It is then easily checked that

$$\sum_{h\geq 1} \left(1 - (1 - \gamma^{\varepsilon h})^{(m(\nu)-1)\lfloor (1-\varepsilon)h \rfloor}\right) < \infty$$

which, via the Borel–Cantelli lemma, entails the almost sure convergence of K_h/h toward 1. Now, as we almost surely have $K_h \leq \#\widehat{\mathcal{S}}_h \cap \mathcal{S}$, the convergence of $\widehat{\mu}_h$ and $\widehat{\mu}_h^\star$ closely follows the proof of Proposition 6.1 below, while the convergence of $\widehat{f_h}$ can be easily deduced from the Law of Large Numbers. □

5. On the rate function of large deviations in sample selection

In Lemma 6.2 below, we show a large deviation-type estimate for the probability that the empirical distribution of some branch of the spinal-structured tree is closer to ν than that of the true spine (in Kullback–Leibler divergence). The purpose of this section is to study the rate

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

function of this estimate, and it is a preliminary to the proof of Theorem 3.2 in the critical and supercritical cases. Throughout this section, we choose some distribution p and q in $\mathcal{M}$ such that $p \neq q$. Our goal is to study the following parametric optimization problem referenced as problem ($P_{\alpha,\varepsilon}$):

$$\begin{aligned}
&(P_{\alpha,\varepsilon}(1)) \quad \min \quad f_\alpha\big(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\big) = (1-\alpha)\big[d_{KL}\big(\mathbf{x}^{(1)}, p\big) + d_{KL}\big(\mathbf{x}^{(2)}, q\big)\big] + \alpha d_{KL}\big(\mathbf{x}^{(3)}, q\big)\\
&(P_{\alpha,\varepsilon}(2)) \quad \text{s.t.} \quad \mathbf{x}_i^{(j)} \geq 0 \quad \text{for all } (i, j) \in \{0, \ldots, N\} \times \{1, 2, 3\},\\
&(P_{\alpha,\varepsilon}(3)) \qquad g_j\big(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\big) = \sum_{i=0}^{N} \mathbf{x}_i^{(j)} - 1 = 0 \quad \text{for all } j \in \{1, 2, 3\},\\
&(P_{\alpha,\varepsilon}(4)) \qquad H_{\alpha,\varepsilon}\big(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\big) \geq 0,
\end{aligned}$$

where

$$H_{\alpha,\varepsilon}\big(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\big) = d_{KL}\big((1-\alpha)\mathbf{x}^{(1)} + \alpha\mathbf{x}^{(3)}, p\big) - d_{KL}\big((1-\alpha)\mathbf{x}^{(2)} + \alpha\mathbf{x}^{(3)}, p\big) + \varepsilon.$$

The value function associated to problem ($P_{\alpha,\varepsilon}$) is denoted $V\colon [0, 1]^2 \ni (\alpha, \varepsilon) \mapsto V(\alpha, \varepsilon) \in \mathbb{R}_+$ and is given by

$$V(\alpha, \varepsilon) = \inf_{(x,y,z)\in\mathcal{M}^3} \{(1-\alpha)(d_{KL}(x, p) + d_{KL}(y, q)) + \alpha d_{KL}(z, q) \mid H_{\alpha,\varepsilon}(x, y, z) \geq 0\}. \tag{5.1}$$

In the particular situation where $\varepsilon = 0$, the value function associated to problem ($P_{\alpha,0}$) is denoted $v\colon [0, 1] \ni \alpha \mapsto v(\alpha) \in \mathbb{R}_+$. Our goal is to show the following theorem.

Theorem 5.1. *The value function V is continuous. In addition, for any $\rho \in (0, 1)$, there exists $\varepsilon^* > 0$ such that*

$$V(\alpha, \varepsilon) \geq v(\alpha) - \rho \quad \textit{for all } \alpha \in [0, 1], \ \varepsilon \in [0, \varepsilon^*],$$

and

$$\frac{v(\alpha)}{1-\alpha} \xrightarrow[\alpha\to 1]{} d_B(p, q),$$

where d_B is the Bhattacharyya divergence defined by

$$d_B(p, q) = -2\log\left(\sum_{i=1}^{N} \sqrt{p_i q_i}\right). \tag{5.2}$$

To show this result, we begin by defining the parameter-dependent Lagrangian associated with problem ($P_{\alpha,\varepsilon}$) by

$$\begin{aligned}
&L\big(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, w, u, \gamma, \alpha, \varepsilon\big)\\
&\quad = (1-\alpha)\big(d_{KL}\big(\mathbf{x}^{(1)}, p\big) + d_{KL}\big(\mathbf{x}^{(2)}, q\big)\big) + \alpha d_{KL}\big(\mathbf{x}^{(3)}, q\big)\\
&\qquad + \sum_{i=1}^{N}\sum_{j=1}^{3} w_{i,j}\mathbf{x}_i^{(j)} + \gamma\,\big(d_{KL}\big((1-\alpha)\mathbf{x}^{(1)} + \alpha\mathbf{x}^{(3)}, p\big) - d_{KL}\big((1-\alpha)\mathbf{x}^{(2)} + \alpha\mathbf{x}^{(3)}, p\big) + \varepsilon\big)\\
&\qquad + \sum_{j=1}^{3} u_j \sum_{i=1}^{N}\big(\mathbf{x}_i^{(j)} - 1\big),
\end{aligned}$$

where $\gamma, u_1, u_2, u_3, (w_{i,j})_{1\leq i\leq N, 1\leq j\leq 3}$ are the Lagrange multipliers. Thus the first-order

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

optimality conditions are given by

$$\begin{cases} (1-\alpha)\left\{\log\left(\dfrac{\mathbf{x}_i^{(1)}}{p_i}\right)+1\right\}+\gamma(1-\alpha)\left\{\log\left(\dfrac{(1-\alpha)\mathbf{x}_i^{(1)}+\alpha\mathbf{x}_i^{(3)}}{p_i}\right)+1\right\}+\lambda=0, \;\; i\in[\![0,N]\!], \\ (1-\alpha)\left\{\log\left(\dfrac{\mathbf{x}_i^{(2)}}{q_i}\right)+1\right\}-\gamma(1-\alpha)\left\{\log\left(\dfrac{(1-\alpha)\mathbf{x}_i^{(2)}+\alpha\mathbf{x}_i^{(3)}}{p_i}\right)+1\right\}+\mu=0, \;\; i\in[\![0,N]\!], \\ \alpha\left\{\log\left(\dfrac{\mathbf{x}_i^{(3)}}{q_i}\right)+1\right\}+\alpha\gamma\left\{\log\left(\dfrac{(1-\alpha)\mathbf{x}_i^{(1)}+\alpha\mathbf{x}_i^{(3)}}{(1-\alpha)\mathbf{x}_i^{(2)}+\alpha\mathbf{x}_i^{(3)}}\right)\right\}+\nu=0, \;\; i\in[[0,N]], \\ \gamma\left(d_{KL}\big((1-\alpha)\mathbf{x}^{(1)}+\alpha\mathbf{x}^{(3)},p\big)-d_{KL}\big((1-\alpha)\mathbf{x}^{(2)}+\alpha\mathbf{x}^{(3)},p\big)\right)=0, \end{cases} \tag{5.3}$$

where λ, μ, ν are the Lagrange multipliers associated with the constraints ($P_{\alpha,\epsilon}$(3)) (corresponding to u in the definition of the Lagrangian).

Let us point out that these optimality conditions do not hold for feasible points such that $\mathbf{x}_j^{(i)}=0$ for some i and j, because our problem is not smooth at these points. It only holds for feasible points in the interior of $\mathbb{R}_+^{3(N+1)}$. In Lemma 5.1, we show that there is no optimal solution in the boundary of $\mathbb{R}_+^{3(N+1)}$ that justifies the use of conditions (5.3). The set of Lagrange multipliers associated with a feasible point (x, y, z) is denoted $\mathbf{L}(x, y, z)$ (and is a subset of $\mathbb{R}_-^{3(N+1)}\times\mathbb{R}^3\times\mathbb{R}_-$). In particular, let us emphasize that due to the inequality constraint ($P_{\alpha,\varepsilon}$(4)), we require $\gamma\le 0$. We let $\mathbf{S}(\alpha, \varepsilon)$ denote the set of solutions of the above problem for given parameters (α, ε) and let $\mathbf{F}(\alpha, \varepsilon)$ be the set of feasible points. In the particular case where $\varepsilon=0$, we use the notation $\mathbf{S}(\alpha)$ and $\mathbf{F}(\alpha)$ for $\mathbf{S}(\alpha, 0)$ and $\mathbf{F}(\alpha, 0)$ respectively. Our first goal is to show that for any $\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\right)\in\mathbf{S}(\alpha, \varepsilon)$, we have $\mathbf{x}_i^{(j)}>0$ for all i,j. This is the point of the following lemma.

Lemma 5.1. *Consider the set* $\mathbf{S}(\alpha, \varepsilon)$ *of solutions of problem* ($P_{\alpha,\varepsilon}$). *Then, for* ε *small enough and any* $\alpha\in(0, 1)$, *we have*

$$\mathbf{S}(\alpha, \varepsilon)\cap\partial\mathbb{R}_+^{3(N+1)}=\emptyset.$$

Proof. The proof has been deferred to Appendix A.1. □

Remark 5.1. In the cases where $\varepsilon=0$, note that one can easily check using the first-order optimally conditions that the inequality constraint ($P_{\alpha,\varepsilon}$(4)) is always saturated. Thus, in the following, we will always assume that $\gamma<0$.

Proof of Theorem5.1.
Step 1: Solving ($P_{0,0}$) *i.e.* $\alpha=\varepsilon=0$. In this case the first-order optimality conditions (5.3) become

$$\begin{cases} \log\left(\dfrac{x_i^{(0)}}{p_i}\right)+1+\gamma\left\{\log\left(\dfrac{x_i^{(0)}}{p_i}\right)+1\right\}+\lambda=0, \quad \text{for all } i\in[\![0,N]\!], & \text{(5.4a)} \\ \log\left(\dfrac{y_i^{(0)}}{q_i}\right)+1-\gamma\left\{\log\left(\frac{y_i^{(0)}}{p_i}\right)+1\right\}+\mu=0, \quad \text{for all } i\in[\![0,N]\!], & \text{(5.4b)} \\ \gamma\left(d_{KL}x^{(0)}p-d_{KL}y^{(0)}p\right)=0, & \text{(5.4c)} \\ \sum_{i=0}^{N}x_i^{(0)}=\sum_{i=0}^{N}y_i^{(0)}=1. & \text{(5.4d)} \end{cases}$$

If we assume that $\gamma \neq -1$, then (5.4a), (5.4c), and (5.4d) lead to

$$x_i^{(0)} = y_i^{(0)} = p_i \quad \text{for all } i \in [\![0, N]\!],$$

which is not compatible with (5.4d) unless $p = q$. In addition, $\gamma = 0$ leads to $x^{(0)} = p$ and $y^{(0)} = q$, which is easily checked to be not feasible. Thus we have $\gamma = -1$, and (5.4b) then gives

$$y_i^{(0)}\, \mathrm{e}^{\mu/2} = \sqrt{p_i q_i} \quad \text{for all } i \in [\![0, N]\!],$$

which, using (5.4d), gives

$$y_i^{(0)} = \frac{\sqrt{p_i q_i}}{\sum_{l=0}^{N} \sqrt{p_l q_l}} \quad \text{for all } i \in [\![0, N]\!].$$

It follows from (5.4c) that $\left(x^{(0)}, y^{(0)}, z^{(0)}\right)$, with

$$\begin{cases} x_i^{(0)} = y_i^{(0)} = \dfrac{\sqrt{p_i q_i}}{\sum_{l=0}^{N} \sqrt{p_l q_l}} & \text{for all } i \in [\![0, N]\!], \\ z^{(0)} = q, \end{cases}$$

is a feasible optimal solution of problem ($P_{0,0}$). In particular,

$$\begin{aligned} &f_0\left(x^{(0)}, y^{(0)}, z^{(0)}\right) \\ &= \sum_{i=0}^{N} \frac{\sqrt{p_i q_i}}{\sum_{l=0}^{N} \sqrt{p_l q_i}} \log\left(\frac{\sqrt{p_i q_i}}{p_i \sum_{l=0}^{N} \sqrt{p_l q_i}}\right) + \sum_{i=0}^{N} \frac{\sqrt{p_i q_i}}{\sum_{l=0}^{N} \sqrt{p_l q_i}} \log\left(\frac{\sqrt{p_i q_i}}{q_i \sum_{l=0}^{N} \sqrt{p_l q_i}}\right) \\ &= -2 \log\left(\sum_{l=0}^{N} \sqrt{p_i q_i}\right) \\ &= d_B(p, q), \end{aligned}$$

where $d_B(\cdot, \cdot)$ is the Bhattacharyya divergence defined in (5.2).

Step 2: Continuity of the value function. The goal of this step is to show that the full value function V is continuous. To do so, we apply Theorem 2.1 in conjunction with Theorem 2.8 of [8]. In view of these theorems, the only point that needs clarification is that

$$\overline{\{(x, y, z) \in \mathcal{M}^3 \mid H_{\alpha,\varepsilon}(x, y, z) > 0\}} = \{(x, y, z) \in \mathcal{M}^3 \mid H_{\alpha,\varepsilon}(x, y, z) \geq 0\}.$$

To do so, it suffices to show that for any $(a, b, c) \in \mathcal{M}^3$ such that $H_{\alpha,\varepsilon}(a, b, c) = 0$ and any $\delta > 0$, there exists an element $(\tilde{a}, \tilde{b}, \tilde{c}) \in \mathcal{M}^3$ such that $H_{\alpha,\varepsilon}(\tilde{a}, \tilde{b}, \tilde{c}) > 0$ with

$$\|(a, b, c) - (\tilde{a}, \tilde{b}, \tilde{c})\|_1 < \delta.$$

As the proof closely follows the ideas of the proof of Lemma 5.1, we do not write down the details. Thus V is continuous. The first statement of Theorem 5.1 now follows from the compactness of [0,1].

Step 3: Limits of $v(\alpha)/(1-\alpha)$ *as* $\alpha \to 1$. For any $\alpha \in [0, 1)$, it is easily seen that

$$\begin{aligned} \frac{v(\alpha)}{1-\alpha} = \inf_{(x,y,z)\in\mathcal{M}^3} \Bigg\{ & d_{KL}(x, p) + d_{KL}(y, q) + \frac{\alpha\, d_{KL}(z, q)}{1-\alpha} \;\Bigg|\; \\ & d_{KL}\left(x + \frac{\alpha}{1-\alpha} z, p\right) - d_{KL}\left(y + \frac{\alpha}{1-\alpha} z, p\right) \geq 0 \Bigg\}. \end{aligned}$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

This is equivalent to studying the behavior of

$$\mathcal{V}(\delta) = \inf_{(x,y,z)\in\mathcal{M}^3} \{d_{KL}(x, p) + d_{KL}(y, q) + \delta d_{KL}(z, q) \mid d_{KL}(x + \delta z, p) - d_{KL}(y + \delta z, p) \geq 0\}, \tag{5.5}$$

as δ goes to infinity. So, let $(\delta_n)_{n\geq 1}$ be some sequence of real numbers such that $\delta_n \xrightarrow[n\to\infty]{} \infty$, and set

$$\mathbf{S}_n := \mathbf{S}\left(\frac{\delta_n}{1+\delta_n}, 0\right).$$

Now, for all $n \geq 1$, choose $(x^{(n)}, y^{(n)}, z^{(n)}) \in \mathbf{S}_n$. As $\cup_{n\geq 1}\mathbf{S}_n \subset \mathcal{M}$ is relatively compact, we may assume, extracting a subsequence if needed, that $(x^{(n)}, y^{(n)}, z^{(n)})$ converges to some element $(x^*, y^*, z^*) \in \mathcal{M}^3$. Now assume that

$$\lim_{n\to\infty} \|z^{(n)} - q\|_1 > 0.$$

However, this would imply that $\liminf_{n\to\infty} d_{KL}(z^{(n)}, q) > 0$, and thus that

$$\liminf_{n\to\infty}\{d_{KL}(x^{(n)}, p) + d_{KL}(y^{(n)}, q) + \delta_n d_{KL}(z^{(n)}, q)\} \geq \liminf_{n\to\infty} \delta_n d_{KL}(z^{(n)}, q) = \infty,$$

but this is impossible, since, according to Step 1, $\mathcal{V}(\delta) \leq d_B(p, q)$ (because the solution given in Step 1 is always feasible). It follows that

$$\lim_{n\to\infty} \|z^{(n)} - q\|_1 = 0$$

and $z^* = q$. Now, for fixed $n \geq 1$, the first-order optimality conditions of problem (5.5) take the form

$$\begin{cases} \log\left(\dfrac{x_i^{(n)}}{p_i}\right) + 1 + \gamma_n\left(\log\left(\dfrac{\delta_n z_i^{(n)} + x_i^{(n)}}{p_i}\right) + 1\right) + \lambda_n = 0, \\ \log\left(\dfrac{y_i^{(n)}}{q_i}\right) + 1 - \gamma_n\left(\log\left(\dfrac{\delta_n z_i^{(n)} + y_i^{(n)}}{p_i}\right) + 1\right) + \mu_n = 0, \\ \delta_n \log\left(\dfrac{z_i^{(n)}}{q_i}\right) + \delta_n + \gamma_n\delta_n\left(\log\left(\dfrac{\delta_n z_i^{(n)} + x_i^{(n)}}{\delta_n z_i^{(n)} + y_i^{(n)}}\right)\right) + \nu_n = 0, \end{cases} \tag{5.6}$$

for some Lagrange multipliers $(\lambda_n, \mu_n, \nu_n, \gamma_n) \in \mathbf{L}(x^{(n)}, y^{(n)}, z^{(n)})$.

We now show that the sequence γ_n must be bounded. For this, let us assume that γ_n is unbounded. So, extracting a subsequence if needed, we may assume that

$$\gamma_n \xrightarrow[n\to\infty]{} -\infty.$$

The second equation of (5.6) implies that, for all i,

$$\frac{y_i^{(n)} p_i}{q_i} = p_i\left(\frac{\delta_n z_i^{(n)} + y_i^{(n)}}{p_i}\right)^{\gamma_n} \exp(-1 + \gamma_n - \mu_n).$$

Summing over i and using Jensen's inequality (since $\gamma_n < 0$), we obtain

$$\sum_{i=1}^N \frac{y_i^{(n)} p_i}{q_i} = \sum_{i=1}^N p_i\left(\frac{\delta_n z_i^{(n)} + x_i^{(n)}}{p_i}\right)^{\gamma_n} \exp(-1 + \gamma_n - \mu_n) \geq (1 + \delta_n)^{\gamma_n} \exp(-1 + \gamma_n - \mu_n).$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

Thus

$$\log\left(\sum_{i=1}^{N}\frac{y_i^{(n)}p_i}{q_i}\right)-\gamma_n\log\left(1+\delta_n\right)+1-\gamma_n+\mu_n\geq 0.$$

Now equations (5.6) also give, for all $1\leq i\leq N$,

$$\begin{aligned}&\log\left(\frac{x_i^{(n)}}{p_i}\right)+\gamma_n\log\left(\frac{z_i^{(n)}+\delta_n^{-1}x_i^{(n)}}{p_i}\right)+\gamma_n\log\left(\delta_n\right)+1+\gamma_n+\lambda_n\\&=\log\left(\frac{x_i^{(n)}}{p_i}\right)+\gamma_n\log\left(\frac{\delta_n z_i^{(n)}+x_i^{(n)}}{(\delta_n+1)p_i}\right)+\gamma_n\log\left(1+\delta_n\right)+1+\gamma_n+\lambda_n\\&=0\end{aligned}\tag{5.7}$$

and

$$\log\left(\frac{y_i^{(n)}}{q_i}\right)-\gamma_n\log\left(\frac{\delta_n z_i^{(n)}+y_i^{(n)}}{(\delta_n+1)p_i}\right)-\gamma_n\log\left(1+\delta_n\right)+1+\gamma_n+\mu_n=0.$$

Hence, for n large enough, we have

$$\log\left(\frac{y_i^{(n)}}{q_i}\right)-\gamma_n\log\left(\frac{\delta_n z_i^{(n)}+y_i^{(n)}}{(\delta_n+1)p_i}\right)-\log\left(\sum_{i=1}^{N}\frac{y_i^{(n)}p_i}{q_i}\right)\leq 0.$$

In particular, this implies that $y_i^*=0$ for all i such that $q_i>p_i$. Similarly to (5.7), we get

$$\log\left(\frac{x_i^{(n)}}{p_i}\right)+\gamma_n\log\left(\frac{\delta_n z_i^{(n)}+x_i^{(n)}}{(\delta_n+1)p_i}\right)\leq 0,$$

and $x_i^*=0$ for all i such that $p_i>q_i$. Now a direct computation gives

$$d_{KL}(x^*,p)+d_{KL}(y^*,q)\geq -\log\left(\sum_{j\in J}p_i\right)-\log\left(\sum_{i\in I}q_i\right),$$

with $I=\{i\in\{1,\ldots,N\}\mid q_i\leq p_i\}$ and $J=\{i\in\{1,\ldots,N\}\mid p_i\leq q_i\}$. However,

$$\left(\sum_{i\in J}p_i\right)\left(\sum_{i\in I}q_i\right)\leq\left(\sum_{i\in J}\sqrt{p_iq_i}\right)\left(\sum_{i\in I}\sqrt{p_iq_i}\right)<\left(\sum_{i=1}^{N}\sqrt{p_iq_i}\right)^2.$$

Thus

$$-\log\left(\sum_{j\in J}p_i\right)-\log\left(\sum_{i\in I}q_i\right)>d_B(p,q).$$

But, as the solution of Step 1 is always an admissible solution, this is absurd since it would imply that, for n large enough, $(x^{(n)},y^{(n)},z^{(n)})$ is not optimal. From the preceding, we conclude that γ_n is bounded. Thus, again extracting a subsequence if needed, we can suppose that there exists some $\gamma_\infty\leq 0$ such that

$$\gamma_n\xrightarrow[n\to\infty]{}\gamma_\infty.$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

In addition, (5.7) implies that the sequence $(\gamma_n \log(\delta_n) + \lambda_n)_{n\geq 1}$ is bounded as well (because there must be at least one i such that $\lim_{n\to\infty} x_i^{(n)} > 0$), which we may also assume to be convergent. From these and (5.7), it follows that

$$x_i^{(n)} = p_i^{1+\gamma_n} q_i^{-\gamma_n}\, \mathrm{e}^{c_n + \gamma_n \mathrm{O}(\delta_n^{-1})} = \frac{p_i^{1+\gamma_\infty} q_i^{-\gamma_\infty}}{\sum_{j=1}^{N} p_j^{1+\gamma_\infty} q_j^{-\gamma_\infty}} + \mathrm{o}(1) \quad \text{for all } 1 \leq i \leq n, \tag{5.8}$$

and similarly, we have

$$y_i^{(n)} = p_i^{-\gamma_n} q_i^{1-\gamma_n}\, \mathrm{e}^{\tilde{c}_n + \gamma_n \mathrm{O}(\delta_n^{-1})} = \frac{p_i^{-\gamma_\infty} q_i^{1+\gamma_\infty}}{\sum_{j=1}^{N} p_j^{-\gamma_\infty} q_j^{1+\gamma_\infty}} + \mathrm{o}(1) \quad \text{for all } 1 \leq i \leq n, \tag{5.9}$$

where $(c_n)_{n\geq 1}$ and $(\tilde{c_n})_{n\geq 1}$ are some convergent sequences. Denoting

$$h(\gamma) = \left(\frac{p_i^{1+\gamma} q_i^{-\gamma}}{\sum_{j=1}^{N} p_j^{1+\gamma} q_j^{-\gamma}} \right)_{1\leq i\leq N},$$

and setting $K(x) = d_{KL}(x, p)$ for any $x \in \mathcal{M}$, it follows that (see Remark 5.1)

$$\begin{aligned}
0 &= K\big(\alpha_n z^{(n)} + (1-\alpha_n) x^{(n)}\big) - K\big(\alpha_n z^{(n)} + (1-\alpha_n) y^{(n)}\big) \\
&= K(\alpha_n q + (1-\alpha_n) h(\gamma_n)) - K(\alpha_n q + (1-\alpha_n) h(-1-\gamma_n)) \\
&\quad + \nabla K(\alpha_n q + (1-\alpha_n) h(\gamma_n)) \cdot \big(\alpha_n (z^{(n)} - q) + (1-\alpha_n)(x^{(n)} - h(\gamma_n))\big) \\
&\quad - \nabla K(\alpha_n q + (1-\alpha_n) h(-1-\gamma_n)) \cdot \big(\alpha_n (z^{(n)} - q) + (1-\alpha_n)(y^{(n)} - h(-1-\gamma_n))\big) \\
&\qquad + \mathrm{O}\left(\frac{1}{\delta_n^2}\right) \\
&= K(\alpha_n q + (1-\alpha_n) h(\gamma_n)) - K(\alpha_n q + (1-\alpha_n) h(-1-\gamma_n)) \\
&\quad + \alpha_n (\nabla K(\alpha_n q + (1-\alpha_n) h(\gamma_n)) - \nabla K(\alpha_n q + (1-\alpha_n) h(-1-\gamma_n))) + \mathrm{o}\left(\frac{1}{\delta_n}\right),
\end{aligned}$$

but since ∇K exists and is continuous in a neighborhood of q, we get

$$\begin{aligned}
&K\big(\alpha_n z^{(n)} + (1-\alpha_n) x^{(n)}\big) - K\big(\alpha_n z^{(n)} + (1-\alpha_n) y^{(n)}\big) \\
&\quad = K(\alpha_n q + (1-\alpha_n) h(\gamma_n)) - K(\alpha_n q + (1-\alpha_n) h(-1-\gamma_n)) + \mathrm{o}\left(\frac{1}{\delta_n}\right) \\
&\quad = (1-\alpha_n) \nabla K(\alpha_n q) \cdot (h(\gamma_n) - h(-1-\gamma_n)) + \mathrm{o}\left(\frac{1}{\delta_n}\right) \\
&\quad = 0.
\end{aligned}$$

Finally, as $1 - \alpha_n \sim 1/\delta_n$, it follows that

$$\nabla K(q) \cdot (h(\gamma_\infty) - h(1 - \gamma_\infty)) = 0.$$

Now, since $\nabla K(q) = (\log(q_i/p_i) + 1)_{1\leq i\leq N}$, we have

$$\sum_{i=1}^{N} \log\left(\frac{q_i}{p_i}\right) (h_i(\gamma_\infty) - h_i(-1-\gamma_\infty)) = 0. \tag{5.10}$$

Since $p_i^{1+\gamma} q_i^{-\gamma} = (p_i/q_i)^{-\gamma-1/2}(p_i q_i)^{1/2}$, (5.10) can be written as

$$F(-\gamma_\infty - 1/2) = F(\gamma_\infty + 1/2), \tag{5.11}$$

where $F(\gamma) = F_1(\gamma)/F_0(\gamma)$ with, for any $k \in \mathbb{N}$,

$$F_k(\gamma) := \sum_{i=1}^{N} \log^k\left(\frac{q_i}{p_i}\right) \cdot \left(\frac{q_i}{p_i}\right)^{\gamma} \sqrt{p_i q_i}.$$

Let $\overline{\gamma_\infty} := -\gamma_\infty - 1/2$, which implies that (5.11) can be rewritten as $F(\overline{\gamma}_\infty) = F(-\overline{\gamma}_\infty)$. We shall show that the only solution of this equation is $\overline{\gamma}_\infty = 0$. One can see that $F(1/2) = d_{KL}(q, p) > 0$ and $F(-1/2) = -d_{KL}(p, q) < 0$. Since $F'_k = F_{k+1}$, we obtain

$$F' = \frac{F'_1 F_0 - F_1 F'_0}{F_0^2} = \frac{F_2 F_0 - F_1^2}{F_0^2} > 0$$

by the Cauchy–Schwarz inequality, because $\log(q_i/p_i)$ is not constant by $p \neq q$. Thus F is a strictly increasing function, which implies that the only solution is $\overline{\gamma}_\infty = 0$. Hence $\gamma_\infty = 1/2$, which finally gives, according to (5.8) and (5.9),

$$x_i^* = y_i^* = \frac{\sqrt{p_i q_i}}{\sum_{j=1}^{N} \sqrt{p_j q_j}} \quad \text{for all } 1 \leq i \leq N.$$

From this, it follows that

$$\mathcal{V}(\delta_n) \xrightarrow[n\to\infty]{} d_B(p, q),$$

which, since the sequence $(\delta_n)_{n\geq 1}$ is arbitrary, implies that

$$\lim_{\alpha\to 1} \frac{v(\alpha)}{1-\alpha} = \lim_{\delta\to\infty} \mathcal{V}(\delta) = d_B(p, q).$$

This ends the proof. □

6. Proof of Theorem 3.2 in the critical and supercritical cases

The purpose of this section is to prove Theorem 3.2 when $m(\mu) \geq 1$.

6.1. Estimation of μ

We aim to prove that $\widehat{\mu}_h$ is always convergent in these cases.

Proposition 6.1. *If $m(\mu) \geq 1$, then the estimators $\widehat{\mu}_h$ and $\widehat{\mu}_h^*$, defined in (3.1) and (3.5) respectively, satisfy*

$$\widehat{\mu}_h \xrightarrow[h\to\infty]{} \mu$$

and

$$\widehat{\mu}_h^* \xrightarrow[h\to\infty]{} \mu$$

almost surely. In addition, for any $\varepsilon > 0$ we have

$$\sum_{h\geq 1} \mathbb{P}(\|\widehat{\mu}_h - \mu\|_1 > \varepsilon) < \infty.$$

Note that the result of Proposition 6.1 is rather intuitive. Indeed, when the normal Galton–Watson subtrees are supercritical, the sample used in (3.1) or (3.5) is a perturbation of size h of a μ i.i.d. sample whose size is of order $m(\mu)^h$. Therefore our primary concern is ensuring that this perturbation is not sufficiently large to hinder the estimation process.

Proof. Recall that the spinal-structured tree T can be decomposed as the grafting of a sequence $(G_{i,j})_{i,j\geq 1}$ of i.i.d. Galton–Watson trees with common birth distribution μ onto the spine. For each of these trees, let us write $\overline{X}_{i,j,h}$, for $i, j, h \in \mathbb{N}$, the random vector defined by

$$\overline{X}_{i,j,h}(k) = \sum_{\{v\in G_{i,j}\colon \mathcal{D}(v)<h\}} \mathbb{1}_{C(v)=k}, \quad 0 \leq k \leq N.$$

We emphasize that i corresponds to generations in the spinal-structured tree whereas j corresponds to indices of the offspring in a given generation. In addition, it is known (see e.g. [7]) that the law of $\overline{X}_{i,j,h}$ conditional on $\#\{v \in G_{i,j}\colon \mathcal{D}(v) < h\}$ is multinomial with parameters μ and $\#\{v \in G_{i,j}\colon \mathcal{D}(v) < h\}$. From this, and from the independence of the $G_{i,j}$, it follows that the random variable $\overline{X}_h$ defined by

$$\overline{X}_h = \sum_{i=1}^{h} \sum_{j=1}^{S_i-1} \overline{X}_{i,j,h-i}$$

is, conditionally on $\#T_h$, a multinomial random variable with parameters $\#T_h - h$ and μ independent of $\mathcal{S}_h$. Now, letting $\overline{\mathcal{S}_h}$ denote the empirical distribution associated with $\mathcal{S}_h$, that is,

$$\overline{\mathcal{S}_h}(k) = \frac{1}{h}\sum_{i=1}^{h} \mathbb{1}_{S_i=k} \quad \text{for all } k \in \{1, \ldots, N\},$$

where $S_1, \ldots, S_h$ denote the numbers of special children of the individuals of the spine, it is easily seen that

$$\widehat{\mu}_h = \left(1 - \frac{h}{\#T_h}\right)\overline{X}_h + \frac{h}{\#T_h}\overline{\mathcal{S}_h}.$$

Now, taking $\varepsilon > 0$, Pinsker's inequality entails that

$$\mathbb{P}\big(\|\widehat{\mu}_h - \mu\|_1 > \sqrt{\varepsilon/2}\big) \leq \mathbb{P}(d_{KL}(\widehat{\mu}_h, \mu) > \varepsilon).$$

The convexity of the Kullback–Leibler divergence gives, with $\alpha_h := h/\#T_h$,

$$\begin{aligned}\mathbb{P}\big(\|\widehat{\mu}_h - \mu\|_1 > \sqrt{\varepsilon/2}\big) &\leq \mathbb{P}\big((1-\alpha_h)d_{KL}\big(\overline{X}_h, \mu\big) + \alpha_h d_{KL}\big(\overline{\mathcal{S}_h}, \mu\big) > \varepsilon\big)\\ &\leq \mathbb{P}\bigg((1-\alpha_h)d_{KL}\big(\overline{X}_h, \mu\big) + \alpha_h d_{KL}\big(\overline{\mathcal{S}_h}, \mu\big) > \varepsilon,\ \alpha_h d_{KL}(\nu, \mu) < \frac{\varepsilon}{2}\bigg)\\ &\quad + \mathbb{P}\bigg(\alpha_h d_{KL}(\nu, \mu) \geq \frac{\varepsilon}{2}\bigg).\end{aligned}$$

Next, using the method of Lemma 6.2, one can show that for any $\delta > 0$ there is a constant $C > 0$ such that

$$\begin{aligned}&\mathbb{P}\bigg((1-\alpha_h)d_{KL}\big(\overline{X}_h, \mu\big) + \alpha_h d_{KL}\big(\overline{\mathcal{S}_h}, \mu\big) > \varepsilon,\ \alpha_h d_{KL}(\mu, \nu) < \frac{\varepsilon}{2}\,\bigg|\,\#T_h\bigg)\\ &\quad \leq C\exp\left(-\#T_h(\Gamma(\alpha_h) - \delta)\right),\end{aligned}$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

with

$$\begin{aligned}\Gamma(\alpha) := &\inf_{(x,y)\in\mathcal{M}^2}\bigg\{(1-\alpha)d_{KL}(x,\,\mu)+(1-\alpha)d_{KL}(y,\,\nu)\,\bigg|\\ &\qquad (1-\alpha)d_{KL}(x,\,\mu)+\alpha d_{KL}(y,\,\mu)>\varepsilon,\,\alpha d_{KL}(\nu,\,\mu)<\frac{\varepsilon}{2}\bigg\}\\ \geq &\inf_{\substack{(x,y)\in\mathcal{M}^2\\ \alpha\in[0,1]}}\bigg\{(1-\alpha)d_{KL}(x,\,\mu)+(1-\alpha)d_{KL}(y,\,\nu)\,\bigg|\\ &\qquad (1-\alpha)d_{KL}(x,\,\mu)+\alpha d_{KL}(y,\,\mu)\geq\varepsilon,\,\alpha d_{KL}(\nu,\,\mu)\leq\frac{\varepsilon}{2}\bigg\}.\end{aligned}$$

As the feasible set of the right-hand side is obviously compact, there exists a feasible point $(x^*,\,y^*,\,\alpha^*)$ such that

$$\begin{aligned}&(1-\alpha^*)d_{KL}(x^*,\,\mu)+(1-\alpha^*)d_{KL}(y^*,\,\nu)\\ &\quad=\inf_{\substack{(x,y)\in\mathcal{M}^2\\ \alpha\in[0,1]}}\bigg\{(1-\alpha)d_{KL}(x,\,\mu)+(1-\alpha)d_{KL}(y,\,\nu)\,\bigg|\\ &\qquad\qquad (1-\alpha)d_{KL}(x,\,\mu)+\alpha d_{KL}(y,\,\mu)\geq\varepsilon,\,\alpha d_{KL}(\nu,\,\mu)\leq\frac{\varepsilon}{2}\bigg\}.\end{aligned}$$

Assume that $(1-\alpha^*)d_{KL}(x^*,\,\mu)+(1-\alpha^*)d_{KL}(y^*,\,\nu)=0$, which readily implies that $x^*=\mu$ and $y^*=\nu$, but it is easily seen that for any $\alpha\in[0,\,1]$, the point $(\mu,\,\nu,\,\alpha)$ is not feasible. Hence there exists a constant $\widetilde{C}>0$ independent of α such that

$$\Gamma(\alpha)\geq\widetilde{C}.$$

Choosing $\delta<\widetilde{C}$, we get

$$\begin{aligned}&\mathbb{P}\bigg((1-\alpha_h)d_{KL}\big(\overline{X}_h,\,\mu\big)+\alpha_h d_{KL}\big(\overline{\mathcal{S}_h},\,\mu\big)>\varepsilon,\ \alpha_h d_{KL}(\mu,\,\nu)<\frac{\varepsilon}{2}\,\bigg|\,\#T_h\bigg)\\ &\quad\leq C\exp\big(-\#T_h(\widetilde{C}-\delta)\big)\\ &\quad\leq \mathrm{e}^{-h(\widetilde{C}-\delta)},\end{aligned}$$

since $\#T_h\geq h$. Then

$$\mathbb{P}\big(\|\widehat{\mu}_h-\mu\|_1>\sqrt{\varepsilon/2}\big)\leq\mathrm{e}^{-hC}+\mathbb{P}\bigg(\alpha_h d_{KL}(\mu,\,\nu)\geq\frac{\varepsilon}{2}\bigg).$$

To ensure that the right-hand side of the previous inequality is summable for $h\geq 1$, it thus remains to check that

$$\sum_{h\geq1}\mathbb{P}\bigg(\alpha_h d_{KL}(\mu,\,\nu)\geq\frac{\varepsilon}{2}\bigg)<\infty.$$

From this point we assume that the birth distribution is critical. The supercritical case is considered below. So, to treat this, we observe that the spinal-structured tree (excluding the spine)

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

can be interpreted as a Galton–Watson tree with immigration with birth distribution μ and immigration $\tilde{\nu}$ given by $\tilde{\nu}_k = \nu_{k+1}$ for $k \geq 0$. It is known (see [14]) that the generating function of $\#T_h$ is given by

$$\mathbb{E}\big[x^{\#T_h}\big] = x^h \prod_{i=0}^{h-1} B(g_i(x)), \tag{6.1}$$

where $B\colon [0,\, 1] \mapsto \mathbb{R}$ is the generating function of the law $\tilde{\nu}$, and g_i is the generating function of the total progeny of a Galton–Watson tree with law μ up to generation i, that is,

$$g_i(x) = \mathbb{E}\big[x^{\sum_{j=0}^{i} Z_j}\big] \quad \text{for all } x \in [0,\, 1],$$

where $(Z_i)_{i\geq 0}$ is a standard Galton–Watson process with birth distribution μ. Now denote

$$v_h = \mathbb{E}\big[x^{\#T_h/h}\big] \quad \text{for all } h \geq 1.$$

We then have (the regularity of B and g_i is easily checked)

$$\log(v_h) = \log(B(g_i(\theta_h))) + (\theta_{h+1} - \theta_h)\frac{g_i'(\eta_h)B'(g_i(\eta_h))}{B(g_i(\eta_h))},$$

with

$$\theta_h = \exp\left(\frac{\log(x)}{h}\right) \text{ and } \eta_h \in (\theta_h,\, \theta_{h+1}) \quad \text{for all } h \geq 1. \tag{6.2}$$

Hence (6.1) entails that

$$\log\left(\frac{v_{h+1}}{v_h}\right) = \log(x) + (\theta_{h+1} - \theta_h)\sum_{i=0}^{h-1} \frac{g_i'(\eta_h)B'(g_i(\eta_h))}{B(g_i(\eta_h))} + \log(B(g_h(\theta_{h+1}))).$$

Now, as the sequence g_i is monotonically decreasing and converging to some proper generating function g (only in the critical case; see again [14]), we obtain

$$\frac{g_i'(\eta_h)B'(g_i(\eta_h))}{B(g_i(\eta_h))} \leq \frac{m(\tilde{\nu})g_i'(\eta_h)}{B(g(\eta_h))}.$$

Now, as for $x \in (0,\, 1)$,

$$g_i'(x) = \mathbb{E}\left[\left(\sum_{j=0}^{i} Z_j\right) x^{\sum_{j=0}^{i} Z_j - 1}\right],$$

we have

$$g_i'(x) \leq -\frac{\mathrm{e}^{-1}}{x\log(x)}.$$

It follows that

$$\begin{aligned}\limsup_{h\to\infty} \log\left(\frac{v_{h+1}}{v_h}\right) &\leq \log(x) - \limsup_{h\to\infty}\left((\theta_{h+1} - \theta_h)\frac{m(\tilde{\nu})}{B(g(\eta_h))}\frac{h\,\mathrm{e}^{-1}}{\eta_h\log(\eta_h)} + \log(B(g_h(\theta_{h+1})))\right)\\ &= \log(x) + m(\tilde{\nu})\,\mathrm{e}^{-1},\end{aligned}$$

where we used (6.2) to get

$$\lim_{h\to\infty} h\eta_h\log(\eta_h) = \log(x).$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

Now, as x is arbitrary, it can always be chosen such that $\log(x) + m(\tilde{\nu})\,\mathrm{e}^{-1} < 0$, which, by the ratio test, implies that, for such x,

$$\sum_{h\geq 1} \mathbb{E}\big[x^{\#T_h/h}\big] < \infty.$$

Finally, we have

$$\mathbb{E}\big[x^{\#T_h/h}\big] \geq \mathbb{E}\big[x^{\#T_h/h}\mathbb{1}_{\#T_h/h\leq c_\varepsilon}\big] \geq x^{c_\varepsilon}\mathbb{P}\bigg(\frac{\#T_h}{h} \leq c_\varepsilon\bigg),$$

where

$$c_\varepsilon = \frac{2d_{KL}(\mu, \nu)}{\varepsilon}.$$

From this, it follows that

$$\sum_{h\geq 1} \mathbb{P}\bigg(\frac{\#T_h}{h} \leq \frac{2d_{KL}(\mu, \nu)}{\varepsilon}\bigg) < \infty. \tag{6.3}$$

We now consider the case where T_h is supercritical. A possible approach is to consider a coupling between the supercritical tree T_h and a critical tree $\widetilde{T}_h$ using a thinning procedure, in order to get the estimate

$$\mathbb{E}\big[x^{\#T_h/h}\big] \leq \mathbb{E}\big[x^{\#\widetilde{T}_h/h}\big]. \tag{6.4}$$

Indeed, now assume that T_h is supercritical. We consider a thinning of T_h where each normal individual (and its descent) is killed independently with probability p. This induces a new tree $\widetilde{T}_h$ with new normal birth distribution $\widetilde{\mu}$ such that $m(\widetilde{\mu}) = pm(\mu)$. So taking $p = m(\mu)^{-1}$ implies that $\widetilde{T}_h$ is a spinal-structured tree with critical birth distribution. Hence, from the first part of the proof, we have

$$\sum_{h\geq 1} \mathbb{E}\big[x^{\#\widetilde{T}_h/h}\big] < \infty,$$

for x such that $\log(x) + m(\tilde{\nu})\,\mathrm{e}^{-1} < 0$. But the thinning procedure used for constructing $\widetilde{T}_h$ directly implies that $\#\widetilde{T}_h \leq \#T_h$ almost surely, which gives (6.4). This implies that

$$\sum_{h\geq 1} \mathbb{E}\big[x^{\#T_h/h}\big] < \infty.$$

The remainder of the proof is the same as for the critical case. This ends the proof of the almost sure convergence of $\widehat{\mu}_h$. Concerning the almost sure convergence of $\widehat{\mu}_h^*$, first note that (6.3) implies that

$$\frac{h}{\#T_h} \xrightarrow[h\to\infty]{} 0 \tag{6.5}$$

almost surely. Now take any $\mathbf{s} \in \mathfrak{S}_h$, where we recall that $\mathfrak{S}_h$ is the set of spine candidates defined in Section 2, and consider the estimator $\mu_h^{\mathbf{s}}$ given by

$$\mu_h^{\mathbf{s}}(i) = \frac{1}{\#T_h - h} \sum_{v\in T_h\setminus\mathbf{s}} \mathbb{1}_{C(v)=i} \quad \text{for all } 0 \leq i \leq N.$$

Thus

$$|\widehat{\mu}_h(i) - \widehat{\mu}_h^{\mathbf{s}}(i)| \leq \widehat{\mu}_h\left|1 - \frac{\#T_h}{\#T_h - h}\right| + \frac{\#\mathbf{s}}{\#T_h - h} = \widehat{\mu}_h\left|1 - \frac{\#T_h}{\#T_h - h}\right| + \frac{h}{\#T_h - h},$$

and the result follows from (6.5) and the almost sure convergence of $\widehat{\mu}_h$. □

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

6.2. Spine recovery

Now, to go further in the proof of Theorem 3.2, we need to understand whether we can recover enough of the spine in order to estimate f. To do so, the idea is to show that the Ugly Duckling $\widehat{\mathcal{S}}_h$ contains a proportion of order h of special individuals. Before this result, we need some preliminary lemmas. The first one concerns Kullback–Leibler divergence.

Lemma 6.1. *Let $p \in \mathcal{M}$ such that $p_- := \inf_{0\le i\le N} p_i > 0$ and $m(p) \ge 1$. Then there exists $\varepsilon_1 > 0$ such that, for any $q, \hat{p} \in \mathcal{M}$, we have*

$$\|p-\hat{p}\|_1 < \varepsilon_1 \Longrightarrow |d_{KL}(q, \mathcal{B}p) - d_{KL}(q, \mathcal{B}\hat{p})| \le C_1\|p-\hat{p}\|_1,$$

where C_1 depends only on p.

Proof. The proof has been deferred to Appendix A.2. □

The following lemma concerns large deviations on the probability of distinguishing two samples.

Lemma 6.2. *Let p and q in $\mathcal{M}$. Let R, M, and S be three independent multinomial random variables with respective parameters $(h-n, p)$, $(h-n, q)$, and (n,q), for some integers h and n such that $h > n$. Then, for any $\delta > 0$, there exists a constant $C > 0$ such that*

$$\mathbb{P}(d_{KL}(M+S, p) + \varepsilon > d_{KL}(R+S, p)) \le C\exp\{h(-(1-\alpha)\mathfrak{D}(p, q) + \varepsilon - \delta)\},$$

where $\mathfrak{D}$ is the divergence defined in (3.7). In addition the constant C depends only on N and δ.

Proof. The proof has been deferred to Appendix A.3.

□

We can finally come to the proof of Theorem 3.2.

Proof of Theorem 3.2, critical and supercritical cases. Let us recall that, for any element $\mathbf{s} \in \mathfrak{S}_h$, $\bar{\mathbf{s}}$ is defined as the random vector given by

$$\bar{\mathbf{s}}_i = \frac{1}{h}\sum_{v\in\mathbf{s}} \mathbb{1}_{C(v)=i} \quad \text{for all } 0 \le i \le N,$$

that is, the empirical distribution of the numbers of children along $\mathbf{s}$. In addition, for any non-negative integer $l \le h$, we let $\mathfrak{S}_h^l$ denote the subset of $\mathfrak{S}_h$ such that

$$\mathfrak{S}_h^l = \{\mathbf{s} \in \mathfrak{S}_h \mid \#(\mathbf{s} \cap \mathcal{S}) \le l\}.$$

From the definition of $\widehat{\mathcal{S}}_h$, we have, for any non-negative integer l,

$$\begin{aligned}
\mathbb{P}(\#(\widehat{\mathcal{S}}_h \cap \mathcal{S}) \le l) &= \mathbb{P}\left(\max_{\mathbf{s}\in\mathfrak{S}_h^l} d_{KL}(\bar{\mathbf{s}}, \mathcal{B}\widehat{\mu}_h) > \max_{\mathbf{s}\in\mathfrak{S}_h\setminus\mathfrak{S}_h^k} d_{KL}(\bar{\mathbf{s}}, \mathcal{B}\widehat{\mu}_h)\right) \\
&\le \mathbb{P}\left(\max_{\mathbf{s}\in\mathfrak{S}_h^l} d_{KL}(\bar{\mathbf{s}}, \mathcal{B}\widehat{\mu}_h) > d_{KL}(\bar{\mathcal{S}}, \mathcal{B}\widehat{\mu}_h)\right) \\
&= \mathbb{P}\left(\bigcup_{\mathbf{s}\in\mathfrak{S}_h^l} \{d_{KL}(\bar{\mathbf{s}}, \mathcal{B}\widehat{\mu}_h) > d_{KL}(\bar{\mathcal{S}}, \mathcal{B}\widehat{\mu}_h)\}\right).
\end{aligned}$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

Now let $\varepsilon > 0$ such that $\varepsilon < \varepsilon_1$, where ε_1 is defined in Lemma 6.1. Thus, according to Lemma 6.1, we have

$$\begin{aligned}\mathbb{P}\left(\bigcup_{\mathbf{s}\in\mathfrak{S}_h^l}\{d_{KL}(\bar{\mathbf{s}}, \mathcal{B}\widehat{\mu}_h) > d_{KL}(\bar{\mathcal{S}}, \mathcal{B}\widehat{\mu}_h)\},\ \|\widehat{\mu}_h - \mu\|_1 \le C_1^{-1}\varepsilon\right)\\ \le \mathbb{P}\left(\bigcup_{\mathbf{s}\in\mathfrak{S}_h^l}\{d_{KL}(\bar{\mathbf{s}}, \mathcal{B}p) + \varepsilon > d_{KL}(\bar{\mathcal{S}}, \mathcal{B}p)\}\right),\end{aligned}$$

where C_1 is also defined in Lemma 6.1. Hence, following the notation of (3.2), we have

$$\begin{aligned}\mathbb{P}(\#(\widehat{\mathcal{S}}_h \cap \mathcal{S}) \le l) &\le \mathbb{E}\left[\sum_{i=0}^{l}\sum_{j=1}^{S_i}\sum_{\{u\in G_{i,j}\colon \mathcal{D}(u)=i\}}\mathbb{1}_{d_{KL}((1-i/h)\overline{\mathcal{A}(u)}+i/h\overline{\mathcal{S}_i},\mathcal{B}p)+\varepsilon>d_{KL}(\overline{\mathcal{S}_h},\mathcal{B}p)}\right]\\ &\quad+\mathbb{P}\big(\|\widehat{\mu}_h - \mu\|_1 > C_1^{-1}\varepsilon\big)\\ &= \mathbb{E}[S]\sum_{i=0}^{l}\mathbb{E}\left[\sum_{\{u\in G\colon \mathcal{D}(u)=i\}}\mathbb{1}_{d_{KL}(i/h\overline{\mathcal{A}(u)}+(1-i/h)\overline{\mathcal{S}_i},\mathcal{B}p)+\varepsilon>d_{KL}(\overline{\mathcal{S}_h},\mathcal{B}p)}\right]\\ &\quad+\mathbb{P}\big(\|\widehat{\mu}_h - \mu\|_1 > C_1^{-1}\varepsilon\big),\end{aligned}$$

where G is some Galton–Watson tree with birth distribution μ and we recall that $\mathcal{S}_i = (S_1, \ldots, S_i)$ are the i first elements of the spine. Now, applying the many-to-one formula (3.2), we get

$$\begin{aligned}\mathbb{E}\left[\sum_{\{u\in G\colon \mathcal{D}(u)=h\}}\mathbb{1}_{d_{KL}(i/h\overline{\mathcal{A}(u)}+(1-i/h)\overline{\mathcal{S}_{h-i}},p)+\varepsilon>d_{KL}(\bar{\mathcal{S}},p)}\right]\\ = m^i\mathbb{P}\left(d_{KL}\left(\frac{i}{h}\overline{X}+\left(1-\frac{i}{h}\right)\overline{\mathcal{S}_{h-i}}, p\right)+\varepsilon > d_{KL}\big(\overline{\mathcal{S}_h}, p\big)\right),\end{aligned}$$

where $\overline{X}$ is the empirical distribution of an i.i.d. sample $X_1, \ldots, X_i$ with law given by $\mathcal{B}\mu$ independent of $\mathcal{S}$.

Now let $\rho > 0$ be such that $\log(m(\mu)) - \mathfrak{D}(\mathcal{B}\mu, \nu) + \rho < 0$. Thus, according to Lemma 6.2 and Theorem 5.1, we can choose $\delta > 0$ and ε small enough such that $V(\alpha, \varepsilon) \ge v(\alpha) - \rho$ and

$$\mathbb{P}\big(\#\big(\widehat{\mathcal{S}}_h \cap \mathcal{S}\big) \le l\big) \le C_\delta\mathbb{E}[S]\sum_{i=0}^{l}m^{h-i}\exp\left(-h\left(v\left(\frac{i}{h}\right)-\rho\right)+h\delta\right)+\mathbb{P}\big(\|\widehat{\mu}_h - \mu\|_1 > C_1^{-1}\varepsilon\big),$$

for some constant C_δ provided by Lemma 6.2. Now let $\eta > 0$; setting $\mathcal{E}_h = h - \#(\widehat{\mathcal{S}}_h \cap \mathcal{S})$, we have

$$\begin{aligned}\mathbb{P}\left(\frac{\mathcal{E}_h}{h} > \eta\right) &= \mathbb{P}\big(\#\big(\widehat{\mathcal{S}}_h \cap \mathcal{S}\big) \le \lfloor(1-\delta)h\rfloor\big)\\ &\le C_\delta\mathbb{E}[S]\sum_{\alpha\in L_h}\exp\left(h((1-\alpha)\log(m) - v(\alpha) + \rho + \delta)\right) + \mathbb{P}(\|\widehat{\mu}_h - \mu\|_1 > \varepsilon),\end{aligned}$$

with

$$L_h = \left\{\frac{i}{h}\mid 0 \le i \le \lfloor(1-\delta)h\rfloor\right\}.$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

One can now easily control the probability by

$$\mathbb{P}\left(\frac{\mathcal{E}_h}{h} > \eta\right) \leq C_\delta \mathbb{E}[S](h+1)\exp\left(h \sup_{\alpha\in[0,(1-\eta)]} \left\{(1-\alpha)\log(m) - V(\alpha) + \rho + \delta\right\}\right).$$

Now let us denote

$$\alpha_\eta := \arg\max{}_{\alpha\in[0,(1-\eta)]}(1-\alpha)\log(m) - V(\alpha).$$

Since $(1-\alpha)\log(m) - V(\alpha) < 0$ for all $\alpha \in [0, 1)$, we have

$$\alpha_\eta \xrightarrow[\eta\to 0]{} 1,$$

by virtue of the continuity of V. Now, according to Theorem 5.1, we have

$$(1-\alpha)\log(m) - v(\alpha) \sim_{\alpha\to 1} (1-\alpha)(\log(m) - d_B(p, q)).$$

Thus, for a fixed $\kappa > 0$ and for η small enough, we have

$$(1-\alpha_\eta)\log(m) - v(\alpha_\eta) \leq (1-\kappa)(1-\alpha_\eta)(\log(m) - d_B(p, q)),$$

which gives

$$\mathbb{P}\left(\frac{\mathcal{E}_h}{h} > \eta\right) \leq C_\delta \mathbb{E}[S](h+1)\exp(h\{(1+\kappa)(1-\alpha_\eta)(\log(m) - d_B(p, q)) + \rho + \delta\}).$$

Thus, for ρ and δ small enough, we find that $\mathbb{P}(\mathcal{E}/h > \eta)$ converges to zero exponentially fast, which entails that

$$\frac{\mathcal{E}_h}{h} \xrightarrow[h\to\infty]{} 0$$

almost surely, and ends the proof. □

7. Simulation study

The numerical results presented in this section were obtained using the `Python` library `treex` [3] dedicated to tree simulation and analysis.

7.1. Consistency of estimators

This part is devoted to the illustration of the consistency result stated in Theorem 3.2 through numerical simulations. For each of three normal birth distributions μ (subcritical, critical, and supercritical), we have simulated 50 spinal-structured trees until generation $h_{\max} + 1$, with $h_{\max} = 125$. The birth distribution of special nodes ν is obtained from μ and f using (1.2), where the only condition imposed on f (in the critical and supercritical regimes) was that the convergence criterion $\mathcal{K}(\mu, \nu) = \log\ m(\mu) - \mathfrak{D}(\mathcal{B}\mu, \nu)$ is negative. The values of the parameters selected for this simulation study are presented in Table 1.

For each of these trees, we have estimated the unknown model parameters for observation windows h between 5 and $h_{\max}$ with a step of 5. The normal birth distribution is estimated twice: by the (biased) maximum likelihood estimator $\widehat{\mu}_h$ given in (3.1) and by the corrected estimator $\widehat{\mu}_h^\star$ defined in (3.5). The transform function f is estimated by $\widehat{f_h}$ defined in (3.6). Finally, the special birth distribution ν is estimated by

$$\widehat{\nu}_h(k) \propto \widehat{f_h}(k)\widehat{\mu}_h^\star(k) \quad \text{for all } 0 \leq k \leq N.$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

TABLE 1. Values of the parameters μ and f selected for the simulation of the spinal-structured trees in the subcritical, critical, and supercritical regimes, as well as the associated convergence criterion $\mathcal{K}(\mu, \nu) = \log\, m(\mu) - \mathfrak{D}(\mathcal{B}\mu, \nu)$.

	Subcritical			Critical			Supercritical		
k	0	1	2	0	1	2	0	1	2
$\mu(k)$	0.35	0.4	0.25	0.4	0.2	0.4	0.29	0.4	0.31
$f(k)$	0	1	3	0	1	3	0	1	4
$\mathcal{K}(\mu, \nu)$	−0.116			−0.017			−0.006		

For these four numerical parameters, we have computed the error in the L^1-norm (since f is identifiable only up to a multiplicative constant, both f and $\widehat{f}_h$ were normalized so that their sum is 1 as before). The spine is estimated by the Ugly Duckling $\widehat{\mathcal{S}}_h$ defined in (3.4). In this case, the estimation error is given by the proportion of special nodes not recovered by $\widehat{\mathcal{S}}_h$.

The average errors computed for each of the five estimators from varying observation windows h are presented for the three growth regimes in Figure 2. First of all, note that the five average errors tend to vanish when h increases (even if it is with different shapes) for any growth regime (the convergence criterion is checked in the three examples). This illustrates the consistency of the estimators stated in Theorem 3.2. However, additional pieces of information can be obtained from these simulations.

- It can be remarked that the correction of $\widehat{\mu}_h$ is useful only in the subcritical regime. In the two other regimes, one can indeed observe that the errors related to $\widehat{\mu}_h$ and $\widehat{\mu}_h^\star$ are almost superimposed. This is due to the fact that, in these growth regimes, the number of normal nodes is sufficiently large (compared to the number of special nodes) so that the bias of the maximum likelihood estimator vanishes.
- The estimators of f and ν are clearly less accurate than $\widehat{\mu}_h^\star$, in particular in the critical and supercritical regimes. A first but likely negligible reason is that $\widehat{f}_h$ is computed from $\widehat{\mu}_h^\star$, which should only add an error to the one associated with the latter. Furthermore, the number of special nodes (used to estimate $\widehat{f}_h$) is smaller than the number of normal nodes (used to estimate $\widehat{\mu}_h^\star$).
- The estimator of the spine seems to converge, but slowly compared to the other estimates. However, we emphasize that, when h increases, the number of unknown node types increases as well, contrary to μ, f, and ν, for which the dimension is fixed. It is thus expected to observe a slower convergence rate.

7.2. Asymptotic test of conditioned Galton–Watson trees

When observing a population modeled by a Galton–Watson tree, it is of primary importance to know whether or not it has been conditioned to survive, in particular when the birth distribution is subcritical. Here we show how the theoretical contributions of this paper can be used to develop an asymptotic test to answer this question.

We observe a subcritical tree T until generation h and would like to test the null hypothesis: T is a Galton–Watson tree conditioned to survive until (at least) generation h. In the framework of spinal-structured trees and approximating conditioned Galton–Watson trees by Kesten's

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

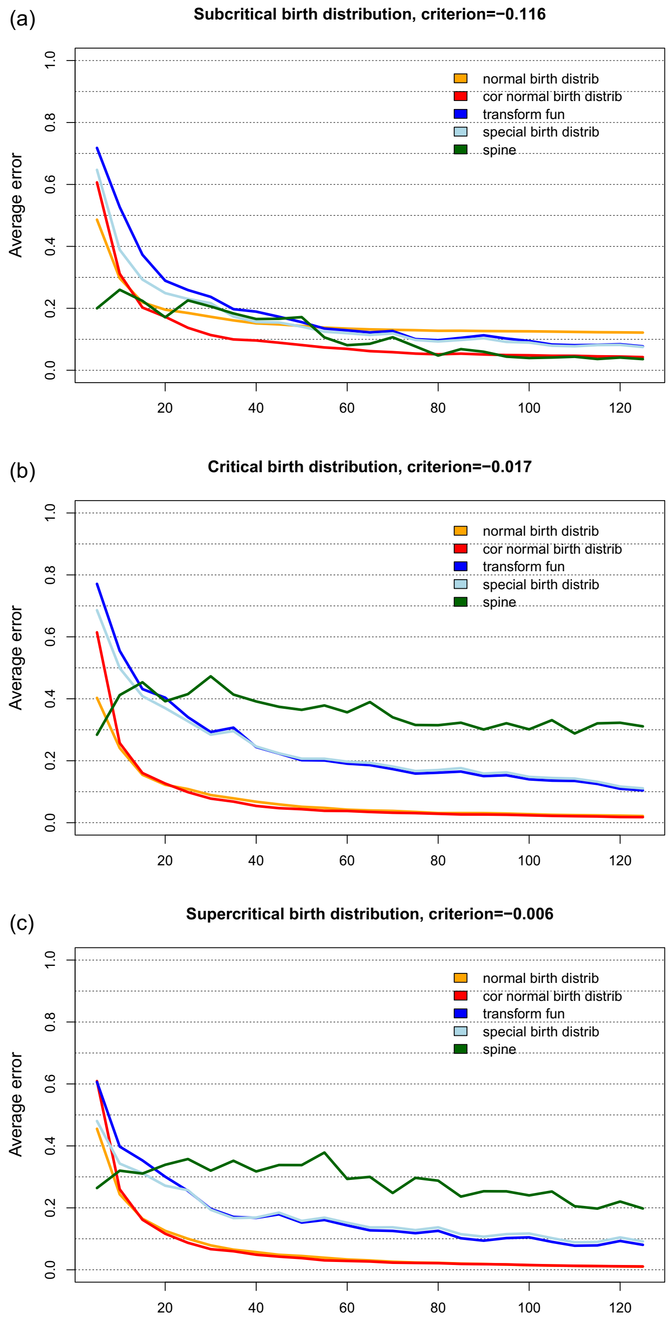


FIGURE 2. Average error as a function of the maximum height observed in the estimation of the unknown parameters (orange and red, μ; blue, f; light blue, ν; green, $\mathcal{S}$) of a spinal-structured tree in the three growth regimes: (a) subcritical, (b) critical, (c) supercritical. Parameter values can be read in Table 1.

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

TABLE 2. Empirical rejection rate measured when testing the null hypothesis $f \propto \mathrm{Id}$ from samples of 100 Kesten's trees with normal birth distribution (0.55, 0.2, 0.25) observed until generation h for various levels of confidence $1-\alpha$.

		Empirical rejection rate			
α	$\mathbf{q}(1-\alpha)$	$h=50$	$h=100$	$h=200$	$h=500$
10%	2.71	19%	8%	15%	5%
5%	3.84	10%	6%	9%	2%
1%	6.63	2%	3%	4%	0%

model [10], this is equivalent to testing $f \propto \mathrm{Id}$, which simplifies the construction of the test but also provides a further motivation for the class of models considered in this paper. As in Section 7.1, we assume that $\sum_{i=0}^{N} f(i) = 1$.

The construction of a test statistic for f requires both a consistent estimator and some knowledge of its asymptotic behavior. The latter is sorely lacking but can be estimated from numerical simulations. Here we restrict ourselves to binary ($f(1)$ is therefore sufficient to know f) spinal-structured trees with $0.5 < m(\mu) < 1$ and $0 < f(1) < 0.5$, that is, f is increasing because $f(1)+f(2)=1$. We suspect that $\widehat{f_h}(1)$ satisfies a central limit theorem with rate $\sqrt{h}$. However, we want to check if this rate seems to be adequate, so we need to estimate its asymptotic variance, possibly as a function of both μ and f. To this end, we have estimated $h\,\mathrm{Var}\,(\widehat{f_h}(1)-f(1))$ from simulated samples of spinal-structured trees from various values of μ and f within the range specified above: the results are presented in Figure 3(a). First, we observe that $h\,\mathrm{Var}\,(\widehat{f_h}(1)-f(1))$ seems to be constant in h for any value of the two parameters, which validates the rate $\sqrt{h}$. In addition, the asymptotic variance clearly depends on the parameters, but can be accurately predicted from them by a linear regression:

$$\sigma^2(\mu, f) = 0.4611141 - 0.5561625 \times m(\mu) + 1.0688165 \times f(1).$$

In Figure 3(b) we display the distribution of $\sqrt{h}(\widehat{f_h}(1)-f(1))/\sigma(\widehat{\mu}_h, \widehat{f_h})$, which is very close to the Gaussian distribution, as expected.

Relying on this short simulation study and recalling that $f(1) = 1/3$ in Kesten's model ($f \propto \mathrm{Id}$ and $f(1)+f(2)=1$), we introduce the test statistic

$$Q_h = \frac{h(\widehat{f_h}(1) - 1/3)^2}{\sigma^2(\widehat{\mu}_h, \widehat{f_h})},$$

which approximately follows a $\chi^2(1)$ distribution when the underlying tree is sampled according to Kesten's model. Denoting $\mathbf{q}(x) = \mathbb{P}(\chi^2(1) > x)$, one rejects the hypothesis $f \propto \mathrm{Id}$ with confidence level $1-\alpha$ when $Q_h > \mathbf{q}(1-\alpha)$. Figure 4(a) illustrates the behavior of Q_h when the tested hypothesis is true. Furthermore, Table 2 shows that the test rejects the null hypothesis with approximately the expected frequency of error α.

To go further, the behavior under the alternative hypothesis needs to be investigated. That is why we propose to apply the test to a population that does not follow Kesten's model. For this purpose, we consider a Galton–Watson model with competition: for any $k \geq 0$, each node of the kth generation gives birth, with distribution μ_s depending on its size s, to a random number of children,

$$\mu_s = \left(\frac{1}{4}\left(1 - \frac{1}{s}\right), \frac{3}{4}\left(1 - \frac{1}{s}\right), \frac{1}{s}\right).$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

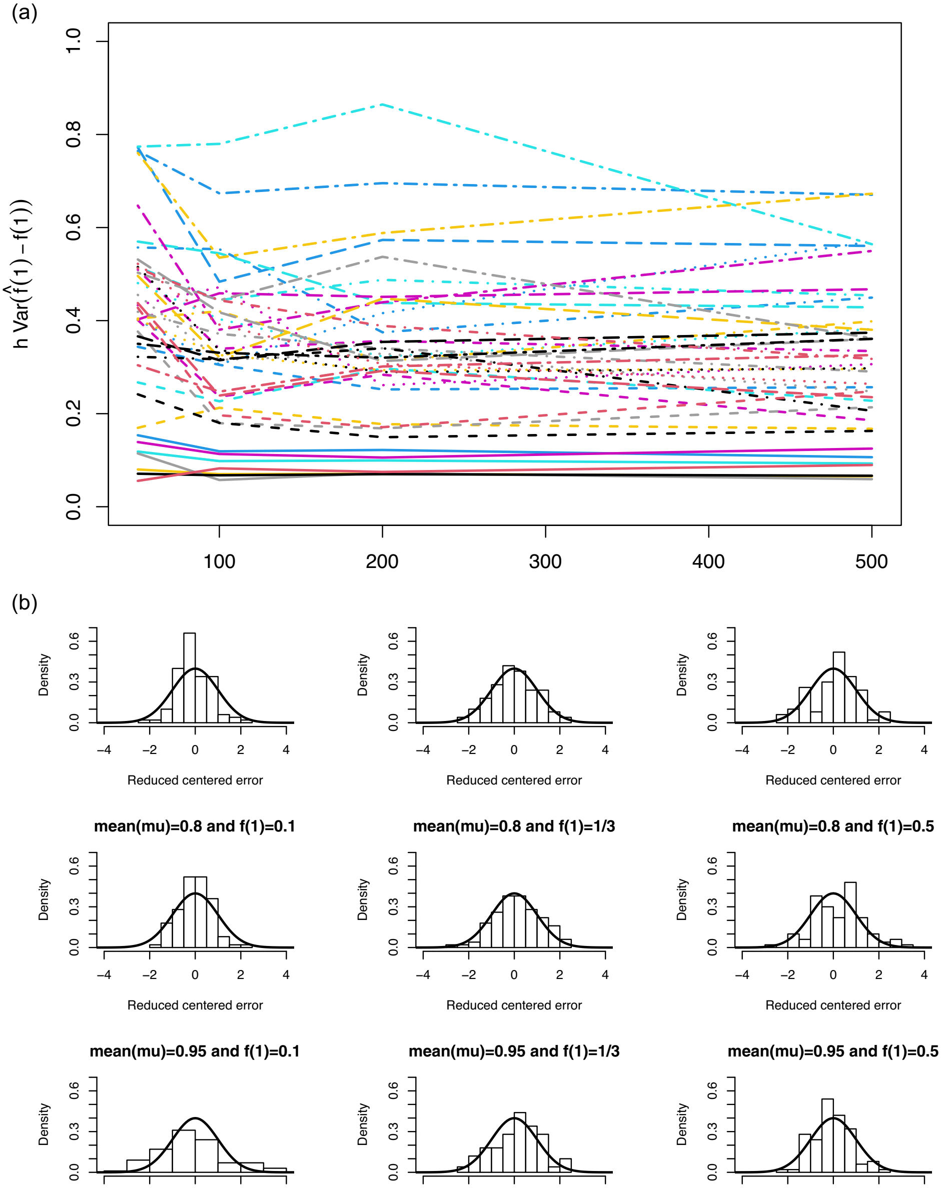
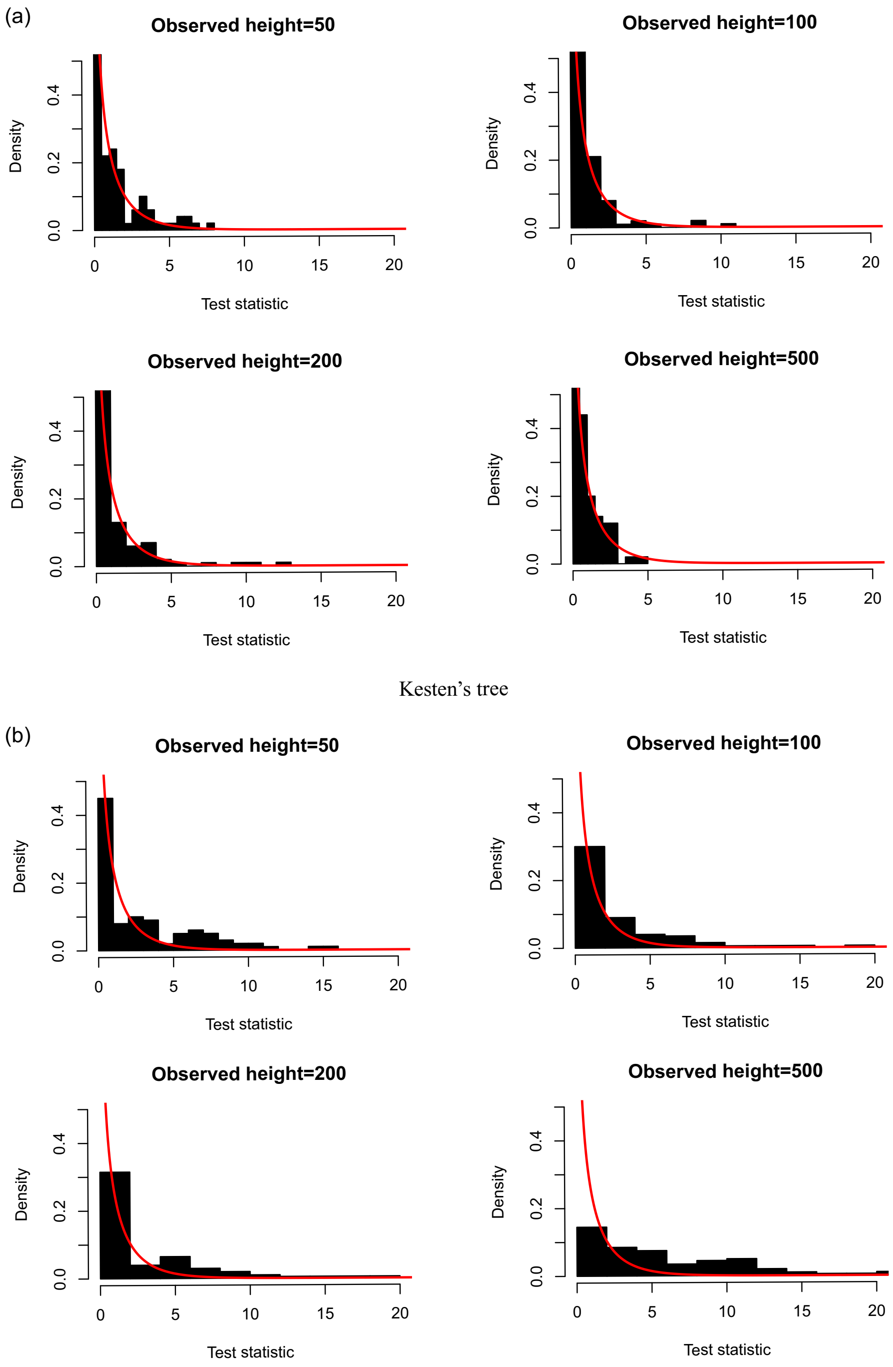


FIGURE 3. (a) Estimates of $h\,\mathrm{Var}\,(\widehat{f}_h(1) - f(1))$ from samples of 100 spinal-structured trees simulated with various parameters μ and f with $0.5 < m(\mu) < 1$ and $0 < f(1) < 0.5$, and (b) empirical distribution of the reduced centered error $\sqrt{h}(\widehat{f}_h(1) - f(1))/\sigma(\widehat{\mu}_h, \widehat{f}_h)$ for some of these parameters with a comparison to the Gaussian distribution (thick black line).

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press



Kesten's tree

population model with competition

FIGURE 4. Empirical distribution of the test statistic Q_h obtained from (a) samples of 100 Kesten's trees with normal birth distribution (0.55, 0.2, 0.25), and (b) samples of 100 Galton–Watson trees with competition, both with a comparison to the $\chi^2(1)$ distribution (red line).

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

It should be noted that $m(\mu_s) = 3/4 + 5/(4s)$, which makes the population growth supercritical when $s \leq 4$, critical when $s = 5$, and subcritical when $s \geq 6$. Oscillating between exponential growth and decay, the population is likely to avoid extinction, without fitting to the behavior of a Galton–Watson tree conditioned on surviving. Figure 4(b) provides the distribution of the test statistic Q_h under this model: it is significantly different from the $\chi^2(1)$ distribution expected under the null hypothesis, even from small populations, and thus differentiates from conditioned Galton–Watson trees.

Appendix. Proofs of intermediate lemmas

A.1. Proof of Lemma 5.1

Let $(x, y, z) \in \mathbf{F}(\alpha, \varepsilon)$. The proof is based on the fact that if x, y, or z has some null coordinate, we may always perturb these points in order to decrease the objective function while still remaining in $\mathbf{F}(\alpha, \varepsilon)$. For the sake of simplicity, we assume throughout the proof that $x_i > 0$, $z_i > 0$, and $y_i > 0$, as soon as $i \neq 0$. The other cases can be treated similarly and are left to the reader. The harder case is when we have $x_0 = z_0 = 0$ and $y_0 > 0$, but to give an example we first treat the case where $x_0 = 0$, $z_0 > 0$ and $y_0 > 0$. Henceforth, for any $X \in \mathbb{R}^{N+1}$, we let $I(X)$ denote the set given by

$$I(X) = \{i \in [\![0, N]\!] \mid X_i = 0\},$$

and let $\nabla^+ d_{KL}(x, p)$ be the vector given by

$$\nabla^+ d_{KL}(x, p) = \begin{cases} 1 + \log(x_i/p_i) & \text{for all } i \in I^c(x), \\ 0 & \text{for all } i \in I(x), \end{cases}$$

which is the vector of directional derivatives of $d_{KL}(\cdot\,, p)$ in the directions where they are well-defined.

Case 1: $x_0 = 0$, $z_0 > 0$ and $y_0 > 0$. We first show that whenever $h \in \mathbb{R}^{N+1}$ satisfies $h_0 > 0$ we have

$$\limsup_{\delta \to 0} \frac{f_\alpha(x + \delta h, y, z) - f_\alpha(x, y, z)}{\delta} < 0. \tag{A.1}$$

For any $i \in I(x)$ and any $\delta > 0$,

$$f_\alpha(x + \delta \mathbf{e}_i, y, z) - f_\alpha(x, y, z) = (1 - \alpha)\delta \log\left(\frac{\delta}{p_i}\right),$$

where $\mathbf{e}_i$ is the ith vector of the canonical basis of $\mathbb{R}^{N+1}$. This easily entails that

$$\lim_{\delta \to 0} \frac{f_\alpha(x + \delta \mathbf{e}_i, y, z) - f_\alpha(x, y, z)}{\delta} = -\infty. \tag{A.2}$$

Now take $h \in \mathbb{R}^{N+1}$ such that $h_0 > 0$. We have that $h = h^{I(x)} + h^{I^c(x)}$, where $h^{I(x)}$ is given by

$$h^{I(x)} = \begin{cases} h_i & \text{if } i \in I(x), \\ 0 & \text{else,} \end{cases}$$

and $h^{I^c(x)} = h - h^{I(x)}$. Thus, because f has well-defined directional derivatives in the direction of the positive coordinate, we have

$$f_\alpha(x + \delta h, y, z) = f_\alpha(x + \delta h^{I(x)}, y, z) + \delta \nabla^+ f_\alpha(x, y, z) \cdot h^{I^c(x)} + \mathrm{o}(\delta). \tag{A.3}$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

Thus (A.3) implies

$$\begin{aligned}&\limsup_{\delta\to 0}\frac{f_\alpha(x+\delta h, y, z)-f_\alpha(x, y, z)}{\delta}\\&\quad=\limsup_{\delta\to 0}\frac{f_\alpha(x+\delta h^{I(x)}, y, z)-f_\alpha(x, y, z)}{\delta}+\nabla^+ f_\alpha(x, y, z)\cdot h^{I^c(x)},\end{aligned}$$

and (A.1) now follows from (A.2). Now let $i^*\in[\![0, N]\!]$ such that

$$\log\left(\frac{(1-\alpha)x_{i^*}+\alpha z_{i^*}}{p_{i^*}}\right)=\min_{0\le i\le N}\log\left(\frac{(1-\alpha)x_i+\alpha z_i}{p_i}\right).$$

Because $(1-\alpha)x+\alpha z\in\mathcal{M}$ and $p\in\mathcal{M}$, we must have

$$\log\left(\frac{(1-\alpha)x_{i^*}+\alpha z_{i^*}}{p_{i^*}}\right)<0.$$

Thus, taking $h=e_0-e_{i^*}$, we get

$$H_{\alpha,\varepsilon}(x+\delta h, y, z)-H_{\alpha,\varepsilon}(x, y, z)=\delta(1-\alpha)\left(\log\left(\frac{\alpha z_0}{p_0}\right)-\log\left(\frac{(1-\alpha)x_{i^*}+\alpha z_{i^*}}{p_{i^*}}\right)\right)+\mathrm{o}(\delta).$$

For δ small enough, we obtain

$$H_{\alpha,\varepsilon}(x+\delta h, y, z)\ge H_{\alpha,\varepsilon}(x, y, z),$$

which implies that $(x+\delta h, y, z)\in\mathbf{F}(\alpha, \varepsilon)$. In addition, (A.1) implies $f(x+\delta h, y, z)<f(x, y, z)$ for δ small enough. Thus (x,y,z) cannot be a solution of problem ($P_{\alpha,\epsilon}$).

Case 2: $x_0=z_0=0$ and $y_0>0$. This particular case raises a new difficulty. Informally, in such a situation a perturbation of type $(x+\delta h, y, z+\delta\tilde{h})$ gives

$$H_{\alpha,\varepsilon}(x+\delta h, y, z+\delta\tilde{h})=H_{\alpha,\varepsilon}(x, y, z)+\delta\log(\delta)+\mathrm{o}(\delta\log(\delta)).$$

It follows that H is decreasing in any direction of type $\delta(h, 0, \tilde{h})$, and $(x+\delta h, y, z+\delta\tilde{h})$ may not be in $\mathbf{F}(\alpha, \varepsilon)$. To overcome this problem, we consider perturbed points of the form

$$\begin{cases}x^\delta=x+\delta\mathbf{e}_0-\delta\mathbf{e}_i+\delta\log(\delta)r,\\ z^\delta=z+\delta\mathbf{e}_0-\delta\mathbf{e}_i,\end{cases}$$

where $r\in\mathbb{R}^{N+1}$ satisfies

$$\begin{cases}r_0=0,\\ \sum_{i=0}^N r_i=0.\end{cases}\tag{A.4}$$

Thus, for sufficiently small δ, we have $z^\delta\in\mathcal{M}$ and $x^\delta\in\mathcal{M}$. Because $x_i>0$ and $z_i>0$ for all $i>0$, we have

$$d_{KL}(x^\delta, p)=d_{KL}(x, p)+\delta\log(\delta)+\delta\log(\delta)r\cdot\nabla^+ d_{KL}(x, p)+\mathrm{o}(\delta\log(\delta))$$

and

$$d_{KL}(z^\delta, q)=d_{KL}(z, q)+\delta\log(\delta)+\mathrm{o}(\delta\log(\delta)),$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

which gives

$$f_\alpha(x^\delta, y, z^\delta) = f(x, y, z) + \delta \log(\delta)(1 + (1-\alpha)r \cdot \nabla^+ d_{KL}(x, p)) + o(\delta \log(\delta)). \tag{A.5}$$

Similarly, we get

$$\begin{aligned} &H_{\alpha,\varepsilon}(x^\delta, y, z^\delta) \\ &\quad = H_{\alpha,\varepsilon}(x, y, z) + \delta \log(\delta)(1 + (1-\alpha)r \cdot \nabla^+ d_{KL}((1-\alpha)x + \alpha z, p)) + o(\delta \log(\delta)). \end{aligned} \tag{A.6}$$

Our next step is to show that for some choice of r and δ small enough, we have $f(x^\delta, y, z^\delta) \leq f(x, y, z)$ and $h(x^\delta, y, z^\delta) \geq h(x, y, z)$, which imply $(x^\delta, y, z^\delta) \in \mathbf{F}(\alpha, \varepsilon)$ and that (x, y, z) is not a minimizer of f among the feasible set $\mathbf{F}(\alpha, \varepsilon)$. To show this, by virtue of (A.5) and (A.6), we only need to find some $r \in \mathbb{R}^{N+1}$ satisfying conditions (A.4) and

$$\begin{cases} -r \cdot \nabla^+ d_{KL}(x, p) \leq 1, \\ r \cdot \nabla^+ d_{KL}((1-\alpha)x + \alpha z, p) \leq -1, \end{cases} \tag{A.7}$$

in particular because $\delta \log(\delta) < 0$ for δ small enough. According to Farkas's lemma, such an r exists as soon as there is no solution (u,v,w) to the problem

$$\begin{cases} -u \log\left(\dfrac{x_i}{p_i}\right) + v \log\left(\dfrac{(1-\alpha)x_i + \alpha z_i}{p_i}\right) + w = 0 \quad \text{for all } i > 0, \\ u - v < 0, \\ u > 0, \ v > 0, \ w > 0. \end{cases} \tag{A.8}$$

Assume that (u,v,w) is a solution to problem (A.8). Thus, for all $i > 0$,

$$x_i = \mathrm{e}^{w/u} p_i \left(\frac{(1-\alpha)x_i + \alpha z_i}{p_i}\right)^{v/u}.$$

Hence, according to Jensen's inequality and the conditions of problem (A.8),

$$1 = \mathrm{e}^{w/u} \sum_{i=1}^{N} p_i \left(\frac{(1-\alpha)x_i + \alpha z_i}{p_i}\right)^{v/u} \geq \mathrm{e}^{w/v} > 1,$$

which is absurd. Thus problem (A.8) has no solution, and Farkas's lemma entails that there exists $r \in \mathbb{R}^{N+1}$ such that conditions (A.4) and (A.7) are satisfied.

The method is similar if we have more than one zero. This ends the proof.

A.2. Proof of Lemma 6.1

We begin the proof by showing that the Kullback–Leibler divergence $(q, p) \mapsto d_{KL}(q, p)$ is locally Lipschitz in the second variable away from 0, and that this holds uniformly with respect to the first variable. We have

$$\nabla_2 d_{KL}(q, p) = \left(-\frac{q_i}{p_i}\right)_{1 \leq i \leq N},$$

where ∇_2 denotes the gradient with respect to the second variable. Hence

$$\|\nabla_2 d_{KL}(q, p)\|_1 \leq \frac{N}{p_-}.$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

Given $0 < \varepsilon < p_-$, as soon as $\|p - \hat{p}\|_1 < \varepsilon$, we have

$$\sup_{\{\hat{p} \in \mathcal{M} \,|\, \|p-\hat{p}\|_1 < \varepsilon\}} \|\nabla_2 d_{KL}(q, \hat{p})\|_1 \le \frac{N}{p_- - \varepsilon},$$

which entails that

$$|d_{KL}(q, p) - d_{KL}(q, \hat{p})| \le \frac{N}{p_- - \varepsilon} \|p - \hat{p}\|_1 \quad \text{for all } \hat{p} \in \mathcal{M} \text{ s.t. } \|p - \hat{p}\|_1 < \varepsilon.$$

To go further, we need to investigate the effect of a perturbation of p on $\mathcal{B}p$, where the operator $\mathcal{B}$ was defined in (3.3). Now one can easily see that on the open set $\{p \in \mathcal{M} \,|\, m(p) > 1/2\}$ we have

$$\|\nabla \mathcal{B}p\|_1 \le 2N.$$

As $m(p) \ge 1$, there exists $\varepsilon > 0$, such that $m(\hat{p}) > 1/2$ for all $\hat{p} \in B_1(p, \varepsilon)$. Thus, for $\hat{p} \in \mathcal{M} \cap B_1(p, \varepsilon)$, we have

$$\|\mathcal{B}\hat{p} - \mathcal{B}p\|_1 \le 2N \|p - \hat{p}\|_1,$$

which ends the proof.

A.3. Proof of Lemma 6.2

First we have

$$\begin{aligned}
&\mathbb{P}(d_{KL}(M + S, p) > d_{KL}(R + S, p)) \\
&\quad = \sum_{r_1+\ldots+r_N=h-n} \; \sum_{m_1+\ldots+m_N=h-n} \; \sum_{s_1+\ldots+s_N=h} ((h-n)!)^2 h! \prod_{i=1}^{N} \frac{p_i^{r_i} q_i^{m_i+s_i}}{r_i! m_i! s_i!} \mathbb{1}_{d_{KL}(\bar{r},p)+\varepsilon > d_{KL}(\bar{r},p)}.
\end{aligned} \tag{A.9}$$

In addition, one can easily check that

$$\begin{aligned}
\prod_{i=1}^{N} p_i^{r_i} q_i^{m_i+s_i} &= \exp\big(-(n-h) d_{KL}(\bar{r}, p) - (n-h) d_{KL}(\bar{m}, p) - h d_{KL}(\bar{s}, q)\big) \\
&\quad \times \prod_{i=1}^{N} \left(\frac{r_i}{n-h}\right)^{r_i} \left(\frac{m_i}{n-h}\right)^{m_i} \left(\frac{s_i}{h}\right)^{s_i}.
\end{aligned}$$

In addition, since

$$(h-n)!(h-n)!h! \prod_{i=1}^{N} (r_i! m_i! s_i!)^{-1} \prod_{i=1}^{N} \left(\frac{r_i}{n-h}\right)^{r_i} \left(\frac{m_i}{n-h}\right)^{m_i} \left(\frac{s_i}{h}\right)^{s_i} \le 1,$$

we obtain from (A.9) that

$$\begin{aligned}
&\mathbb{P}(d_{KL}(M + S, p) > d_{KL}(R + S, p)) \\
&\quad \le \sum_{\substack{{}_1+\ldots+r_N=h-n \\ m_1+\ldots+m_N=h-n \\ s_1+\ldots+s_N=h}} \mathrm{e}^{-(n-h) d_{KL}(\bar{r},p) - (n-h) d_{KL}(\bar{m},p) - h d_{KL}(\bar{s},q)} \mathbb{1}_{d_{KL}(\bar{r},p)+\varepsilon > d_{KL}(\bar{r},p)},
\end{aligned}$$

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

which leads to

$$\mathbb{P}(d_{KL}(M+S,p) > d_{KL}(R+S,p)) \leq \binom{h-n+N-1}{N-1}^2 \binom{h+N-1}{N-1} \exp\,(-hV(\alpha,\varepsilon)),$$

where $V(\alpha,\varepsilon)$ is defined in (5.1) and $\alpha = h/n$. Now, as

$$\binom{h+N-1}{N-1} \leq \mathrm{e}^{N-1}\left(\frac{h+N-1}{N-1}\right)^{N-1} \leq C_N h^N$$

for some constant C_N, we get

$$\mathbb{P}(d_{KL}(M+S,p) > d_{KL}(R+S,p)) \leq C_N^3 h^{3N} \exp\,(-hV(\alpha,\varepsilon)).$$

For any $\delta > 0$, one can always find some constant (independent of h and $V(\alpha,\varepsilon)$) such that

$$C_N^3 h^{3N} \exp\,(-hV(\alpha,\varepsilon)) \leq C \exp\,(-h(V(\alpha,\varepsilon)-\delta)).$$

This ends the proof.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments that helped them to considerably improve the preliminary version of the paper. In particular, one of the reviewers pointed out an error in the proof of Theorem 5.1, which we corrected by following his/her suggestions.

Funding information

There are no funding bodies to thank relating to the creation of this article.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] Abraham, R. and Delmas, J.-F. (2014). Local limits of conditioned Galton–Watson trees: the infinite spine case. *Electron. J. Prob.* **19**, 56.

[2] Athreya, K. B. and Ney, P. E. (2004). *Branching Processes*. Dover Publications, Mineola, NY. Reprint of the 1972 original.

[3] Azaïs, R., Cerutti, G., Gemmerlé, D. and Ingels, F. (2019). treex: a Python package for manipulating rooted trees. *J. Open Source Softw.* **4**, 1351.

[4] Bhat, B. R. and Adke, S. R. (1981). Maximum likelihood estimation for branching processes with immigration. *Adv. Appl. Prob.* **13**, 498–509.

[5] Carvalho, M. L. (1997). A joint estimator for the eigenvalues of the reproduction mean matrix of a multitype Galton–Watson process. *Linear Algebra Appl.* **264**, 189–203.

[6] Cloez, B., Daufresne, T., Kerioui, M. and Fontez, B. (2019). Galton–Watson process and Bayesian inference: a turnkey method for the viability study of small populations. Available at arXiv:1901.09562.

[7] Devroye, L. (2012). Simulating size-constrained Galton–Watson trees. *SIAM J. Comput.* **41**, 1–11.

[8] Fiacco, A. V. and Ishizuka, Y. (1990). Sensitivity and stability analysis for nonlinear programming. *Ann. Operat. Res.* **27**, 215–235.

[9] González, M., Martín, J., Martínez, R. and Mota, M. (2008). Non-parametric Bayesian estimation for multitype branching processes through simulation-based methods. *Comput. Statist. Data Anal.* **52**, 1281–1291.

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press

[10] Kesten, H. (1986). Subdiffusive behavior of random walk on a random cluster. *Ann. Inst. H. Poincaré Prob. Statist.* **22**, 425–487.

[11] Khaịrullin, R. (1992). On estimating parameters of a multitype Galton–Watson process by ϕ-branching processes. *Siberian Math. J.* **33**, 703–713.

[12] Lyons, R., Pemantle, R. and Peres, Y. (1995). Conceptual proofs of $L \log L$ criteria for mean behavior of branching processes. *Ann. Prob.* **23**, 1125–1138.

[13] Maaouia, F., Touati, A., et al. (2005). Identification of multitype branching processes. *Ann. Statist.* **33**, 2655–2694.

[14] Pakes, A. G. (1972). Further results on the critical Galton–Watson process with immigration. *J. Austral. Math. Soc.* **13**, 277–290.

[15] Qi, J., Wang, J. and Sun, K. (2017). Efficient estimation of component interactions for cascading failure analysis by EM algorithm. *IEEE Trans. Power Systems* **33**, 3153–3161.

[16] Staneva, A. and Stoimenova, V. (2020). EM algorithm for statistical estimation of two-type branching processes: a focus on the multinomial offspring distribution. *AIP Conference Proceedings* **2302**, 030003.

https://doi.org/10.1017/jpr.2024.56 Published online by Cambridge University Press