Limit distribution of the quartet balance index for Aldous's $\beta \geq 0$–model

Krzysztof Bartoszek

March 6, 2018

Abstract

This note accompanies T. Martínez–Coronado, A. Mir, F. Rosselló and G. Valiente's work "A balance index for phylogenetic trees based on quartets", introducing a new balance index for trees. We show here that this balance index, in the case of Aldous's $\beta \geq 0$–model, convergences weakly to a contraction–type distribution.

Keywords : Balance index; contraction type distribution; phylogenetic tree; tree shape; weak convergence

1 Introduction

Phylogenetic trees are key to evolutionary biology. However, they are not easy to summarize or compare as their topology cannot be directly tackled by basic mathematical methods. Therefore, many summary indices have been proposed in order to "project" a tree into $\mathbb{R}$. Such indices have as their aim to quantify some property of the tree and one of the most studied properties is the symmetry of the tree. Tree symmetry is commonly captured by a balance index. Multiple balance indices have been proposed, Sackin's (Sackin, 1972), Colless' (Colless, 1982) or the total cophenetic index (Mir et al., 2013). This work accompanies a newly proposed balance index—the quartet index (QI Martínez-Coronado et al., 2018b).

One of the reasons for introducing summary indices for trees is to use them for significance testing—does a tree come from a given probabilistic model. Obtaining the distribution (for a given n–number of tip species or

in the limit $n \to \infty$) of indices is usually difficult and often is done only for the "simplest" Yule (pure–birth) tree case. In this work here, using the contraction method, we show convergence in law of the (scaled and centred) quartet index and derive a representation (as a contraction type distribution) of the weak–limit. Remarkably, this is possible not only for the Yule tree case but for Aldous's more general β–model (in the $\beta \geq 0$ regime).

The paper is organized as follows. In Section 2 we introduce Aldous's β–model and the quartet index. In Section 3, where we introduce contraction–type distributions, we prove our main result—Thm. 3.1. When studying the limit behaviour of recursive–type indices for pure–birth binary trees one has that for each internal node the tips inside its clade are uniformly split into to sub–clades as the node splits. However, in Aldous's β–model this is not the case, the split is according to a BetaBinomial distribution, and a much finer analysis is required to show weak–convergence, with n, of the recursive–type index to the fixed point of the appropriate contraction. Theorem 3.1 is not specific for the quartet index but covers a more general class of models, where each internal node split divides its tip descendants according to a BetaBinomial distribution (with $\beta \geq 0$). In Section 4 we apply Thm. 3.1 to the quartet index and characterize its weak limit. Finally, in the appendix we provide R code used to simulate from this weak limit.

2 Preliminaries

2.1 Aldous's β–model for phylogenetic trees

Birth–death models are popular choices for modelling the evolution of phylogenetic trees. However, Aldous (1996, 2001) proposed a different class of models—the so–called β–model for binary phylogenetic trees

The main idea behind this model is to consider a (suitable) family $\{q_n\}_{n=2}^{\infty}$ of symmetric, $q_n(i) = q_n(n-i)$, probability distributions on the natural numbers. In particular $q_n : \{1, \ldots, n-1\} \to [0, 1]$. The tree grows in a natural way. The root node of a n–tip tree defines a partition of the n nodes into two sets of sizes i and $n-i$ ($i \in \{1, \ldots, n-1\}$). We randomly choose the number of tips of the left subtree, $L_n = i$, according to the distribution q_n and this induces the number of tips, $n - L_n$, in the right subtree. We then repeat recursively in the left and right subtrees, i.e. splitting according to the distributions q_{L_n} and q_{n-L_n} respectively. Notice that due to q_n's symmetry

the terms left and right do not have any particular meaning attached.

Aldous (1996) proposed to consider a one–parameter, $-2 \leq \beta \leq \infty$, family of probability distributions,

$$q_n(i) = \frac{1}{a_n(\beta)} \frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)}, \qquad 1 \leq i \leq n-1, \tag{1}$$

where $a_n(\beta)$ is the normalizing constant and $\Gamma(\cdot)$ the gamma function. We may actually recognize this as the BetaBinomial$(n, \beta+1, \beta+1)$ distribution and represent

$$q_n(i) = B(\beta+1, \beta+1)^{-1} \int_0^1 \left(\binom{n}{i} \tau^i (1-\tau)^{n-i} \right) \tau^\beta (1-\tau)^\beta \mathrm{d}\tau, \tag{2}$$

where $B(a, b)$ is the Beta function with parameters a and b. Writing informally, from the form of the probability distribution function, Eq. (2), we can see that if we would condition under the integral on τ, then we obtain a binomial distributed random variable. This is a key observation that is the intuition for the analysis presented here.

Particular values of β correspond to some well known models. The uniform tree model is represented by $\beta = -3/2$, and the pure birth, Yule, model by $\beta = 0$.

Of particular importance to our work is the limiting behaviour of the scaled size of the left (and hence right) subtree, $n^{-1}L_n$. Aldous (1996) characterizes the asymptotics in his Lemma 3,

Lemma 2.1 *(Aldous, 1996, Lemma 3)*

1. $\beta = \infty$, $n^{-1}L_n \xrightarrow{\mathcal{D}} \frac{1}{2}$;

2. $-1 < \beta < \infty$, $n^{-1}L_n \xrightarrow{\mathcal{D}} \tau_\beta$, *where* τ_β *has the beta distribution*

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} x^\beta (1-x)^\beta, \quad 0 < x < 1; \tag{3}$$

3. $\beta = -1$

$$\frac{\log(\min\{L_n, n - L_n\})}{\log n} \xrightarrow{\mathcal{D}} U \sim \text{Unif}[0,1];$$

4. $-2 < \beta < -1$ *for each fixed* $i \geq 1$

$$Pr(\min\{L_n, n - L_n\} = i) \to \pi^{-1} \left(\Gamma(\beta) \sin(\pi(-\beta - 1))\right) \frac{\Gamma(i + 1 + \beta)}{\Gamma(i + 1)}.$$

2.2 Quartet index

Martínez-Coronado et al. (2018b) now introduce a new type of balance index for discrete (i.e. without branch lengths) phylogenetic trees—the quartet index. This index is based on considering the number of quartets of each type made up by the leaves of the tree. A quartet is the induced subtree from choosing some four tips. For a given tree T, let $\mathcal{P}_4(T)$ be the set of quartets of the tree. Then, the quartet index of T is defined as

$$QI(T) = \sum_{\mathcal{P}_4(T)} QI(Q), \tag{4}$$

where $QI(Q)$ assigns a predefined value to a specific quartet. When the tree is a binary one (as here) there are only two possible topologies on four tips (see Fig. 1). Following Martínez-Coronado et al. (2018b) (their Table 1) we assign the value 0 to K_4 quartets and 1 to B_4 quartets. Therefore the QI for a binary tree (QIB) will be

$$QIB(T) = \text{number of } B_4 \text{ quartets of } T. \tag{5}$$

Importantly for us Martínez-Coronado et al. (2018b) show in their Lemma 7 that the quartet index has a recursive representation as

$$QIB(T_n) = QIB(T_{L_n}) + QIB(T_{n-L_n}) + \binom{L_n}{2}\binom{n - L_n}{2}, \tag{6}$$

where T_n is the tree on n leaves.

Martínez-Coronado et al. (2018b) considered various models of tree growth, Aldous's β–model, Ford's α–model (Ford, 2005, but see also Martínez-Coronado et al. (2018a)) and Chen–Ford–Winkel's α–γ–model (Chen et al., 2009). In

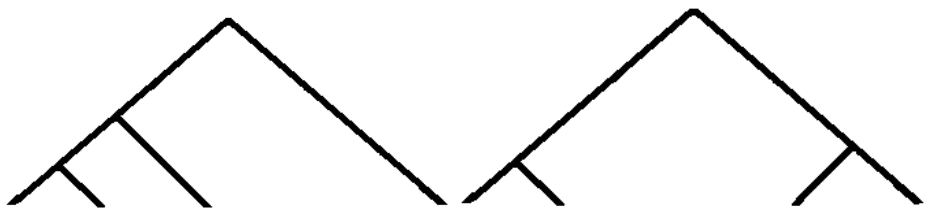

Figure 1: The two possible quartets for a binary tree. Left: K_4 the four tip caterpillar tree, right: B_4 the fully balanced tree on four nodes.

this work we will focus on the Aldous's $\beta \geq 0$–model of tree growth and characterize the limit distribution, as the number of tips n grows to infinity, of the QI. We will take advantage of the recursive representation of Eq. (6) that allows for the usage of powerful techniques developed for the so–called contraction type distributions.

We require knowledge of the mean and variance of the QI and Martínez-Coronado et al. (2018b) show for Aldous's β–model that these are (their Corollaries 20 and 24)

$$\begin{aligned} \mathrm{E}\left[QIB(T_n)\right] &= \tfrac{3\beta+6}{7\beta+18}\tbinom{n}{4} \\ \mathrm{Var}\left[QIB(T_n)\right] &= \tfrac{(\beta+2)\left(2\beta^2+9\beta+12\right)}{2(7\beta+18)^2(127\beta^3+1383\beta^2+4958\beta+5880)}n^8 + O(n^7) \end{aligned} \tag{7}$$

3 Contraction type distributions

Consider the space D of distribution functions with finite second moment and first moment equalling 0. On D we define the Wasserstein metric

$$d(F,G) = \inf \|X - Y\|_2$$

where $\|\cdot\|_2$ denotes the L_2 norm and the infimum is over all $X \sim F$, $Y \sim G$. Notice that convergence in d induces convergence in distribution. For $r \in \mathbb{N}_+$ define the transformation $S : D \to D$ as

$$S(F) = \mathcal{L}\left(\tau^r Y' + (1-\tau)^r Y'' + C(\tau)\right), \tag{8}$$

where Y', Y'', τ are independent, $Y', Y'' \sim F$, $\tau \in [0,1]$ whose distribution is not a Dirac δ at 0 nor 1, satisfying, for all n,

$$2\sum_{i=1}^{n} p_{n,i}\left(\frac{i}{n}\right)^{2r} < 1, \tag{9}$$

where $p_{n,i} = P((i-1)/n < \tau \leq i/n)$ and the function $C(\cdot)$ is of the form

$$C(\tau) = \sum_{r_1+r_2\leq r} C_{r_1,r_2}\tau^{r_1}(1-\tau)^{r_2} \tag{10}$$

for some constants C_{r_1,r_2} and furthermore satisfies $\mathrm{E}\left[C(\tau)\right] = 0$. By Rösler (1992)'s Thms. 3 and 4 S is well defined, has a unique fixed point and for any $F \in D$ the sequence $S^n(F)$ converges exponentially fast in the d metric to S's fixed point. Using the exact arguments to show Rösler (1991)'s Thm. 2.1 one can show that the map S is a contraction. Only the Lipschitz constant of convergence will differ being $\sqrt{C_\tau}$, where $C_\tau = \max\{\mathrm{E}\left[\tau^{2r}\right], \mathrm{E}\left[(1-\tau)^{2r}\right]\}$ in our case. Notice that as $\tau \in [0,1]$ and is non–degenerate at the edges, then $C_\tau < 1$ and we have a contraction.

Theorem 3.1 *(cf. Rösler, 1991, Thm. 3.1) For $n \geq 2$, $\beta > 0$ let $L_n \in \{1,\ldots,n-1\}$ be such that $(L_n - 1)$ is* BetaBinomial$(n-2, \beta+1, \beta+1)$ *distributed and $\tau \sim$* Beta$(\beta+1,\beta+1) =: F_\tau$ *distributed. Starting from the Dirac δ at 0, i.e. $Y_1 = 0$ and convention* BetaBinomial$(0, \beta+1, \beta+1) = \delta_0$, *for $r \in \mathbb{N}_+$ such that the condition of Eq. (9) is met with the previous choice of F_τ, define recursively the sequence of random variables,*

$$Y_n = \left(\frac{L_n}{n}\right)^r Y_{L_n} + \left(1 - \frac{L_n}{n}\right)^r Y_{n-L_n} + C_n(L_n),$$

where the function $C_n(\cdot)$ is of the form

$$C_n(i) = n^{-r}\left(\sum_{r_1+r_2+r_3\leq r} C_{r_1,r_2,r_3} i^{r_1}(n-i)^{r_2} n^{r_3} + h_n(i)\right), \tag{11}$$

where $\mathrm{E}\left[C_n(L_n)\right] = 0$ and $\sup_i n^{-r}h_n(i) \to 0$. If $\mathrm{E}\left[Y_n\right] = 0$ and $\mathrm{E}\left[Y_n^2\right]$ is uniformly bounded then, the random variable Y_n converges in the Wasserstein d–metric to the random variable Y_∞ whose distribution satisfies the unique fixed point of S (Eq. 8).

Notice that $L_n/n \xrightarrow{D} \tau$. It would be tempting to suspect that Thm. 3.1 should be the conclusion of a general result on contraction type distributions.

However, to the best of my knowledge, general results assume L_2 convergence of L_n/n (e.g. Thm. 8.6, p. 354 Drmota, 2009), while in our phylogenetic balance index case we will have only convergence in distribution. In such a case it seems that convergence has to be proved case by case (e.g. examples in Rachev and Rüschendorf, 1995). Here we show the convergence of Thm. 3.1 along the lines of Rösler (1991).

Lemma 3.1 *(cf. Rösler, 1991, Prop. 3.2) Let* $C_n : \{1, \ldots, n-1\} \to \mathbb{R}$ *be as in Eq.* (11)*. Then*

$$\sup_{x\in[0,1)} \Big| C_n(\lfloor (n-1)x \rfloor + 1) - C(x) \Big| \leq \sup_i n^{-r} h_n(i) + O(n^{-1}). \tag{12}$$

Proof For $1 \leq \lfloor (n-1)x \rfloor + 1 \leq n-1$ and writing $i = \lfloor (n-1)x \rfloor + 1$ we have

$$\Big| C_n(\lfloor (n-1)x \rfloor + 1) - C(x) \Big| \leq \max\{C_{r_1,r_2}\} \left(\Big| \left(\tfrac{i}{n}\right)^r - x^r \Big| + \Big| \left(1 - \tfrac{i}{n}\right)^r - (1-x)^r \Big| + \sum_{r_1+r_2\leq r} \Big| \left(\tfrac{i}{n}\right)^{r_1} \left(1 - \tfrac{i}{n}\right)^{r_2} - x^{r_1}(1-x)^{r_2} \Big| \right) + \sup_i n^{-r} h_n(i).$$

Bounding the individual components

$$\Big| \left(\tfrac{i}{n}\right)^r - x^r \Big| \leq \sup_{|y-z|<1/n} \Big| y^r - z^r \Big| \leq \sum_{0\leq y\leq 1} \Big| y^r - (y+n^{-1})^r \Big| = O(n^{-1})$$

$$\Big| \left(1 - \tfrac{i}{n}\right)^r - (1-x)^r \Big| \leq \sup_{|y-z|<1/n} \Big| (1-y)^r - (1-z)^r \Big| \leq \sum_{0\leq y\leq 1} \Big| y^r - (y+n^{-1})^r \Big| = O(n^{-1})$$

$$\begin{aligned}
&\Big| \left(\tfrac{i}{n}\right)^{r_1} \left(1 - \tfrac{i}{n}\right)^{r_2} - x^{r_1}(1-x)^{r_2} \Big| \leq \sup_{\epsilon\in[-n^{-1},n^{-1}]} \Big| y^{r_1}(1-y)^{r_2} - (y+\epsilon)^{r_1}(1-y-\epsilon)^{r_2} \Big| \\
&= \Big| y^{r_1}(1-y)^{r_2} - \left(\sum_{k_1=0}^{r_1} \tbinom{r_1}{k_1} y^{r_1-k_1} \epsilon^{k_1} \right) \left(\sum_{k_2=0}^{r_2} \tbinom{r_2}{k_2} y^{r_2-k_2} \epsilon^{k_2} \right) \Big| \\
&= \Big| \left(\sum_{k_1=1}^{r_1} \tbinom{r_1}{k_1} y^{r_1-k_1} \epsilon^{k_1} \right) + \left(\sum_{k_2=1}^{r_2} \tbinom{r_2}{k_2} y^{r_2-k_2} \epsilon^{k_2} \right) + \left(\sum_{k_1=1}^{r_1} \tbinom{r_1}{k_1} y^{r_1-k_1} \epsilon^{k_1} \right) \left(\sum_{k_2=1}^{r_2} \tbinom{r_2}{k_2} y^{r_2-k_2} \epsilon^{k_2} \right) \Big| \\
&= O(n^{-1})
\end{aligned}$$

□

Lemma 3.2 *(cf. Rösler, 1991, Prop. 3.3) Let a_n, b_n, $p_{n,i}$, $n \in \mathbb{N}$ be three sequences such that $0 \leq b_n \to 0$ with n, $0 \leq p_{n,i} \leq 1$,*

$$0 \leq a_{n+1} \leq 2 \sum_{i=1}^{n} p_{n,i} \left(\frac{i}{n}\right)^R \left(\sup_{i \in \{1,\ldots,n\}} a_i\right) + b_n. \tag{13}$$

and

$$0 < 2 \sum_{i=1}^{n} p_{n,i} \left(\frac{i}{n}\right)^R = C < 1.$$

Then $\lim_{n\to\infty} a_n = 0$.

Proof The proof is exactly the same as Rösler (1991)'s proof of his Proposition 3.3. In the last step we will have with $a := \limsup a_n < \infty$ the sandwiching for all $\epsilon > 0$

$$0 \leq a \leq C(a + \epsilon).$$

□

Proof[of Thm. 3.1] Denote the law of Y_n as $\mathcal{L}(Y_n) = G_n$. We take Y_∞ and Y'_∞ independent and distributed as G_∞, the fixed point of S. Then, for $i = 1, \ldots, n-1$ we choose independent versions of Y_i and Y'_i. We need to show $d^2(G_n, G_\infty) \to 0$. As the metric is the infimum over all pairs of random variables that have marginal distributions G_n and G_∞ the obvious choice is to take Y_n, Y_∞ such that L_n/n will be close to τ for large n. Rösler (1991) was considering the Yule model ($\beta = 0$) and there $\tau \sim \text{Unif}[0,1]$ and L_n is uniform on $\{1, \ldots, n-1\}$. Hence, $\lfloor (n-1)\tau \rfloor + 1$ will be uniform on $\{1, \ldots, n-1\}$, remember $P(\tau = 1) = 0$, and $L_n/n \stackrel{D}{=} (\lfloor (n-1)\tau \rfloor + 1)/n$. However, when $\beta > 0$ the situation complicates. For a given n, $(L_n - 1)$ is BetaBinomial$(n-2, \beta+1, \beta+1)$ distributed (cf. Eq. 1 and Aldous (1996) Eqs. 1 and 3). Hence, if $\tau \sim \text{Beta}(\beta+1, \beta+1)$ and $(L_n - 1) \sim$ BetaBinomial$(n-2, \beta+1, \beta+1)$ we do not have $L_n/n \stackrel{D}{=} (\lfloor (n-1)\tau \rfloor + 1)/n$ exactly. We may bound the Wasserstein metric by any coupling that retains the marginal distributions of the two random variables. Therefore, from now on we will be considering a version, where conditional on τ, the random variable $(L_n - 1)$ is Binomial$(n-2, \tau)$ distributed. Let r_n be any sequence

such that $r_n/n \to 0$ and $n/r_n^2 \to 0$, e.g. $r_n = n \ln^{-1} n$. Then, by Chebyshev's inequality

$$P\left(|L_n - \mathrm{E}\left[L_n|\tau\right]| \geq r_n \middle| \tau\right) \leq \frac{n\tau(1-\tau)}{r_n^2} \leq \frac{n}{4r_n^2} \to 0.$$

Let $\tilde{r}_n$ be a (appropriately) random integer in $\{\pm 1, \ldots, \pm r_n\}$ and we may write (with the chosen coupling of L_n and τ),

$$\begin{aligned}
&\mathrm{E}\left[\left(\left(\tfrac{L_n}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2\right] = \mathrm{E}\left[\mathrm{E}\left[\left(\left(\tfrac{L_n}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2 \middle| \tau\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\left(\left(\tfrac{\lfloor (n-1)\tau\rfloor + 1 + \tilde{r}_n}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2 \middle| |L_n - \mathrm{E}\left[L_n\right]| \leq r_n, \tau\right] P(|L_n - \mathrm{E}\left[L_n\right]| \leq r_n \middle| \tau)\right] \\
&+ \mathrm{E}\left[\mathrm{E}\left[\left(\left(\tfrac{L_n}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2 \middle| |L_n - \mathrm{E}\left[L_n\right]| \geq r_n, \tau\right] P(|L_n - \mathrm{E}\left[L_n\right]| \geq r_n \middle| \tau)\right] \\
&\leq \mathrm{E}\left[\left(\left(\tfrac{\lfloor (n-1)\tau\rfloor + 1 + \tilde{r}_n}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2\right] + \tfrac{n}{4r_n^2}\,\mathrm{E}\left[\left(\left(\tfrac{L_n}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2\right] \\
&= \mathrm{E}\left[\left(\left(\left(\tfrac{\lfloor (n-1)\tau\rfloor + 1}{n}\right)^r + r\tfrac{\tilde{r}_n}{n}\left(\tfrac{\lfloor (n-1)\tau\rfloor + 1 + \xi_{\tilde{r}_n}}{n}\right)^{r-1}\right) Y_{L_n} - \tau^r Y_\infty\right)^2\right] \\
&+ \tfrac{n}{4r_n^2}\,\mathrm{E}\left[\left(\left(\tfrac{L_n}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2\right] \\
&= \mathrm{E}\left[\left(\left(\tfrac{\lfloor (n-1)\tau\rfloor + 1}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2\right] + r^2 n^{-2}\,\mathrm{E}\left[\tilde{r}_n^2\left(\tfrac{\lfloor (n-1)\tau\rfloor + 1 + \xi_{\tilde{r}_n}}{n}\right)^{2(r-1)}\right] \\
&+ 2rn^{-1}\,\mathrm{E}\left[\tilde{r}_n\left(\tfrac{\lfloor (n-1)\tau\rfloor + 1 + \xi_{\tilde{r}_n}}{n}\right)^{r-1}\left(\left(\tfrac{\lfloor (n-1)\tau\rfloor + 1 + \xi_{\tilde{r}_n}}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)\right] \\
&+ \tfrac{n}{4r_n^2}\,\mathrm{E}\left[\left(\left(\tfrac{L_n}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2\right],
\end{aligned}$$

where $\xi_{\tilde{r}_n} \in (0, \tilde{r}_n)$ is (a random variable) such that the mean value theorem holds. As Y_n, Y_∞ have uniformly bounded second moments, $0 \leq \xi_{\tilde{r}_n} \leq \tilde{r}_n \leq r_n \leq n$, by assumption $r_n/n \to 0$ and $n/r_n^2 \to 0$, we have that

$$\begin{aligned}
&\left(r\tfrac{r_n}{n}\right)^2 \mathrm{E}\left[\left(\tfrac{\lfloor (n-1)\tau\rfloor + 1 + \xi_{\tilde{r}_n}}{n}\right)^{2(r-1)}\right] + \tfrac{n}{4r_n^2}\,\mathrm{E}\left[\left(\left(\tfrac{L_n}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2\right] \\
&+ 2r\tfrac{r_n}{n}\,\mathrm{E}\left[\left(\tfrac{\lfloor (n-1)\tau\rfloor + 1 + \xi_{\tilde{r}_n}}{n}\right)^{r-1}\left(\left(\tfrac{\lfloor (n-1)\tau\rfloor + 1 + \xi_{\tilde{r}_n}}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)\right] \to 0
\end{aligned}$$

and hence, $\mathrm{E}\left[\left(\left(\frac{L_n}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2\right] \to 0$.

Remember that $\mathrm{E}\left[Y_i\right] = \mathrm{E}\left[Y_\infty\right] = 0$ so that the expectation of the cross products disappears. Remembering the assumption $\sup_i n^{-r} h_n(i) \to 0$, all components can be treated in the same way as $\mathrm{E}\left[\left(\left(\frac{L_n}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2\right]$ with conditioning on τ and then controlling by r_n and Chebyshev's inequality L_n's deviation from its expected value. We therefore have for some sequence $v_n \to 0$

$$d^2(G_n, G_\infty) \leq \mathrm{E}\left[\left(\left(\frac{\lfloor (n-1)\tau \rfloor + 1}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2\right] + \mathrm{E}\left[(C_n(\lfloor (n-1)x \rfloor + 1) - C(\tau))^2\right]$$
$$+ \mathrm{E}\left[\left(\left(\frac{n - \lfloor (n-1)\tau \rfloor - 1}{n}\right)^r Y_{n-L_n} - (1-\tau)^r Y'_\infty\right)^2\right] + v_n.$$

Consider the first term of the right–hand side of the inequality and denote $d^2_{n-1} := \sup_{i \in \{1,\ldots,n-1\}} d^2(G_i, G_\infty)$

$$\mathrm{E}\left[\left(\left(\frac{\lfloor (n-1)\tau \rfloor + 1}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2\right] = \mathrm{E}\left[\sum_{i=1}^{n-1} 1_{(i-1)/(n-1) < \tau \leq i/(n-1)} \left(\left(\frac{i}{n}\right)^r Y_{L_n} - \tau^r Y_\infty\right)^2\right]$$
$$\leq \sum_{i=1}^{n-1} p_{n-1,i} \left(\frac{i}{n}\right)^{2r} \mathrm{E}\left[(Y_{L_n} - Y_\infty)^2\right] = \sum_{i=1}^{n-1} p_{n-1,i} \left(\frac{i}{n}\right)^{2r} d^2_{n-1},$$

where $p_{n,i} = P((i-1)/(n-1) < \tau \leq i/(n-1))$. Invoking Lemmata 3.1, 3.2 and using the assumption of Eq. (9) with $R = 2r$ we have

$$d^2(G_n, G_\infty) \leq 2 \sum_{i=1}^{n-1} p_{n-1,i} \left(\frac{i}{n}\right)^{2r} d^2_{n-1} + \left(n^{-r} \sup_i h_n(i)\right)^2 + v_n + O(n^{-2}) \to 0.$$

□

4 Limit distribution of the quartet index for Aldous's $\beta \geq 0$–model trees

4.1 Contraction–distribution type limit

We show here that the QIB of Aldous's $\beta \geq 0$–model trees satisfies the conditions of Thm. 3.1 with $r = 4$ and hence the QIB has a well characterized

limit distribution. We define a centred and scaled version of the QIB for Aldous's $\beta \geq 0$–model tree on n tips

$$Y_n^Q = n^{-4}\left(QIB(T_n) - \frac{3\beta+6}{7\beta+18}\binom{n}{4}\right) \tag{14}$$

Theorem 4.1 *The sequence of random variables Y_n^Q random variable for trees generated by Aldous's β–model with $\beta \geq 0$ converges with $n \to \infty$ in the Wasserstein d–metric (and hence in distribution) to a random variable $Y_Q \sim \mathcal{Q} \equiv G_\infty$ satisfying the following equality in distribution*

$$Y_Q \stackrel{D}{=} \tau^4 Y_Q' + (1-\tau)^4 Y_Q'' + \frac{3\beta+6}{24(7\beta+18)}\left(\tau^4 + (1-\tau)^4\right) - \frac{3\beta+6}{24(7\beta+18)} + \frac{1}{4}\tau^2(1-\tau)^2, \tag{15}$$

where $\tau \sim F_\tau$ is distributed as the Beta distribution of Eq. (3), $Y_Q, Y_Q', Y_Q'' \sim \mathcal{Q}$ and Y_Q', Y_Q'', τ are all independent.

Proof Denote by $P_3(x,y)$ a polynomial of degree at most three in terms of the variables x, y. From the recursive representation of Eq. (6)

$$\begin{aligned}
Y_n^Q &= n^{-4}\Big(QIB(T_{L_n}) - \tfrac{3\beta+6}{(7\beta+18)}\tbinom{L_n}{4} + QIB(T_{n-L_n}) - \tfrac{3\beta+6}{(7\beta+18)}\tbinom{n-L_n}{4} + \tbinom{L_n}{2}\tbinom{n-L_n}{2} \\
&+ \tfrac{3\beta+6}{(7\beta+18)}\tbinom{L_n}{4} + \tfrac{3\beta+6}{(7\beta+18)}\tbinom{n-L_n}{4} - \tfrac{3\beta+6}{(7\beta+18)}\tbinom{n}{4}\Big) \\
&= \left(\tfrac{L_n}{n}\right)^4 Y_{L_n}^Q + \left(1-\tfrac{L_n}{n}\right)^4 Y_{n-L_n}^Q + \tfrac{1}{4}\left(\tfrac{L_n}{n}\right)^2\left(1-\tfrac{L_n}{n}\right)^2 + \tfrac{3\beta+6}{24(7\beta+18)}\left(\tfrac{L_n}{n}\right)^4 \\
&+ \tfrac{3\beta+6}{24(7\beta+18)}\left(1-\tfrac{L_n}{n}\right)^4 - \tfrac{3\beta+6}{24(7\beta+18)} + n^{-4}P_3(n, L_n).
\end{aligned}$$

We therefore have $r = 4$ and

$$\begin{aligned}
C_n(i) &= \tfrac{1}{4}\left(\tfrac{i}{n}\right)^2\left(1-\tfrac{i}{n}\right)^2 + \tfrac{3\beta+6}{24(7\beta+18)}\left(\left(\tfrac{i}{n}\right)^4 + \left(1-\tfrac{i}{n}\right)^4\right) - \tfrac{3\beta+6}{24(7\beta+18)} \\
&+ n^{-4}P_3(n,i).
\end{aligned}$$

By the scaling and centring we know that $EY_n^Q = 0$ and $E\left(Y_n^Q\right)^2$ is uniformly bounded by Eq. (7). For the Beta law of τ we need to examine for all i

$$p_{n,i} := P\left(\frac{i-1}{n} \leq \tau < \frac{i}{n}\right) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} \int_{(i-1)/n}^{i/n} x^\beta (1-x)^\beta \mathrm{d}x.$$

We consider two cases

1. $\beta > 0$, we have to check if the condition of Eq. (9) is satisfied. Let

$$B_x(\beta+1,\beta+1) = \int_0^x u^\beta (1-u)^\beta \mathrm{d}u$$

be the incomplete Beta function. Then,

$$\begin{aligned} p_{n,i} &= \tfrac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} \left(B_{i/n}(\beta+1,\beta+1) - B_{(i-1)/n}(\beta+1,\beta+1) \right) \\ &= n^{-1} \tfrac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} B'_\xi(\beta+1,\beta+1) \end{aligned}$$

by the mean value theorem for some $\xi \in ((i-1)/n, i/n)$. Obviously

$$B'_\xi(\beta+1,\beta+1) = \xi^\beta (1-\xi)^\beta \le \left(\frac{i}{n}\right)^\beta \left(1 - \frac{i-1}{n}\right)^\beta$$

and now

$$\sum_{i=1}^n p_{n,i} \left(\tfrac{i}{n}\right)^R \le \tfrac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} n^{-1} \sum_{i=1}^n \left(\tfrac{i}{n}\right)^R \left(\tfrac{i}{n}\right)^\beta \left(1 - \tfrac{i-1}{n}\right)^\beta \to \tfrac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} \int_0^1 u^{\beta+R}(1-u)^\beta \mathrm{d}u$$

$$= \tfrac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} \tfrac{\Gamma(\beta+R+1)\Gamma(\beta+1)}{\Gamma(2\beta+R+2)} = \tfrac{\Gamma(2\beta+2)}{\Gamma(\beta+1)} \tfrac{\Gamma(\beta+R+1)}{\Gamma(2\beta+R+2)}. \tag{16}$$

Take $1 < R_1 < R_2$ and consider the ratio

$$A = \tfrac{\Gamma(2\beta+2)}{\Gamma(\beta+1)} \tfrac{\Gamma(\beta+R_2+1)}{\Gamma(2\beta+R_2+2)} \left(\tfrac{\Gamma(2\beta+2)}{\Gamma(\beta+1)} \tfrac{\Gamma(\beta+R_1+1)}{\Gamma(2\beta+R_1+2)} \right)^{-1} = \tfrac{\Gamma(\beta+R_2+1)}{\Gamma(2\beta+R_2+2)} \tfrac{\Gamma(\beta)}{\Gamma(\beta)} \tfrac{\Gamma(2\beta+R_1+1)}{\Gamma(\beta+R_1+1)}$$

$$= \tfrac{B(\beta+1+R_2,\beta)}{B(\beta+1+R_1,\beta)}.$$

The ratio $A < 1$ as the Beta function is decreasing in its arguments—hence the derived upper bound in Eq. (16) is decreasing in R. For $R = 1$ the bound equals

$$\frac{\Gamma(2\beta+2)}{\Gamma(\beta+1)}\frac{\Gamma(\beta+2)}{\Gamma(2\beta+3)} = \frac{\Gamma(2\beta+2)}{\Gamma(\beta+1)}\frac{(\beta+1)\Gamma(\beta+1)}{(2\beta+2)\Gamma(2\beta+2)} = \frac{(\beta+1)}{2(\beta+1)} = \frac{1}{2}$$

and hence for all $R > 1$ and all $\beta > 0$

$$\sum_{i=1}^{n} p_{n,i}\left(\frac{i}{n}\right)^R < \frac{1}{2}.$$

As in our case we have $r \geq 1$, then for $R = 2r \geq 2$ the assumptions of Lemma 3.2 are satisfied and the statement of the theorem follows through.

2. $\beta = 0$, then directly $p_{n,i} = n^{-1}$, Eq. (9) and assumptions of Lemma 3.2 are immediately satisfied and the statement of the theorem follows through.

□

Remark 4.1 *When $\beta < 0$ the process L_n/n seems to have a more involved asymptotic behaviour (cf. Lemma 3 Aldous, 1996). Furthermore, the bounds applied do not hold for $\beta < 0$. Therefore, this family of tree models (including the important uniform model) deserves a separate study with respect to its quartet index.*

4.2 Comparing with simulations

To verify the results we compared the simulated values from the limiting theoretical distribution of Y_Q with scaled and centred values of Yule tree QI values. The 500–tip Yule trees were simulated using the `rtreeshape()` of the apTreeshape R package and Tomás Martínez–Coronado's in–house Python code. Then, for each tree the QI value was calculated by Gabriel Valiente's and Tomás Martínez–Coronado's in–house programs. The raw values $QIB(\text{Yule}_{500})$ were scaled and centred as

$$Y_n^Q = 500^{-4}\left(QIB(\text{Yule}_{500}) - \frac{1}{3}\binom{500}{4}\right).$$

The Y_Q values were simulated using the proposed by Bartoszek (2017) heuristic Algorithm 2 (R code in Appendix). The results of the simulation are presented in Fig. 2.

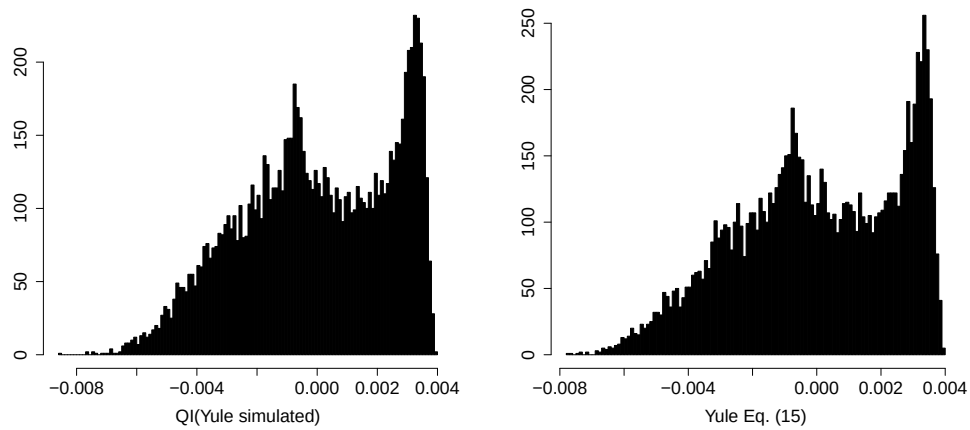

Figure 2: Left: histogram of scaled and centred simulated values of the QIB for the Yule tree, Y_n^Q, right: histogram of Y_Q for the Yule model, $\beta = 0$. The mean, variance, skewness and excess kurtosis of the simulated values are $-3.177 \cdot 10^{-6}$, $6.321 \cdot 10^{-6}$, -0.308, -0.852 (left, simulated values) and $1.682 \cdot 10^{-5}$, $6.38 \cdot 10^{-6}$, -0.317, -0.834 (right, theoretical values by Bartoszek (2017)'s heuristic Algorithm 2 with recursion depth 15). For $\beta = 0$ the leading constant of the variance in Eq. (7) is $5/(24 * 33075) \approx 6.299 \cdot 10^{-6}$.

Acknowledgments

I would like to thank the whole Computational Biology and Bioinformatics Research Group of the Balearic Islands University for hosting me on multiple occasions, introducing me to problems related with the quartet index and for valuable comments on this manuscript. The simulated values of the quartet index for the Yule tree were provided by Gabriel Valiente and Tomás Martínez–Coronado. I was supported by the Knut and Alice Wallenberg Foundation and am by the Swedish Research Council (Vetenskapsrådet) grant no. 2017–04951. My visits to the Balearic Islands University were partially supported by the the G S Magnuson Foundation of the Royal Swedish Academy of Sciences (grants no. MG2015–0055, MG2017–0066) and The Foundation for Scientific Research and Education in Mathematics (SVeFUM).

References

D. Aldous. Probability distributions on cladograms. In *Random Discrete Structures*, pages 1–18. Springer–Verlag, 1996.

D. Aldous. Stochastic models and descriptive statistics for phylogenetic trees. *Statist. Sci.*, 16:23–34, 2001.

K. Bartoszek. Exact and approximate limit behaviour of the Yule tree's cophenetic index. *ArXiv e-prints*, 2017.

B. Chen, D. Ford, and M. Winkel. A new family of Markov branching trees: the alpha–gamma model. *Electron. J. Probab.*, 14:400–430, 2009.

D. H. Colless. Review of "Phylogenetics: the theory and practise of phylogenetic systematics". *Syst. Zool.*, 31:100–104, 1982.

M. Drmota. *Random Trees: an Interplay between Combinatorics and Probability*. Springer–Verlag, 2009.

D. J. Ford. Probabilities on cladograms: introduction to the alpha model. *ArXiv e-prints*, 2005.

T. Martínez-Coronado, A. Mir, and F. Rosselló. The probabilities of trees and cladograms under Ford's α–model. *ArXiv e-prints*, 2018a.

T. Martínez-Coronado, A. Mir, F. Rosselló, and G. Valiente. A balance index for phylogenetic trees based on quartets. *ArXiv e-prints*, 2018b.

A. Mir, F. Rosselló, and L. Rotger. A new balance index for phylogenetic trees. *Math. Biosci.*, 241(1):125–136, 2013.

S. T. Rachev and L. Rüschendorf. Probability metrics and recursive algorithms. *Adv. Appl. Probab.*, 27:770–779, 1995.

U. Rösler. A limit theorem for "Quicksort". *Theor. Inf. Applic.*, 25(1): 85–100, 1991.

U. Rösler. A fixed point theorem for distributions. *Stoch. Proc. Appl.*, 42: 195–214, 1992.

M. J. Sackin. "Good" and "bad" phenograms. *Syst. Zool.*, 21:225–226, 1972.

Appendix: R code for simulating from the limit distribution of the quartet index

```
fCtau_QIB<-function(x,beta=0){
    PB4beta<-(3*beta+6)/(7*beta+18);
    PB4beta*(x^4)/24+PB4beta*((1-x)^4)/24-PB4beta/24+0.25*x*x*(1-x)^2
}

fdistribution_limitQIB<-function(num.iter=10,popsize=10000,Y0=0){
    replicate(popsize,fdraw_limitQIB(num.iter,Y0))
}

fdraw_limitQIB<-function(num.iter=15,Y0=0){
    res<-0
    if (num.iter==0){
        Y1<-Y0
        Y2<-Y0
    }
    else{
        Y1<-fdraw_limitQIB(num.iter-1,Y0)
        Y2<-fdraw_limitQIB(num.iter-1,Y0)
    }
    tau<-runif(1)
    res<-((tau)^4)*Y1+((1-tau)^4)*Y2+fCtau_QIB(tau)
    res
}

popsize<-10000 ## size of sample for histogram
num.iter<-15 ## depth of the recursion
Y0 <- 0 ## initial value

vlimitQIB<-fdistribution_limitQIB(num.iter=num.iter,popsize=popsize)
```