Balance indices for phylogenetic trees under well-known probability models
Universitat de les Illes Balears

Tomás M. Coronado
1 What is a phylogenetic tree?
   Balance
2 Probabilistic models for phylogenetic trees
   The Yule model
   The Uniform model
   The $\alpha$ and $\alpha-\gamma$ models
   The $\beta$-model
3 Balance indices
   The Colless index
   The Sackin index
   The Cophenetic index
   The Quadratic Colless index
   The rooted Quartet index
4 Conclusions
5 References
1 What is a phylogenetic tree?
   Balance

2 Probabilistic models for phylogenetic trees
   The Yule model
   The Uniform model
   The $\alpha$ and $\alpha-\gamma$ models
   The $\beta$-model

3 Balance indices
   The Colless index
   The Sackin index
   The Cophenetic index
   The Quadratic Colless index
   The rooted Quartet index

4 Conclusions

5 References
What is a phylogenetic tree?

source:
https://microbenotes.com/how-to-construct-a-phylogenetic-tree/
A phylogenetic tree depicts the joint evolutionary history of a set of species.
A phylogenetic tree depicts the joint evolutionary history of a set of species. Two main aspects are interesting to biologists:

- The length of the branches of a phylogenetic tree: the timing of speciation events.
- The *shape*, or *topology*, of the tree: differences in diversification rates among subtrees.
A phylogenetic tree depicts the joint evolutionary history of a set of species. Two main aspects are interesting to biologists:

- The length of the branches of a phylogenetic tree: the timing of speciation events.
- The shape, or topology, of the tree: differences in diversification rates among subtrees.

Reconstructing the former is usually harder than reconstructing the latter [Drummond et al. 2006], since many reconstructing methods agree on the shape.
What is a phylogenetic tree (Mathematically)?

- Let $T$ be a rooted tree, and understand it as a directed graph.
- Let $L(T)$ be the set of leaves of $T$; i.e., of the nodes with out-degree 0. Conversely, call $\mathring{V}(T) = V(T) \setminus L(T)$ the set of internal nodes of $T$.
- Let $\Lambda$ be a set of labels, and $\lambda : L(T) \to \Lambda$ a map.

The pair $(T, \lambda)$ is a phylogenetic tree if $\lambda$ is injective.
If $\lambda$ is not injective, it is a multilabelled tree.
Balance

• A popular way to assess the underlying shape of a phylogenetic tree is to consider a quantitative measure over it.
Balance

- A popular way to assess the underlying shape of a phylogenetic tree is to consider a quantitative measure over it.
- The “balance” of a phylogenetic tree is a pre-theoretic, intuitive concept reflecting its shape.
- It measures the propensity of internal nodes to have the same number of descendants.
Balance

- A popular way to assess the underlying shape of a phylogenetic tree is to consider a quantitative measure over it.
- The “balance” of a phylogenetic tree is a pre-theoretic, intuitive concept reflecting its shape.
- It measures the propensity of internal nodes to have the same number of descendants. Sort of.
Three families of trees

The caterpillar
Caterpillars are bifurcating trees all of whose internal nodes are parents to at least one leaf.

They are considered to be “the least balanced family of trees”, because

• they are completely one-sided,
• they minimize the number of automorphisms of a tree.
Three families of trees

The maximally balanced tree
Maximally balanced trees are bifurcating trees all of whose internal nodes have children whose subtrees have numbers of leaves that differ in at most 1.

They are considered to be “the most balanced family of bifurcating trees”, because it splits “as evenly as possible” the number of descendant leaves at each step.
Three families of trees

The star
Stars are usually non bifurcating trees all of whose leaves pend from the root.

They are considered to be “the most balanced family of trees”, because

• there is only one internal node,
• they maximize the number of automorphisms of a tree.
1 What is a phylogenetic tree?
   Balance

2 Probabilistic models for phylogenetic trees
   The Yule model
   The Uniform model
   The $\alpha$ and $\alpha-\gamma$ models
   The $\beta$-model

3 Balance indices
   The Colless index
   The Sackin index
   The Cophenetic index
   The Quadratic Colless index
   The rooted Quartet index

4 Conclusions

5 References
Why are we interested in probabilistic models?

• Be able to produce new trees to test evolutionary hypothesis against the trees appearing in the bibliography.
Why are we interested in probabilistic models?

- Be able to produce new trees to test evolutionary hypothesis against the trees appearing in the bibliography.
- If we know the first moments of balance indices, to test reconstructed trees against the null hypothesis “this tree is obtained under the model $P_n$”.
How to create a phylogenetic tree?

A probabilistic model \( (P_n) \) on phylogenetic trees is a family of functions \( P_n : \mathcal{T}_n \rightarrow [0, 1] \) assigning a probability \( P_n(T) \) to each \( T \in \mathcal{T}_n \) such that \( \sum_{T \in \mathcal{T}_n} P_n(T) = 1 \).

- Most models in this section only deal with bifurcating trees.
- That means that the probability of multifurcating trees is 0.
Three properties

Three properties that a probabilistic model for phylogenetic trees can have and that ease the computations are that of *Markovianity*, *shape invariance* and *sampling consistency*:

- **Markovianity** (bifurcating version): A probabilistic model \((P_n)\) of phylogenetic trees is sampling consistent if there exists a family \(q(k, n - k)\) in \([0, 1]\) such that \(\sum_{k=1}^{n-1} q(k, n - k) = 1\) and
  \[
P_n(T_k * T_{n-k}) = q(k, n - k) P_k(T_k) P_{n-k}(T_{n-k}),
\]
where \(T_k * T_{n-k}\) is the root join of \(T_k \in \mathcal{T}_k\) and \(T_{n-k} \in \mathcal{T}_{n-k}\): the tree whose root has \(T_k\) and \(T_{n-k}\) as children.

![Diagram](image)

Figur: The tree \(T_1 * \cdots * T_k\), with maximal pending subtrees \(T_1, \ldots, T_k\).
Three properties

- **Shape invariance:** If $T_1, T_2$ have the same shape but possibly different labelling, $P_n(T_1) = P_n(T_2)$. 
Three properties

- **Shape invariance**: If $T_1, T_2$ have the same shape but possibly different labelling, $P_n(T_1) = P_n(T_2)$.

- **Sampling consistency**: Given a tree $T_{n-1}$ leaves, we have

  $$P_{n-1}(T_{n-1}) = \sum_{T_n \in \mathcal{T}_n} \sum_{x \in L(T)} P_n(T_n)$$

  where $T_n(-n)$ is the tree resulting after removing the leaf labelled $n$ from $T_n$. 
The Yule model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
The Yule model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
2. For every step $m$, add a new leaf by choosing uniformly between pending arcs
The Yule model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
2. For every step $m$, add a new leaf by choosing uniformly between pending arcs
The Yule model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
2. For every step $m$, add a new leaf by choosing uniformly between pending arcs
3. Until the number of leaves $n$ is reached
The Yule model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
2. For every step $m$, add a new leaf by choosing uniformly between pending arcs
3. Until the number of leaves $n$ is reached
4. Label the tree uniformly
The Yule model

The Yule model explicitly assumes that, at each speciation event, all the current species are equally likely to speciate.
The Yule model

The Yule model explicitly assumes that, at each speciation event, all the current species are equally likely to speciate.

The Yule model is

- Markovian with $q(k, n - k) = \frac{1}{n-1}$ [Semple and Steel 2003].
- Shape invariant by construction.
- Sampling consistent [Ford 2005].
The Uniform model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
The Uniform model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
2. For every step $m$, add a new leaf by choosing uniformly between any arc
The Uniform model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
2. For every step $m$, add a new leaf by choosing uniformly between any arc

\[
\frac{1}{2(m-1)}
\]
The Uniform model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
2. For every step \( m \), add a new leaf by choosing uniformly between any arc
3. Until the number of leaves \( n \) is reached
The Uniform model

Recursive model of tree growth for bifurcating trees:

1. Start with a single node
2. For every step $m$, add a new leaf by choosing uniformly between any arc
3. Until the number of leaves $n$ is reached
4. Label the tree uniformly
The Uniform model

Equivalently: Uniformly choose a tree with \( n \) leaves from the set of all phylogenetic trees with \( n \) leaves.
The Uniform model

Equivalently: Uniformly choose a tree with $n$ leaves from the set of all phylogenetic trees with $n$ leaves. Therefore, it assumes that all the joint evolutive histories are equally likely.
The Uniform model

Equivalently: Uniformly choose a tree with $n$ leaves from the set of all phylogenetic trees with $n$ leaves. Therefore, it assumes that all the joint evolutive histories are equally likely.

- There are $(2n - 3)!!$ trees with $n$ leaves [Schröder 1870].
- Therefore, each tree has probability $\frac{1}{(2n-3)!!}$. 
The Uniform model

Equivalently: Uniformly choose a tree with $n$ leaves from the set of all phylogenetic trees with $n$ leaves. Therefore, it assumes that all the joint evolutive histories are equally likely.

- There are $(2n - 3)!!$ trees with $n$ leaves [Schröder 1870].
- Therefore, each tree has probability $\frac{1}{(2n-3)!!}$.

As a result, the Uniform model is

- Markovian with $q(k, n - k) = C_{k,n-k} = \frac{1}{2} \binom{n}{k}^{-1} \frac{(2k-3)!!(2(n-k)-3)!!}{(2n-3)!!}$ [Semple and Steel 2003], where $n!! = n(n-2)(n-4) \cdots 1$ if $n$ is odd and $n!! = n(n-2)(n-4) \cdots 2$ if it is even.
- Shape invariant by construction.
- Sampling consistent [Ford 2005].
The \(\alpha\)-model

Recursive and parametric model of tree growth for bifurcating trees with \(0 \leq \alpha \leq 1\):

1. Start with a single node labelled
The $\alpha$-model

Recursive and parametric model of tree growth for bifurcating trees with $0 \leq \alpha \leq 1$:

1. Start with a single node labelled
2. For every step $m$, add a new leaf by choosing randomly between:
The $\alpha$-model

Recursive and parametric model of tree growth for bifurcating trees with $0 \leq \alpha \leq 1$:

1. Start with a single node labelled
2. For every step $m$, add a new leaf by choosing randomly between:
   - pending arc

\[ \frac{1 - \alpha}{n - \alpha} \]
The $\alpha$-model

Recursive and parametric model of tree growth for bifurcating trees with $0 \leq \alpha \leq 1$:

1. Start with a single node labelled
2. For every step $m$, add a new leaf by choosing randomly between:
   - pending arc
   - internal arc

\[
\frac{\alpha}{n-\alpha}
\]
The $\alpha$-model

Recursive and parametric model of tree growth for bifurcating trees with $0 \leq \alpha \leq 1$:

1. Start with a single node labelled
2. For every step $m$, add a new leaf by choosing randomly between:
   - pending arc
   - internal arc (including a new root)

\[
\frac{\alpha}{n-\alpha}
\]
The $\alpha$-model

Recursive and parametric model of tree growth for bifurcating trees with $0 \leq \alpha \leq 1$:

1. Start with a single node labelled
2. For every step $m$, add a new leaf by choosing randomly between:
   - pending arc
   - internal arc (including a new root)
3. Until number of leaves $n$ is reached
The $\alpha$-model

Recursive and parametric model of tree growth for bifurcating trees with $0 \leq \alpha \leq 1$:

1. Start with a single node labelled
2. For every step $m$, add a new leaf by choosing randomly between:
   - pending arc
   - internal arc (including a new root)
3. Until number of leaves $n$ is reached
4. Label the tree uniformly
The $\alpha$-model

- Markovian [Ford 2005].
- Shape invariant by construction.
- Sampling consistent [Ford 2005].
The $\alpha$-model

- Equal to the Yule model if $\alpha = 0$ [Ford 2005].
- Equal to the Uniform model if $\alpha = 1/2$ [Ford 2005].
The $\alpha$-$\gamma$-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
The $\alpha$-$\gamma$-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
2. For every step $m$, add a new leaf by choosing randomly between:

---

```
1
o---o
    |
  2---3---4---5
```
The $\alpha$-$\gamma$-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
2. For every step $m$, add a new leaf by choosing randomly between:
   
   - pending arc

\[
\frac{1-\alpha}{n-\alpha}
\]
The $\alpha$-$\gamma$-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
2. For every step $m$, add a new leaf by choosing randomly between:
   - pending arc
   - internal node

\[
\frac{(\text{deg}(v) - 1)\alpha - \gamma}{n - \alpha}
\]
The $\alpha$-$\gamma$-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
2. For every step $m$, add a new leaf by choosing randomly between:
   - pending arc
   - internal node
   - internal arc

\[ \frac{\gamma}{n-\alpha} \]
The $\alpha$-$\gamma$-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
2. For every step $m$, add a new leaf by choosing randomly between:
   - pending arc
   - internal node
   - internal arc (including a new root)

$\frac{\gamma}{n-\alpha}$
The $\alpha$-$\gamma$-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
2. For every step $m$, add a new leaf by choosing randomly between:
   - pending arc
   - internal node
   - internal arc (including a new root)
and label it $m$
The $\alpha$-$\gamma$-model

Recursive and parametric model of tree growth for multifurcating trees with $0 \leq \gamma \leq \alpha \leq 1$:

1. Start with a single node labelled 1
2. For every step $m$, add a new leaf by choosing randomly between:
   - pending arc
   - internal node
   - internal arc (including a new root)
   and label it $m$
3. Until number of leaves $n$ is reached
The $\alpha$-$\gamma$-model

The only probabilistic model presented here of multifurcating trees.

- Markovian [Chen, Ford, and Winkel 2009].
- Not shape invariant in general.
- Sampling consistent [Chen, Ford, and Winkel 2009].
The $\alpha$-$\gamma$-model

- Equal to the $\alpha$-model when $\alpha = \gamma$ if we relabel each leaf uniformly [Chen, Ford, and Winkel 2009].
The $\beta$-model

1. Start with $n$ dots uniformly distributed over the interval $[0,1]$
The $\beta$-model

1. Start with $n$ dots uniformly distributed over the interval $[0, 1]$
2. Choose a point in $[0, 1]$ with beta density

$$f(x) = \frac{\Gamma(2\beta + 2)}{\Gamma^2(\beta + 1)} x^\beta (1 - x)^\beta, \quad 0 < x < 1.$$
The $\beta$-model

1. Start with $n$ dots uniformly distributed over the interval $[0, 1]$
2. Choose a point in $[0, 1]$ with beta density

$$f(x) = \frac{\Gamma(2\beta + 2)}{\Gamma^2(\beta + 1)} x^\beta (1 - x)^\beta, \quad 0 < x < 1.$$
The $\beta$-model

1. Start with $n$ dots uniformly distributed over the interval $[0, 1]$
2. Choose a point in $[0, 1]$ with beta density

$$f(x) = \frac{\Gamma(2\beta + 2)}{\Gamma^2(\beta + 1)} x^\beta (1 - x)^\beta, \quad 0 < x < 1.$$
The $\beta$-model

1. Start with $n$ dots uniformly distributed over the interval $[0, 1]$
2. Choose a point in $[0, 1]$ with beta density

\[ f(x) = \frac{\Gamma(2\beta + 2)}{\Gamma^2(\beta + 1)} x^\beta (1 - x)^\beta, \quad 0 < x < 1. \]

3. Until each pair of leaves is separated by at least one point
The $\beta$-model

1. Start with $n$ dots uniformly distributed over the interval $[0, 1]$
2. Choose a point in $[0, 1]$ with beta density

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

3. Until each pair of leaves is separated by at least one point
4. Construct the tree accordingly
The $\beta$-model

1. Start with $n$ dots uniformly distributed over the interval $[0, 1]$
2. Choose a point in $[0, 1]$ with beta density
   \[ f(x) = \frac{\Gamma(2\beta + 2)}{\Gamma^2(\beta + 1)} x^\beta (1 - x)^\beta, \quad 0 < x < 1. \]
3. Until each pair of leaves is separated by at least one point
4. Construct the tree accordingly
5. Label the tree uniformly
The $\beta$-model

- It is Markovian [Aldous 1996].
- Shape invariant by construction.
- Sampling consistent [Aldous 1996].
The $\beta$-model

- Equal to the Yule model if $\beta = 0$ [Aldous 1996].
- Equal to the Uniform model if $\beta = -3/2$ [Aldous 1996].
The $\beta$-model

- Equal to the Yule model if $\beta = 0$ [Aldous 1996].
- Equal to the Uniform model if $\beta = -3/2$ [Aldous 1996].
- Therefore, the $\alpha$ and $\beta$ models intersect at these points...
The \( \beta \)-model

- Equal to the Yule model if \( \beta = 0 \) [Aldous 1996].
- Equal to the Uniform model if \( \beta = -3/2 \) [Aldous 1996].
- Therefore, the \( \alpha \) and \( \beta \) models intersect at these points...
- ... and these are the only points at which they intersect (Theorem 43 at [Ford 2005]).
1 What is a phylogenetic tree?
   Balance
2 Probabilistic models for phylogenetic trees
   The Yule model
   The Uniform model
   The $\alpha$ and $\alpha-\gamma$ models
   The $\beta$-model
3 Balance indices
   The Colless index
   The Sackin index
   The Cophenetic index
   The Quadratic Colless index
   The rooted Quartet index
4 Conclusions
5 References
Balance indices: What do we know?

- Most balance indices have only been studied under the models of Yule and Uniform.
Balance indices: What do we know?

- Most balance indices have only been studied under the models of Yule and Uniform.
- The only index presented here of which we know both the first and second moments under every probabilistic model presented is the rooted Quartet index.
The Colless index

- Introduced in [Colless 1982].
- Only sound for bifurcating trees.
- Let \( u \in \mathcal{V}(T) \), and call \( u_1, u_2 \) its two children. Let \( \kappa(u_i) \) be the number of leaves of \( T \) under \( u_i \).
- Then,

\[
C(T) = \sum_{u \in \mathcal{V}(T)} |\kappa(u_1) - \kappa(u_2)|.
\]
The Colless index

- Introduced in [Colless 1982].
- Only sound for bifurcating trees.
- Let $u \in \mathcal{V}(T)$, and call $u_1, u_2$ its two children. Let $\kappa(u_i)$ be the number of leaves of $T$ under $u_i$.
- Then,

$$C(T) = \sum_{u \in \mathcal{V}(T)} |\kappa(u_1) - \kappa(u_2)|.$$

In other words, the sum over all internal nodes of the absolute difference of numbers of leaves of each pair of subtrees rooted at the same internal node.
The Colless index

The Colless index has the undeniable quality of being intuitive, as it sums up all the “local imbalances” of a tree.

• Its maximum value for a tree with \( n \) leaves is \( \binom{n-1}{2} \) and it is attained exactly by the caterpillars [Mir, Rotger, and Rosselló 2013].

• Its minimum value is \( \sum_{i=0}^{\ell-1} 2^{m_i} (m_\ell - m_i - 2(\ell - i - 1)) \), where \( \sum_{i=0}^{\ell} 2^{m_i} \), with \( m_i < m_{i+1} \), is the binary decomposition of \( n \). It is attained by the maximally balanced trees, among other trees [Coronado, Fischer, et al. 2020].
The Colless index

The Colless index has the undeniable quality of being intuitive, as it sums up all the “local imbalances” of a tree.

- Its maximum value for a tree with \( n \) leaves is \( \binom{n-1}{2} \) and it is attained exactly by the caterpillars [Mir, Rotger, and Rosselló 2013].
- Its minimum value is \( \sum_{i=0}^{\ell-1} 2^{m_i} (m_\ell - m_i - 2(\ell - i - 1)) \), where \( \sum_{i=0}^{\ell} 2^{m_i} \), with \( m_i < m_{i+1} \), is the binary decomposition of \( n \). It is attained by the maximally balanced trees, among other trees [Coronado, Fischer, et al. 2020].
- By far, the most popular balance index in the literature.
The Colless index: what do we know?

<table>
<thead>
<tr>
<th>index</th>
<th>$E_{\text{Yule}}$</th>
<th>$\sigma_{\text{Yule}}^2$</th>
<th>$E_{\text{unif}}$</th>
<th>$\sigma_{\text{unif}}^2$</th>
<th>$E_\alpha$</th>
<th>$\sigma_\alpha^2$</th>
<th>$E_\beta$</th>
<th>$\sigma_\beta^2$</th>
</tr>
</thead>
</table>

[3] Blum, François, and Janson 1996
The Colless index: what do we know?

<table>
<thead>
<tr>
<th>index</th>
<th>$E_{\text{Yule}}$</th>
<th>$\sigma^2_{\text{Yule}}$</th>
<th>$E_{\text{unif}}$</th>
<th>$\sigma^2_{\text{unif}}$</th>
<th>$E_\alpha$</th>
<th>$\sigma^2_\alpha$</th>
<th>$E_\beta$</th>
<th>$\sigma^2_\beta$</th>
</tr>
</thead>
</table>

[3]  Blum, François, and Janson 1996

- If we knew the expected value or the variance under the $\beta$ or $\alpha$ model, we would know it under the Uniform model.
The Sackin index

- Introduced in [Sokal 1983].
- Can be defined for all trees, but we usually study it only for bifurcating trees.
- Defined as

\[ S(T) = \sum_{x \in L(T)} \delta(x), \]

where \( \delta(x) \) is the depth of \( x \); i.e., the length of the shortest path from the root to \( x \).
The Sackin index

- Introduced in [Sokal 1983].
- Can be defined for all trees, but we usually study it only for bifurcating trees.
- Defined as

\[ S(T) = \sum_{x \in L(T)} \delta(x), \]

where \( \delta(x) \) is the depth of \( x \); i.e., the length of the shortest path from the root to \( x \).

In other words, the sum of the depths of all the leaves of \( T \).
The Sackin index

Also intuitive: the caterpillar has more different depths than the maximally balanced tree does.

- Its maximum value for a tree with \( n \) leaves is \( \frac{(n-1)(n+2)}{2} \) and it is attained exactly by the caterpillars [Fischer 2018].
- Its minimum value is \( 2^m m + 2s(m + 1) \), where \( n = 2^m + s \), with \( s < 2^m \). It is attained exactly by the maximally balanced trees and the trees depth-equivalent to them [Fischer 2018].
The Sackin index

Also intuitive: the caterpillar has more different depths than the maximally balanced tree does.

- Its maximum value for a tree with \( n \) leaves is \( \frac{(n-1)(n+2)}{2} \) and it is attained exactly by the caterpillars [Fischer 2018].
- Its minimum value is \( 2^m m + 2s(m+1) \), where \( n = 2^m + s \), with \( s < 2^m \). It is attained exactly by the maximally balanced trees and the trees depth-equivalent to them [Fischer 2018].
- The second most popular balance index in the literature.
The Sackin index: what do we know?

<table>
<thead>
<tr>
<th>index</th>
<th>$E_{\text{Yule}}$</th>
<th>$\sigma^2_{\text{Yule}}$</th>
<th>$E_{\text{unif}}$</th>
<th>$\sigma^2_{\text{unif}}$</th>
<th>$E_\alpha$</th>
<th>$\sigma^2_\alpha$</th>
<th>$E_\beta$</th>
<th>$\sigma^2_\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colless</td>
<td>✓</td>
<td>✓</td>
<td>✓ O</td>
<td>X ×</td>
<td>✓ ×</td>
<td>✓ ×</td>
<td>✓ ×</td>
<td></td>
</tr>
</tbody>
</table>

The Sackin index

This last result is known thanks to the proof in the Supplementary Material of [Coronado, Mir, Rosselló, and Rotger 2020] of Proposition 6 thereof: the solution of the family of recurrences

\[ X_n = 2 \sum_{k=1}^{n-1} C_k X_k + \sum_{l=1}^{r} a_l \binom{n}{l} + \frac{(2n-2)!!}{(2n-3)!!} \sum_{i=1}^{s} b_l \binom{n}{l}, \]

with initial condition \( X_1 \) and \( a_l, b_l \) real numbers.
The Sackin index

This last result is known thanks to the proof in the Supplementary Material of [Coronado, Mir, Rosselló, and Rotger 2020] of Proposition 6 thereof: the solution of the family of recurrences

\[ X_n = 2 \sum_{k=1}^{n-1} C_k X_k + \sum_{l=1}^{r} a_l \binom{n}{l} + \frac{(2n-2)!!}{(2n-3)!!} \sum_{i=1}^{s} b_l \binom{n}{l}, \]

with initial condition \( X_1 \) and \( a_l, b_l \) real numbers.

As a further note, the term \( \frac{(2n-2)!!}{(2n-3)!!} \) appears when dealing with the expected value or the variance of recursive shape indices under the Uniform model.
The Colless and Sackin indices

In [Blum, François, and Janson 1996], we find the following results

- The Pearson correlation under the Yule model of the Sackin and Colless indices tends to
  \[
  \text{cor}_{\text{Yule}}(C_n, S_n) \sim \frac{27 - 2\pi^2 - 6 \log 2}{\sqrt{2(18 - \pi^2 - 6 \log 2)(21 - 2\pi^2)}} \sim 0.98,
  \]
as \( n \) goes to \( \infty \).

- Under the Uniform model,
  \[
  \frac{S_n - C_n}{n^{3/2}} \rightarrow 0
  \]
in probability as \( n \) tends to \( \infty \).

- Let \( A \) be the Airy distribution [Flajolet and Louchard 2001]. Under the Uniform model,
  \[
  \frac{S_n}{n^{3/2}} \rightarrow A
  \]
in distribution as \( n \) tends to \( \infty \).
The Cophenetic index

- Introduced in [Mir, Rotger, and Rosselló 2013].
- Can be defined for all trees, but we usually study it only for bifurcating trees.
- Defined as

\[ \Phi(T) = \sum_{x,y \in L(T)} \phi(x, y), \]

where \( \phi(x, y) \) is the cophenetic value of \( x \) and \( y \); i.e., depth of the lowest common ancestor of both \( x \) and \( y \).
The Cophenetic index

- Introduced in [Mir, Rotger, and Rosselló 2013].
- Can be defined for all trees, but we usually study it only for bifurcating trees.
- Defined as

\[
\Phi(T) = \sum_{x,y \in L(T)} \phi(x, y),
\]

where \(\phi(x, y)\) is the cophenetic value of \(x\) and \(y\); i.e., depth of the lowest common ancestor of both \(x\) and \(y\).

In other words, the sum over all pairs of leaves of the length of their shared evolutive history.
The Cophenetic index

- Its maximum value for a tree with $n$ leaves is $\binom{n}{3}$ and it is attained exactly by the caterpillars [Mir, Rotger, and Rosselló 2013].
- Its minimum value for a multifurcating tree with $n$ leaves is $\binom{n}{2}$ and is attained exactly at the stars.
- Its minimum value for a bifurcating tree with $n$ leaves is

$$\binom{n}{2} - \sum_{j=1}^{s_n} 2^{m_j(n)-1}(m_j(n) + 2(s_n - j))$$

, where $\sum_{j=0}^{\ell}$ is the binary decomposition of $n$, $m_i < m_{i+1}$ [to be submitted]. It is attained exactly by the maximally balanced trees [Mir, Rotger, and Rosselló 2013].
The Cophenetic index: what do we know?

<table>
<thead>
<tr>
<th>index</th>
<th>$E_{\text{Yule}}$</th>
<th>$\sigma^2_{\text{Yule}}$</th>
<th>$E_{\text{unif}}$</th>
<th>$\sigma^2_{\text{unif}}$</th>
<th>$E_{\alpha}$</th>
<th>$\sigma^2_{\alpha}$</th>
<th>$E_{\beta}$</th>
<th>$\sigma^2_{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colless</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Sackin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Cophenetic</td>
<td>✓ [1]</td>
<td>✓ [2]</td>
<td>✓ [1]</td>
<td>✓ [3]</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>

[3] Coronado, Mir, Rosselló, and Rotger 2020
The Cophenetic index: limit behaviour under the Yule model

We can extend the definition of the Cophenetic index continuously taking into account edge lengths [Bartoszek 2018a], call it $\hat{\Phi}$.

- For the continuos Cophenetic index, $\left(\binom{n}{2}\right)^{-1}\hat{\Phi}_n$ is a positive submartingale that converges almost surely and in $L^2$ to a finite first and second moment random variable [Bartoszek 2018a] under the Yule model.

- For the (discrete) Cophenetic index, it can be shown that $\left(\binom{n}{2}\right)^{-1}\Phi_n$ is an almost surely and $L^2$ convergent submartingale [Bartoszek 2018a] under the Yule model.
The Sackin and Cophenetic indices

• The covariance of the Sackin and Cophenetic indices under the Uniform model is known [Coronado, Mir, Rosselló, and Rotger 2020]:

\[ \text{cov}_{\text{unif}}(S_n, \Phi_n) = \binom{n}{2} \frac{26n^2 - 5n - 4}{15} - \frac{3n + 2}{8} \binom{n}{2} \frac{(2n - 2)!!}{(2n - 3)!!} \]

\[ - \frac{n}{2} \binom{n}{2} \left( \frac{(2n - 2)!!}{(2n - 3)!!} \right)^2. \]
The Sackin and Cophenetic indices

- The covariance of the Sackin and Cophenetic indices under the Uniform model is known [Coronado, Mir, Rosselló, and Rotger 2020]:

\[
\text{cov}_{\text{unif}}(S_n, \Phi_n) = \binom{n}{2} \frac{26n^2 - 5n - 4}{15} - \frac{3n + 2}{8} \binom{n}{2} \frac{(2n - 2)!!}{(2n - 3)!!} \\
- \frac{n}{2} \binom{n}{2} \left( \frac{(2n - 2)!!}{(2n - 3)!!} \right)^2.
\]

- The Pearson correlation of the Sackin and Cophenetic under the Uniform model is estimated [Coronado, Mir, Rosselló, and Rotger 2020]:

\[
\text{cor}_{\text{unif}}(S_n, \Phi_n) = \frac{52 - 15\pi}{60} \sqrt{\frac{10 - 3\pi}{3} \frac{56 - 15\pi}{240}} \sim 0.965.
\]
The Quadratic Colless index

- Introduced in [Bartoszek et al. 2020].
- Only sound for bifurcating trees.
- Let $u \in \mathcal{V}(T)$, and call $u_1, u_2$ its two children. Let $\kappa(u_i)$ be the number of leaves of $T$ under $u_i$.
- Then,

$$C^{(2)}(T) = \sum_{u \in \mathcal{V}(T)} (\kappa(u_1) - \kappa(u_2))^2.$$
The Quadratic Colless index

- Introduced in [Bartoszek et al. 2020].
- Only sound for bifurcating trees.
- Let $u \in \mathcal{V}(T)$, and call $u_1, u_2$ its two children. Let $\kappa(u_i)$ be the number of leaves of $T$ under $u_i$.
- Then,

$$C^{(2)}(T) = \sum_{u \in \mathcal{V}(T)} (\kappa(u_1) - \kappa(u_2))^2.$$  

In other words, it has the same intuitive justification as the Colless index, but the square instead of the absolute value makes it much more easy to manipulate.
The Quadratic Colless index

The Quadratic Colless index has the undeniable quality of being intuitive, as it sums up all the “local imbalances” of a tree.

- Its maximum value for a tree with $n$ leaves is $\frac{n(n-1)(2n-1)}{6}$ and it is attained exactly by the caterpillars [Bartoszek et al. 2020].
- Its minimum value is the same of the Colless index. It is attained exactly by the maximally balanced trees [Bartoszek et al. 2020].
The Quadratic Colless index

The Quadratic Colless index has the undeniable quality of being intuitive, as it sums up all the “local imbalances” of a tree.

• Its maximum value for a tree with $n$ leaves is $\frac{n(n-1)(2n-1)}{6}$ and it is attained exactly by the caterpillars [Bartoszek et al. 2020].

• Its minimum value is the same of the Colless index. It is attained exactly by the maximally balanced trees [Bartoszek et al. 2020]. In contrast with the difficult characterization of the trees attaining the minimum Colless index.
The Quadratic Colless index: what do we know?

<table>
<thead>
<tr>
<th>index</th>
<th>$E_{\text{Yule}}$</th>
<th>$\sigma^2_{\text{Yule}}$</th>
<th>$E_{\text{unif}}$</th>
<th>$\sigma^2_{\text{unif}}$</th>
<th>$E_{\alpha}$</th>
<th>$\sigma^2_{\alpha}$</th>
<th>$E_{\beta}$</th>
<th>$\sigma^2_{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colless</td>
<td>✓</td>
<td>✓</td>
<td>O</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Sackin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Cophenetic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Q. Colless</td>
<td>✓ [1]</td>
<td>✓ [1]</td>
<td>✓ [1]</td>
<td>✓ [4]</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

[1] Bartoszek et al. 2020
The Quadratic Colless index: limit behaviour under the Yule model

Set $Y := \frac{C_n^{(2)} - E_{Yule}(C_n^{(2)})}{n^2}$.

As $n \to \infty$, the distribution under the Yule model of $Y$ is such that

$$Y \to \tau^2 Y' + (1 - \tau)^2 Y'' + (1 + 6\tau^2 - 6\tau),$$

in distribution, where $\tau \sim \text{Unif}[0, 1]$ and $Y', Y''$ are independent and distributed according to the same law as the limit of $Y$ [Bartoszek et al. 2020].
The rooted Quartet index

There are five different trees with five leaves.

Figur: The five tree shapes in $\mathcal{T}_4$. 
The rooted Quartet index

There are five different trees with five leaves.

They are ordered according to their number of automorphisms, and assigned a number $q_i$ increasing on it.
The rooted Quartet index

- Introduced in [Coronado, Mir, Rosselló, and Valiente 2019].
- Can be defined (and makes sense) for all trees.
- Defined as

\[
QI(T) = \sum_{i=0}^{4} |\{Q \in \text{Part}_4(L(T)) : T(Q) = Q_i\}| \cdot q_i.
\]
The rooted Quartet index

- Introduced in [Coronado, Mir, Rosselló, and Valiente 2019].
- Can be defined (and makes sense) for all trees.
- Defined as

\[ QI(T) = \sum_{i=0}^{4} |\{Q \in \text{Part}_4(L(T)) : T(Q) = Q_i\}| \cdot q_i. \]

Notice that, in this case, the value “increases with balance”, where in the other cases more “balanced” trees had smaller index values.
The rooted Quartet index

- Its maximum value for a multifurcating tree with $n$ leaves is $\binom{n}{4}q_4$ and it is attained exactly by the stars [Coronado, Mir, Rosselló, and Valiente 2019].

- Its maximum value for a bifurcating tree with $n$ leaves can be found in Sloane’s Encyclopedia of Integer Sequences [Sloane 1964], seq. A300445. It is attained exactly by the maximally balanced trees [Coronado, Mir, Rosselló, and Valiente 2019].

- Its minimum value is 0, and it is attained exactly by the caterpillars.
The rooted Quartet index: what do we know?

<table>
<thead>
<tr>
<th>index</th>
<th>$E_{\text{Yule}}$</th>
<th>$\sigma^2_{\text{Yule}}$</th>
<th>$E_{\text{unif}}$</th>
<th>$\sigma^2_{\text{unif}}$</th>
<th>$E_{\alpha}$</th>
<th>$\sigma^2_{\alpha}$</th>
<th>$E_{\beta}$</th>
<th>$\sigma^2_{\beta}$</th>
<th>$E_{\alpha,\gamma}$</th>
<th>$\sigma^2_{\alpha,\gamma}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colless</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Sackin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Cophenetic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Q. Colless</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>r. Quartet</td>
<td>✓ [1]</td>
<td>✓ [1]</td>
<td>✓ [1]</td>
<td>✓ [1]</td>
<td>✓ [1]</td>
<td>✓ [1]</td>
<td>✓ [1]</td>
<td>✓ [1]</td>
<td>✓ [1]</td>
<td>✓ [1]</td>
</tr>
</tbody>
</table>

[1] Coronado, Mir, Rosselló, and Valiente 2019
The rooted Quartet index

- The rooted Quartet index is the only balance index presented whose first and second moments are known under all the probabilistic models presented so far.
The rooted Quartet index

• The rooted Quartet index is the only balance index presented whose first and second moments are known under all the probabilistic models presented so far.
• The backbone of the above proofs is that the $\alpha$-$\gamma$-model is sampling consistent.
The rooted Quartet index

- The rooted Quartet index is the only balance index presented whose first and second moments are known under all the probabilistic models presented so far.
- The backbone of the above proofs is that the $\alpha$-$\gamma$-model is sampling consistent.
- Indeed: for any sampling consistent probabilistic model $(P^*_n)$ of trees [Coronado, Mir, Rosselló, and Valiente 2019],

\[
E_P(QI_n) = \binom{n}{4} \sum_{i=1}^{4} P^*_4(Q_i)q_i.
\]

\[
\sigma^2_P(QI_n) = \binom{n}{4} \sum_{i=1}^{4} q_i^2 P^*_4(Q_i) - \left( \binom{n}{4} \sum_{i=1}^{4} q_i P^*_4(Q_i) \right)^2
\]

\[
\sum_{i=1}^{4} \sum_{j=1}^{4} q_i q_j \left( \sum_{k=5}^{8} \binom{n}{k} \sum_{T \in T_k} \Theta_{ij}(T) P^*_k(T) \right),
\]
The rooted Quartet index

**Under the $\alpha$-$\gamma$-model**

- The expected value of the rooted Quartet index under the $\alpha$-$\gamma$-model is known [Coronado, Mir, Rosselló, and Valiente 2019]:

$$E_{\alpha,\gamma}(QI_n) = \left( \frac{(2\alpha - \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)}q_4 + \frac{(1 - \alpha)(2(1 - \alpha) + \gamma)}{(3 - \alpha)(2 - \alpha)}q_3 \right) \frac{2(1 - \alpha + \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)}q_2 + \frac{(5(1 - \alpha) + \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)}q_1 \binom{n}{4}.$$
The rooted Quartet index

**Under the \(\alpha-\gamma\)-model**

- The expected value of the rooted Quartet index under the \(\alpha-\gamma\)-model is known [Coronado, Mir, Rosselló, and Valiente 2019]:

\[
E_{\alpha,\gamma}(QI_n) = \left(\frac{(2\alpha - \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)} q_4 + \frac{(1 - \alpha)(2(1 - \alpha) + \gamma)}{(3 - \alpha)(2 - \alpha)} q_3 \right) \left(\frac{2(1 - \alpha + \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)} q_2 + \frac{5(1 - \alpha) + \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)} q_1\right) \binom{n}{4}.
\]

- When \(\alpha = \gamma\) (\(\alpha\)-model), we get \(E_\alpha(QI_n) = \frac{(1-\alpha)(2-\alpha)}{(3-\alpha)(2-\alpha)} \binom{n}{4} q_3\).
The rooted Quartet index

**Under the $\alpha$-$\gamma$-model**

- The expected value of the rooted Quartet index under the $\alpha$-$\gamma$-model is known [Coronado, Mir, Rosselló, and Valiente 2019]:

\[
E_{\alpha,\gamma}(\text{QI}_n) = \left( \frac{(2\alpha - \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)} q_4 + \frac{(1 - \alpha)(2(1 - \alpha) + \gamma)}{(3 - \alpha)(2 - \alpha)} q_3 
+ \frac{2(1 - \alpha + \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)} q_2 + \frac{5(1 - \alpha + \gamma)(\alpha - \gamma)}{(3 - \alpha)(2 - \alpha)} q_1 \right) \binom{n}{4}.
\]

- When $\alpha = \gamma$ ($\alpha$-model), we get $E_{\alpha}(\text{QI}_n) = \frac{(1-\alpha)(2-\alpha)}{(3-\alpha)(2-\alpha)} \binom{n}{4} q_3$.

- Yule model: $\alpha = 0$. Uniform model: $\alpha = 1/2$. 
The rooted Quartet index

- The variance under the $\alpha$-$\gamma$ model is also known, but the formula is too long! [Coronado, Mir, Rosselló, and Valiente 2019]
The rooted Quartet index

Under the $\beta$-model

- The $\beta$-model is also sampling consistent.
The rooted Quartet index

**Under the $\beta$-model**

- The $\beta$-model is also sampling consistent.
- That gives us the expected value of $QI_n$ under the $\beta$-model, too [Coronado, Mir, Rosselló, and Valiente 2019]:

$$E_\beta(QI_n) = \frac{3\beta + 6}{7\beta + 18}.$$
The rooted Quartet index

**Under the \( \beta \)-model**

- The \( \beta \)-model is also sampling consistent.
- That gives us the expected value of \( \text{QI}_n \) under the \( \beta \)-model, too [Coronado, Mir, Rosselló, and Valiente 2019]:

\[
E_\beta(\text{QI}_n) = \frac{3\beta + 6}{7\beta + 18}.
\]

and its variance (again, too long!) [Coronado, Mir, Rosselló, and Valiente 2019].
The rooted Quartet index

**Under the β-model**

- The β-model is also sampling consistent.
- That gives us the expected value of QI\(_n\) under the β-model, too [Coronado, Mir, Rosselló, and Valiente 2019]:
  \[
  E_\beta(QI_n) = \frac{3\beta + 6}{7\beta + 18}.
  \]
  and its variance (again, too long!) [Coronado, Mir, Rosselló, and Valiente 2019].
- Yule model: \(\beta = 0\). Uniform model \(\beta = -3/2\).
An interesting result about the limit distribution of the Quartet index under the $\beta$-model, $\beta \geq 0$, can be found in [Bartoszek 2018b]. It shows that it converges weakly to a distribution that can be characterized as the fixed point of a contraction operator on a class of distributions.
1 What is a phylogenetic tree?
   Balance
2 Probabilistic models for phylogenetic trees
   The Yule model
   The Uniform model
   The $\alpha$ and $\alpha-\gamma$ models
   The $\beta$-model
3 Balance indices
   The Colless index
   The Sackin index
   The Cophenetic index
   The Quadratic Colless index
   The rooted Quartet index
4 Conclusions
5 References
Conclusions

• We know some things
Conclusions

- We know some things
- but we ignore some other things.
The rooted Quartet index: what do we know?

<table>
<thead>
<tr>
<th>index</th>
<th>$E_{\text{Yule}}$</th>
<th>$\sigma_{\text{Yule}}^2$</th>
<th>$E_{\text{unif}}$</th>
<th>$\sigma_{\text{unif}}^2$</th>
<th>$E_{\alpha}$</th>
<th>$\sigma_{\alpha}^2$</th>
<th>$E_{\beta}$</th>
<th>$\sigma_{\beta}^2$</th>
<th>$E_{\alpha,\gamma}$</th>
<th>$\sigma_{\alpha,\gamma}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colless</td>
<td>✓ [1]</td>
<td>✓ [2]</td>
<td>✓ [3]</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>r. Quartet</td>
<td>✓ [8]</td>
<td>✓ [8]</td>
<td>✓ [8]</td>
<td>✓ [8]</td>
<td>✓ [8]</td>
<td>✓ [8]</td>
<td>✓ [8]</td>
<td>✓ [8]</td>
<td>✓ [8]</td>
<td>✓ [8]</td>
</tr>
</tbody>
</table>

[3] Blum, François, and Janson 1996
[8] Coronado, Mir, Rosselló, and Valiente 2019
1 What is a phylogenetic tree?
   Balance

2 Probabilistic models for phylogenetic trees
   The Yule model
   The Uniform model
   The $\alpha$ and $\alpha-\gamma$ models
   The $\beta$-model

3 Balance indices
   The Colless index
   The Sackin index
   The Cophenetic index
   The Quadratic Colless index
   The rooted Quartet index

4 Conclusions

5 References


– (2018b). “Limit distribution of the quartet index for Aldous’s $\beta \geq 0$-model”. In: *biorxiv*.


Bartoszek, K. et al. (2020). “Squaring within the Colless index yields a better balance index”. In: *arXiv*.

Tack för idag!

www.liu.se