Testing Topologies and Splits

Edward Susko

Department of Mathematics and Statistics
Dalhousie University

Outline

- Focus on Inference for Likelihood Methods and Bootstrapping

1. Topology Testing
   - Formulating the Problem
   - The K-H Test
   - Adjusting for Selection Bias
2. Testing Splits
   - Bootstrap Support for Splits
   - Adjusting for Selection Bias

Outline

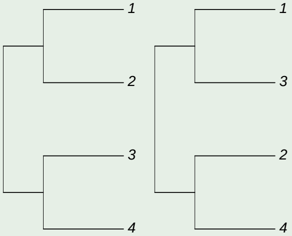
1. Topology Testing
   - Formulating the Problem
   - The K-H Test
   - Adjusting for Selection Bias
2. Testing Splits
   - Bootstrap Support for Splits
   - Adjusting for Selection Bias

Outline

1. Topology Testing
   - Formulating the Problem
   - The K-H Test
   - Adjusting for Selection Bias
2. Testing Splits
   - Bootstrap Support for Splits
   - Adjusting for Selection Bias

Two Tree Problem

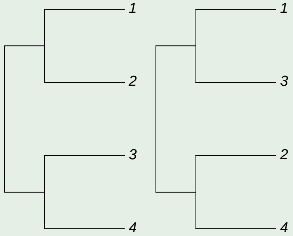
Tree 1 vs Tree 2



- One-sided: Is tree 1 significantly better than tree 2?
- Two-sided: Is there significant evidence for tree 1 or tree 2?

Two Tree Problem

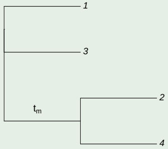
Tree 1 vs Tree 2



- One-sided: Is tree 1 significantly better than tree 2?
- Two-sided: Is there significant evidence for tree 1 or tree 2?
- Two-sided more natural (usually) for a priori trees
- One-sided more frequently reported by software

Null Hypothesis: Tree 2 correct

1. Tree 2 Correct

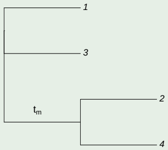


- α-level test should satisfy:

$$P_{T_2}(\text{reject } H_0) \leq \alpha$$

Null Hypothesis: Tree 2 correct

1. Tree 2 Correct



- α-level test should satisfy:

$$P_{T_2}(\text{reject } H_0) \leq \alpha$$

- For almost any test

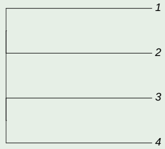
$$P_{T_2}(\text{reject } H_0; t_m = 0.1) < P_{T_2}(\text{reject } H_0; t_m = 0)$$

Need

$$P_{T_2}(\text{reject } H_0; t_m = 0) = \alpha$$

Null Hypothesis: Alternative Argument

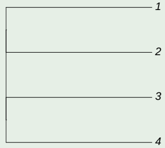
Star tree



- If tree 2 is estimated, we do not reject.
- If not, star tree is the least distant tree from estimated to Tree 2.

Null Hypothesis: Alternative Argument

Star tree



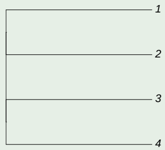
- If tree 2 is estimated, we do not reject.
- If not, star tree is the least distant tree from estimated to Tree 2.

- Need

$$P_{T_2}(\text{reject } H_0; t_m = 0) = \alpha$$

Null Hypothesis: Alternative Argument

Star tree



- If tree 2 is estimated, we do not reject.
- If not, star tree is the least distant tree from estimated to Tree 2.

- Need

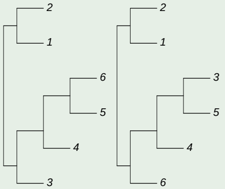
$$P_{T_2}(\text{reject } H_0; t_m = 0) = \alpha$$

- Analogous to testing mean: $H_0 : \mu_2 \leq \mu_1$, $H_A : \mu_2 > \mu_1$, p-values, Type I error evaluated under $H_0 : \mu_2 = \mu_1$

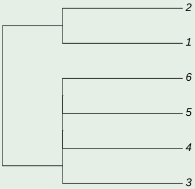
Null Hypothesis - Two Trees

- Collapse as many branches as needed to make the trees equivalent.
- Don’t collapse more.

Tree 1 vs Tree 2



Null Tree



Outline

1. Topology Testing
   - Formulating the Problem
   - The K-H Test
   - Adjusting for Selection Bias
2. Testing Splits
   - Bootstrap Support for Splits
   - Adjusting for Selection Bias

Kishino & Hasegawa (1989)

- Log likelihoods for tree 1: l_1 is like a sample mean $l_1/n = n^{-1}\sum_{i=1}^{n} \log p_{T_1}(x_i; \hat{\boldsymbol{t}}_1) = n^{-1}\sum_{i=1}^{n} l_{1i}$.
- Comparing l_1 and l_2 is like comparing two sample means.

Kishino & Hasegawa (1989)

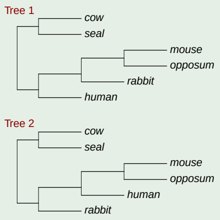
- Log likelihoods for tree 1: l_1 is like a sample mean $l_1/n = n^{-1}\sum_{i=1}^{n} \log p_{T_1}(x_i; \hat{\boldsymbol{t}}_1) = n^{-1}\sum_{i=1}^{n} l_{1i}$.
- Comparing l_1 and l_2 is like comparing two sample means.
- l_{1i} and l_{2i} are dependent: paired on the same observation i.
- Paired z-test adjusts for dependence of l_{1i} and l_{2i}.
- $d_i = l_{1i} - l_{2i}$.

$$z = \frac{\bar{d}}{(s_d/\sqrt{n})}$$

p-value=$P(Z > z)$, $Z \sim N(0,1)$ (One-sided H_A)

K-H Example - Mammalian Mitochondrial Data

Mammal Trees

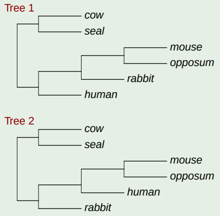


- mtREV24, 8 Gamma rate categories

site i	l_{1i}	l_{2i}	d_i
1	-8.533	-8.556	0.023
2	-3.775	-3.776	-0.001
⋮		⋮	
3414	-14.053	-14.158	0.105
	-21765.04	-21766.23	1.190

K-H Example - Mammalian Mitochondrial Data

Mammal Trees



- mtREV24, 8 Gamma rate categories

site i	l_{1i}	l_{2i}	d_i
1	-8.533	-8.556	0.023
2	-3.775	-3.776	-0.001
$\vdots$		$\vdots$	
3414	-14.053	-14.158	0.105
	-21765.04	-21766.23	1.190

$$z = (1.190/3414) \;\; / \;\; (s_d/\sqrt{3414}) = 0.132$$

One sided p-value = $P(Z > 0.132) = 0.44$

K-H Test Motivation in More Detail

- If $l_{1i} = \log p_{T_1}(x_i; \hat{\boldsymbol{t}}_1)$ are independent and identically distributed,
  CLT $\Rightarrow$ $\bar{d}$ is approximately normal.

K-H Test Motivation in More Detail

- If $l_{1i} = \log p_{T_1}(x_i; \hat{\boldsymbol{t}}_1)$ are independent and identically distributed,
  CLT $\Rightarrow \bar{d}$ is approximately normal.
- Usual models: sites evolve independently
- But sites $1, \ldots, n$ all contribute to $\hat{\boldsymbol{t}}_1$
- So $\log p_{T_1}(x_i; \hat{\boldsymbol{t}}_1)$ are not independent
  whereas $\log p_{T_1}(x_i; \boldsymbol{t}_1)$ are independent

K-H Test Motivation in More Detail

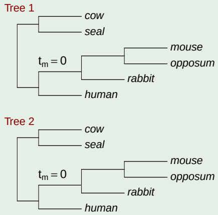
- If $l_{1i} = \log p_{T_1}(x_i; \hat{\boldsymbol{t}}_1)$ are independent and identically distributed,
  CLT $\Rightarrow$ $\bar{d}$ is approximately normal.
- Usual models: sites evolve independently
- But sites $1, \ldots, n$ all contribute to $\hat{\boldsymbol{t}}_1$
- So $\log p_{T_1}(x_i; \hat{\boldsymbol{t}}_1)$ are not independent
  whereas $\log p_{T_1}(x_i; \boldsymbol{t}_1)$ are independent
- Argument by approximation: $\hat{\boldsymbol{t}}_1 \approx \boldsymbol{t}_1$,

$$\begin{aligned} l_1/n &= n^{-1}\sum_{i=1}^{n} \log p_{T_1}(x_i; \hat{\boldsymbol{t}}_1) \\ &\approx n^{-1}\sum_{i=1}^{n} \log p_{T_1}(x_i; \boldsymbol{t}_1) + r_{1n}(\boldsymbol{t}_1) \end{aligned}$$

$r_{1n}(\boldsymbol{t})$ is relatively small.

Null Distribution of dLnL=$l_1 - l_2$

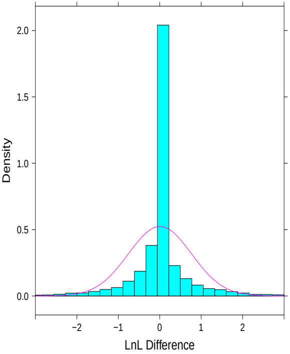
Mammal Trees

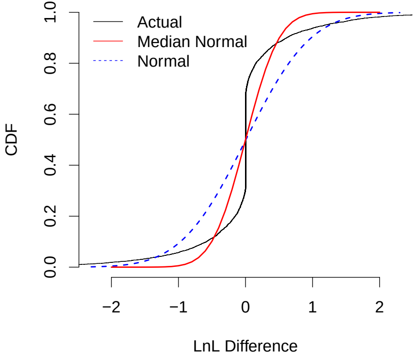


- Simulate 5000 data sets under mtREV24 model
- $\alpha = 0.44$, 8 Gamma categories
- Tree 1, with $t_m = 0$

Null Distribution of dLnL

- mixed continuous and discrete distribution





K-H motivation difficulty

$$l_1/n \approx n^{-1} \sum_{i=1}^{n} \log p_{T_1}(x_i; \boldsymbol{t}_1) + r_{1n}(\boldsymbol{t}_1)$$

and

$$l_2/n \approx n^{-1} \sum_{i=1}^{n} \log p_{T_2}(x_i; \boldsymbol{t}_2) + r_{2n}(\boldsymbol{t}_2)$$

K-H motivation difficulty

$$l_1/n \approx n^{-1} \sum_{i=1}^{n} \log p_{T_1}(x_i; \boldsymbol{t}_1) + r_{1n}(\boldsymbol{t}_1)$$

and

$$l_2/n \approx n^{-1} \sum_{i=1}^{n} \log p_{T_2}(x_i; \boldsymbol{t}_2) + r_{2n}(\boldsymbol{t}_2)$$

but $T_1 = T_2$ under the null so first order terms cancel:

$$l_1/n - l_2/n \approx r_{1n}(\boldsymbol{t}_1) - r_{2n}(\boldsymbol{t}_2)$$

Bootstrapping: Motivation

Setting: $d_1, \ldots, d_n$ independent and identically distributed (P)
Need distribution of $\bar{d} - \mu$.

$$\begin{aligned} \hat{P}(A) &:= \text{Proportion of } d_i \text{ in } A \\ &\approx P(D \text{ in } A) \end{aligned}$$

$\hat{P}$ assigns mass $1/n$ to each observed d_i (empirical distribution)

Bootstrapping: Motivation

Setting: $d_1, \ldots, d_n$ independent and identically distributed (P)
Need distribution of $\bar{d} - \mu$.

$$\begin{aligned}\hat{P}(A) &:= \text{Proportion of } d_i \text{ in } A \\ &\approx P(D \text{ in } A)\end{aligned}$$

$\hat{P}$ assigns mass $1/n$ to each observed d_i (empirical distribution)

$$E_{\hat{P}}[d] = \sum_{d_i} p(d_i) d_i = \bar{d}$$

Bootstrapping: Motivation

Setting: $d_1, \ldots, d_n$ independent and identically distributed (P)
Need distribution of $\bar{d} - \mu$.

$$\begin{aligned}\hat{P}(A) &:= \text{Proportion of } d_i \text{ in } A \\ &\approx P(D \text{ in } A)\end{aligned}$$

$\hat{P}$ assigns mass $1/n$ to each observed d_i (empirical distribution)

$$E_{\hat{P}}[d] = \sum_{d_i} p(d_i) d_i = \bar{d}$$

Suggests: If $d_1^*, \ldots, d_n^*$ are generated from $\hat{P}$

$$\text{distribution of } \bar{d}^* - \bar{d} \approx \text{distribution of } \bar{d} - \mu$$

Bootstrapping

- Select sites with replacement.
  eg. $i_1 = 32, i_2 = 32, \ldots, i_n = 3$
  Then $d_1^* = d_{32}, \ldots, d_n^* = d_3$ give a sample from $\hat{P}$.

$$\Rightarrow \bar{d}^* - \bar{d} \text{ gives a realization from } \hat{P}$$

Bootstrapping

- Select sites with replacement.
  eg. $i_1 = 32, i_2 = 32, \ldots, i_n = 3$
  Then $d_1^* = d_{32}, \ldots, d_n^* = d_3$ give a sample from $\hat{P}$.

$$\Rightarrow \bar{d}^* - \bar{d} \text{ gives a realization from } \hat{P}$$

- Repeat a large number (B) of times

$$\text{Proportion of } \bar{d}^* - \bar{d} \leq x \approx \hat{P}(\bar{d}^* - \bar{d} \leq x) \approx P(\bar{d} - \mu \leq x)$$

KH test - RELL Version - Kishino, Miyata & Hasegawa

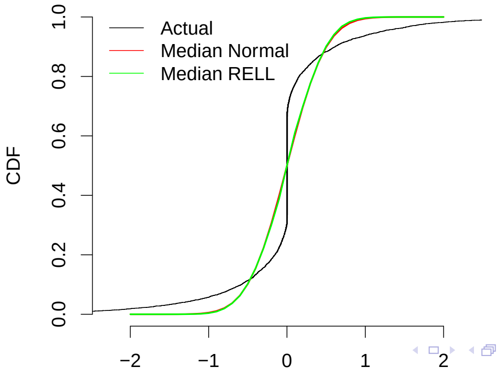
- $\bar{d}$ still used as a test statistic
- $N(0, s_d^2/\sqrt{n})$ is replaced by bootstrap distribution of $\bar{d}^* - \bar{d}$

KH test - RELL Version - Kishino, Miyata & Hasegawa

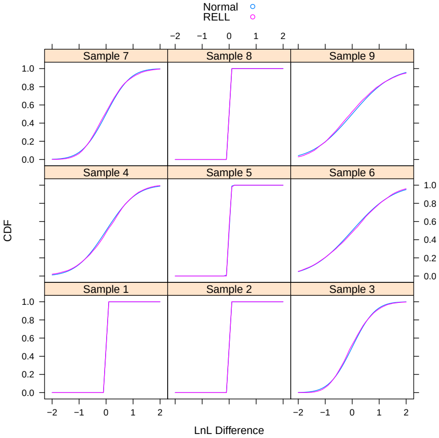
- $\bar{d}$ still used as a test statistic
- $N(0, s_d^2/\sqrt{n})$ is replaced by bootstrap distribution of $\bar{d}^* - \bar{d}$
- minor adjustment: $\bar{d}$ replaced by $\text{ave}_b \bar{d}^*$

Null Distribution of dLnL

- parameter settings from mammal data: mtREV24, $\alpha = 0.4$, $n = 3415$
5000 simulated data sets. B=5000.



Normal vs RELL resampling



- For a given data set, $d_1^*, \ldots, d_n^*$ i.i.d. from $\hat{P}$ (fixed) CLT $\Rightarrow \bar{d}^* - \bar{d} \sim N(0, s_d^2/\sqrt{n})$.
- Main source of variation: $\hat{t}_m = 0$ implies point mass at 0.

Full Bootstrapping - KH setting

- Bootstrapping so far has been RELL
- Bootstrap principle: Bootstrapping should mimic what is done with original data.

Full Bootstrapping - KH setting

- Bootstrapping so far has been RELL
- Bootstrap principle: Bootstrapping should mimic what is done with original data.
- Original data: Estimate $\hat{\boldsymbol{t}}_1$ & $\hat{\boldsymbol{t}}_2$ from $x_1, \ldots x_n$.

$$l_1 - l_2 = l_1(\hat{\boldsymbol{t}}_1) - l_2\hat{\boldsymbol{t}}_2)$$

Full Bootstrapping - KH setting

- Bootstrapping so far has been RELL
- Bootstrap principle: Bootstrapping should mimic what is done with original data.
- Original data: Estimate $\hat{\boldsymbol{t}}_1$ & $\hat{\boldsymbol{t}}_2$ from $x_1, \ldots x_n$.

$$l_1 - l_2 = l_1(\hat{\boldsymbol{t}}_1) - l_2\hat{\boldsymbol{t}}_2)$$

- Bootstrap principle: Estimate $\hat{\boldsymbol{t}}_1^*$ & $\hat{\boldsymbol{t}}_2^*$ from $x_1^*, \ldots x_n^*$.

$$l_1^* - l_2^* = l_1^*(\hat{\boldsymbol{t}}_1^*) - l_2^*(\hat{\boldsymbol{t}}_2^*)$$

Full Bootstrapping - KH setting

- Bootstrapping so far has been RELL
- Bootstrap principle: Bootstrapping should mimic what is done with original data.
- Original data: Estimate $\hat{\boldsymbol{t}}_1$ & $\hat{\boldsymbol{t}}_2$ from $x_1, \ldots x_n$.

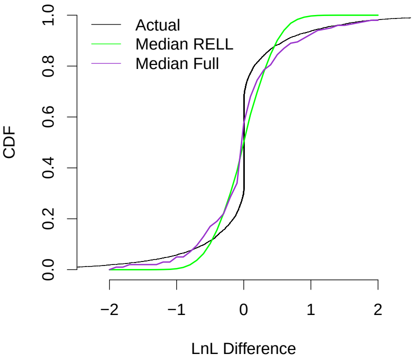
$$l_1 - l_2 = l_1(\hat{\boldsymbol{t}}_1) - l_2\hat{\boldsymbol{t}}_2)$$

- Bootstrap principle: Estimate $\hat{\boldsymbol{t}}_1^*$ & $\hat{\boldsymbol{t}}_2^*$ from $x_1^*, \ldots x_n^*$.

$$l_1^* - l_2^* = l_1^*(\hat{\boldsymbol{t}}_1^*) - l_2^*(\hat{\boldsymbol{t}}_2^*)$$

- By contrast, RELL uses $l_1^*(\hat{\boldsymbol{t}}_1) - l_2^*(\hat{\boldsymbol{t}}_2)$

Full Bootstrapping - Mammal Example

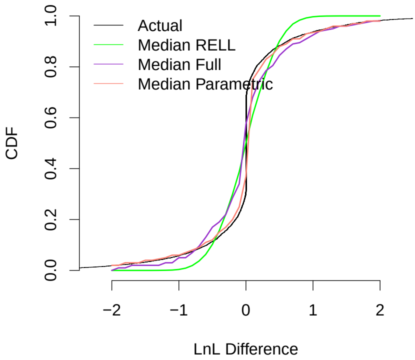


- B=100 and 100 simulations for full

Parametric Bootstrapping

- Generate from $\hat{P}_{\hat{\theta}}$ instead of $\hat{P}$. eg. mtREV24 on estimated Tree 2, α
  Generate from $\hat{P}_{\hat{\theta}}$
  eg. mtREV24 on ML tree, $\hat{\alpha} = 0.4$

KH with parametric boostrapping - Mammal Example



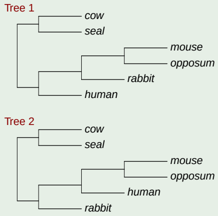
- $B = 100$ and 100 simulations for parametric

Outline

1. Topology Testing
   - Formulating the Problem
   - The K-H Test
   - Adjusting for Selection Bias
2. Testing Splits
   - Bootstrap Support for Splits
   - Adjusting for Selection Bias

SH Test - Shimodaira & Hasegawa (1999)

Mammal Trees

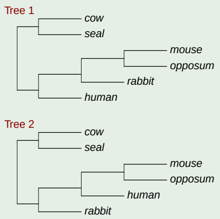


- T_1 and T_2 fixed a priori:

  Q_{KH} : T_1 significantly better than T_2?

SH Test - Shimodaira & Hasegawa (1999)

Mammal Trees



- T_1 and T_2 fixed a priori:

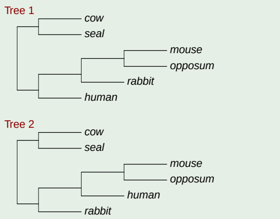
  Q_{KH} : T_1 significantly better than T_2?

- If instead, only T_2 is fixed a priori,

  Q_{SH} : ML tree significantly better than T_2?

SH Test - Shimodaira & Hasegawa (1999)

Mammal Trees



- T_1 and T_2 fixed a priori:

  Q_{KH} : T_1 significantly better than T_2?

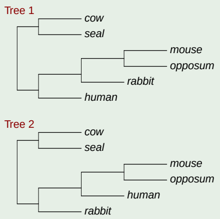
- If instead, only T_2 is fixed a priori,

  Q_{SH} : ML tree significantly better than T_2?

- $P(l_1 - l_2 > 0) < 1$
- $P(l_{MLE} - l_2 > 0) = 1.$

SH Test - Shimodaira & Hasegawa (1999)

Mammal Trees



- T_1 and T_2 fixed a priori:

  Q_{KH} : T_1 significantly better than T_2?

- If instead, only T_2 is fixed a priori,

  Q_{SH} : ML tree significantly better than T_2?

- $P(l_1 - l_2 > 0) < 1$
- $P(l_{MLE} - l_2 > 0) = 1$.
- Paradox: tree 1 could be both a fixed tree of interest and ML tree.

SH Adjustment to Bootstrap

- **Setting**: T_1 and T_2 become $T_1, \ldots, T_M \Rightarrow l_1, \ldots, l_M$
    - Mammal data. 6 taxa $\Rightarrow M = 105$ trees.
- Test statistic $l_1 - l_2$ replaced by $l_m - l_1$
  m indice of MLE.

SH Adjustment to Bootstrap

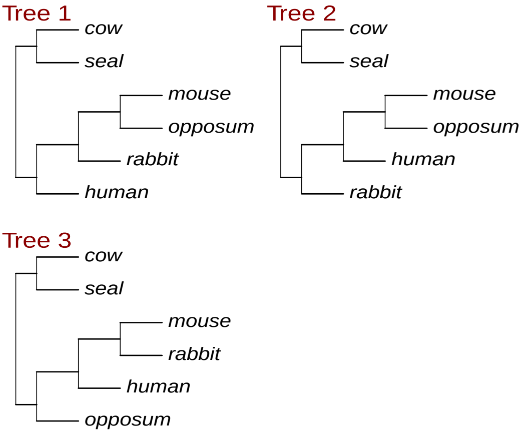
- **Setting**: T_1 and T_2 become $T_1, \ldots, T_M \Rightarrow l_1, \ldots, l_M$
    - Mammal data. 6 taxa $\Rightarrow M = 105$ trees.
- Test statistic $l_1 - l_2$ replaced by $l_m - l_1$

  m indice of MLE.
- **Bootstrapping**
    - Replace $l_1^*, \ldots, l_M^*$ by

      $l_1^* - \text{ave}_b l_1^*, \ldots, l_M^* - \text{ave}_b l_M^*$
    - Use observed $l_{m^*}^* - l_2^*$ from bootstrapping for null distribution.

      m^*: indice of MLE for bootstrap sample.

Mammal Data Example - Three trees



- Tree 1 was the ML tree for this data

Mammal Data Example - Three trees

- $B = 5000$, $M = 3$
- $l_1 - l_2 = 1.19$

Mammal Data Example - Three trees

- $B = 5000$, $M = 3$
- $l_1 - l_2 = 1.19$
- l_j^* (after centering), first three bootstrap samples

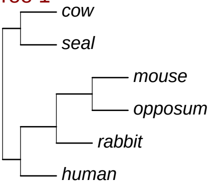
l_1^*	l_2^*	l_3^*	m^*	$l_1^* - l_2^*$	$l_{m^*} - l_2^*$
-359.78	-360.62	-352.52	3	0.84	8.10
-84.45	-94.44	-95.87	1	9.99	9.99
-65.93	-58.62	-62.19	2	-7.31	0.00

$$pKH = \text{proportion of } l_1^* - l_2^* > 1.19 = 0.44$$

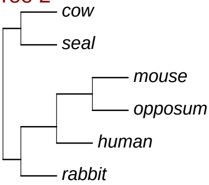
$$pSH = \text{proportion of } l_{m^*}^* - l_2^* > 1.19 = 0.59$$

Mammal Data Example - Four trees

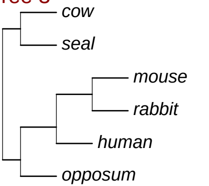
Tree 1



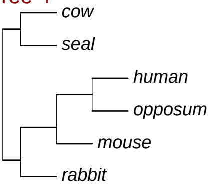
Tree 2



Tree 3



Tree 4



Mammal Data Example - Four trees

- $B = 5000$, $M = 4$
- $l_1 - l_2 = 1.19$

$$pKH = \text{proportion of } l_1^* - l_2^* > 1.19 = 0.44$$

$$pSH_3 = \text{proportion of } l_{m_3^*}^* - l_2^* > 1.19 = 0.59$$

$$pSH_4 = \text{proportion of } l_{m_4^*}^* - l_2^* > 1.19 = 0.74$$

SH test - Choice of Trees

- The larger M is, the larger p_{SH} is.

SH test - Choice of Trees

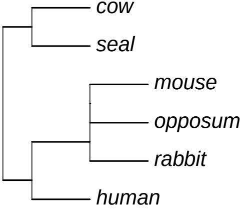
- The larger M is, the larger p_{SH} is.
- Because of the SH centering procedure, H_0 depends on M:

$$H_0 : \mu_1 = \cdots \mu_M$$

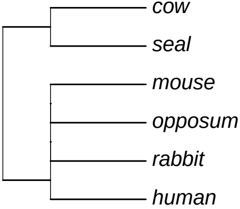
  μ_i - mean log likelihood ith tree
- H_0 is only possible if edge-length set to 0 to make all trees same

Mammal Data Example - Null trees

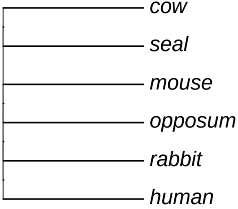
Trees 1 and 2



Cow, Seal Constraint



All 105 trees



SH test - Choice of Trees

- With large n, even when ML is over 105 trees
  - $l_m - l_2$ is effectively maximum of three $l_i - l_2$ when one $t = 0$ in generating tree.
  - $l_m - l_2$ is effectively maximum of 105 $l_i - l_2$ under star tree
- Much more likely to see large $l_m - l_2$ for star tree $\Rightarrow$ harder to reject a tree

SH test - Choice of Trees

- With large n, even when ML is over 105 trees
  - $l_m - l_2$ is effectively maximum of three $l_i - l_2$ when one $t = 0$ in generating tree.
  - $l_m - l_2$ is effectively maximum of 105 $l_i - l_2$ under star tree
- Much more likely to see large $l_m - l_2$ for star tree $\Rightarrow$ harder to reject a tree
- Bootstrap Principle: Bootstrapping should mimic what is being done with original data.
- If exhaustive search for ML tree, $M = 105$

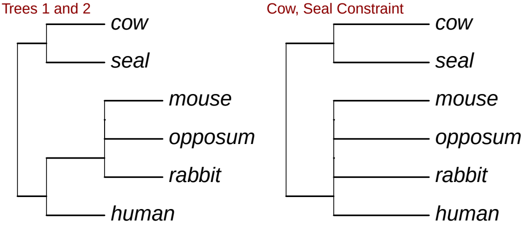
SOWH Test - Goldman, Anderson and Rodrigo (2000)

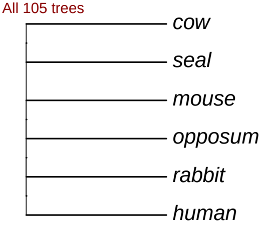
- SH test with following modifications:
  - Full parametric bootstrapping from tree under H_0 instead of RELL to get l_i^*
  - No centering.
    - SH replaces l_i^* with $l_i^* - \text{ave}_b l_i^*$.
    - SOWH does not
  - $p_{SOWH} << p_{SH}$

SOWH Test - Goldman, Anderson and Rodrigo (2000)

- SH test with following modifications:
    - Full parametric bootstrapping from tree under H_0 instead of RELL to get l_i^*
    - No centering.
        - SH replaces l_i^* with $l_i^* - \text{ave}_b l_i^*$.
        - SOWH does not
    - $p_{SOWH} << p_{SH}$
    - Sometimes, SOWH will generate from a fully-resolved trees
    - But, under null, it will tend to give trees for bootstrapping that are close to true.

Mammal Data Example - Null trees





Concluding Remarks - Testing Topologies

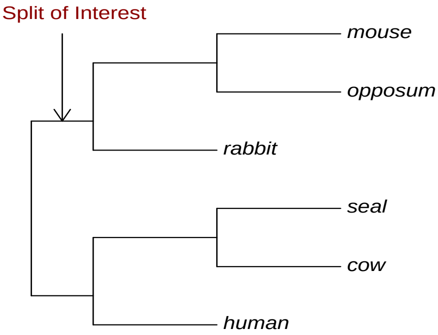
- Two fixed trees a priori
  - K-H test and variations
  - RELL vs Full:
    RELL is fast, Full is accurate
- Adjusting for Selection Bias
  - SOWH/SH
  - Choice of Null is major performance issue for SH
  - SH is fast, SOWH is accurate

Outline

1. Topology Testing
   - Formulating the Problem
   - The K-H Test
   - Adjusting for Selection Bias
2. Testing Splits
   - Bootstrap Support for Splits
   - Adjusting for Selection Bias

Formulating the Problem

- Tree inference: Is Tree 1 correct?



- Is a tree with opposum, mouse and rabbit split from human, cow and seal correct?

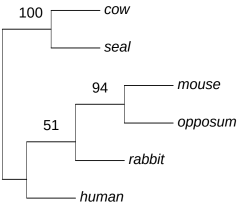
Outline

1. Topology Testing
   - Formulating the Problem
   - The K-H Test
   - Adjusting for Selection Bias

2. Testing Splits
   - Bootstrap Support for Splits
   - Adjusting for Selection Bias

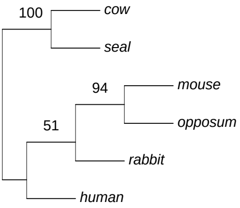
Bootstrap Support

- For each bootstrap sample $x_1^*, \ldots, x_n^*$ obtain $\hat{T}^*$
- BP for opposum, mouse and rabbit = proportion of T^* with that split.



Bootstrap Support

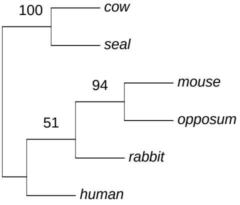
- For each bootstrap sample $x_1^*, \ldots, x_n^*$ obtain $\hat{T}^*$
- BP for opposum, mouse and rabbit = proportion of T^* with that split.



- Can be applied to any estimation procedure
- By far the most frequent measure of uncertainty

Bootstrap Support

- For each bootstrap sample $x_1^*, \ldots, x_n^*$ obtain $\hat{T}^*$
- BP for opposum, mouse and rabbit = proportion of T^* with that split.



- Can be applied to any estimation procedure
- By far the most frequent measure of uncertainty
- How large of BP is large?

History

- Felsenstein (1985): Bootstrap Support (BP) introduced

History

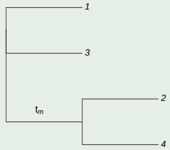
- Felsenstein (1985): Bootstrap Support (BP) introduced
- Hillis and Bull (1993): BP is probability split is correct. 70% is large.
- Felsenstein and Kishino (1993): 1-BP is p-value for hypothesis that split is not present. 95% is large.

History

- Felsenstein (1985): Bootstrap Support (BP) introduced
- Hillis and Bull (1993): BP is probability split is correct. 70% is large.
- Felsenstein and Kishino (1993): 1-BP is p-value for hypothesis that split is not present. 95% is large.
- Efron, Halloran and Holmes (1996) [EHH] and Efron and Tibshirani (1998) [ET]: 1-BP is first order correct.

Null Hypothesis

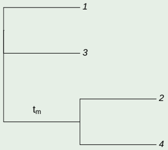
H_0



- H_0 : Split 12|34 not present

Null Hypothesis

H_0



- H_0 : Split 12|34 not present
- For almost any test,

$$P_{H_0}(\text{reject } H_0; t_m = 0.1) < P_{H_0}(\text{reject } H_0; t_m = 0)$$

To guarantee

$$P_{H_0}(\text{reject } H_0) \leq \alpha$$

need

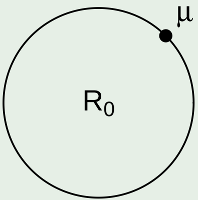
$$P_{H_0}(\text{reject } H_0; t_m = 0) = \alpha$$

P-Value Interpretation

- A valid p-value should have a uniform distribution under H_0
- 1-BP has uniform distribution with $t_m = 0$
- **First Order Correctness**: BP has a uniform limiting distribution in this setting

Problem of Regions: Normal Form

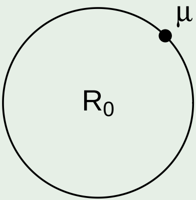
Example Region



- Setting: $\bar{\boldsymbol{Y}} \sim N(\boldsymbol{\mu}, n^{-1}\Sigma)$
- H_0 : $\boldsymbol{\mu}$ not in R_0 (on boundary).

Problem of Regions: Normal Form

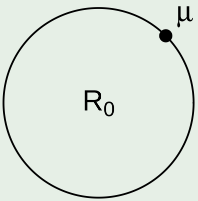
Example Region



- Setting: $\bar{\boldsymbol{Y}} \sim N(\boldsymbol{\mu}, n^{-1}\Sigma)$
- $H_0 : \boldsymbol{\mu}$ not in R_0 (on boundary).
- Parametric Bootstrap: Generate $\bar{\boldsymbol{Y}}^* \sim N(\bar{\boldsymbol{Y}}, n^{-1}\Sigma)$.
- Asymptotically equivalent to nonparametric bootstrap

Problem of Regions: Normal Form

Example Region



- Setting: $\bar{\boldsymbol{Y}} \sim N(\boldsymbol{\mu}, n^{-1}\Sigma)$
- $H_0 : \boldsymbol{\mu}$ not in R_0 (on boundary).
- Parametric Bootstrap: Generate $\bar{\boldsymbol{Y}}^* \sim N(\bar{\boldsymbol{Y}}, n^{-1}\Sigma)$.
- Asymptotically equivalent to nonparametric bootstrap
- BP is percentage of time $\bar{\boldsymbol{Y}}^* \in R_0$.

Problem of Regions: Normal Form

- Reparameterize: $\boldsymbol{Z} = n^{1/2}(\bar{\boldsymbol{Y}} - \boldsymbol{\mu})$

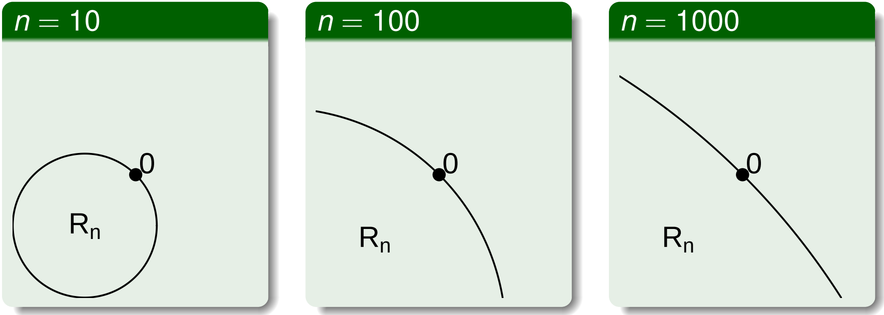
Problem of Regions: Normal Form

- Reparameterize: $\boldsymbol{Z} = n^{1/2}(\bar{\boldsymbol{Y}} - \boldsymbol{\mu})$
- Setting: $\boldsymbol{Z} \sim N(\mathbf{0}, \Sigma)$
- Bootstrap: Generate $\boldsymbol{Z}^* \sim N(\boldsymbol{Z}, \Sigma)$.

Problem of Regions: Normal Form

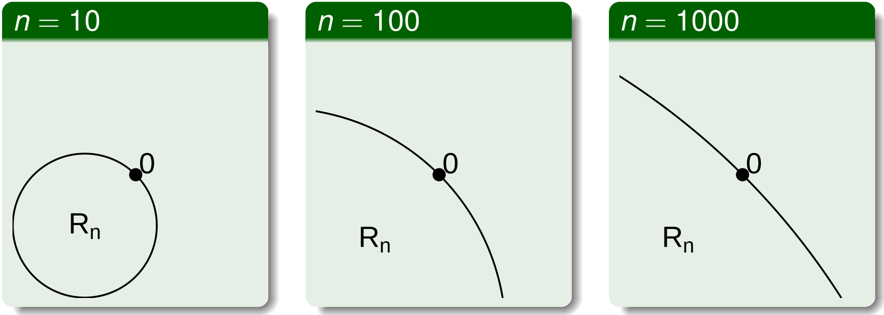
- Reparameterize: $\boldsymbol{Z} = n^{1/2}(\bar{\boldsymbol{Y}} - \boldsymbol{\mu})$
- Setting: $\boldsymbol{Z} \sim N(\mathbf{0}, \Sigma)$
- Bootstrap: Generate $\boldsymbol{Z}^* \sim N(\boldsymbol{Z}, \Sigma)$.
- $R_n = \{n^{1/2}(\boldsymbol{x} - \boldsymbol{\mu}) : \boldsymbol{x} \in R_0\}$
- BP is percentage of time $\boldsymbol{Z}^* \in R_n$.

Problem of Regions: Normal Form



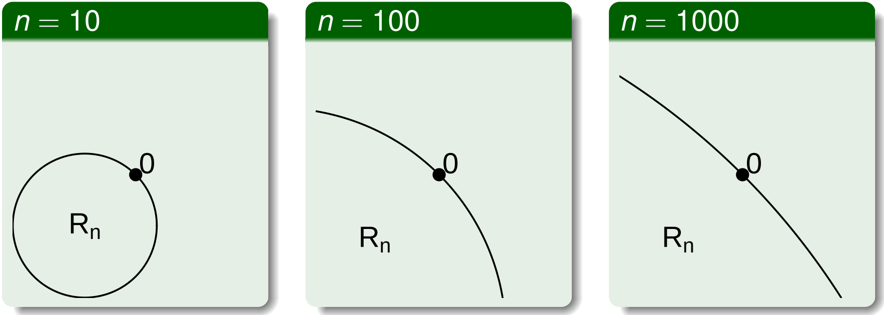
- Asymptotic Setting: $\boldsymbol{Z} \sim N(\mathbf{0}, \Sigma)$
- Bootstrap: Generate $\boldsymbol{Z}^* \sim N(\boldsymbol{Z}, \Sigma)$.

Problem of Regions: Normal Form



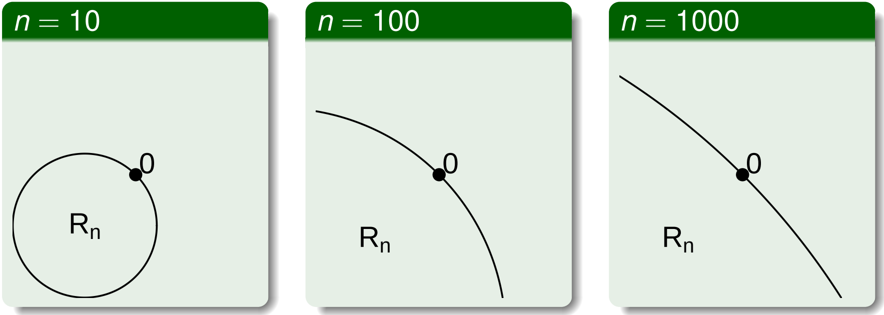
- Asymptotic Setting: $\boldsymbol{Z} \sim N(\mathbf{0}, \Sigma)$
- Bootstrap: Generate $\boldsymbol{Z}^* \sim N(\boldsymbol{Z}, \Sigma)$.
- BP is percentage of time $\boldsymbol{Z}^*$ in a half-space

Problem of Regions: Normal Form



- Asymptotic Setting: $\boldsymbol{Z} \sim N(\mathbf{0}, \Sigma)$
- Bootstrap: Generate $\boldsymbol{Z}^* \sim N(\boldsymbol{Z}, \Sigma)$.
- BP is percentage of time $\boldsymbol{Z}^*$ in a half-space
- ET result: BP is uniformly distributed

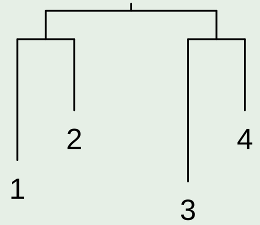
Problem of Regions: Normal Form



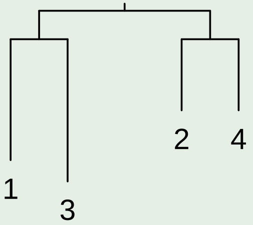
- Asymptotic Setting: $\boldsymbol{Z} \sim N(\mathbf{0}, \Sigma)$
- Bootstrap: Generate $\boldsymbol{Z}^* \sim N(\boldsymbol{Z}, \Sigma)$.
- BP is percentage of time $\boldsymbol{Z}^*$ in a half-space
- ET result: BP is uniformly distributed
- Smooth boundary needed for half-space approximation

Problem of regions: Alternate Version

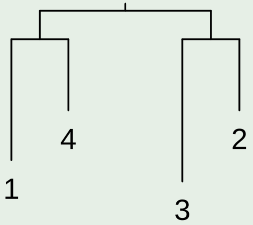
Tree 1



Tree 2



Tree 3



Notation: Middle edge-length estimated alone

- $l_j(t)$ log likelihood for the jth topology; t middle edge-length.
- $I_j = E[-l_j''(0)]/n$
- $V_{jn} = I_j^{-1/2} l_j'(0)/\sqrt{n}$ (standardized score)

Asymptotic Approximations

- $l_j(\hat{t}) \approx l_j(0) + V_{jn}^2 I\{V_{jn} > 0\}/2$

Asymptotic Approximations

- $l_j(\hat{t}) \approx l_j(0) + V_{jn}^2 I\{V_{jn} > 0\}/2$
- $\boldsymbol{V}_n = [V_{1n}, V_{2n}, V_{3n}] \sim N(\mathbf{0}, \Sigma)$

Asymptotic Approximations

- $l_j(\hat{t}) \approx l_j(0) + V_{jn}^2 I\{V_{jn} > 0\}/2$
- $\boldsymbol{V}_n = [V_{1n}, V_{2n}, V_{3n}] \sim N(\mathbf{0}, \Sigma)$
- $l_j(0)$: likelihood under star tree $\Rightarrow$ independent of j.
- Topology j preferred to k if $V_{jn} > 0$ and $V_{jn} > V_{kn}$.
- R_0 is $\boldsymbol{v}$-space where split 12|34 estimated:

$$\{\boldsymbol{v} : v_1 > 0, v_1 > v_2, v_1 > v_3\}$$

Asymptotic Approximations

- $l_j(\hat{t}) \approx l_j(0) + V_{jn}^2 I\{V_{jn} > 0\}/2$
- $\boldsymbol{V}_n = [V_{1n}, V_{2n}, V_{3n}] \sim N(\mathbf{0}, \Sigma)$
- $l_j(0)$: likelihood under star tree $\Rightarrow$ independent of j.
- Topology j preferred to k if $V_{jn} > 0$ and $V_{jn} > V_{kn}$.
- R_0 is $\boldsymbol{v}$-space where split 12|34 estimated:

$$\{\boldsymbol{v} : v_1 > 0, v_1 > v_2, v_1 > v_3\}$$

- For trees with split 12|34: $\mu = E[\boldsymbol{V}_n]$ is in R_0

Asymptotic Approximations

- $l_j^*(\hat{t}) \approx l_j^*(0) + V_{jn}^{*2} I\{V_{jn}^{*2} > 0\}/2$ where V_{jn}^* is standardized score for the bootstrap sample
- Approximate Bootstrap: Generate $\boldsymbol{V}_n^* \sim N(\boldsymbol{V}_n, \Sigma)$

Problem of regions: 3-D version

- Setting: $\boldsymbol{V} \sim N(\boldsymbol{\mu}, \Sigma)$
- Bootstrap: Generate $\boldsymbol{V}^* \sim N(\boldsymbol{V}, \Sigma)$
- $H_0 : \boldsymbol{\mu} = 0$ on boundary of $R_0 = \{\boldsymbol{\mu} : \mu_1 > 0, \mu_1 > \mu_2, \mu_1 > \mu_3\}$
- BP is proportion of $\boldsymbol{V}^*$ in R_0.

Problem of regions: 3-D version

- Setting: $\boldsymbol{V} \sim N(\boldsymbol{\mu}, \Sigma)$
- Bootstrap: Generate $\boldsymbol{V}^* \sim N(\boldsymbol{V}, \Sigma)$
- $H_0 : \boldsymbol{\mu} = 0$ on boundary of
  $R_0 = \{\boldsymbol{\mu} : \mu_1 > 0, \mu_1 > \mu_2, \mu_1 > \mu_3\}$
- BP is proportion of $\boldsymbol{V}^*$ in R_0.
- A half space, $R_0' = \{\boldsymbol{\mu} : \mu_1 > \mu_2\}$, contains the region R_0.

Problem of regions: 3-D version

- Setting: $\boldsymbol{V} \sim N(\boldsymbol{\mu}, \Sigma)$
- Bootstrap: Generate $\boldsymbol{V}^* \sim N(\boldsymbol{V}, \Sigma)$
- $H_0 : \boldsymbol{\mu} = 0$ on boundary of
  $R_0 = \{\boldsymbol{\mu} : \mu_1 > 0, \mu_1 > \mu_2, \mu_1 > \mu_3\}$
- BP is proportion of $\boldsymbol{V}^*$ in R_0.
- A half space, $R_0' = \{\boldsymbol{\mu} : \mu_1 > \mu_2\}$, contains the region R_0.
- BP for R_0' is larger than BP for R_0

Problem of regions: 3-D version

- Setting: $\boldsymbol{V} \sim N(\boldsymbol{\mu}, \Sigma)$
- Bootstrap: Generate $\boldsymbol{V}^* \sim N(\boldsymbol{V}, \Sigma)$
- $H_0 : \boldsymbol{\mu} = 0$ on boundary of
  $R_0 = \{\boldsymbol{\mu} : \mu_1 > 0, \mu_1 > \mu_2, \mu_1 > \mu_3\}$
- BP is proportion of $\boldsymbol{V}^*$ in R_0.
- A half space, $R_0' = \{\boldsymbol{\mu} : \mu_1 > \mu_2\}$, contains the region R_0.
- BP for R_0' is larger than BP for R_0
- Problem of regions theory: BP for R_0' is uniform

Problem of regions: 3-D version

- Setting: $\boldsymbol{V} \sim N(\boldsymbol{\mu}, \Sigma)$
- Bootstrap: Generate $\boldsymbol{V}^* \sim N(\boldsymbol{V}, \Sigma)$
- $H_0 : \boldsymbol{\mu} = 0$ on boundary of
  $R_0 = \{\boldsymbol{\mu} : \mu_1 > 0, \mu_1 > \mu_2, \mu_1 > \mu_3\}$
- BP is proportion of $\boldsymbol{V}^*$ in R_0.
- A half space, $R_0' = \{\boldsymbol{\mu} : \mu_1 > \mu_2\}$, contains the region R_0.
- BP for R_0' is larger than BP for R_0
- Problem of regions theory: BP for R_0' is uniform
- BP is stochastically smaller than uniform
- Under H_0, BP larger than 95% less than 5% of the time.

Simulation to obtain BP distribution

- Generate V from $N(0, \Sigma) \sim$ Generate alignment
- $X_j = V_j^2 I\{V_j > 0\}/2 \sim l_j(\hat{t}_j) - l_j(0)$
- Generate $V^* \sim N(0, \Sigma) \sim$ Generate bootstrap alignment

Simulation to obtain BP distribution

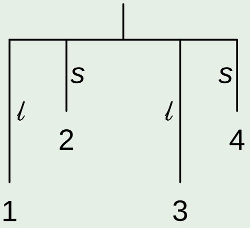
- Generate V from $N(0,\Sigma) \sim$ Generate alignment
- $X_j = V_j^2 I\{V_j > 0\}/2 \sim l_j(\hat{t}_j) - l_j(0)$
- Generate $V^* \sim N(0,\Sigma) \sim$ Generate bootstrap alignment

.

1. Repeatedly generate $V \sim N(0,\Sigma)$,
    1. For each V, repeatedly generate V^* from $N(V,\Sigma)$
    2. Set $BP =$ proportion of V^* in R_0
2. $P(BP > x) \approx$ proportion of $BP > x$

Example

Generating Tree



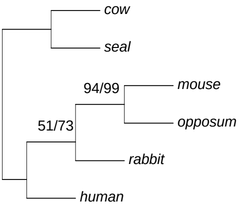
- Tree 1: 12|34 LB-apart

LB-apart tree: $P(BP > t)$

		t	
s	l	0.70	0.90
0.01	0.01	0.13	0.03
0.01	0.50	0.14	0.03
0.10	0.10	0.13	0.03
0.10	0.50	0.13	0.03
0.50	0.50	0.12	0.02
0.50	1.00	0.13	0.03
1.00	1.00	0.12	0.02
1.00	1.50	0.12	0.02

Adjusted Bootstrap Support

- F =CDF of BP can be obtained through fast normal simulation
- Define aBP as $F(BP)$
- Then $P(aBP > 0.95) = 0.05$

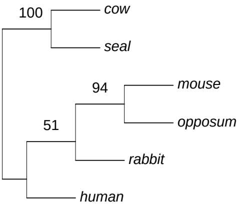


Outline

1. Topology Testing
   - Formulating the Problem
   - The K-H Test
   - Adjusting for Selection Bias
2. Testing Splits
   - Bootstrap Support for Splits
   - Adjusting for Selection Bias

Selection Bias and Splits

- In many? cases *BP* is on a priori hypothesized trees of interest
- Frequently BP is on the ML tree



- Even if H_0 is true, highly unlikely that $BP \approx 10\%$, otherwise it wouldn't be ML tree.

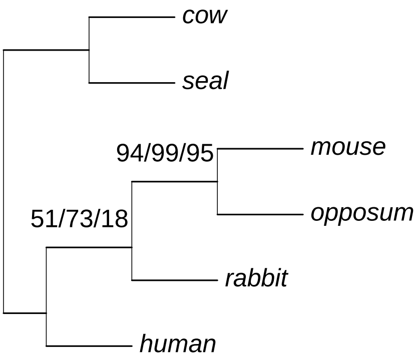
aBP adjusting for selection bias

Simulation for usual distribution of BP:

1. Repeatedly generate $V \sim N(0, \Sigma)$,
    1. For each V, repeatedly generate V^* from $N(V, \Sigma)$
    2. Set $BP =$ proportion of V^* in R_0

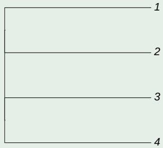
- Generate V from $N(0, \Sigma) \sim$ Generate alignment
- $X_j = V_j^2 I\{V_j > 0\}/2 \sim l_j(\hat{t}_j) - l_j(0)$
- Generate $V^* \sim N(0, \Sigma) \sim$ Generate bootstrap alignment
- Adjustment: Only consider cases where largest X_j is at $j = 1$ (12|34)

aBP adjusting for selection bias - mammal example

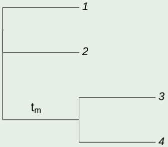


Likelihood Ratio Test for splits

H_0



H_A

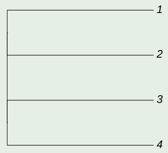


- $$2\{l_j(\hat{t}_m) - l_j(0)\} \approx V_j^2 I\{V_j \geq 0\}$$

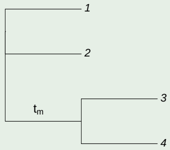
  $$V_j \sim N(0,1).$$

Likelihood Ratio Test for splits

H_0



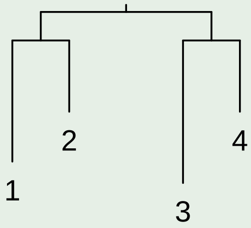
H_A



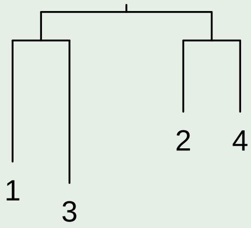
- $$2\{l_j(\hat{t}_m) - l_j(0)\} \approx V_j^2 I\{V_j \geq 0\}$$
  $$V_j \sim N(0,1).$$
- $$2\{l(\hat{t}_m) - l(0)\} \sim \frac{1}{2}\delta_0 + \frac{1}{2}\chi_1^2$$
- 1/2 of the time $\hat{t}_m = 0$
  otherwise
  usual behaviour.

alrt - Anisimova & Gascuel (2006)

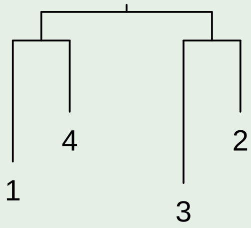
Tree 1



Tree 2



Tree 3



- If ML tree, then $2\{l_{ML}(\hat{t}) - l_1(0)\}$ is being used in place of

$$X_1 := 2\{l_1(\hat{t}) - l_1(0)\}$$

- Alternatively $T = \max\{X_1, X_2, X_3\}$

Distribution of T

- We know
  - $T = \max\{X_1, X_2, X_3\}$
  - $X_i \sim \frac{1}{2}\delta_0 + \frac{1}{2}\chi^2_1$
- Don't know dependence structure of X_1, X_2 and X_3.

Distribution of T

- We know
  - $T = \max\{X_1, X_2, X_3\}$
  - $X_i \sim \frac{1}{2}\delta_0 + \frac{1}{2}\chi_1^2$
- Don't know dependence structure of X_1, X_2 and X_3.
- Bonferroni correction ($t > 0$):

$$\begin{aligned}\text{true p-value} &= P(\max\{X_1, X_2, X_3\} > t)\\ &= P(X_1 > t \text{ or } X_2 > t \text{ or } X_3 > t)\\ &\leq P(X_1 > t) + P(X_2 > t) + P(X_3 > t)\\ &= \frac{3}{2}P(\chi_1^2 > t) = p_T\end{aligned}$$

Distribution of T

- We know
  - $T = \max\{X_1, X_2, X_3\}$
  - $X_i \sim \frac{1}{2}\delta_0 + \frac{1}{2}\chi_1^2$
- Don't know dependence structure of X_1, X_2 and X_3.
- Bonferroni correction ($t > 0$):

$$\begin{aligned}\text{true p-value} &= P(\max\{X_1, X_2, X_3\} > t) \\ &= P(X_1 > t \text{ or } X_2 > t \text{ or } X_3 > t) \\ &\leq P(X_1 > t) + P(X_2 > t) + P(X_3 > t) \\ &= \frac{3}{2}P(\chi_1^2 > t) = p_T\end{aligned}$$

- Conservative p-value

$$p_T \geq \text{true p-value}$$

$\Rightarrow P(\text{Type I error}) \leq \alpha$

alrt T'

- Order X_1, X_2, X_3 as $X_{(1)} < X_{(2)} < X_{(3)}$
- alrt replaces $T = \max\{X_1, X_2, X_3\}$ with $T' = T - X_{(2)}$
- Still uses

$$\text{alrt p-value} = \frac{3}{2} P(\chi^2_1 > t')$$

alrt T'

- Order X_1, X_2, X_3 as $X_{(1)} < X_{(2)} < X_{(3)}$
- alrt replaces $T = \max\{X_1, X_2, X_3\}$ with $T' = T - X_{(2)}$
- Still uses

$$\text{alrt p-value} = \frac{3}{2} P(\chi^2_1 > t')$$

- Since $t' \leq t$, conservative p-value:

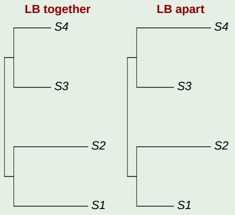
$$\begin{aligned} \text{alrt p-value} &= \frac{3}{2} P(\chi^2_1 > t') \\ &\geq \frac{3}{2} P(\chi^2_1 > t) \geq \text{true p-value} \end{aligned}$$

alrt with correction

- $X_i =_d V_i^2 I\{V_i \geq 0\}$, $V \sim N(0, \Sigma)$
- So quick simulation approximation to T' is possible
  1. Generate $V_1, \ldots, V_B \sim N(0, \Sigma)$
  2. Calculate $X_{bi} = V_{bi}^2 I\{V_{bi} \geq 0\}$
  3. $T'_b = \max\{X_{b1}, X_{b2}, X_{b3}\} - X_{b(2)}$
- p-value = proportion of $T'_b \geq t$

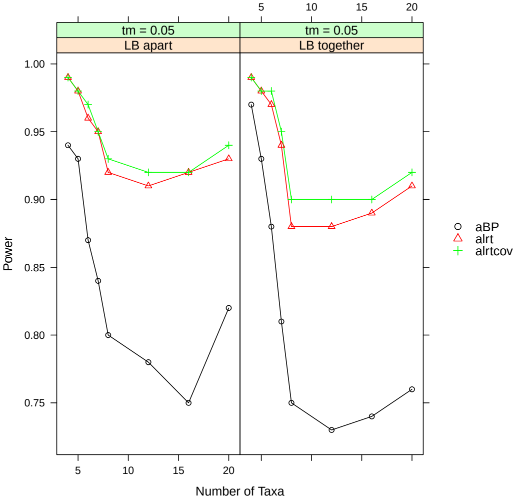
aLRT Simulation

Generating Trees



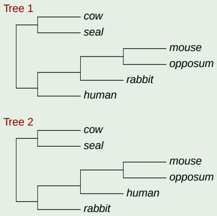
- $S_1, \ldots, S_4$ caterpillar trees with distance 0.15 between nodes
- Long branch 0.2, short 0.1
- Middle branch $t_m = 0.05$
- HKY, $\kappa = 4.5$, $\alpha = 1$
- A, C, G, T: 0.18, 0.24, 0.32, 0.26

alrt Simulation Results



Adjusting the K-H test

Mammal Trees



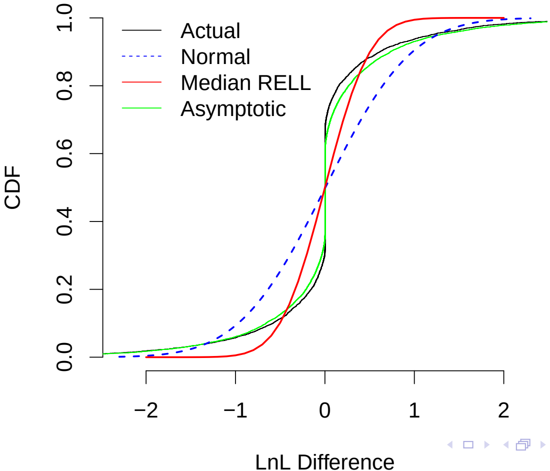
- Single Split Difference between Trees
- Asymptotic Theory indicates

$$l_1 - l_2 \approx V_1^2 I\{V_1 > 0\} - V_2^2 I\{V_2 > 0\}$$

$$\boldsymbol{V} \sim N(\mathbf{0}, \Sigma).$$

- Can be used via normal simulation to obtain p-value

Asymptotic K-H distribution - Mammal Example



Concluding Remarks - Inference for Splits

- aBP, $(1 - alrt) \times 100$ more interpretable that BP
- Selection Bias: useful to report both w/ and w/o adjustment.
- Asymptotic theory can be useful

Acknowledgements

Andrew Roger
Matt Spencer

SEB group at Dalhousie