Semialgebraic Hypothesis Testing with Incomplete U-Statistics: A Case Study with Biologically-Motivated Models

Max Hill (UH Manoa, mhill@math.hawaii.edu);
Joint work with David Barnhill, Marina Garrote-López, Elizabeth Gross, John Rhodes, Bryson Kagy, and Joy Zhang

July 25, 2025

Overview

- This talk is about doing hypothesis testing with semialgebraic statistical models.
  → *Methodological considerations for semialgebraic hypothesis testing with incomplete U-statistics* (`https://arxiv.org/abs/2507.13531`)
- Recently, Sturma, Drton, and Leung (2024) [2] introduced a remarkably general stochastic method, *the SDL method*, for doing such tests.
- In this talk, I'll discuss work implementing this method and applying it to a number of biologically-motivated models.
  - Our goal was to evaluate how this method performed in practice, and also to develop best practices for using the method.
  - Along the way, we uncovered a number of surprising methodological issues.

Semialgebraic Statistical Model

- Our main object of study are *semialgebraic statistical models*—defined by polynomial equalities and inequalities.
- More precisely:

$$\mathcal{M} = \{P_\theta : P_\theta \text{ is a probability measure and } \theta \in \Theta_0\},$$

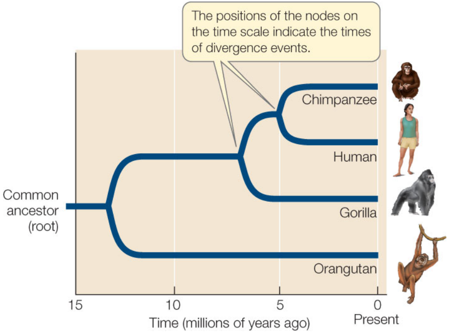
where the parameter space Θ_0 is a basic semi-algebraic set of the form

$$\Theta_0 = \left\{\theta \in \mathbb{R}^d : f_i(\theta) \leq 0, \text{ for } i = 1, \ldots, p\right\}$$

where $f_1, \ldots, f_p$ are polynomials.

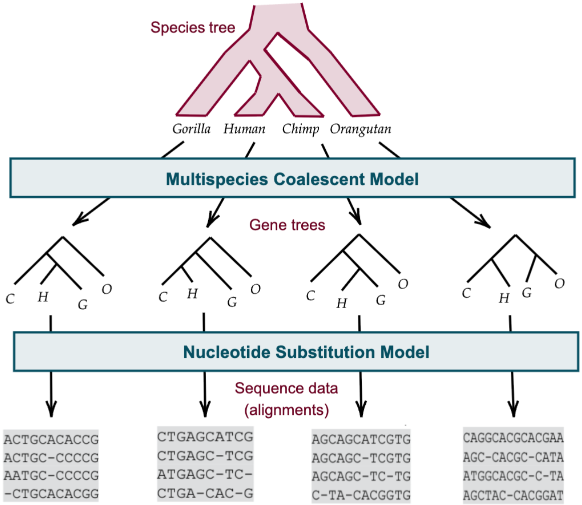
A First Example: Gene Evolution

Given a set of taxa, a *species tree* is a labeled tree representing their **population-level** evolutionary history:


The positions of the nodes on the time scale indicate the times of divergence events.

However, individual genes in the genome may have treelike histories which look different from that of the population. For great apes, about 23% of genes have treelike histories whose topologies do not match the species tree, while 77% do [1].

Larger Context: Two-step Model of Evolution



Multispecies Coalescent Model

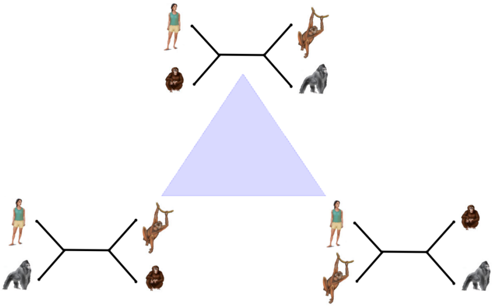
Nucleotide Substitution Model

```
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
```

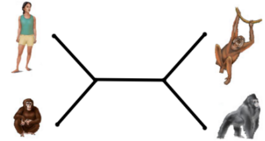
```
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
```

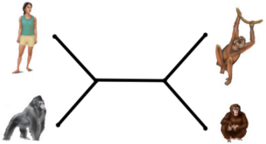
```
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
```

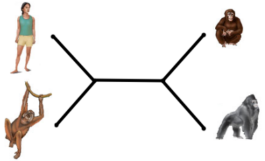
```
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT
```

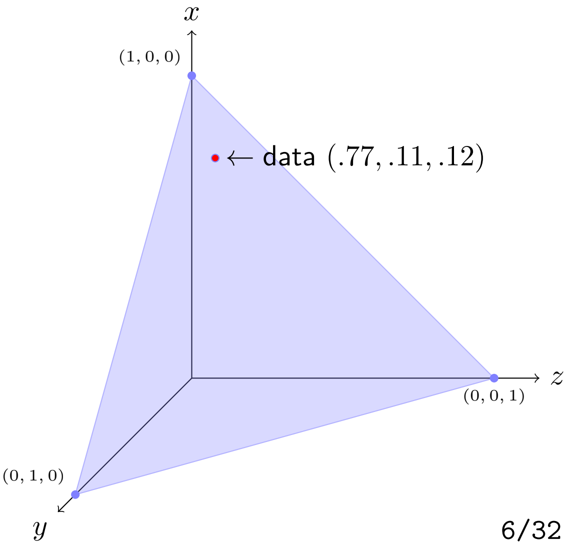


Three “Gene Tree” Toplogies

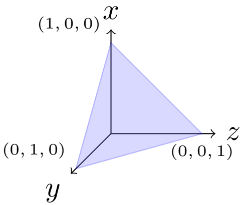
$x = P\Bigg($  $\Bigg) \approx .77$

$y = P\Bigg($  $\Bigg) \approx .11$

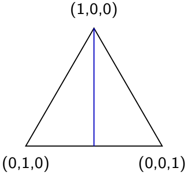
$z = P\Bigg($  $\Bigg) \approx .12$



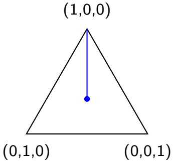
What is the semialgebraic model??



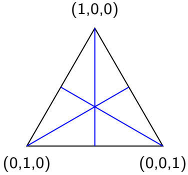
Trinomial



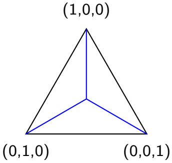
cut1



T1



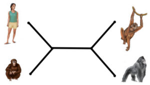
cut



T3

Model T1: $\Theta_0 = \left\{(x,y,z) \in \Delta^2 : y - z \leq 0, z - y \leq 0, \frac{1}{3} - x \leq 0\right\}$

Under a standard model of gene evolution, Model T1 represents the evolutionary hypothesis

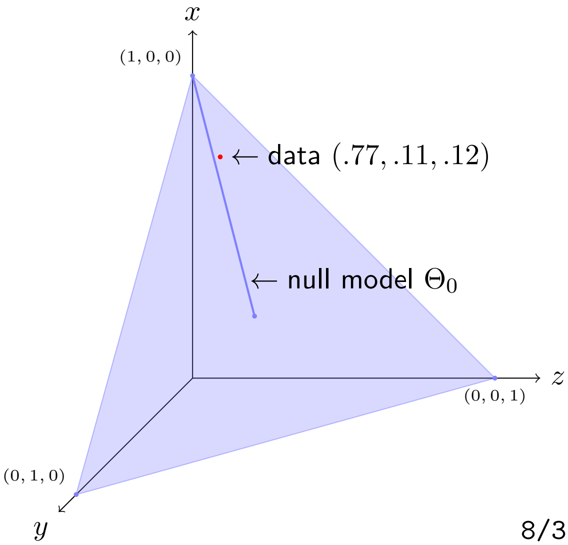


The Hypothesis Test

Data: $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$, for some unknown $\theta \in \Theta$.

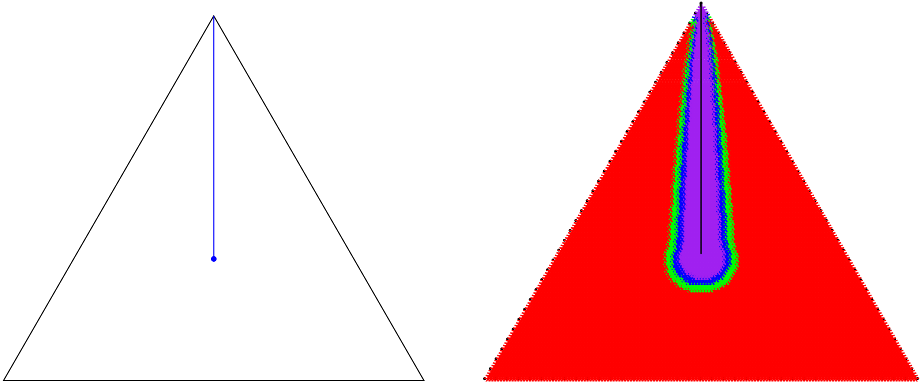
The problem: Given a semialgebraic subset $\Theta_0 \subseteq \Theta$ distinguish between

- (Null hypothesis) $H_0 : \theta \in \Theta_0$
- (Alternative hypothesis) $H_1 : \theta \notin \Theta_0$


x
$(1, 0, 0)$
$\leftarrow$ data $(.77, .11, .12)$
$\leftarrow$ null model Θ_0
z
$(0, 0, 1)$
$(0, 1, 0)$
y

Classical Approach: The Likelihood Ratio Test

The likelihood ratio test gives a measure of the distance of the data from the submodel Θ_0.



🟥 $p < 0.01$ 🟩 $0.01 \leq p < 0.05$ 🟦 $0.05 \leq p < 0.10$ 🟪 $p \geq 0.10$

The likelihood ratio test runs into trouble near irregularities, i.e., singular points and certain boundaries.

A Brief Introduction to the SDL Method

Given a null model

$$\Theta_0 = \left\{\theta \in \mathbb{R}^d : f_i(\theta) \leq 0 \text{ for polynomials } f_i \text{ with } i = 1, \ldots, p\right\}$$

- **Subsample:** $S = \{X_{i_1}, \ldots, X_{i_m}\}$, a set m data points drawn from $X_1, \ldots, X_n$
- **Kernel function:** A symmetric function $h : \mathbb{R}^m \to \mathbb{R}^p$ such that
  - $h(S)$ is an unbiased estimator of $f(\theta) := (f_1(\theta), \ldots, f_p(\theta))$
- We are going to take random subsamples and plug them into the kernel function
  - do this once and you get a poor estimate of the polynomial constraints
  - but we'll do this many times and take the average.

The Test Statistic

- **Incomplete U-statistic:** Take the average of the value of $h(S)$ over many randomly-chosen subsamples S:

  $$U := \frac{1}{|\mathcal{I}|} \sum_{S \in \mathcal{I}} h(S),$$

  where $\mathcal{I}$ is a *random* collection of subsamples of the data.
- The **SDL Test Statistic:**

  $$\mathcal{T} = \max_{1 \leq j \leq p} \frac{\sqrt{n} U_j}{\widehat{\sigma}_j}$$

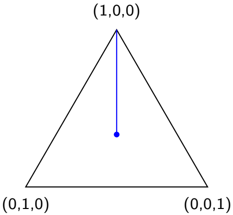
  where $\widehat{\sigma}_j$ is an approximation of the standard deviation of U_j obtained by Gaussian bootstrapping.
- A large value of $\mathcal{T}$ is interpreted as evidence against H_0.
  - → How we interpret "large" can be formalized through the use of a Gaussian bootstrap approximation of the distribution of a related statistic, allowing us to compute "p-values".
- This is a **stochastic test:** the SDL "p-values" are actually estimates of p-values.

What's the big picture?

- The holy grail of semialgebraic hypothesis testing is the ability to do valid a hypothesis test for any semialgebraic model using only the defining inequalities—without knowing anything else about the model geometry.
  - → We've seen that classical methods may fail near singularities of the model. And other existing algebraic methods in phylogenetics are often ad hoc, tailored to specific models.
- By contrast, the SDL method offers a rigorous and fully general statistical framework for hypothesis testing.
  - → Has good statistical guarantees, even near model singularities.
- But understanding how best to implement the SDL test is not trivial. In the remainder of the talk, I will discuss two of the methodological challenges that arose, and how we were able to deal with them.

Challenge #1

There is more than one way to represent a semialgebraic set



There is more than one way to represent this model with polynomial inequalities:

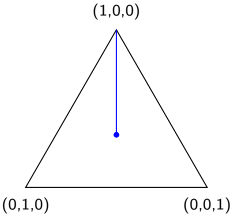
Representation A

$$y - z \leq 0$$
$$z - y \leq 0$$
$$\frac{1}{3} - x \leq 0$$

Representation B

$$y - z \leq 0$$
$$z - y \leq 0$$
$$\frac{2}{3} - y - z \leq 0$$

There is more than one way to represent a semialgebraic set



There is more than one way to represent this model with polynomial inequalities:

Representation A

$$\begin{aligned} y - z &\leq 0 \\ z - y &\leq 0 \\ \frac{1}{3} - x &\leq 0 \end{aligned}$$

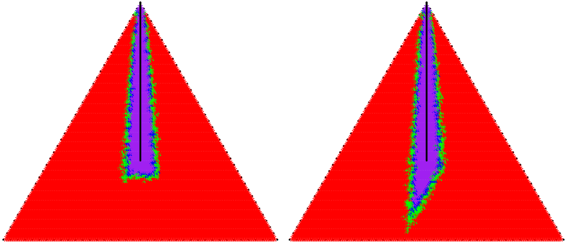
Representation B

$$\begin{aligned} y - z &\leq 0 \\ z - y &\leq 0 \\ \frac{2}{3} - y - z &\leq 0 \end{aligned}$$

Representation C

$$\begin{aligned} y - z &\leq 0 \\ z - y &\leq 0 \\ \tfrac{1}{3} - x &\leq 0 \\ .038\overline{6} - .116x - .214y + .214z &\leq 0 \\ .031\overline{3} - .094x - .291y + .291z &\leq 0 \\ .277\overline{6} - .833x - .143y + .143z &\leq 0 \\ .210\overline{3} - .631x + .175y - .175z &\leq 0 \\ .082\overline{6} - .248x + .225y - .225z &\leq 0 \\ .147\overline{3} - .442x + .064y - .064z &\leq 0 \\ .040\overline{6} - .122x + .865y - .865z &\leq 0 \\ .216\overline{3} - .649x + .342y - .342z &\leq 0 \\ .079\overline{3} - .238x + .118y - .118z &\leq 0 \\ .114 - .342x - .474y + .474z &\leq 0 \end{aligned}$$

The SDL test is affected by the choice of model constraints (1/2)



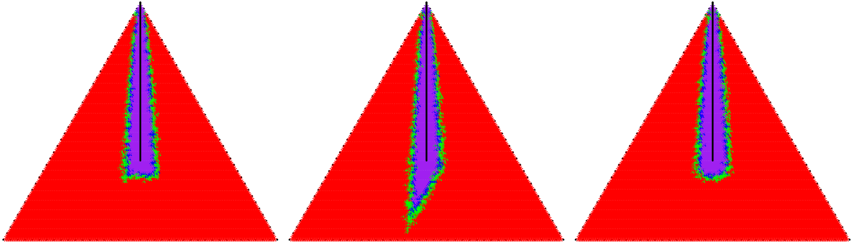
Representation A

Representation B

🟥 $p < 0.01$ 🟩 $0.01 \leq p < 0.05$ 🟦 $0.05 \leq p < 0.10$ 🟪 $p \geq 0.10$

The rejection regions are shaped differently!

The SDL test is affected by the choice of model constraints (2/2)

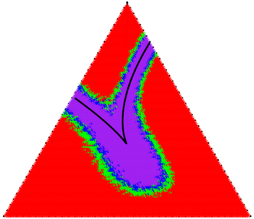


Representation A   Representation B   Representation C

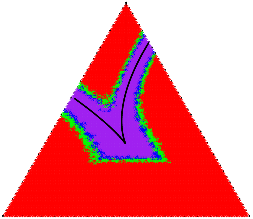
🟥 $p < 0.01$ 🟩 $0.01 \leq p < 0.05$ 🟦 $0.05 \leq p < 0.10$ 🟪 $p \geq 0.10$

The rejection regions are shaped differently!

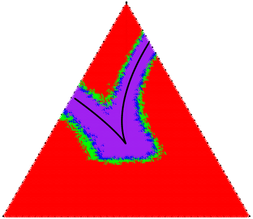
Here's another example, involving the cuspidal cubic model:



$$(y - \tfrac{1}{3})^2 - (x - \tfrac{1}{3})^3 = 0$$

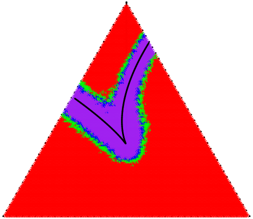


$$(y - \tfrac{1}{3})^2 - (x - \tfrac{1}{3})^3 = 0$$
$$\tfrac{1}{3} - x \leq 0$$



$$(y - \tfrac{1}{3})^2 - (x - \tfrac{1}{3})^3 = 0$$
$$\tfrac{1}{3} - x \leq 0$$

$+$ 10 random convex combinations



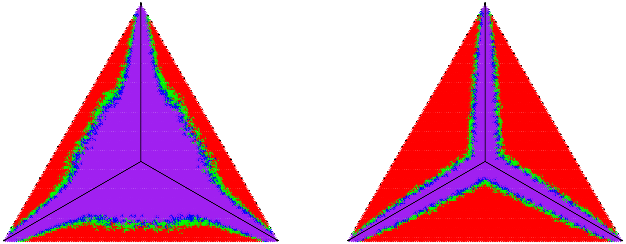
$$(y - \tfrac{1}{3})^2 - (x - \tfrac{1}{3})^3 = 0$$
$$\tfrac{1}{3} - x \leq 0$$

$+$ 10 random convex combinations

$+$ 2 well-chosen linear inequalities

Adding redundant constraints tends to improve test performance. And being smart about how you choose your additional constraints even more so!

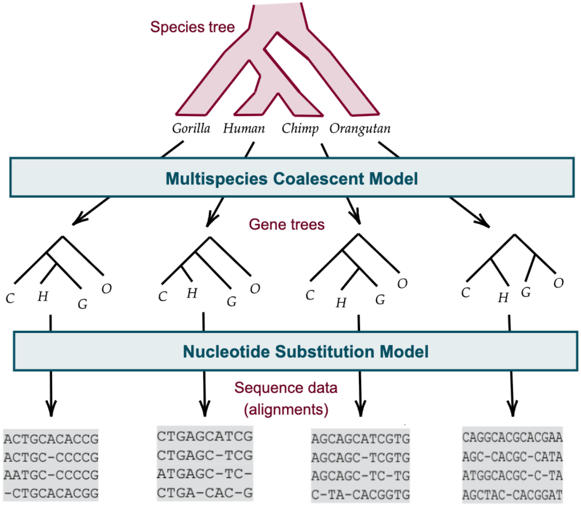
One more example, that knowing model geometry can be valuable:



$$
\begin{aligned}
(x-y)(x-z)(y-z) &= 0\\
(x-z)^2(y-z)^2(1/3-x) &\leq 0\\
(x-y)^2(y-z)^2(1/3-y) &\leq 0\\
(x-y)^2(x-z)^2(1/3-z) &\leq 0
\end{aligned}
$$

- **Left:** Rejection region using defining constraints and 10 random convex combinations.
- **Right:** Using an Intersection-Union test using the SDL tests for the 3 irreducible components of the model.

Now, let's transition to a second class of models



Species tree

Multispecies Coalescent Model

Gene trees

Nucleotide Substitution Model

Sequence data (alignments)

```
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
```

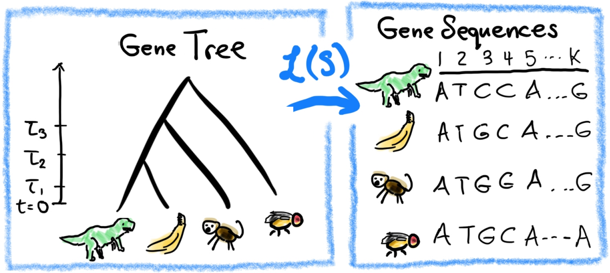
```
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
```

```
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
```

```
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT
```

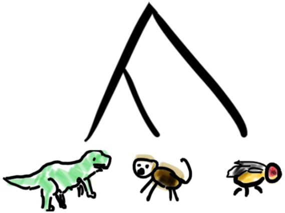
Nucleotide Evolution

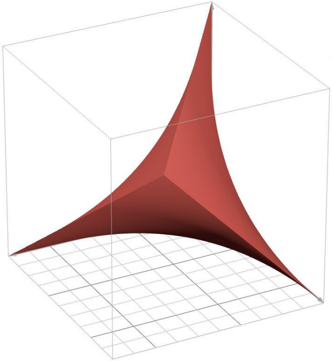
- Another class of semialgebraic models arise when modeling DNA mutation on macroevolutionary timescales.
- **Standard approach:** DNA sequences are produced from a stochastic process parameterized by an evolutionary tree:



- We consider the simplest such model, the Cavendar-Farris-Neyman (CFN) model.

The CFN Model, Visualized





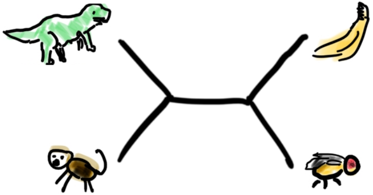
- For a 3-leaf tree, the CFN model is a 3-dimensional semialgebraic subset of Δ^3:

$$\begin{cases} x, y, z \geq 0 \\ (1-2x-2y)(1-2x-2y) \leq 1-2y-2z \\ (1-2x-2y)(1-2y-2y) \leq 1-2x-2z \\ (1-2y-2z)(1-2x-2z) \leq 1-2x-2y \end{cases}$$

- For 4-leaf trees, we get a 5-dimensional set (next slide).

CFN Model for a 4-leaf tree

The CFN model for a 4-leaf tree is a 5-dimensional subset of $\Delta^7 = \{p \in \mathbb{R}^8 : p_i \geq 0, \text{ and } p_1 + \ldots + p_8 - 1 = 0\}$ satisfying



$$\begin{aligned} p_3p_5 - p_4p_6 - p_1p_7 + p_2p_8 &= 0 \\ p_2p_5 - p_1p_6 - p_4p_7 + p_3p_8 &= 0 \\ p_2p_3 - p_1p_4 - p_6p_7 + p_5p_8 &\leq 0 \end{aligned}$$

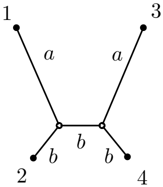
along with additional inequalities

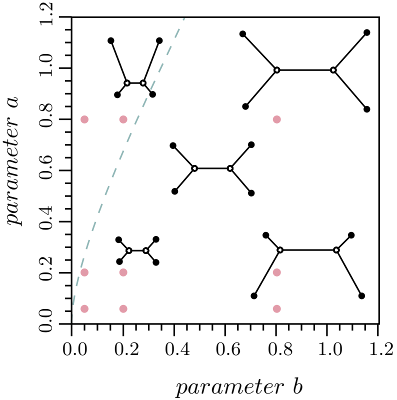
$$\begin{aligned} (p_3+p_4)(p_5+p_6) - (p_1+p_2)(p_7+p_8) &\leq 0, & (p_2+p_6)(p_3+p_7) - (p_1+p_5)(p_4+p_8) &\leq 0 \\ (p_2+p_4)(p_6+p_8) - (p_1+p_3)(p_5+p_7) &\leq 0, & (p_2+p_7)(p_3+p_6) - (p_4+p_5)(p_1+p_8) &\leq 0 \\ (p_5+p_6)(p_7+p_8) - (p_1+p_2)(p_3+p_4) &\leq 0, & (p_2+p_6)(p_4+p_8) - (p_1+p_5)(p_3+p_7) &\leq 0 \\ (p_5+p_7)(p_6+p_8) - (p_1+p_3)(p_2+p_4) &\leq 0, & (p_3+p_6)(p_4+p_5) - (p_1+p_8)(p_2+p_7) &\leq 0 \end{aligned}$$

(There are other ways to represent this set using polynomial inequalities.)

What phylogenetic trees did we look at?

We tested the SDL method for a range of parameters:



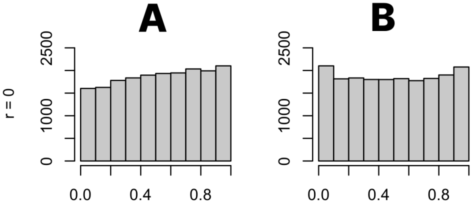


Our tests focused on the topology of the tree (the true topology is always $12|34$).

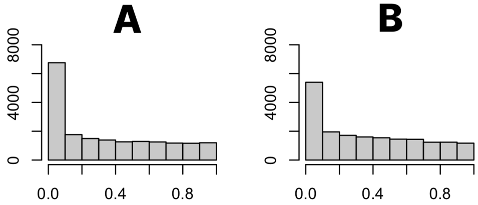
Again, we find the choice of generating polynomials matters:

Let's compare SDL p-values from two different representations A and B of the polynomial constraints:

Tests of the true null hypothesis $(H_{12|34})$



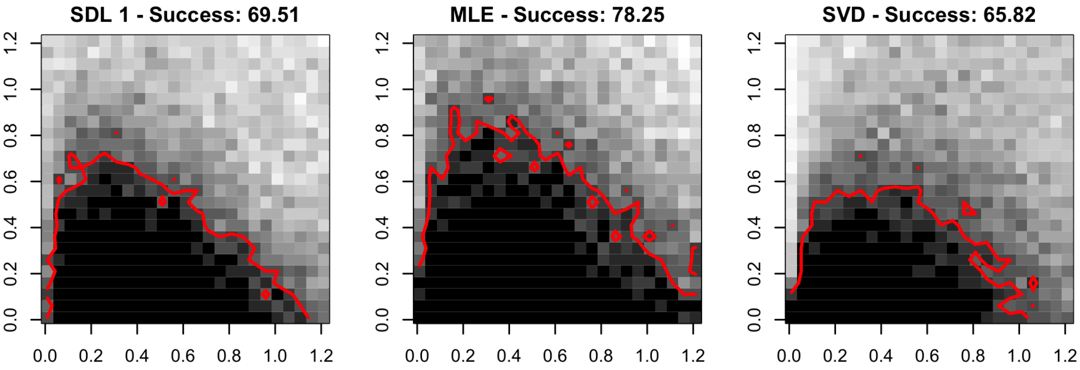
Tests of a false null hypothesis $(H_{13|24})$



- Conclusion: A is better than B!
- However, when we added 20 random convex combination constraints, differences in performance were substantially reduced.

Going Beyond Hypothesis Testing: A New Direction

We introduced a new method to infer the a 4-leaf tree topology.



SDL 1 - Success: 69.51

MLE - Success: 78.25

SVD - Success: 65.82

Challenge #2

The SDL test requires a symmetric kernel

- Recall the kernel function h used to define the incomplete U-statistic:

$$U := \frac{1}{|\mathcal{I}|} \sum_{S \in \mathcal{I}} h(S),$$

- The theory requires that h must be a *symmetric function*: i.e., for any permutation π,

$$h(x_1, \ldots, x_m) = h(x_{\pi(1)}, \ldots, x_{\pi(m)}).$$

- On one hand, if your kernel is not symmetric, maybe that's okay... a non-symmetric kernel h can always be symmetrized by averaging over all permutations of its m arguments:

$$h_{\text{sym}}(x_1, \ldots, x_m) := \frac{1}{m!} \sum_{\pi \in \mathcal{S}_m} h(x_{\pi(1)}, \ldots, x_{\pi(m)})$$

A simple example of what I'm talking about:

- Suppose we have the non-symmetric kernel

$$h(x_1, x_2, x_3) = x_1 x_2 + x_3^2.$$

  (This is not symmetric because, e.g., $h(1,2,0) = 2$ but $h(2,0,1) = 1$.)
- But we can symmetrize it:

$$\begin{aligned} h_{\text{sym}}(x) &= \frac{1}{3!} \sum_{\pi \in S_3} h(x_{\pi(1)}, x_{\pi(2)}, x_{\pi(3)}) \\ &= \frac{1}{6} \left(h(x_1,x_2,x_3) + h(x_1,x_3,x_2) + h(x_2,x_1,x_3) + h(x_2,x_3,x_1) + h(x_3,x_1,x_2) + h(x_3,x_2,x_1)\right) \\ &= \frac{1}{6} \left((x_1x_2 + x_3^2) + (x_1x_3 + x_2^2) + (x_2x_1 + x_3^2) + (x_2x_3 + x_1^2) + (x_3x_1 + x_2^2) + (x_3x_2 + x_1^2)\right) \\ &= \frac{1}{3} \left(x_1x_2 + x_1x_3 + x_2x_3 + x_1^2 + x_2^2 + x_3^2\right) \end{aligned}$$

- The sum has $3! = 6$ terms.

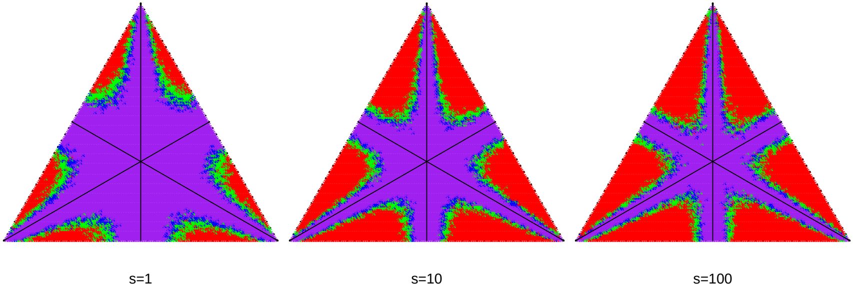
The Problem with Symmetrization (and a Partial Solution)

- After some consideration, the symmetrization procedure

$$h_{\text{sym}}(x_1, \ldots, x_m) := \frac{1}{m!} \sum_{\pi \in \mathcal{S}_m} h(x_{\pi(1)}, \ldots, x_{\pi(m)})$$

  is unsatisfactory because it is not computationally feasible when m is large.
- **Partial random symmetrization:** each time h is evaluated, average over s randomly-chosen permutations to “partially symmetrize” it.
  - → Here, $s \in \mathbb{N}$ is fixed, e.g., $s = 100$.

Partial Symmetrization Works in Practice



Rejection regions obtained using $s = 1,\ 10,$ and 100 random permutations. For all, $m = 15$.

Open problem: are the statistical properties of the SDL test preserved when partial symmetrization is used?

- How many permutations s are sufficient to approximate the fully symmetric kernel?
- And how does s scale with dimension and degree?

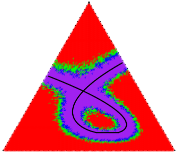
Conclusion

- In this talk, I've focused on two methodological challenges that we faced in implementing this test.
  1. The question of **how the test is affected by the choice of model representation**, and how that can often (but not always) be mitigated by adding redundant constraints.
  2. The difficulty of **constructing a symmetric kernel function**. We resolved this by implementing random, partial symmetrization, but additional theoretical work is necessary.
- Other methodological considerations:
  - How to go about choosing various other user-specified parameters—need to balance the validity, statistical power, as well as the stability of the stochastic p-values.
  - Knowing features of model geometry often enabled us to improve test performance.
- As a general-purpose framework, the SDL method performed remarkably well, and with thoughtful implementation it was able to match performance of traditional deterministic tests, at least for the low-dimensional models that we considered.

Questions?

[1] Cécile Ané. “Reconstructing concordance trees and testing the coalescent model from genome-wide data sets”. In: *Estimating Species Trees: Practical and Theoretical Aspects.* Ed. by Lacey Knowles and Laura Kubatko. Wiley-Blackwell, 2010, pp. 35–36.

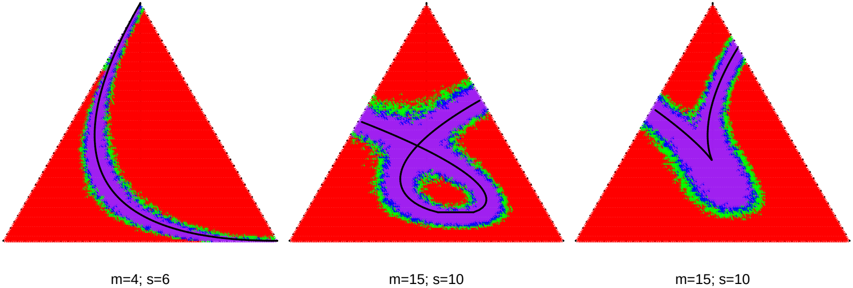
[2] Nils Sturma, Mathias Drton, and Dennis Leung. “Testing many constraints in possibly irregular models using incomplete U-statistics”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* (Mar. 2024), qkae022. ISSN: 1369-7412. DOI: 10.1093/jrsssb/qkae022.



$$(y-1/3)^2-6(x-2/5)^2(x-1/9)=0$$

- Slide 4 Figure: David Savada, et. al, *Principles of Life* (2014)
- Slide 5 Figure: Marina Garrote-López (modified)

Higher degree irreducible models



Rejection regions for SDL tests of (L-R) (a) the Hardy-Weinberg 2-allele model defined by $y^2 - 4xz = 0$, (b) a nodal cubic model defined by $(y - 1/3)^2 - 6(x - 2/5)^2(x - 1/9) = 0$, (c) a cuspidal cubic model, defined by $(y - 1/3)^2 - (x - 1/3)^3 = 0$.