Asymptotic joint normality of counts of uncorrelated motifs in recursive trees

Mohan Gopaldesikan[1], Hosam Mahmoud[2], and Mark Daniel Ward[1]

Department of Statistics, Purdue University, West Lafayette, IN[1]
Department of Statistics, The George Washington University, Washington, D.C.[2]

Center for Science of Information
NSF Science and Technology Center

Introduction

- A random recursive tree is a rooted nonplanar tree that grows by the successive insertion of nodes labelled **1**,**2**,**3**, ...
- A new node chooses any of the existing nodes at random as its parent.
- After n insertions there are $(n-1)!$ trees, which are equally likely.
- *Motif*: a specific nonplanar unlabelled rooted tree shape of finite size.
- A motif occurs on the *fringe* if the subtree rooted at the root of the motif is the motif itself.
- *Uncorrelated collection of motifs*: For any two motif in the collection, neither appears as a subtree on the fringe of the other.

Illustrations

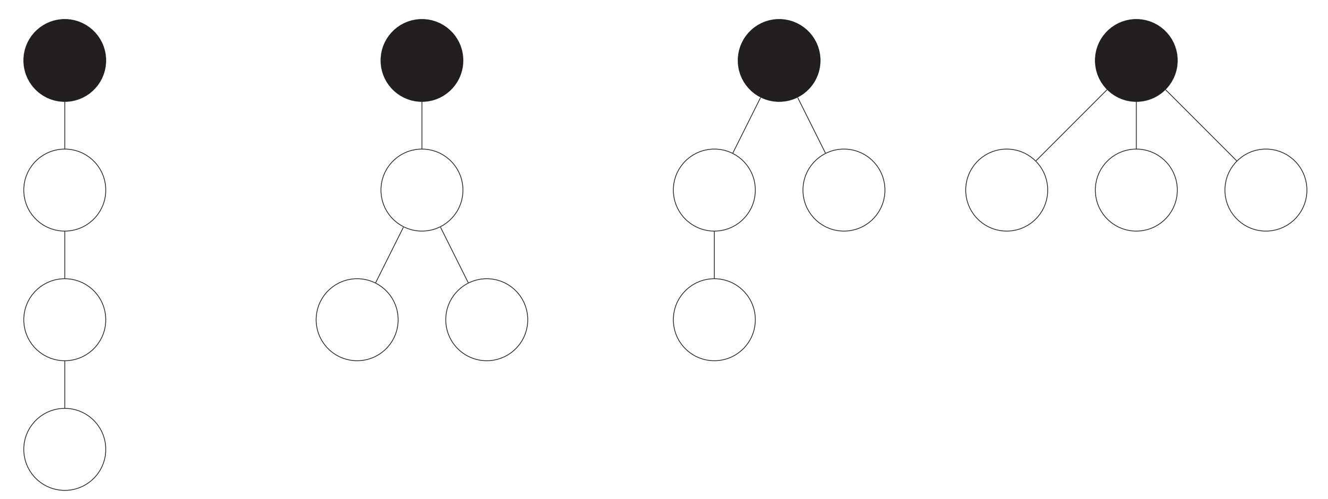


Illustration-I: All motifs of size 4. When generating a recursive tree of size 4, these motifs occur with probabilities $\frac{1}{6}, \frac{1}{6}, \frac{3}{6}$ and $\frac{1}{6}$, from left to right respectively.

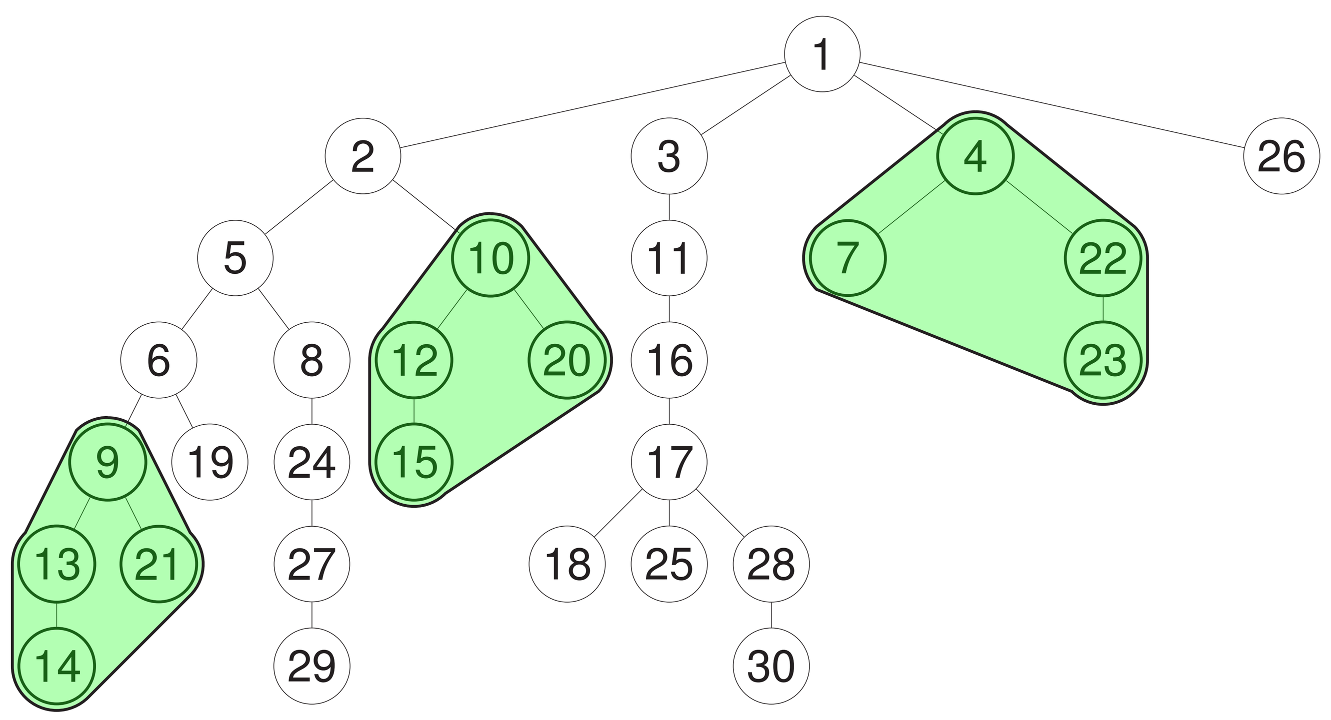


Illustration-II:Example of a recursive tree of size 30 with three occurrences of a motif on the fringe.

Applications to data compression

- Instead of storing a relatively large motif many times in a tree, we can store the content with only one nexus pointing to the motif to realize the shape in the recursive tree.
- The content itself should be stored in an appropriate canonical order to fit its original position in the recursive tree.
- In a plain practical implementation not utilizing data compression ideas, each of these nodes would carry a number of pointers (equal to the number of its children), that can be eliminated.

Research question

We want to characterize the asymptotic joint distribution of the counts of the occurrences of the motifs on the fringe.

Theorem-I

Let $\mathscr{I}$ be a countable set (finite or infinite). Let $\mathscr{C} = \{\Gamma_i \mid i \in \mathscr{I}\}$ be an uncorrelated collection of nonplanar, unlabeled, rooted trees, each of a finite size (motifs). Let $X_{n,\Gamma}$ be the number of occurrences of the motif Γ, of size γ, on the fringe of a random recursive tree of size n. Then, we have

$$\mathrm{Cov}[\mathbf{X}_{n,\mathscr{C}}] = \Sigma_{\mathscr{C}}\, n,$$

with

$$(\Sigma_{\mathscr{C}})_{i,j} = \begin{cases} \left(\dfrac{(\gamma_i+1)(2\gamma_i+1) - (3\gamma_i+2)\,\mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)^2(2\gamma_i+1)}\right)\mathcal{C}(\Gamma_i) \times \mathbf{1}_{\{n>2\gamma_i\}}, & \text{if } i=j; \\[2ex] \dfrac{1}{2}\left(\dfrac{2\mathbf{E}[X_{2\gamma^*_{i,j}+1,\Gamma_i} X_{2\gamma^*_{i,j}+1,\Gamma_j}]}{2\gamma^*_{i,j}+1} + \dfrac{\mathscr{W}(2\gamma^*_{i,j}+2, \mathscr{C}, \mathbf{b}_{i,j})}{(2\gamma^*_{i,j}+2)(2\gamma^*_{i,j}+1)} - \dfrac{\mathcal{C}^2(\Gamma_i)}{\gamma_i^2(\gamma_i+1)^2} - \dfrac{\mathcal{C}^2(\Gamma_j)}{\gamma_j^2(\gamma_j+1)^2} - \dfrac{2(2\gamma^*_{i,j}+2)\,\mathcal{C}(\Gamma_i)\,\mathcal{C}(\Gamma_j)}{\gamma_i(\gamma_i+1)\gamma_j(\gamma_j+1)}\right)\mathbf{1}_{\{n>2\gamma^*_{i,j}+1\}}, & \text{if } i \neq j; \end{cases}$$

where $\mathbf{X}_{n,\mathscr{C}}$ is the vector with components X_{n,Γ_i}, $\gamma^*_{i,j} = \max\{\gamma_i, \gamma_j\}$, $\mathscr{W}(.,.,.)$ is a function of the collection, and $\mathbf{b}_{i,j}$ is a vector of $|\mathscr{I}|$ dimensions with all entries being zero except positions i and j, where these entries are $\mathbf{1}$.

Theorem-II

Let $\mathscr{I}$ be a countable set (finite or infinite). Let $\mathscr{C} = \{\Gamma_i \mid i \in \mathscr{I}\}$ be an uncorrelated collection of nonplanar, unlabeled, rooted trees, each of finite size (motifs). Let $X_{n,\Gamma}$ be the number of occurrences of the motif Γ, of size γ, on the fringe of a random recursive tree of size n. Then, we have

$$\frac{\mathbf{X}_{n,\mathscr{C}} - \mu_{\mathscr{C}} n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}_{|\mathscr{I}|}(\mathbf{0}, \Sigma_{\mathscr{C}}),$$

where $\mathbf{X}_{n,\mathscr{C}}$, is the vector with components X_{n,Γ_i}, and $\mu_{\mathscr{C}}$ is the vector with components

$$(\mu_{\mathscr{C}})_i = \frac{\mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)},$$

for $i \in \mathscr{I}$, and $\mathcal{C}(\Gamma_i)$ is the shape functional of the motif Γ_i, $\mathcal{N}_{|\mathscr{I}|}(\mathbf{0}, \Sigma_{\mathscr{C}})$ is the jointly multivariate normally distributed random vector in $|\mathscr{I}|$ dimensions with mean vector $\mathbf{0}$ (of $|\mathscr{I}|$ components) and $|\mathscr{I}| \times |\mathscr{I}|$ covariance matrix $\Sigma_{\mathscr{C}}$.

Methodology

- We used the decomposition into special and nonspecial trees as in [3].
- As in [2] for $n > \gamma$
$$X_{n,\Gamma} \overset{\mathcal{D}}{=} X_{U_n,\Gamma} + \tilde{X}_{n-U_n,\Gamma} - \mathbf{1}_{\{n-U_n=\gamma\}}\,\mathrm{Ber}(\mathcal{C}(\Gamma));$$
where U_n is the size of the subtree(special) rooted at node **2**.
- We define $Y_{n,\mathscr{C},\alpha} = \alpha \mathbf{X}_{n,\mathscr{C}} = \sum_{i\in\mathscr{I}} \alpha_i X_{n,\Gamma_i}$ where α is any real vector of $|\mathscr{I}|$ dimensions.
- Evaluate the expectation and variance of $Y_{n,\mathscr{C},\alpha}$ which are both $\Theta(n)$.
- Prove $Y_{n,\mathscr{C},\alpha}$ satisfies the criterions given by [5] for the application of the contraction method.
- Hence under under the Maejimam-Rachev metric [4] $Y_{n,\mathscr{C},\alpha}$, under appropriate scaling, converges in distribution to the standard normal distribution.
- Invoke the Cramér-Wold device [1] to claim the asymptotic joint multivariate normality of $\mathbf{X}_{n,\mathscr{C}}$ from the asymptotic univariate normality of $Y_{n,\mathscr{C},\alpha}$.

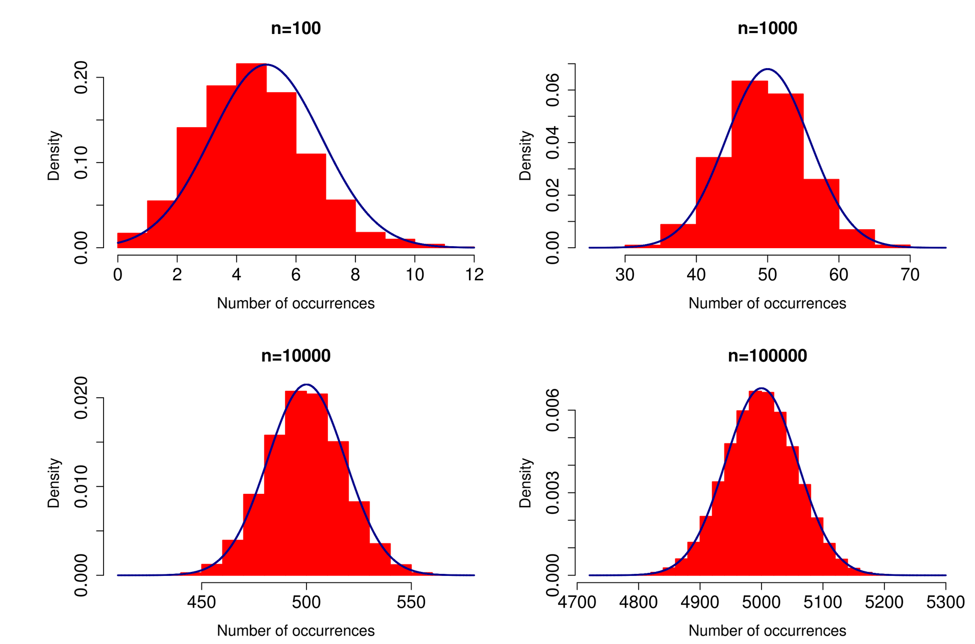
Example

Applying Theorem-II on Illustration I we have the following asymptotic result:

$$\frac{\mathbf{X}_{n,\mathscr{C}} - \begin{pmatrix}1\\1\\3\\1\end{pmatrix}\frac{n}{120}}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}_4\left(\mathbf{0}, \frac{1}{16200}\begin{pmatrix}128 & -7 & -21 & -7\\ -7 & 128 & -21 & -7\\ -21 & -21 & 342 & -21\\ -7 & -7 & -21 & 128\end{pmatrix}\right).$$

Simulations

We simulated $\mathbf{10}n$ samples of recursive trees for $n = \mathbf{100}, \mathbf{1000}, \mathbf{10000}, \mathbf{100000}$ and counted the sum of occurrences of the motifs in Illustration I. We compared them to the asymptotic normal probability predicted from Theorem-II.



Plots showing sum of occurrences of the motifs in Illustration I converging to normality

Future work

- The same question could be extended to correlated motifs.
- Count the occurrences of a single motif *everywhere* in the recursive tree.
- Characterize the probability of forbidden motifs in the fringe and the interior.

References

- Billingsley, P.: Probability and Measure, 3 edn.
Wiley-Interscience (1995)
- Feng, Q., Mahmoud, H.M.: On the variety of shapes on the fringe of a random recursive tree.
Journal of Applied Probability **47**, 191–200 (2010)
- van der Hofstad, R., Hooghiemstra, G., Van Mieghem, P.: On the covariance of the level sizes in random recursive trees.
Random Structures & Algorithms **20**, 519–539 (2002)
- Maejima, M., Rachev, S.T.: An ideal metric and the rate of convergence to a self-similar process.
The Annals of Probability **15**, 708–727 (1987)
- Rachev, S.T., Rüschendorf, L.: Probability metrics and recursive algorithms.
Advances in Applied Probability **27**, 770–799 (1995)

Acknowledgments

M. Gopaladesikan's & M. D. Ward's research is supported by NSF Science & Technology Center for Science of Information Grant CCF-0939370. This research was done while the second author was visiting Purdue Statistics. The authors thank Anirban DasGupta (Purdue) and Robert Smythe (Oregon State) for advice on some technical points.

mohang@purdue.edu,hosam@gwu.edu,mdw@purdue.edu