Theoretical Population Biology 143 (2022) 1–13

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

A compendium of covariances and correlation coefficients of coalescent tree properties

Egor Alimpiev *, Noah A. Rosenberg

Department of Biology, Stanford University, Stanford, CA 94305, United States of America

ARTICLE INFO

Article history:
Received 2 March 2021
Available online 30 October 2021

Keywords:
Coalescent
External branches
Genealogies
Internal branches
Time to the most recent common ancestor

ABSTRACT

Gene genealogies are frequently studied by measuring properties such as their height (H), length (L), sum of external branches (E), sum of internal branches (I), and mean of their two basal branches (B), and the coalescence times that contribute to the other genealogical features (T). These tree properties and their relationships can provide insight into the effects of population-genetic processes on genealogies and genetic sequences. Here, under the coalescent model, we study the 15 correlations among pairs of features of genealogical trees: H_n, L_n, E_n, I_n, B_n, and T_k for a sample of size n, with $2 \leq k \leq n$. We report high correlations among H_n, L_n, I_n, and B_n, with all pairwise correlations of these quantities having values greater than or equal to $\sqrt{6[6\zeta(3)+6-\pi^2]/(\pi\sqrt{18+9\pi^2-\pi^4})} \approx 0.84930$ in the limit as $n \to \infty$, where ζ is the Riemann zeta function. Although E_n has expectation 2 for all n and H_n has expectation 2 in the $n \to \infty$ limit, their limiting correlation is 0. The results contribute toward understanding features of the shapes of coalescent trees.

1. Introduction

In coalescent theory, features of gene genealogies are investigated in relation to the evolutionary processes that are included in population-genetic models (Hein et al., 2005; Wakeley, 2009). For example, comparing a constant-sized and an exponentially growing population, exponential growth increases the total length of the branches of a gene genealogy in relation to its height (Slatkin and Hudson, 1991; Slatkin, 1996; Sano and Tachida, 2005). Coalescences are rare in recent generations, when the population is large, and they occur primarily in the period deep in the past when the population was small.

Several tree features have been used for measuring effects of population-genetic processes on gene genealogies (Slatkin, 1996; Uyenoyama, 1997; Schierup and Hein, 2000; Rosenberg, 2006). For a binary ultrametric tree of n lineages, these features (Fig. 1) include the tree height from the tips to the root (H_n), the total length of all the branches (L_n), the total length of external branches connecting tips to the nearest internal node (E_n), the total length of internal branches connecting internal nodes to other internal nodes (I_n), and the mean length of the two basal branches incident to the root node (B_n).

These tree features can all be expressed as linear combinations, random linear combinations in some cases, of the same underlying random variables — the coalescence times T_k for coalescence of k to $k-1$ lineages, with $2 \leq k \leq n$. Hence, the quantities are correlated. For example, the tree height H_n includes the mean basal branch length B_n, and the total length L_n is the sum of the length E_n of the external branches and the length I_n of the internal branches; an increase in L_n necessarily increases E_n, I_n, or both.

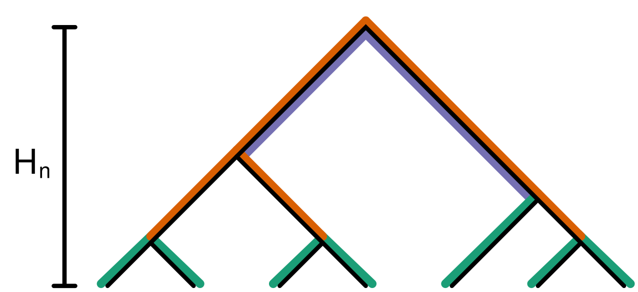

Fig. 1. Tree features. The tree height is H_n and the sum of the lengths of all tree branches is L_n. External branches, with total length E_n, appear in green; internal branches, with total length I_n, appear in red; basal branches, with mean length B_n, appear in purple.

Analyses of coalescent models have examined some of the correlations between tree features, notably the relationship between H_n and L_n (Fu, 1996; Griffiths and Tavaré, 1996; Rosenberg and Hirsh, 2003; Arbisser et al., 2018). Here, we perform a detailed investigation of correlations among H_n, L_n, E_n, I_n, and B_n. For each pair, under the coalescent, assuming a constant-sized population,

* Corresponding author.
E-mail address: alimpiev@stanford.edu (E. Alimpiev).

https://doi.org/10.1016/j.tpb.2021.09.008
0040-5809/

we evaluate their covariance and correlation. We explore limiting values as $n \to \infty$. The approach follows Arbisser et al. (2018), who obtained the covariance and correlation of H_n and L_n; we perform analogous calculations for all 10 pairs among $\{H_n, L_n, E_n, I_n, B_n\}$, as well as for the five pairs involving one of $\{H_n, L_n, E_n, I_n, B_n\}$ and T_k.

2. Tree properties

We consider the standard coalescent model of a constant-sized population of size N haploids. Time is measured in units of the population size, with one time unit representing N generations. For sample size $n \geq 2$, we examine tree properties H_n, L_n, E_n, I_n, and B_n, as well as the coalescence time T_k, $2 \leq k \leq n$. In this section, we recall basic features of the various quantities.

For convenience, for a mathematical expression we will use frequently, we write

$$S_{p,n} = \sum_{k=1}^{n} \frac{1}{k^p}. \tag{1}$$

The limit $S_{p,\infty} = \lim_{n\to\infty} S_{p,n}$ is the Riemann zeta function $\zeta(p)$. The harmonic sum $S_{1,\infty}$ diverges, and the sum of the reciprocals of squares is $S_{2,\infty} = \pi^2/6 \approx 1.64493$. The sum of the reciprocals of cubes is Apéry's constant, $S_{3,\infty} = \zeta(3) \approx 1.20206$.

2.1. T_k

T_k is a random variable representing the time during which k lineages coalesce to $k-1$ lineages. The T_k, $2 \leq k \leq n$, are independent and exponentially distributed with probability density function $f_{T_k}(t_k) = \binom{k}{2} e^{-\binom{k}{2} t_k}$ (Wakeley, 2009, p. 60). The expectation and variance of T_k are then

$$\mathbb{E}[T_k] = \frac{2}{k(k-1)}, \tag{2}$$

$$\mathrm{Var}[T_k] = \frac{4}{k^2(k-1)^2}. \tag{3}$$

As $n, k \to \infty$ with $k \leq n$, both $\mathbb{E}[T_k]$ and $\mathrm{Var}[T_k]$ have limit 0.

2.2. H_n

For $n \geq 2$, the height H_n of a tree from root to leaves can be written

$$H_n = \sum_{k=2}^{n} T_k. \tag{4}$$

The expectation and variance of H_n are then found using the expectation and variance of T_k (Eqs. (2) and (3)), noting that the T_k are independent:

$$\mathbb{E}[H_n] = \sum_{k=2}^{n} \mathbb{E}[T_k] = \frac{2(n-1)}{n}, \tag{5}$$

$$\mathrm{Var}[H_n] = 8\left(\sum_{k=2}^{n} \frac{1}{k^2}\right) - 4\left(\frac{n-1}{n}\right)^2. \tag{6}$$

The variance can be written $\mathrm{Var}[H_n] = 4(2S_{2,n}n^2 - 3n^2 + 2n - 1)/n^2$. The limits are $\lim_{n\to\infty} \mathbb{E}[H_n] = 2$ and $\lim_{n\to\infty} \mathrm{Var}[H_n] = 4\pi^2/3 - 12 \approx 1.15947$ (Wakeley, 2009, p. 76).

2.3. L_n

For $n \geq 2$, the total tree length, summing the lengths of all branches of a tree, is

$$L_n = \sum_{k=2}^{n} kT_k. \tag{7}$$

By Eqs. (2) and (3) and the independence of the T_k, we have

$$\mathbb{E}[L_n] = \sum_{k=2}^{n} k\mathbb{E}[T_k] = 2\sum_{k=1}^{n-1} \frac{1}{k}, \tag{8}$$

$$\mathrm{Var}[L_n] = 4\sum_{k=1}^{n-1} \frac{1}{k^2}. \tag{9}$$

In terms of $S_{p,n}$ (Eq. (1)), these expressions are $\mathbb{E}[L_n] = 2S_{1,n-1}$ and $\mathrm{Var}[L_n] = 4S_{2,n-1}$. The limits are $\lim_{n\to\infty} \mathbb{E}[L_n] = \infty$ and $\lim_{n\to\infty} \mathrm{Var}[L_n] = 2\pi^2/3 \approx 6.57974$ (Wakeley, 2009, p. 76).

2.4. E_n

The external branches of a tree are the branches that connect leaves to their nearest internal nodes. Denoting the individual external branch lengths $e_1^{(n)}, e_2^{(n)}, \ldots, e_n^{(n)}$, the sum of external branch lengths is

$$E_n = e_1^{(n)} + e_2^{(n)} + \cdots + e_n^{(n)}.$$

The $e_k^{(n)}$ are identically distributed, and we write e_n for the length of a randomly chosen external branch of a tree of n lineages. The sum of the external branches has expectation

$$\mathbb{E}[E_n] = n\mathbb{E}[e_n]. \tag{10}$$

The random variable e_n can be written recursively as (Fu and Li, 1993, eq. 7)

$$e_n = \begin{cases} e_{n-1} + T_n, & \text{with probability } \frac{n-2}{n}, \\ T_n, & \text{with probability } \frac{2}{n}. \end{cases} \tag{11}$$

Expressions for $\mathbb{E}[e_n]$, $\mathbb{E}[E_n]$, and $\mathrm{Var}[E_n]$ can then be obtained by solving recurrence equations (Fu and Li, 1993). We have

$$\mathbb{E}[e_n] = \frac{2}{n}. \tag{12}$$

For the mean and variance of E_n, we obtain Fu and Li (1993, Eqs. 10 and 14)

$$\mathbb{E}[E_n] = 2, \tag{13}$$

$$\mathrm{Var}[E_n] = \begin{cases} 4, & n = 2, \\ \frac{8}{(n-1)(n-2)}\left[n\left(\sum_{k=1}^{n-1} \frac{1}{k}\right) - 2(n-1)\right], & n > 2. \end{cases} \tag{14}$$

$\mathbb{E}[E_n]$ is equal to 2 irrespective of the choice of n, so that $\lim_{n\to\infty} \mathbb{E}[E_n] = 2$. The limit of the variance is $\lim_{n\to\infty} \mathrm{Var}[E_n] = 0$ (Fu and Li, 1993).

2.5. I_n

The internal branches connect internal nodes to other internal nodes. Their total length is I_n, with

$$I_n = L_n - E_n. \tag{15}$$

The mean and variance of the sum of internal branches are (Fu and Li, 1993, Eqs. 12 and 17)

$$\mathbb{E}[I_n] = \mathbb{E}[L_n] - \mathbb{E}[E_n] = 2\left(\sum_{k=1}^{n-1} \frac{1}{k}\right) - 2, \tag{16}$$

$$\mathrm{Var}[I_n] = 4\left[\frac{2[S_{1,n-1}n - 2(n-1)]}{(n-1)(n-2)} - \frac{2S_{1,n-1}}{n-1} + S_{2,n-1}\right]. \tag{17}$$

The limits are $\lim_{n\to\infty} \mathbb{E}[I_n] = \infty$ and $\lim_{n\to\infty} \mathrm{Var}[I_n] = 2\pi^2/3 \approx 6.57974$, the same as for L_n (Section 2.3).

2.6. B_n

Finally, we consider the basal branches, the two branches that extend from the root. We define B_n as the mean of the two branch lengths. One of the branches has length T_2, and we denote the other length b_n. We assume here that $n \geq 4$ for calculations involving B_n. The appendix of Uyenoyama (1997) gives

$$B_n = \frac{T_2 + b_n}{2}, \tag{18}$$

with

$$b_n = \left[\sum_{j=3}^{n-1} \sum_{k=2}^{j} \frac{T_k}{j} \prod_{i=3}^{j-1} \left(1 - \frac{1}{i}\right)\right] + \left[\sum_{k=2}^{n} T_k \prod_{i=3}^{n-1} \left(1 - \frac{1}{i}\right)\right]$$

for $n \geq 4$. A convenient form for b_n encodes the fact that with probability $2/[j(j-1)]$, $b_n = H_j$ for $j = 3, 4, \ldots, n-1$, and with probability $2/(n-1)$, $b_n = H_n$:

$$b_n = \left[\sum_{j=3}^{n-1} \sum_{k=2}^{j} \frac{2}{j(j-1)} T_k\right] + \left(\sum_{k=2}^{n} \frac{2}{n-1} T_k\right). \tag{19}$$

Assuming $n \geq 4$, the branch length b_n has expectation (Uyenoyama, 1997):

$$\mathbb{E}[b_n] = \frac{4}{n} + 4 \sum_{k=3}^{n-1} \frac{1}{k^2}. \tag{20}$$

The expectation and variance of B_n then equal

$$\mathbb{E}[B_n] = \frac{2}{n} + 2 \sum_{k=2}^{n-1} \frac{1}{k^2}, \tag{21}$$

$$\mathrm{Var}[B_n] = \frac{2(3S_{2,n-1}n^2 - 2S_{2,n-1}^2 n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}{n^2}. \tag{22}$$

The expectation appears in the appendix of Uyenoyama (1997). We calculate the expression for the variance in Section 3.11. Taking limits of these equations, we obtain $\lim_{n\to\infty} \mathbb{E}[B_n] = \pi^2/3 - 2 \approx 1.28987$ and $\lim_{n\to\infty} \mathrm{Var}[B_n] = 2 + \pi^2 - \pi^4/9 \approx 1.04637$.

3. Theoretical results

For pairs of variables among $\{H_n, L_n, E_n, I_n, B_n, T_k\}$, we apply results from Section 2 to compute covariances and correlations. First, for each pair, we compute their covariance. The covariance together with the variances of the two quantities from Section 2 provides their correlation. We obtain the limiting correlation for large trees by taking $n \to \infty$. Among the 15 pairs, our analyses for 13 are exact; for (E_n, B_n) and (I_n, B_n), we offer approximate covariances and correlations. We also provide the derivation of Eq. (22) for $\mathrm{Var}[B_n]$.

Note that correlations in pairs involving E_n have distinct forms for $n = 2$ and $n \geq 3$, owing to the piecewise definition of $\mathrm{Var}[E_n]$ in Eq. (14). We exclude the case of $n = 2$ for pairs involving I_n, as $I_2 = 0$ with $\mathrm{Var}[I_2] = 0$. We also assume that B_n is defined only for $n \geq 4$.

We present a summary of our mathematical results in Tables 1 and 2. Table 1 shows covariances of pairs of variables and their limits as $n \to \infty$. Table 2 shows correlations and their $n \to \infty$ limits.

3.1. H_n and T_k

We calculate the covariance of H_n and T_k using $\mathrm{Cov}[H_n, T_k] = \mathbb{E}[H_n T_k] - \mathbb{E}[H_n]\mathbb{E}[T_k]$. Recalling that T_i and T_j are independent for $i \neq j$ (Section 2.1), we have $\mathbb{E}[T_i T_j] = \mathbb{E}[T_i]\mathbb{E}[T_j]$ for $i \neq j$. Hence, inserting Eq. (4) for H_n and Eq. (2) for $\mathbb{E}[T_i]$, we have

$$\begin{aligned}\mathrm{Cov}[H_n, T_k] &= \mathbb{E}\left[T_k \sum_{i=2}^{n} T_i\right] - \mathbb{E}\left[\sum_{i=2}^{n} T_i\right]\mathbb{E}[T_k] \\ &= \sum_{i=2}^{n} \mathbb{E}[T_i T_k] - \sum_{i=2}^{n} \mathbb{E}[T_i]\mathbb{E}[T_k] \\ &= \mathrm{Var}[T_k] + \sum_{i=2,\, i\neq k}^{n} \mathbb{E}[T_i]\mathbb{E}[T_k] \\ &\quad - \sum_{i=2,\, i\neq k}^{n} \mathbb{E}[T_i]\mathbb{E}[T_k] = \mathrm{Var}[T_k] = \frac{4}{k^2(k-1)^2},\end{aligned} \tag{23}$$

where the last step uses $\mathrm{Var}[T_k]$ from Eq. (3). We observe that the covariance is independent of n.

For the correlation coefficient $\mathrm{Corr}[H_n, T_k] = \mathrm{Cov}[H_n, T_k]/\sqrt{\mathrm{Var}[H_n]\mathrm{Var}[T_k]}$, applying Eq. (6) for $\mathrm{Var}[H_n]$, Eq. (3) for $\mathrm{Var}[T_k]$, and Eq. (23) for $\mathrm{Cov}[H_n, T_k]$, we have

$$\mathrm{Corr}[H_n, T_k] = \frac{n}{\sqrt{2S_{2,n}n^2 - 3n^2 + 2n - 1}} \frac{1}{k(k-1)}. \tag{24}$$

Taking a limit as $n \to \infty$, we obtain

$$\lim_{n\to\infty} \mathrm{Corr}[H_n, T_k] = \frac{\sqrt{3}}{\sqrt{\pi^2 - 9}} \frac{1}{k(k-1)}. \tag{25}$$

The limiting correlation decreases to 0 with k from an initial value of $\frac{1}{2}[\sqrt{3/(\pi^2 - 9)}] \approx 0.92869$ at $k = 2$.

3.2. L_n and T_k

For L_n and T_k, applying Eqs. (7), (8) and (2), we have for the covariance

$$\begin{aligned}\mathrm{Cov}[L_n, T_k] &= \mathbb{E}[L_n T_k] - \mathbb{E}[L_n]\mathbb{E}[T_k] \\ &= \mathbb{E}\left[T_k \sum_{i=2}^{n} i T_i\right] - \mathbb{E}\left[\sum_{i=2}^{n} i T_i\right]\mathbb{E}[T_k] \\ &= \mathbb{E}\left[k T_k^2\right] - k\mathbb{E}[T_k]^2 = k\,\mathrm{Var}[T_k] = \frac{4}{k(k-1)^2},\end{aligned} \tag{26}$$

where the last step uses Eq. (3). The covariance of L_n and T_k, like $\mathrm{Cov}[H_n, T_k]$ (Eq. (23)), is independent of n.

Now we calculate the correlation coefficient from Eqs. (26), (9) and (3):

$$\mathrm{Corr}[L_n, T_k] = \frac{1}{\sqrt{S_{2,n-1}}} \frac{1}{k-1}. \tag{27}$$

If we let $n \to \infty$, then this quantity becomes

$$\lim_{n\to\infty} \mathrm{Corr}[L_n, T_k] = \frac{\sqrt{6}}{\pi} \frac{1}{k-1}. \tag{28}$$

The limiting correlation decreases to 0 with k, starting for $k = 2$ at $\sqrt{6}/\pi \approx 0.77970$.

3.3. H_n and L_n

Arbisser et al. (2018) reported the covariance and correlation of H_n and L_n. By Eq. (4) and the linearity of the covariance,

$$\mathrm{Cov}[H_n, L_n] = \sum_{k=2}^{n} \mathrm{Cov}[L_n, T_k].$$

Table 1
Covariances of pairs of variables. Expressions involving E_n or I_n apply for $n \geq 3$ and expressions involving B_n apply for $n \geq 4$.

(X_n, Y_n)	$\mathrm{Cov}[X_n, Y_n]$	$\lim_{n\to\infty} \mathrm{Cov}[X_n, Y_n]$	Reference
H_n, T_k	$\frac{4}{k^2(k-1)^2}$	$\frac{4}{k^2(k-1)^2}$	3.1, Eq. (23)
H_n, L_n	$4S_{2,n-1} - 4 + \frac{4}{n}$	$\frac{2\pi^2}{3} - 4 \approx 2.57974$	3.3, Eqs. (29), (30)
H_n, E_n	$\frac{4}{n}$	0	3.4, Eqs. (37), (38)
H_n, I_n	$4S_{2,n-1} - 4$	$\frac{2\pi^2}{3} - 4 \approx 2.57974$	3.7, Eqs. (49), (50)
H_n, B_n	$\frac{4[S_{3,n-1}n^2 - 3S_{2,n-1}n^2 + (n-1)(4n+1)]}{n^2}$	$4\zeta(3) + 16 - 2\pi^2 \approx 1.06902$	3.13, Eqs. (71), (72)
L_n, T_k	$\frac{4}{k(k-1)^2}$	$\frac{4}{k(k-1)^2}$	3.2, Eq. (26)
L_n, E_n	$\frac{4S_{1,n-1}}{n-1}$	0	3.6, Eqs. (45), (46)
L_n, I_n	$4S_{2,n-1} - \frac{4S_{1,n-1}}{n-1}$	$\frac{2\pi^2}{3} \approx 6.57974$	3.9, Eqs. (57), (58)
L_n, B_n	$\frac{4[S_{3,n-1}n - S_{2,n-1}n + n - 1]}{n}$	$4\zeta(3) + 4 - \frac{2\pi^2}{3} \approx 2.22849$	3.14, Eqs. (75), (76)
E_n, T_k	$\frac{4}{k(k-1)(n-1)}$	0	3.5, Eqs. (41), (42)
E_n, I_n	$\frac{4S_{1,n-1}}{n-1} - \frac{8S_{1,n-1}n}{(n-1)(n-2)} + \frac{16}{n-2}$	0	3.10, Eqs. (61), (62)
E_n, B_n	$\frac{4\left(S_{2,n-1}n - n + 1\right)}{n(n-1)}$	0	3.15, Eqs. (84), (85)
I_n, T_k	$\frac{4(n-k)}{k(k-1)^2(n-1)}$	$\frac{4}{k(k-1)^2}$	3.8, Eqs. (53), (54)
I_n, B_n	$\frac{4(S_{3,n-1}n - S_{2,n-1}n + n - S_{3,n-1} - 1)}{n-1}$	$4\zeta(3) + 4 - \frac{2\pi^2}{3} \approx 2.22849$	3.16, Eqs. (88), (89)
B_n, T_k	$\frac{4}{k^2(k-1)^3}$	$\frac{4}{k^2(k-1)^3}$	3.12, Eq. (68)

Applying Eq. (26), we obtain

$$\mathrm{Cov}[H_n, L_n] = 4S_{2,n-1} - 4 + \frac{4}{n}. \tag{29}$$

The limit of the covariance is

$$\lim_{n\to\infty} \mathrm{Cov}[H_n, L_n] = \frac{2\pi^2}{3} - 4 \approx 2.57974. \tag{30}$$

Dividing the covariance in Eq. (29) by the square root of the product of Eqs. (6) and (9), we obtain

$$\mathrm{Corr}[H_n, L_n] = \frac{S_{2,n-1}n - n + 1}{\sqrt{S_{2,n-1}(2S_{2,n}n^2 - 3n^2 + 2n - 1)}}. \tag{31}$$

The limit is

$$\lim_{n\to\infty} \mathrm{Corr}[H_n, L_n] = \frac{\pi^2 - 6}{\pi\sqrt{2(\pi^2 - 9)}} \approx 0.93399. \tag{32}$$

3.4. H_n and E_n

For the covariance $\mathrm{Cov}[H_n, E_n] = \mathbb{E}[H_nE_n] - \mathbb{E}[H_n]\mathbb{E}[E_n]$, we first note that by Eqs. (5) and (13), the second term is simply $4\left(1 - \frac{1}{n}\right)$. Expanding $\mathbb{E}[H_nE_n]$ by using the definition of H_n (Eq. (4)) gives us

$$\mathbb{E}[H_nE_n] = \sum_{i=2}^{n} \mathbb{E}[E_nT_i] = n\sum_{i=2}^{n} \mathbb{E}[e_nT_i],$$

as all external branch lengths are identically distributed (Eq. (10)).

For integers k, i with $2 \leq k, i, \leq n$, the external branch length e_k, representing the length of a randomly chosen external branch for a tree with k leaves, and the coalescence time T_i, satisfy (Eq. (11))

$$e_kT_i = \begin{cases} e_{k-1}T_i + T_kT_i, & \text{with probability } \frac{k-2}{k}, \\ T_kT_i, & \text{with probability } \frac{2}{k}, \end{cases} \tag{33}$$

where for convenience, we write $e_1 = 0$.

Note that e_k and T_i are independent for $i > k$, as the recurrence for e_k is constructed only using coalescence times $T_2, T_3, \ldots, T_k$ (Eq. (11)); each of these times is independent of T_i for $i > k$ (Section 2.1). We solve to find $\mathbb{E}[e_nT_i]$ by computing $\mathbb{E}[e_kT_i]$, incrementing k from 2 to n. The calculations are similar to those of the Appendix of Fu and Li (1993).

$\mathbb{E}[e_2T_2]$ is trivial, with $e_2 = T_2$, and $\mathbb{E}[e_2T_2] = \mathbb{E}\left[T_2^2\right] = 2$ by Eqs. (2) and (3). By Eqs. (12) and (2) and the independence of e_k and T_i for $i > k$, for $i \geq 3$,

$$\mathbb{E}[e_{i-1}T_i] = \mathbb{E}[e_{i-1}]\mathbb{E}[T_i] = \frac{4}{i(i-1)^2}.$$

Noting $\mathbb{E}\left[T_i^2\right] = \mathrm{Var}[T_i] + \mathbb{E}[T_i]^2 = 2\mathbb{E}[T_i]^2$ by Eqs. (2) and (3), we use Eq. (33) to write an expression for $\mathbb{E}[e_iT_i]$:

$$\mathbb{E}[e_iT_i] = \frac{i-2}{i}\mathbb{E}[e_{i-1}T_i] + \mathbb{E}\left[T_i^2\right] = \frac{4}{i(i-1)^2}. \tag{34}$$

Next, incrementing Eq. (34), we have

$$\mathbb{E}[e_{i+1}T_i] = \frac{i-1}{i+1}\mathbb{E}[e_iT_i] + \mathbb{E}[T_{i+1}]\mathbb{E}[T_i] = \frac{4}{i^2(i-1)}, \tag{35}$$

by Eqs. (2) and (12).

The final step is to solve the recurrence equation

$$\mathbb{E}[e_nT_i] = \frac{n-2}{n}\mathbb{E}[e_{n-1}T_i] + \mathbb{E}[T_n]\mathbb{E}[T_i],$$

Table 2
Correlation coefficients of pairs of variables. Expressions involving E_n or I_n apply for $n \geq 3$ and expressions involving B_n apply for $n \geq 4$.

(X_n, Y_n)	$\mathrm{Corr}[X_n, Y_n]$	$\lim_{n\to\infty} \mathrm{Corr}[X_n, Y_n]$	Reference
H_n, T_k	$\frac{n}{\sqrt{2n^2 S_{2,n} - 3n^2 + 2n - 1}} \frac{1}{k(k-1)}$	$\frac{\sqrt{3}}{\sqrt{\pi^2 - 9}} \frac{1}{k(k-1)} \approx \frac{1.85738}{k(k-1)}$	3.1, Eqs. (24), (25)
H_n, L_n	$\frac{S_{2,n-1}n - n + 1}{\sqrt{S_{2,n-1}\left(2S_{2,n}n^2 - 3n^2 + 2n - 1\right)}}$	$\frac{\pi^2 - 6}{\pi\sqrt{2(\pi^2 - 9)}} \approx 0.93399$	3.3, Eqs. (31), (32)
H_n, E_n	$\frac{\sqrt{(n-1)(n-2)}}{\sqrt{2(2S_{2,n}n^2 - 3n^2 + 2n - 1)\left(S_{1,n-1}n - 2n + 2\right)}}$	0	3.4, Eqs. (39), (40)
H_n, I_n	$\frac{\left(S_{2,n-1} - 1\right) n\sqrt{(n-1)(n-2)}}{\left(2S_{2,n}n^2 - 3n^2 + 2n - 1\right)\left[4S_{1,n-1} + S_{2,n-1}(n-1)(n-2) - 4(n-1)\right]}$	$\frac{\pi^2 - 6}{\pi\sqrt{2(\pi^2 - 9)}} \approx 0.93399$	3.7, Eqs. (51), (52)
H_n, B_n	$\frac{\sqrt{2}[S_{3,n-1}n^2 - 3S_{2,n-1}n^2 + (n-1)(4n+1)]}{\sqrt{(2S_{2,n}n^2 - 3n^2 + 2n - 1)(3S_{2,n-1}n^2 - 2S_{2,n-1}^2 n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}}$	$\frac{3\sqrt{3}[2\zeta(3) + 8 - \pi^2]}{\sqrt{(\pi^2 - 9)(18 + 9\pi^2 - \pi^4)}} \approx 0.97054$	3.13, Eqs. (73), (74)
L_n, T_k	$\frac{1}{\sqrt{S_{2,n-1}}} \frac{1}{k-1}$	$\frac{\sqrt{6}}{\pi} \frac{1}{k-1} \approx \frac{0.77970}{k-1}$	3.2, Eqs. (27), (28)
L_n, E_n	$\frac{S_{1,n-1}\sqrt{n-2}}{\sqrt{2S_{2,n-1}(n-1)\left(S_{1,n-1}n - 2n + 2\right)}}$	0	3.6, Eqs. (47), (48)
L_n, I_n	$\frac{[S_{2,n-1}(n-1) - S_{1,n-1}]\sqrt{n-2}}{(n-1)\sqrt{S_{2,n-1}\left[4S_{1,n-1} + S_{2,n-1}(n-1)(n-2) - 4(n-1)\right]}}$	1	3.9, Eqs. (59), (60)
L_n, B_n	$\frac{\sqrt{2}(S_{3,n-1}n - S_{2,n-1}n + n - 1)}{\sqrt{S_{2,n-1}(3S_{2,n-1}n^2 - 2S_{2,n-1}^2 n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}}$	$\frac{\sqrt{6}[6\zeta(3) + 6 - \pi^2]}{\pi\sqrt{18 + 9\pi^2 - \pi^4}} \approx 0.84930$	3.14, Eqs. (77), (78)
E_n, T_k	$\frac{\sqrt{n-2}}{\sqrt{2(n-1)(S_{1,n-1}n - 2n + 2)}}$	0	3.5, Eqs. (43), (44)
E_n, I_n	$\frac{4(n-1) - S_{1,n-1}(n+2)}{\sqrt{2\left(S_{1,n-1}n - 2n + 2\right)\left[4S_{1,n-1} + S_{2,n-1}(n-1)(n-2) - 4(n-1)\right]}}$	0	3.10, Eqs. (63), (64)
E_n, B_n	$\frac{(S_{2,n-1}n - n + 1)\sqrt{n-2}}{\sqrt{(n-1)(S_{1,n-1}n - 2n + 2)(3S_{2,n-1}n^2 - 2S_{2,n-1}^2 n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}}$	0	3.15, Eqs. (86), (87)
I_n, T_k	$\frac{(n-k)\sqrt{n-2}}{(k-1)\sqrt{(n-1)\left[4S_{1,n-1} + S_{2,n-1}(n-1)(n-2) - 4(n-1)\right]}}$	$\frac{\sqrt{6}}{\pi} \frac{1}{k-1} \approx \frac{0.77970}{k-1}$	3.8, Eqs. (55), (56)
I_n, B_n	$\frac{\sqrt{2}(S_{3,n-1}n - S_{2,n-1}n + n - S_{3,n-1} - 1)n\sqrt{n-2}}{\sqrt{(n-1)[4S_{1,n-1} + S_{2,n-1}(n-1)(n-2) - 4(n-1)](3S_{2,n-1}n^2 - 2S_{2,n-1}^2 n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}}$	$\frac{\sqrt{6}[6\zeta(3) + 6 - \pi^2]}{\pi\sqrt{18 + 9\pi^2 - \pi^4}} \approx 0.84930$	3.16, Eqs. (90), (91)
B_n, T_k	$\frac{\sqrt{2}n}{\sqrt{3S_{2,n-1}n^2 - 2S_{2,n-1}^2 n^2 + n^2 - 4S_{2,n-1}n + 3n - 4}} \frac{1}{k(k-1)^2}$	$\frac{6}{\sqrt{18 + 9\pi^2 - \pi^4}} \frac{1}{k(k-1)^2} \approx \frac{1.95518}{k(k-1)^2}$	3.12, Eqs. (69), (70)

with initial condition Eq. (35). Recalling the case of $i = n = 2$, with $2 \leq i \leq n$, we obtain solution

$$\mathbb{E}[e_n T_i] = \frac{4}{i(i-1)(n-1)}. \tag{36}$$

Applying Eq. (10), the expression for $\mathrm{Cov}[H_n, E_n]$ becomes

$$\mathrm{Cov}[H_n, E_n] = n\sum_{i=2}^{n} \frac{4}{i(i-1)(n-1)} - 4\left(1 - \frac{1}{n}\right) = \frac{4}{n}. \tag{37}$$

The limit of the covariance as $n \to \infty$ is

$$\lim_{n\to\infty} \mathrm{Cov}[H_n, E_n] = 0. \tag{38}$$

Dividing Eq. (37) by the square root of the product of variances from Eqs. (6) and (14), the correlation is

$$\mathrm{Corr}[H_n, E_n] = \begin{cases} 1, & n = 2, \\ \frac{\sqrt{(n-1)(n-2)}}{\sqrt{2\left(2S_{2,n}n^2 - 3n^2 + 2n - 1\right)\left(S_{1,n-1}n - 2n + 2\right)}}, & n > 2. \end{cases} \tag{39}$$

The limit of the correlation is

$$\lim_{n\to\infty} \mathrm{Corr}[H_n, E_n] = 0. \tag{40}$$

3.5. E_n and T_k

In the process of computing $\mathrm{Cov}[H_n, E_n]$, we have obtained an expression for $\mathbb{E}[e_n T_i]$ (Eq. (36)), from which we can obtain $\mathrm{Cov}[E_n, T_k] = n\mathbb{E}[e_n T_k] - \mathbb{E}[E_n]\,\mathbb{E}[T_k]$. Applying Eqs. (13) and (2), we have

$$\mathrm{Cov}[E_n, T_k] = \frac{4}{k(k-1)(n-1)}. \tag{41}$$

Irrespective of the value of k, we have

$$\lim_{n\to\infty} \mathrm{Cov}[E_n, T_k] = 0. \tag{42}$$

Applying Eqs. (3) and (14), the correlation coefficient is

$$\mathrm{Corr}[E_n, T_k] = \begin{cases} 1, & n = 2, \\ \frac{\sqrt{n-2}}{\sqrt{2(n-1)(S_{1,n-1}n - 2n + 2)}}, & n \geq 3. \end{cases} \tag{43}$$

The correlation coefficient is independent of k, and it has limit

$$\lim_{n\to\infty} \mathrm{Corr}[E_n, T_k] = 0. \tag{44}$$

3.6. L_n and E_n

Fu and Li (1993) provided the expression for $\mathbb{E}[L_n E_n]$ (see also p. 167 of Durrett (2008), with all values scaled by $2N_e$). The main result is the following expression, obtained by solving recurrence equations:

$$\mathbb{E}[L_n E_n] = \frac{4S_{1,n-1}n}{n-1}.$$

We can use this result to calculate the covariance of L_n and E_n by $\mathrm{Cov}[L_n, E_n] = \mathbb{E}[L_nE_n] - \mathbb{E}[L_n]\mathbb{E}[E_n]$ with Eqs. (8) and (13). The covariance can also be quickly obtained from Eqs. (7) and (41),

$$\mathrm{Cov}[L_n, E_n] = \sum_{k=2}^{n} k\,\mathrm{Cov}[E_n, T_k] = \frac{4S_{1,n-1}}{n-1}. \quad (45)$$

The limit is

$$\lim_{n\to\infty} \mathrm{Cov}[L_n, E_n] = 0. \quad (46)$$

Applying Eqs. (45), (9) and (14), the correlation coefficient of L_n and E_n is

$$\mathrm{Corr}[L_n, E_n] = \begin{cases} 1, & n = 2, \\ \frac{S_{1,n-1}\sqrt{n-2}}{\sqrt{2S_{2,n-1}(n-1)(S_{1,n-1}n-2n+2)}}, & n \geq 3, \end{cases} \quad (47)$$

with the limit

$$\lim_{n\to\infty} \mathrm{Corr}[L_n, E_n] = 0. \quad (48)$$

3.7. H_n and I_n

For the pair H_n and I_n, we exploit results obtained for other pairs to quickly obtain the covariance. Remembering that $I_n = L_n - E_n$ (Eq. (15)), we use Eqs. (29) and (37) to obtain for $n \geq 3$

$$\begin{aligned}\mathrm{Cov}[H_n, I_n] &= \mathrm{Cov}[H_n, L_n] - \mathrm{Cov}[H_n, E_n] \\ &= 4S_{2,n-1} - 4. \end{aligned} \quad (49)$$

For this covariance, we have

$$\lim_{n\to\infty} \mathrm{Cov}[H_n, I_n] = \frac{2\pi^2}{3} - 4 \approx 2.57974. \quad (50)$$

From the covariance in Eq. (49) and variances in Eqs. (6) and (17), we compute the correlation coefficient:

$$\mathrm{Corr}[H_n, I_n] = \frac{\left(S_{2,n-1} - 1\right) n\sqrt{(n-1)(n-2)}}{\sqrt{\left(2S_{2,n}n^2 - 3n^2 + 2n - 1\right)\left[4S_{1,n-1} + S_{2,n-1}(n-1)(n-2) - 4(n-1)\right]}}. \quad (51)$$

The limit is the same as that of $\mathrm{Corr}[H_n, L_n]$, or (Eq. (32))

$$\lim_{n\to\infty} \mathrm{Corr}[H_n, I_n] = \frac{\pi^2 - 6}{\pi\sqrt{2\left(\pi^2 - 9\right)}} \approx 0.93399. \quad (52)$$

3.8. I_n and T_k

By Eqs. (15), (26) and (41), assuming $n \geq 3$, we have

$$\begin{aligned}\mathrm{Cov}[I_n, T_k] &= \mathrm{Cov}[L_n, T_k] - \mathrm{Cov}[E_n, T_k] \\ &= \frac{4(n-k)}{k(k-1)^2(n-1)}. \end{aligned} \quad (53)$$

The limit of this expression is a rapidly decreasing function of k,

$$\lim_{n\to\infty} \mathrm{Cov}[I_n, T_k] = \frac{4}{k(k-1)^2}. \quad (54)$$

Using the variances in Eqs. (17) and (3), the correlation coefficient is

$$\mathrm{Corr}[I_n, T_k] = \frac{(n-k)\sqrt{n-2}}{(k-1)\sqrt{(n-1)\left[4S_{1,n-1} + S_{2,n-1}(n-1)(n-2) - 4(n-1)\right]}}, \quad (55)$$

with limit

$$\lim_{n\to\infty} \mathrm{Corr}[I_n, T_k] = \frac{\sqrt{6}}{\pi}\frac{1}{k-1}. \quad (56)$$

The limit of $\mathrm{Corr}[I_n, T_k]$ as $n \to \infty$ is equal to that of $\mathrm{Corr}[L_n, T_k]$ (Eq. (28)).

3.9. L_n and I_n

By Eq. (15), we can apply Eqs. (9) and (45) to obtain for $n \geq 3$

$$\begin{aligned}\mathrm{Cov}[L_n, I_n] &= \mathrm{Var}[L_n] - \mathrm{Cov}[L_n, E_n] \\ &= 4S_{2,n-1} - \frac{4S_{1,n-1}}{n-1}. \end{aligned} \quad (57)$$

The limit as $n \to \infty$ is

$$\lim_{n\to\infty} \mathrm{Cov}[L_n, I_n] = \frac{2\pi^2}{3} \approx 6.57974. \quad (58)$$

For the correlation coefficient, applying Eqs. (57), (9) and (17), we get

$$\mathrm{Corr}[L_n, I_n] = \frac{\left[S_{2,n-1}(n-1) - S_{1,n-1}\right]\sqrt{n-2}}{(n-1)\sqrt{S_{2,n-1}\left[4S_{1,n-1} + S_{2,n-1}(n-1)(n-2) - 4(n-1)\right]}}, \quad (59)$$

with

$$\lim_{n\to\infty} \mathrm{Corr}[L_n, I_n] = 1. \quad (60)$$

3.10. E_n and I_n

For this pair, with $n \geq 3$, the covariance was reported by Fu and Li (1993):

$$\mathrm{Cov}[E_n, I_n] = \frac{4S_{1,n-1}}{n-1} - \frac{8S_{1,n-1}n}{(n-1)(n-2)} + \frac{16}{n-2}. \quad (61)$$

We can also obtain this result quickly from Eqs. (15), (57) and (17), as $\mathrm{Cov}[E_n, I_n] = \mathrm{Cov}[L_n - I_n, I_n] = \mathrm{Cov}[L_n, I_n] - \mathrm{Var}[I_n]$. In the limit, we have

$$\lim_{n\to\infty} \mathrm{Cov}[E_n, I_n] = 0. \quad (62)$$

For the correlation coefficient, we divide Eq. (61) by the product of the square roots of Eqs. (14) and (17):

$$\mathrm{Corr}[E_n, I_n] = \frac{4(n-1) - S_{1,n-1}(n+2)}{\sqrt{2\left(S_{1,n-1}n - 2n + 2\right)\left[4S_{1,n-1} + S_{2,n-1}(n-1)(n-2) - 4(n-1)\right]}}, \quad (63)$$

with the limit

$$\lim_{n\to\infty} \mathrm{Corr}[E_n, I_n] = 0. \quad (64)$$

This result is equal to the limit for $\mathrm{Corr}[L_n, E_n]$ (Eq. (48)).

3.11. $\mathrm{Var}[B_n]$

To obtain correlation coefficients involving B_n, assuming $n \geq 4$, we first verify the expression for $\mathrm{Var}[B_n]$ in Eq. (22). By definition of B_n in Eq. (18), we have

$$\begin{aligned}\mathrm{Var}[B_n] &= \mathbb{E}\left[\frac{1}{4}(T_2 + b_n)^2\right] - \mathbb{E}\left[\frac{1}{2}(T_2 + b_n)\right]^2 \\ &= \frac{1}{4}\mathrm{Var}[b_n] + \frac{1}{2}\mathrm{Cov}[b_n, T_2] + \frac{1}{4}, \end{aligned} \quad (65)$$

where we have used $\mathrm{Var}[T_2] = 1$ (Eq. (3)).

$$\mathrm{Corr}[H_n, B_n] = \frac{\sqrt{2}[S_{3,n-1}n^2 - 3S_{2,n-1}n^2 + (n-1)(4n+1)]}{\sqrt{(2S_{2,n}n^2 - 3n^2 + 2n - 1)(3S_{2,n-1}n^2 - 2S_{2,n-1}^2 n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}}. \quad (73)$$

Box I.

To calculate $\mathrm{Var}[b_n]$, we first recall that for $j = 3, 4, \ldots, n-1$, with probability $2/[j(j-1)]$, we have $b_n = H_j$; with probability $2/(n-1)$ we have $b_n = H_n$. Hence, applying Eq. (19) and $\mathbb{E}[H_k^2] = \mathrm{Var}[H_k] + \mathbb{E}[H_k]^2$ with Eqs. (5) and (6), we have

$$\mathbb{E}[b_n^2] = \left[\sum_{j=3}^{n-1} \frac{2}{j(j-1)} \mathbb{E}[H_j]^2\right] + \frac{2}{n-2}\mathbb{E}[H_n]^2 = \frac{-16S_{2,n-1}n^2 + 30n^2 - 16n - 16}{n^2}.$$

Using the expression for $\mathbb{E}[b_n]$ from Eq. (20), we obtain

$$\mathrm{Var}[b_n] = \frac{24S_{2,n-1}n^2 - 16S_{2,n-1}^2 n^2 + 5n^2 - 32S_{2,n-1}n + 24n - 32}{n^2}. \quad (66)$$

Next, we compute $\mathrm{Cov}[b_n, T_k]$ and insert $k = 2$. By Eq. (19), applying the independence of the T_i (Section 2.1) and inserting Eq. (3), we have

$$\begin{aligned}
&\mathrm{Cov}[b_n, T_k] \\
&= \left[\sum_{j=3}^{n-1}\sum_{i=2}^{j} \frac{2}{j(j-1)} \mathrm{Cov}[T_i, T_k]\right] + \left(\sum_{i=2}^{n} \frac{2}{n-1}\mathrm{Cov}[T_i, T_k]\right) \\
&= \left[\sum_{i=2}^{n-1} \sum_{\substack{j=i \text{ if } i\geq 3 \\ j=3 \text{ if } i=2}}^{n-1} \frac{2}{j(j-1)} \mathrm{Cov}[T_i, T_k]\right] + \left(\frac{2}{n-1}\mathrm{Var}[T_k]\right) \\
&= \begin{cases} \left[\sum_{j=3}^{n-1} \frac{2}{j(j-1)} \mathrm{Var}[T_k]\right] + \left(\frac{2}{n-1}\mathrm{Var}[T_k]\right), & k = 2, \\ \left[\sum_{j=k}^{n-1} \frac{2}{j(j-1)} \mathrm{Var}[T_k]\right] + \left(\frac{2}{n-1}\mathrm{Var}[T_k]\right), & k = 3, 4, \ldots, n-1, \\ \frac{2}{n-1}\mathrm{Var}[T_k], & k = n, \end{cases} \\
&= \begin{cases} 1, & k = 2, \\ \frac{8}{k^2(k-1)^3}, & k = 3, 4, \ldots, n. \end{cases} \quad (67)
\end{aligned}$$

Inserting $\mathrm{Var}[b_n]$ from Eq. (66) and $\mathrm{Cov}[b_n, T_2]$ from Eq. (67) into Eq. (65), we confirm Eq. (22).

3.12. B_n and T_k

We extract $\mathrm{Cov}[B_n, T_k]$ from Section 3.11, as $\mathrm{Cov}[B_n, T_k] = \mathrm{Cov}[b_n, T_k]/2 + \mathrm{Cov}[T_2, T_k]/2$ by the definition in Eq. (18), and $\mathrm{Cov}[T_2, T_k] = \delta_{k,2}$, where δ is the Kronecker delta (Section 2.1). By Eq. (67), recalling $n \geq 4$,

$$\mathrm{Cov}[B_n, T_k] = \frac{4}{k^2(k-1)^3}. \quad (68)$$

The covariance is independent of n.

For the correlation coefficient, using Eqs. (68), (22) and (3), we have

$$\mathrm{Corr}[B_n, T_k] = \frac{\sqrt{2}n}{\sqrt{3S_{2,n-1}n^2 - 2S_{2,n-1}^2 n^2 + n^2 - 4S_{2,n-1}n + 3n - 4}} \times \frac{1}{k(k-1)^2}, \quad (69)$$

with the limit

$$\lim_{n\to\infty} \mathrm{Corr}[B_n, T_k] = \frac{6}{\sqrt{18 + 9\pi^2 - \pi^4}} \frac{1}{k(k-1)^2}. \quad (70)$$

The limit begins at $3/\sqrt{18 + 9\pi^2 - \pi^4} \approx 0.97759$ for $k = 2$ and quickly decreases to 0 as k increases.

3.13. H_n and B_n

To obtain $\mathrm{Cov}[H_n, B_n]$ with $n \geq 4$, we begin from Eq. (18):

$$\mathrm{Cov}[H_n, B_n] = \frac{1}{2}\mathrm{Cov}[H_n, b_n] + \frac{1}{2}\mathrm{Cov}[H_n, T_2].$$

The second term was computed in Eq. (23), $\mathrm{Cov}[H_n, T_2] = 1$.

For the first term, $\mathrm{Cov}[H_n, b_n]$, we decompose H_n (Eq. (4)) and apply Eq. (67) to obtain

$$\mathrm{Cov}[H_n, b_n] = \sum_{k=2}^{n} \mathrm{Cov}[b_n, T_k] = 1 + \sum_{k=3}^{n} \frac{8}{k^2(k-1)^3}.$$

We use a partial fraction decomposition to sum the series, obtaining

$$\mathrm{Cov}[H_n, B_n] = 1 + \sum_{k=3}^{n} \frac{4}{k^2(k-1)^3} = \frac{4[S_{3,n-1}n^2 - 3S_{2,n-1}n^2 + (n-1)(4n+1)]}{n^2}. \quad (71)$$

The asymptotic limit of $\mathrm{Cov}[H_n, B_n]$ is

$$\lim_{n\to\infty} \mathrm{Cov}[H_n, B_n] = 4\zeta(3) + 16 - 2\pi^2 \approx 1.06902. \quad (72)$$

The correlation coefficient is then equal to Eq. (73) given in Box I. The limit of the correlation coefficient is:

$$\lim_{n\to\infty} \mathrm{Corr}[H_n, B_n] = \frac{3\sqrt{3}[2\zeta(3) + 8 - \pi^2]}{\sqrt{(\pi^2 - 9)(18 + 9\pi^2 - \pi^4)}} \approx 0.97054. \quad (74)$$

3.14. L_n and B_n

In a manner similar to that used in Section 3.13, with $n \geq 4$, we expand $\mathrm{Cov}[L_n, B_n]$ using Eq. (18):

$$\mathrm{Cov}[L_n, B_n] = \frac{1}{2}\mathrm{Cov}[L_n, b_n] + \frac{1}{2}\mathrm{Cov}[L_n, T_2].$$

The second term is $\mathrm{Cov}[L_n, T_2] = 2$ by Eq. (26). The first term is decomposable by Eq. (7); applying Eq. (67),

$$\mathrm{Cov}[L_n, b_n] = \sum_{k=2}^{n} k\,\mathrm{Cov}[b_n, T_k] = 2 + \sum_{k=3}^{n} \frac{8}{k(k-1)^3}.$$

$$\widetilde{\mathrm{Corr}}\,[E_n, B_n] = \frac{(S_{2,n-1}n - n + 1)\sqrt{n-2}}{\sqrt{(n-1)(S_{1,n-1}n - 2n + 2)(3S_{2,n-1}n^2 - 2S_{2,n-1}^2 n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}}. \quad (86)$$

Box II.

Summing the series, we have

$$\mathrm{Cov}[L_n, B_n] = 2 + \sum_{k=3}^{n} \frac{4}{k(k-1)^3} = \frac{4[S_{3,n-1}n - S_{2,n-1}n + n - 1]}{n}. \quad (75)$$

The limiting covariance is

$$\lim_{n\to\infty} \mathrm{Cov}[L_n, B_n] = 4\zeta(3) + 4 - \frac{2\pi^2}{3} \approx 2.22849. \quad (76)$$

Using Eqs. (75), (9) and (22), we now obtain an expression for the correlation coefficient:

$$\mathrm{Corr}[L_n, B_n] = \frac{\sqrt{2}(S_{3,n-1}n - S_{2,n-1}n + n - 1)}{\sqrt{S_{2,n-1}(3S_{2,n-1}n^2 - 2S_{2,n-1}^2 n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}}. \quad (77)$$

The limit is

$$\lim_{n\to\infty} \mathrm{Corr}[L_n, B_n] = \frac{\sqrt{6}[6\zeta(3) + 6 - \pi^2]}{\pi\sqrt{18 + 9\pi^2 - \pi^4}} \approx 0.84930. \quad (78)$$

3.15. E_n and B_n

For $\mathrm{Cov}[E_n, B_n]$, we obtain an approximate rather than exact answer. Decomposing B_n by Eq. (18), we have

$$\mathrm{Cov}[E_n, B_n] = \frac{1}{2}\mathrm{Cov}[E_n, T_2] + \frac{1}{2}\mathrm{Cov}[E_n, b_n]. \quad (79)$$

Recall that b_n can be defined conditionally, in terms of a random variable J that characterizes the coalescence times that it contains (Section 2.6). More precisely, we say that for a random variable J, $b_n = H_j$ with probability p_j, where $p_j = \mathbb{P}[J = j] = 2/[j(j-1)]$ for $J = 3, 4, \ldots, n-1$ and $p_j = 2/(j-1)$ for $J = n$. We can then decompose the covariance $\mathrm{Cov}[E_n, b_n]$ by the conditional covariance formula, conditioning on J:

$$\mathrm{Cov}[E_n, b_n] = \mathbb{E}\Big[\mathrm{Cov}[E_n, b_n|J]\Big] + \mathrm{Cov}\Big[\mathbb{E}[E_n|J], \mathbb{E}[B_n|J]\Big]. \quad (80)$$

We next perform an approximation by ignoring the second term in the covariance decomposition. Noting that $\mathrm{Cov}[E_n, T_2] = \frac{2}{n-1}$ by Eq. (41), we use Eq. (79) together with Eq. (80) to write approximations

$$\widetilde{\mathrm{Cov}}\,[E_n, b_n] = \mathbb{E}\Big[\mathrm{Cov}[E_n, b_n|J]\Big] \quad (81)$$

$$\widetilde{\mathrm{Cov}}\,[E_n, B_n] = \frac{1}{n-1} + \frac{1}{2}\widetilde{\mathrm{Cov}}\,[E_n, b_n]. \quad (82)$$

Weighting each $\mathrm{Cov}[H_j, E_n]$ by the associated probability p_j, and decomposing H_j by Eq. (4), Eq. (81) gives

$$\begin{aligned}\mathbb{E}\Big[\mathrm{Cov}[E_n, b_n|J]\Big] &= \sum_{j=3}^{n} p_j\,\mathrm{Cov}[E_n, b_n|J = j] \\ &= \left[\sum_{j=3}^{n-1} \frac{2}{j(j-1)}\mathrm{Cov}[H_j, E_n]\right] + \frac{2}{n-1}\mathrm{Cov}[H_n, E_n] \\ &= \left[\sum_{j=3}^{n-1} \frac{2}{j(j-1)}\sum_{k=2}^{j}\mathrm{Cov}[E_n, T_k]\right] + \frac{2}{n-1}\sum_{k=2}^{j}\mathrm{Cov}[E_n, T_k].\end{aligned}$$

We can then insert the result of Eq. (41) and simplify to obtain

$$\widetilde{\mathrm{Cov}}\,[E_n, b_n] = \frac{2(4S_{2,n-1}n - 5n + 4)}{n(n-1)}. \quad (83)$$

Finally, inserting Eq. (83) into Eq. (82), we have

$$\widetilde{\mathrm{Cov}}\,[E_n, B_n] = \frac{4(S_{2,n-1}n - n + 1)}{n(n-1)}. \quad (84)$$

The limit is

$$\lim_{n\to\infty} \widetilde{\mathrm{Cov}}\,[E_n, B_n] = 0. \quad (85)$$

For the approximate correlation coefficient $\widetilde{\mathrm{Corr}}\,[E_n, B_n] = \widetilde{\mathrm{Cov}}\,[E_n, B_n]\,/\sqrt{\mathrm{Var}\,[E_n]\,\mathrm{Var}\,[B_n]}$, we use Eqs. (84), (14) and (22) to obtain Eq. (86) given in Box II, with limit

$$\lim_{n\to\infty} \widetilde{\mathrm{Corr}}\,[E_n, B_n] = 0. \quad (87)$$

3.16. I_n and B_n

We use Eq. (15) and results involving L_n (Eq. (75)) and E_n (Eq. (84)) to obtain

$$\widetilde{\mathrm{Cov}}\,[I_n, B_n] = \mathrm{Cov}[L_n, B_n] - \widetilde{\mathrm{Cov}}\,[E_n, B_n] = \frac{4(S_{3,n-1}n - S_{2,n-1}n + n - S_{3,n-1} - 1)}{n-1}, \quad (88)$$

with limit

$$\lim_{n\to\infty} \widetilde{\mathrm{Cov}}\,[I_n, B_n] = 4\zeta(3) + 4 - \frac{2\pi^2}{3} \approx 2.22849. \quad (89)$$

Dividing Eq. (88) by the product of the square roots of Eqs. (17) and (22), the approximate correlation is Eq. (90) given in Box III, with limit

$$\lim_{n\to\infty} \widetilde{\mathrm{Corr}}\,[I_n, B_n] = \frac{\sqrt{6}[6\zeta(3) + 6 - \pi^2]}{\pi\sqrt{18 + 9\pi^2 - \pi^4}} \approx 0.84930. \quad (91)$$

4. Numerical and simulation-based analysis

4.1. Analysis methods

We examine the results of Section 3 summarized in Tables 1 and 2 numerically and by coalescent simulation. For 13 of 15 covariances and correlations, the theoretical results are exact, and simulations merely verify that the mathematics has proceeded without error. For the covariances and correlations involving (E_n, B_n) and (I_n, B_n), the theoretical results are approximate, and the simulations assess the accuracy of the approximations.

We simulated the coalescent process for a series of values of n beginning with $n = 2$, at each value of n performing 100,000

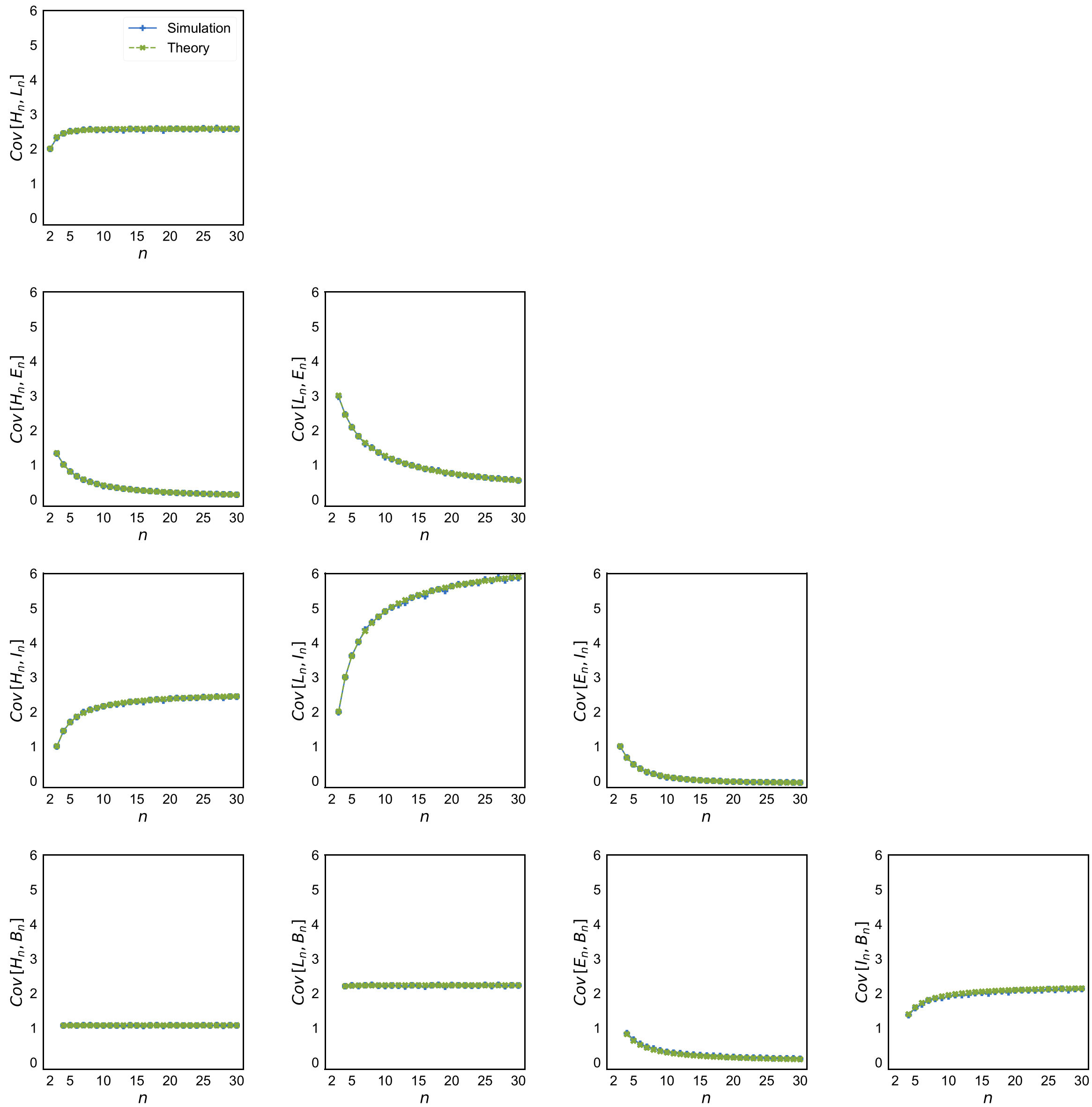

Fig. 2. Simulated and theoretical values of covariances of pairs of variables, plotted as functions of sample size n. Expressions for theoretical values are taken from Table 1.

$$\widetilde{\mathrm{Corr}}\,[I_n, B_n] = \frac{\sqrt{2}(S_{3,n-1}n - S_{2,n-1}n + n - S_{3,n-1} - 1)n\sqrt{n-2}}{\sqrt{(n-1)[4S_{1,n-1} + S_{2,n-1}(n-1)(n-2) - 4(n-1)](3S_{2,n-1}n^2 - 2S_{2,n-1}^2 n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}}. \tag{90}$$

Box III.

replicate simulations. To generate the simulated replicates, we employed `ms` (Hudson, 2002), using the command `ms n 100000 -T`, with `n` taken from $\{2, 3, \ldots, 50\}$. In the set of simulated replicates, we evaluated simulated covariances and correlation coefficients for pairs of quantities.

We plot the mathematical results of Tables 1 and 2 together with simulation values in Figs. 2–5. Figs. 2 and 3 show covariances of pairs of variables; Figs. 4 and 5 show correlations.

4.2. Accuracy of approximations

Fig. 2 shows the analytical and simulated covariances, and Fig. 4 shows the analytical and simulated correlations, for pairs of variables among $\{H_n, L_n, I_n, E_n, B_n\}$. For pairs of variables for which no approximations were needed in obtaining covariances—all except (E_n, B_n) and (I_n, B_n)—the simulated and analytical values produce plots that are nearly indistinguishable.

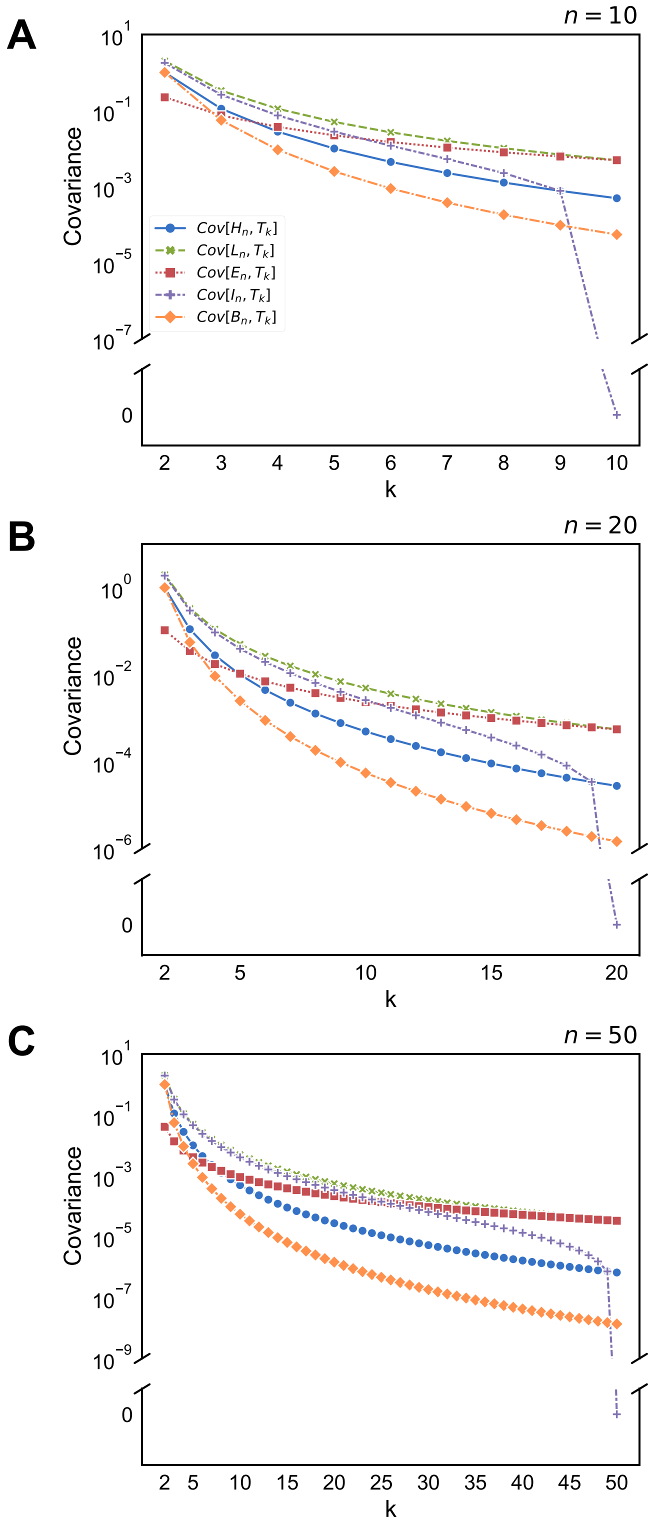

Fig. 3. Theoretical values of covariances $\mathrm{Cov}[X, T_k]$ for variables X in $\{H_n, L_n, E_n, I_n, B_n\}$, plotted as functions of k for $n = 10$, $n = 20$, and $n = 50$. The plots appear on a logarithmic scale.

For (E_n, B_n) and (I_n, B_n), the approximate and simulated correlations are close, but noticeably different (Fig. 4); the mean absolute difference between the analytical and simulated values across choices of n from 4 to 30 is 0.02458 for (E_n, B_n) and 0.01089 for (I_n, B_n). For covariance, which unlike the correlation coefficient is not standardized to lie in $[-1, 1]$, the approximate and simulated values are quite close, with corresponding mean absolute deviations of 0.02372 for (E_n, B_n) and 0.03101 for (I_n, B_n).

4.3. Properties of correlations

We observe that H_n, L_n, I_n, and B_n all remain strongly correlated as n increases, with the six limiting correlations among these four quantities lying between 0.84930 for $\mathrm{Corr}[L_n, B_n]$ and $\mathrm{Corr}[I_n, B_n]$ and 1 for $\mathrm{Corr}[L_n, I_n]$ (Table 2). The high limiting $\mathrm{Corr}[H_n, L_n]$ of approximately 0.93399 reflects the strong influence of times T_k with small k on both H_n and L_n (Figs. 3 and 5). As n increases, $\mathbb{E}[I_n]$ increases without bound (Eq. (16)), whereas $\mathbb{E}[E_n]$ remains constant (Eq. (13)); the contribution of E_n to the total tree length L_n becomes negligible, and $\mathrm{Corr}[L_n, I_n]$ approaches 1. $\mathrm{Corr}[H_n, I_n]$ has the same limiting value as $\mathrm{Corr}[H_n, L_n]$, and H_n, L_n, and I_n all have limiting correlation 0 with E_n. Interestingly, although H_n and E_n have the same limiting expectation of 2, the limit of their correlation $\mathrm{Corr}[H_n, E_n]$ is 0.

The correlations of H_n, L_n, and I_n with B_n, like their correlations with each other, are relatively high. $\mathrm{Corr}[H_n, B_n]$ is nearly constant in n, with limit approximately 0.97054; both H_n and B_n are determined in large part by the T_k with small k (Eqs. (5) and (21)), so that little change occurs in the correlation as n increases. Because $\mathrm{Corr}[H_n, B_n]$ is high and $\mathrm{Corr}[H_n, L_n]$ is also high, the constraint on a correlation $\mathrm{Corr}[Y, Z]$ given $\mathrm{Corr}[X, Y]$ and $\mathrm{Corr}[X, Z]$, or (Wickens, 2014, eq. 7.1)

$$\mathrm{Corr}[Y,Z] \geq \mathrm{Corr}[X,Y]\,\mathrm{Corr}[X,Z] - \sqrt{1-\mathrm{Corr}[X,Y]^2}\sqrt{1-\mathrm{Corr}[X,Z]^2} \tag{92}$$

$$\mathrm{Corr}[Y,Z] \leq \mathrm{Corr}[X,Y]\,\mathrm{Corr}[X,Z] + \sqrt{1-\mathrm{Corr}[X,Y]^2}\sqrt{1-\mathrm{Corr}[X,Z]^2}, \tag{93}$$

forces a high value for $\mathrm{Corr}[L_n, B_n]$ as well. In particular, placing H_n, L_n, B_n in the roles of X, Y, Z, with $\lim_{n\to\infty} \mathrm{Corr}[H_n, L_n] \approx 0.93399$ and $\lim_{n\to\infty} \mathrm{Corr}[H_n, B_n] \approx 0.97054$, we obtain an interval $0.82037 \leq \lim_{n\to\infty} \mathrm{Corr}[L_n, B_n] \leq 0.99256$ from Eqs. (92) and (93); $\lim_{n\to\infty} \mathrm{Corr}[L_n, B_n] \approx 0.84930$ lies near its lower end. Eqs. (92) and (93) similarly force a high value for $\lim_{n\to\infty} \mathrm{Corr}[I_n, B_n]$, using H_n, I_n, B_n as X, Y, Z.

Next, for correlations involving the T_k, we observe that for fixed n, as k increases from 2 to n, $\mathrm{Corr}[H_n, T_k]$ decreases (Fig. 5). At fixed n and k, $\mathrm{Corr}[L_n, T_k]$ generally exceeds $\mathrm{Corr}[H_n, T_k]$; k copies of the branch length T_k contribute to tree length L_n (Eq. (7)), whereas only one copy contributes to the tree height H_n (Eq. (4)), giving rise to a greater value for the correlation of T_k with L_n than with H_n. For $k > 2$, $\mathrm{Corr}[B_n, T_k]$ is generally smaller than $\mathrm{Corr}[H_n, T_k]$; because B_n is determined to a larger extent by T_2 than is H_n, the correlations of B_n with T_k for $k > 2$ are generally smaller. Finally, because tree length L_n consists primarily of internal branches for large n, the correlation $\mathrm{Corr}[I_n, T_k]$ is similar to $\mathrm{Corr}[L_n, T_k]$ (Fig. 5), approaching the same limit as $n \to \infty$ (Table 2); the correlation of E_n and T_k is a constant that does not depend on k.

5. Discussion

We have examined relationships between pairs of tree features under the coalescent model by deriving expressions for their covariances and correlation coefficients (Tables 1 and 2). For 13 of 15 pairs examined, we obtained exact expressions for the covariances and correlation coefficients, and for the remaining two pairs, we obtained quantities observed in simulations to closely approximate the desired quantities (Figs. 2 and 4). The results provide a compendium of basic relationships among

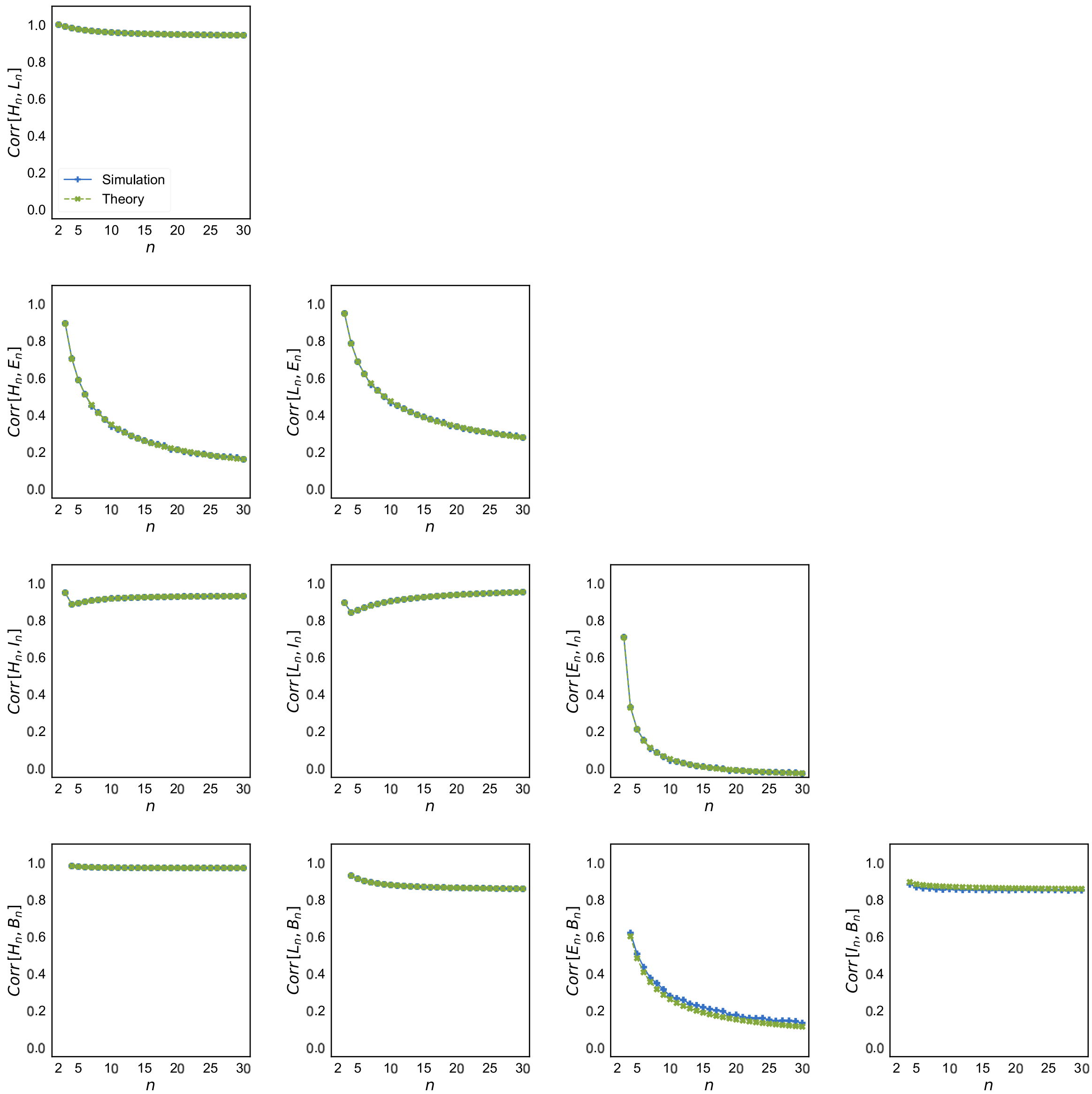

Fig. 4. Simulated and theoretical values of correlation coefficients of pairs of variables, plotted as functions of sample size n. Expressions for theoretical values are taken from Table 2.

coalescent tree features, contributing to a more precise understanding of the way in which the properties of coalescent trees relate to each other.

In most cases, the covariances have relatively simple expressions, comparable to the simplicity of most expressions for expectations and variances (Table 1). Expressions for the correlation coefficients are somewhat more complex, in many cases with $n \to \infty$ limits that contain terms resulting from the limit $\sum_{k=1}^{\infty} 1/k^2 = \pi^2/6$.

Numerically, we obtain tight correlations between H_n, L_n, I_n, and B_n as n grows large, with all of these quantities possessing limiting correlations of 0.84930 or greater (Table 2). In the limit, L_n and I_n are perfectly correlated, and all limiting correlations of other quantities with I_n are equal to their corresponding correlations with L_n. Decreasing correlations are observed for H_n, L_n, I_n, and B_n with E_n, with limits of 0 observed in all cases (Table 2). Although H_n and E_n both have limiting expectation 2 (Eqs. (5) and (13)), their limiting correlation coefficient is 0. The correlations among H_n, L_n, and B_n are all large; however, the limiting correlation for (L_n, B_n) is near the lower end of the interval suggested by the larger limiting correlations for (H_n, L_n) and (H_n, B_n) (Eqs. (92) and (93)). This result suggests that L_n and B_n capture relatively distinct features of coalescent trees in relation to the constraints placed on a pair of correlated variables that are each highly correlated with a third variable (H_n). A similar observation can be made concerning I_n and B_n, as L_n and I_n are asymptotically fully correlated.

Although tree properties such as H_n, L_n, E_n, I_n, and B_n are not themselves observable in genetic sequences, interest in these quantities arises in part from their relationship to statistical tests that assess the fit of the coalescent model to data on genetic variation. Features of tree shape underlie predictions of the coalescent regarding allele frequencies; in particular, tree properties contribute to predictions for the unfolded site-frequency spectrum (SFS) of a genomic region, the vector that for a sample of size n tabulates how many variable (biallelic) sites in the region possess allele frequencies $1/n, 2/n, \ldots, (n-1)/n$ for the derived allele (e.g. Fu, 1995; Ferretti et al., 2017). Test statistics then assess agreement of site-frequency spectra with the predictions (e.g. Zeng et al., 2006; Achaz, 2009; Ferretti et al.,

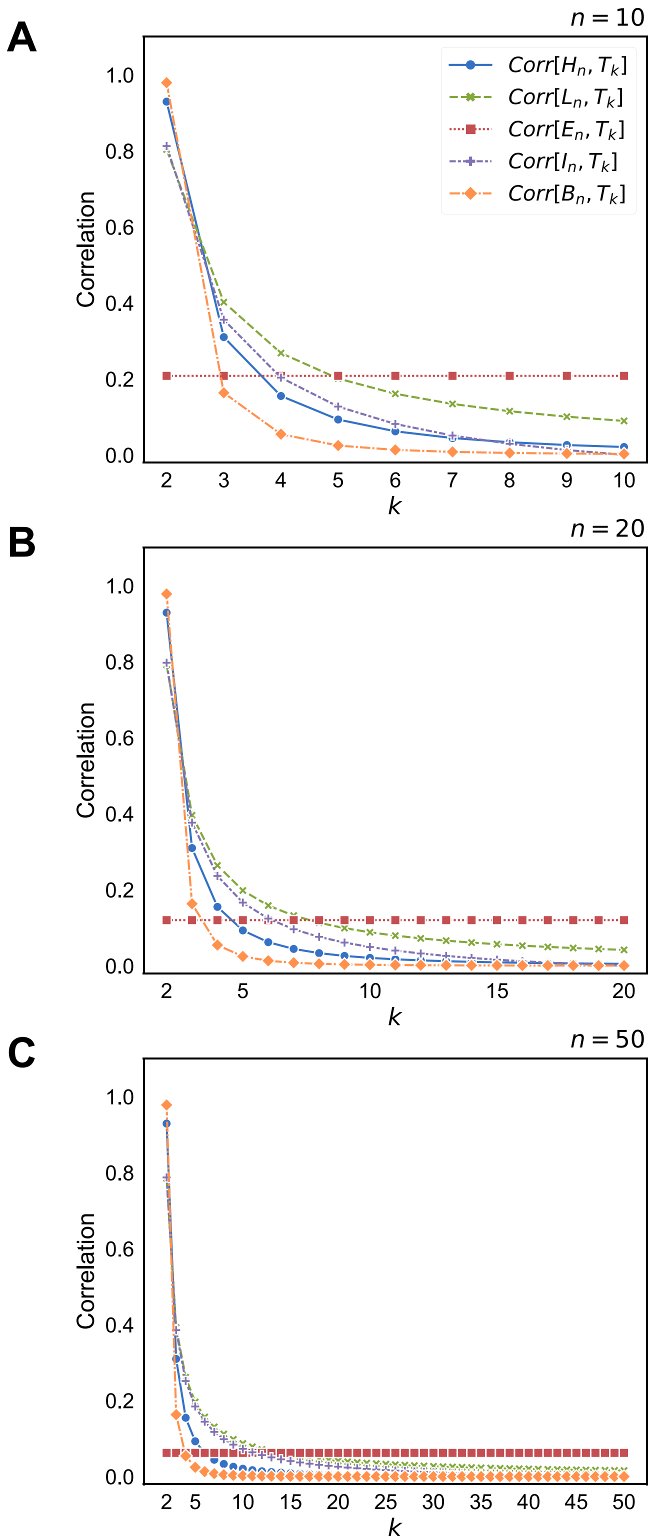

Fig. 5. Theoretical values of correlation coefficients $\mathrm{Corr}[X, T_k]$ for variables X in $\{H_n, L_n, E_n, I_n, B_n\}$, plotted as functions of k for $n = 10$, $n = 20$, and $n = 50$.

2010; Ronen et al., 2013), so that correlations among statistics emphasizing different aspects of site-frequency spectra emerge from dependence on correlated tree features. In this context, further understanding of correlations among tree properties can assist in understanding the joint behavior of SFS-based tests of the coalescent model.

Our computations augment earlier calculations concerning quantities associated with coalescent trees. The pairs (H_n, L_n) (Arbisser et al., 2018) and (L_n, E_n) and (E_n, I_n) (Fu and Li, 1993) have been studied in detail. Results for pairs (H_n, T_k), (L_n, T_k), and (L_n, I_n) follow trivially from the derivations and results of Arbisser et al. (2018) and Fu and Li (1993), but were not highlighted in those studies. Results for pairs (H_n, E_n), (H_n, I_n), (E_n, T_k), and (I_n, T_k) follow from derivations similar to those of Fu and Li (1993), but to our knowledge, they have not been previously reported.

The least-studied of the variables we consider, B_n, was introduced by Uyenoyama (1997) in the context of balancing selection and self-incompatibility alleles in plants. Under balancing selection, the mean B_n of the two basal branches is expected to be long in relation to the tree length L_n, so that $2B_n/L_n$ predicts the fraction of segregating sites that distinguish two long-separated sets of lineages. For B_n, which gives a portion of the height H_n—but which, unlike H_n, is obtained from a sum with a random length—we derived the variance (Eq. (22)), as well as exact covariances and correlations with H_n, L_n, and T_k and approximate covariances and correlations with E_n and I_n. Several studies have extended the work of Fu and Li (1993) on features of the external and internal branch lengths (Blum and François, 2005; Caliebe et al., 2007; Janson and Kersting, 2011; Dahmer and Kersting, 2015, 2017; Disanto and Wiehe, 2020); it may be possible to seek exact rather than approximate covariances and correlations for (E_n, B_n) and (I_n, B_n) by building on these studies.

When examining joint distributions of H_n and L_n, Arbisser et al. (2018) used computations of the expectations and variances of H_n and L_n and the covariance of H_n and L_n to obtain approximations for the expectation and variance of H_n/L_n. Following the approach of Arbisser et al. (2018), our results could be used to obtain similar approximate expressions for expectations and variances of ratios of additional pairs.

Acknowledgments

We thank Yun-Xin Fu and John Wakeley for comments on the manuscript. We acknowledge support from NIH, United States grants R01 GM131404 and R01 HG005855 and NSF, United States grant BCS-2116322.

References

Achaz, G., 2009. Frequency spectrum neutrality tests: one for all and all for one. Genetics 183 (1), 249–258.
Arbisser, I.M., Jewett, E.M., Rosenberg, N.A., 2018. On the joint distribution of tree height and tree length under the coalescent. Theor. Popul. Biol. 122, 46–56.
Blum, M.G.B., François, O., 2005. Minimal clade size and external branch length under the neutral coalescent. Adv. Appl. Probab. 37, 647–662.
Caliebe, A., Neininger, R., Krawczak, M., Rösler, U., 2007. On the length distribution of external branches in coalescence trees: genetic diversity within species. Theor. Popul. Biol. 72, 245–252.
Dahmer, I., Kersting, G., 2015. The internal branch lengths of the Kingman coalescent. Ann. Appl. Probab. 25 (3), 1325–1348.
Dahmer, I., Kersting, G., 2017. The total external length of the evolving Kingman coalescent. Probab. Theory Related Fields 167 (3–4), 1165–1214.
Disanto, F., Wiehe, T., 2020. Mapping the external branges of a Kingman tree: a discrete approach. Theor. Popul. Biol. 134, 92–105.
Durrett, R., 2008. Probability Models for DNA Sequence Evolution, second ed. Springer-Verlag, New York.
Ferretti, L., Ledda, A., Wiehe, T., Achaz, G., Ramos-Onsins, S., 2017. Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests. Genetics 207 (1), 229–240.
Ferretti, L., Perez-Enciso, M., Ramos-Onsins, S., 2010. Optimal neutrality tests based on the frequency spectrum. Genetics 186 (1), 353–365.
Fu, Y.-X., 1995. Statistical properties of segregating sites. Theor. Popul. Biol. 48 (2), 172–197.
Fu, Y.-X., 1996. Estimating the age of the common ancestor of a DNA sample using the number of segregating sites. Genetics 144 (2), 829–838.

Fu, Y.-X., Li, W.-H., 1993. Statistical tests of neutrality of mutations. Genetics 133 (3), 693–709.
Griffiths, R., Tavaré, S., 1996. Monte Carlo inference methods in population genetics. Math. Comput. Modelling 23 (8–9), 141–158.
Hein, J., Schierup, M., Wiuf, C., 2005. Gene Genealogies, Variation and Evolution. Oxford University Press, Oxford.
Hudson, R.R., 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18 (2), 337–338.
Janson, S., Kersting, G., 2011. On the total external length of the Kingman coalescent. Electron. J. Probab. 16, 2203–2218.
Ronen, R., Udpa, N., Halperin, E., Bafna, V., 2013. Learning natural selection from the site frequency spectrum. Genetics 195 (1), 181–193.
Rosenberg, N.A., 2006. Gene genealogies. In: Fox, C.W., Wolf, J.B. (Eds.), Evolutionary Genetics: Concepts and Case Studies. Oxford University Press, Oxford, pp. 173–189.
Rosenberg, N.A., Hirsh, A.E., 2003. On the use of star-shaped genealogies in inference of coalescence times. Genetics 164 (4), 1677–1682.
Sano, A., Tachida, H., 2005. Gene genealogy and properties of test statistics of neutrality under population growth. Genetics 169 (3), 1687–1697.
Schierup, M.H., Hein, J., 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics 156 (2), 879–891.
Slatkin, M., 1996. Gene genealogies within mutant allelic classes. Genetics 143 (1), 579–587.
Slatkin, M., Hudson, R.R., 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129 (2), 555–562.
Uyenoyama, M.K., 1997. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. Genetics 147 (3), 1389–1400.
Wakeley, J., 2009. Coalescent Theory. Roberts & Company, Greenwood Village, CO.
Wickens, T.D., 2014. The Geometry of Multivariate Statistics. Psychology Press, New York.
Zeng, K., Fu, Y.-X., Shi, S., Wu, C.-I., 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174 (3), 1431–1439.