May 12, 2017

Evolution of traits along the tree of life: statistical contrasts and their estimation error

Qing (Sabrina) Yu[1]

under the supervision of
Professor Cécile Ané[2]

[1] *Departments of Statistics and Computer Sciences, University of Wisconsin, Madison, Wisconsin, 53706, USA;*

[2] *Departments of Statistics and Botany, University of Wisconsin, Madison, Wisconsin, 53706, USA*

Abstract

Phylogenetic contrast is used to explore how different organisms have adapted to the environment and changed to the current state. It has good statistical characteristics and is able to transform species' trait values (correlated) to independent identically distributed values. Those values can be analyzed with linear regression model and used for bootstrapping.

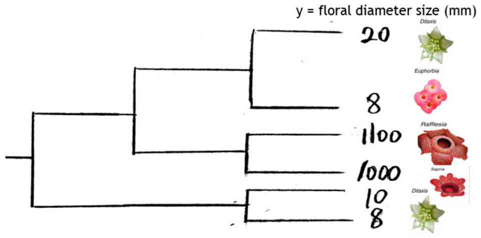
The main goal of the project is to explore the distribution of contrasts and see if estimated parameters affect the behavior of contrasts and contrasts are under or over-estimated if the shift configuration is estimated with error.

Keywords: Phylogenetic; contrasts; bootstrap; Ornstein-Uhlenbeck process

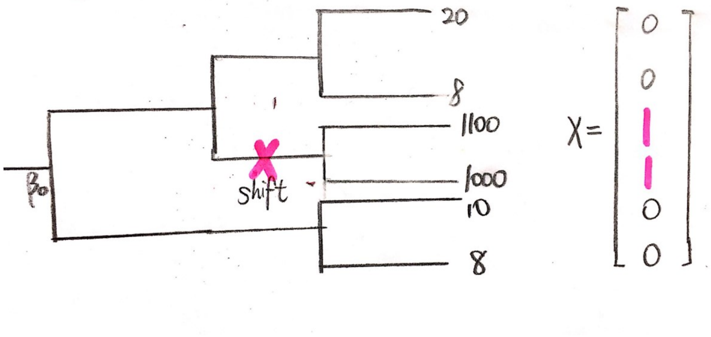
1 Background of phylogenetic tree and contrasts

Phylogenetic (genealogical) tree is a diagram showing the evolutionary relationships among different species. A shift indicates a place on the phylogenetic tree where large morphological changes occurred in the past.

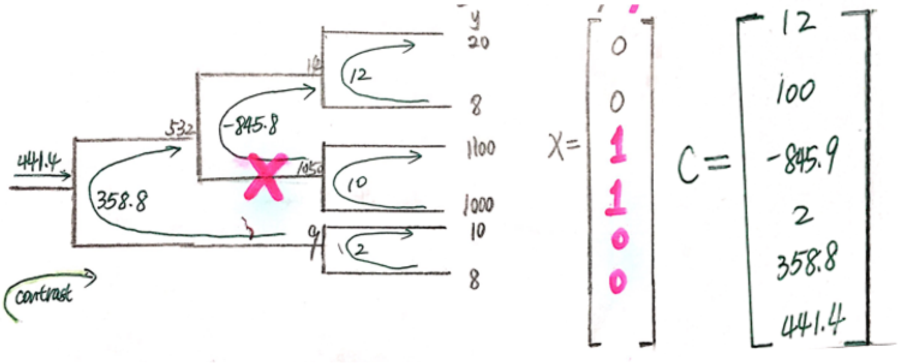
The plot used below describes sizes of flowers, which have some significant variations of flower sizes. Those flowers are current species whose trait values clearly indicate that there may be some shifts in the past which cause those variations.



(a) artificial example inspired by paper of Davis



(b) linear regression model



(c) contrasts of internal nodes

The pink label represents a shift which causes the significant variations of flower sizes. So, X with values of 1 indicate two present species with relatively large morphological changes. Others are zero because they do not have significant changes from their ancestors. It is worthwhile to mention here that Y represent values of tips. However, contrasts shown above are the dynamic difference of bifurcate edges corresponding to internal nodes, which is shown by the arrow.

The reason I am interested in studying contrasts is that there are many statistical methods which require independent data points. One thing is for linear regression, which requires each data point to be independent. We can plot the linear regression on the contrast values instead of trait values which are correlated. The other use is for bootstrapping which used to get confidence intervals. We need to resample the contrast values to build bootstrap data sets.

We can use Ornstein-Uhlenbeck process to model trait values (y) over time (t).

$$dy_t = -\alpha(y_t - \mu) + \sigma dB_t$$

In this formula, dy_t is the changes of y values, α and σ are constants across the whole tree. μ is constant among each edge and dB_t is the variation from standard Brownian Motion. Since we know dy_t and root values, then we will know all trait values.

We can use a linear model to fit trait values at leaves and measure shifts, which is:

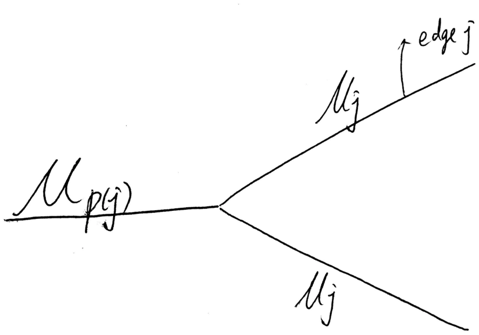
$$Y = \beta_0 \mathbb{1} + X^{(\alpha)}\beta + \varepsilon$$

where

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 V_a)$$

Y is flower size. β_0 is overall mean of flower size and β is the size changes from the shift. $\mathbb{1}$ is a vector of ones.

In addition, β values are shift differences. From the picture below, β_j=shift differences $= \mu_j - \mu_{p(j)}$ which p(j) is the parent of edge j.



X matrix has rows corresponding to taxa and columns corresponding to branches.

$$X_{ib}^{(\alpha)} = \begin{cases} 1 - e^{-\alpha a_b} & \text{if } b \text{ is on the path from the root to taxon } i \\ 0 & \text{if taxon } i \text{ is not a descendant of } b \end{cases}$$

as a_b to be the age of b's parent node.

Covariance matrix derived from Ornstein-Uhlenbeck process is shown below where t_{ij} is the evolutionary time from the root to the common ancestor of species i and j and d_{ij} is their distance on the tree (Khabbazian, 2016)

$$\Sigma_{ij}^{(\alpha)} = \begin{cases} \sigma^2 e^{-\alpha d_{ij}}(1 - e^{-2\alpha t_{ij}})/(2\alpha) & \text{if the root value is fixed} \\ \sigma^2 e^{-\alpha d_{ij}}/(2\alpha) & \text{if the root value has the stationary distribution} \end{cases}$$

We can calculate the contrasts based on:

$$C = {V_\alpha}^{-1/2}(y - \mu)$$

is estimated by

$$\widehat{C} = \widehat{V}_{\widehat{\alpha}}^{-1/2}(y - \widehat{\mu})$$

which $\widehat{\mu}$ depends on $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2 \cdots$. V_α determined by the tree and closely related species have high V_α term. It has rows which are tips and columns which are internal nodes. In addition,

$$\mu_0 = \beta_0 = e^{-\alpha t} y_0 + (1 - e^{-\alpha t})\mu_0$$

if we assume

$$\mu_0 = \beta_0 = y_0$$

V_a also has a root state as a parameter and alpha is an adaption rate. When alpha is too large, there is no correlation between tips. We used ${V_\alpha}^{-1/2}$ to calculate contrasts. There are many ways to calculate ${V_\alpha}^{-1/2}$ However, Khabbazian used one way to calculate ${V_\alpha}^{-1/2}$ which they implemented a linear-time algorithm (Stone 2011) and please refer to the paper of Mohammad Khabbazian.

Methods

1.1 Data Simulation

Lizard data used for this project are from ι1ou package. I repeat the simulation of y for 100,000 times on a tree with 100 species, 8 shifts. Then, I estimate unknown parameters (β, α, and/or shifts). The next step is to calculate square root of covariance matrix using 'sqrtOUcovariance' function from ι1ou package and further calculate contrasts.

ι1ou : R package to estimate where the shifts are (shift configuration), shift values (β), and correlation parameter (α), and variance parameter(σ^2).

We want to explore the distribution of contrasts calculated under four scenarios.

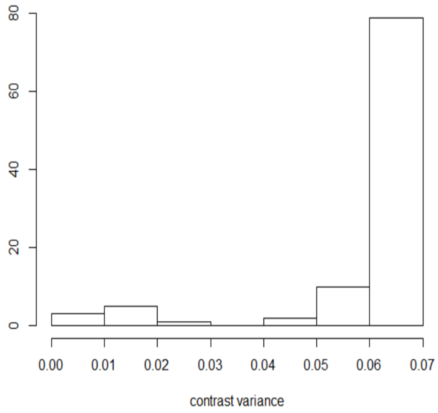
1.2 Scenario 1

For Scenario 1, all parameters are known.

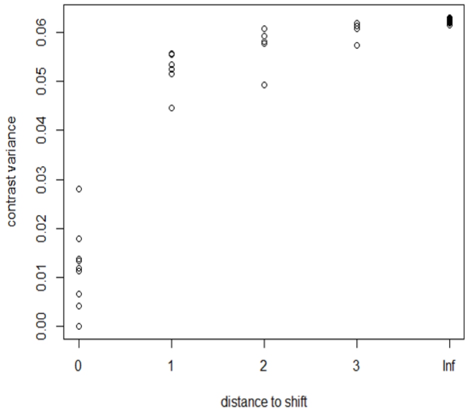
I expected contrasts to be centered with mean 0, variance σ^2. At first, they were not the same as what I expected, so I discovered bugs(in appendix) in the ι1ou package and helped fix them. After that, the contrast means were 0 and the contrast variances were σ^2.

1.3 Scenario 2

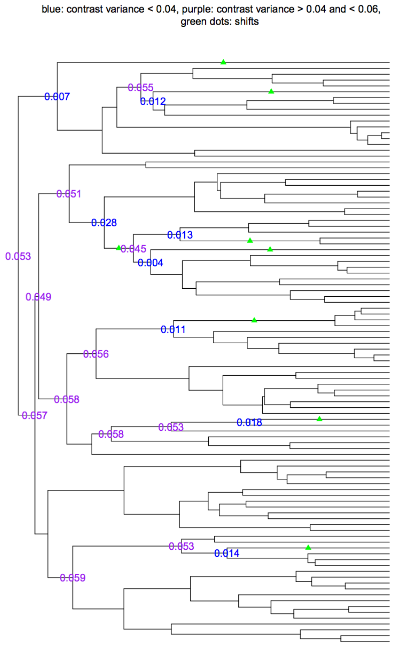
For Scenario 2, α and shifts known, but we estimate β (shift values) (2 plots)



(d)



(e)



(f)

Plot (d) shows most of contrast variances are close to the true σ^2, around 0.0628. However, there are still quite a few contrasts with small variances below 0.04. Then I plot (f) to see where those nodes with small contrast variances are located. It is clear to see those nodes with small contrast variances are exactly before each shift.

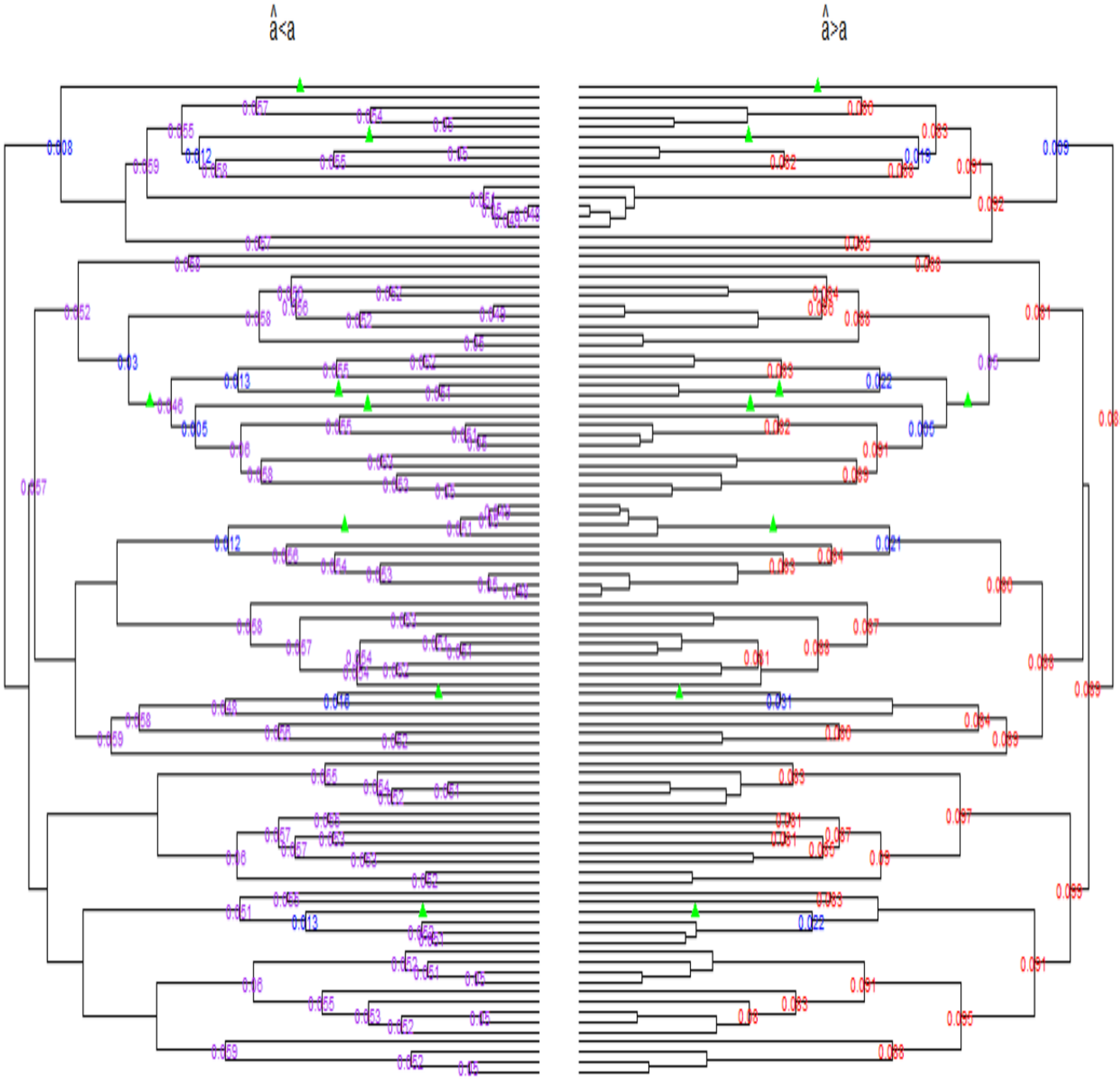
Key observation 1: root contrast is exactly zero. It may be because $\widehat{\mu}$ is too close to trait values (y) and make variance of root edge to be zero.

Key observation 2: Contrast variance at nodes before shifts are underestimated, which show strong bias just before the shift, less bias at nodes most distant from a shift. As it is shown from the last graph, the contrast variances are getting larger as their distances to a shift is further away.

1.4 Scenario 3

For scenario 3, shifts are known, but we estimate α and β (shift values)

Conclusion of Scenario 3: Besides inheriting same conclusion from Scenario 2, Scenario 3 clearly shows that contrast variances near tips are underestimated if α is underestimated (first plot) and contrast variances near the root are over-estimated when α is over-estimated. These two plots reveal patterns of contrasts if we take the subset of whole data based on underestimated or overestimated α values. In the real world, true α value is unknown so all simulations are grouped together and two effects described above compensate each other and those nodes may not show unusual patterns as it is shown below.

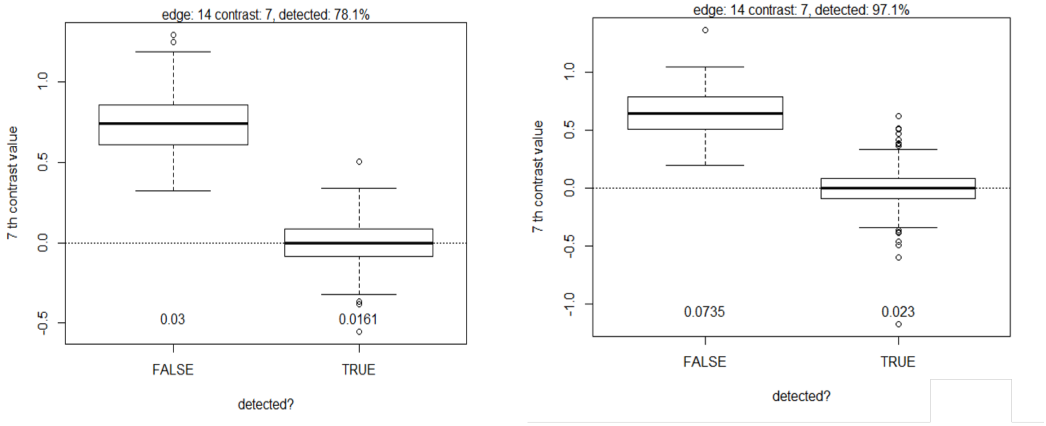


(g) blue: contrast variance < 0.04,purple: contrast variance > 0.04 and < 0.06, green dots: shift

(h) red: contrast variance > 0.08

1.5 Scenario 4

For Scenario 4, all parameters are estimated: β, α and shift configuration. Two criteria can be used to estimate the number of shifts: AICc (liberal) or pBIC (phylogenetic BIC: conservative). This is related to my previous research project which is "the procedure for accurate bootstrap support that measures shifts on a genealogical tree". I will describe this in the later section. The point is from my previous report, I further prove that AICc is more liberal, meaning that AICc can help detect all shifts but have many false positives. pBIC cannot detect all shifts but it can guarantee the correctness of shifts being detected. So, from the two plots we can see if we use AICc, the true shift has higher chance being detected. However, it is clear to see from the two plots that the contrast mean is biased when the shift is undetected. The contrast mean should be all zero whether shift has been detected or not. This needs further investigation.

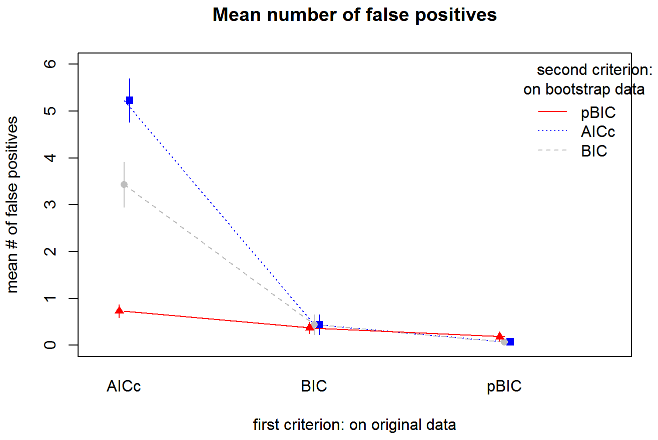


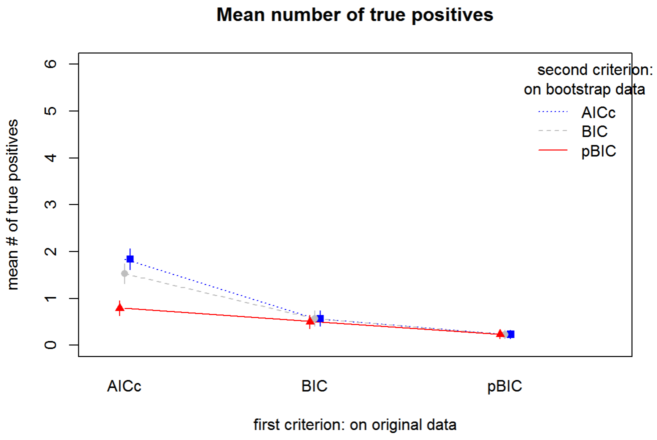
(i) pBIC

(j) AICc

Previous Research

The goal of my previous project is to find the procedure that gives "most accurate" support values for shifts. I first estimate the shift configuration on a dataset based on one of 3 criteria: "AICc","BIC","pBIC" (first criterion). Then, this model is used by the bootstrap procedure to simulate bootstrap data. Each bootstrap data is analyzed with AICc, BIC or pBIC (second criterion). There are nine combinations of criteria (first + second). I compared the bootstrap support values from these nine procedures, on simulated data described below. I found that by using AIC + pBIC, I got some of the 8 shifts with high bootstrap support and rare high support for wrong edges (edges where no shifts were simulated). AICc + pBIC gave high support at an acceptable rate Other procedures gave high support as well but they gave low support for the true shifts. So, it seems that AICc+pBIC is the "most accurate" procedure in a conservative way. Two plots shown below are false-positives and true-positives of shifts calculated under nine combinations. We define a false positive as a non-shift edge that had the bootstrap support of 0.4 or higher (high support). A true positive is a shift edge with the bootstrap support of 0.4 or higher (low support).





We want to find the "most accurate" procedure to ensure small number of false-positives and large number of true-positives. And combining the two plots above, we can tell AICc+pBIC is a relatively good choice. When AICc is the first criterion, then the second criterion can change the mean number of false positives and true positives significantly: with AICc as the second criterion again, the mean number of false positive and of true positives is larger than with BIC or pBIC. When BIC or pBIC is the first criterion, the second criterion has little impact on the number of false/true positives. There are fewer true positives with pBIC than with BIC (as first criterion), for a similar number of false positives.

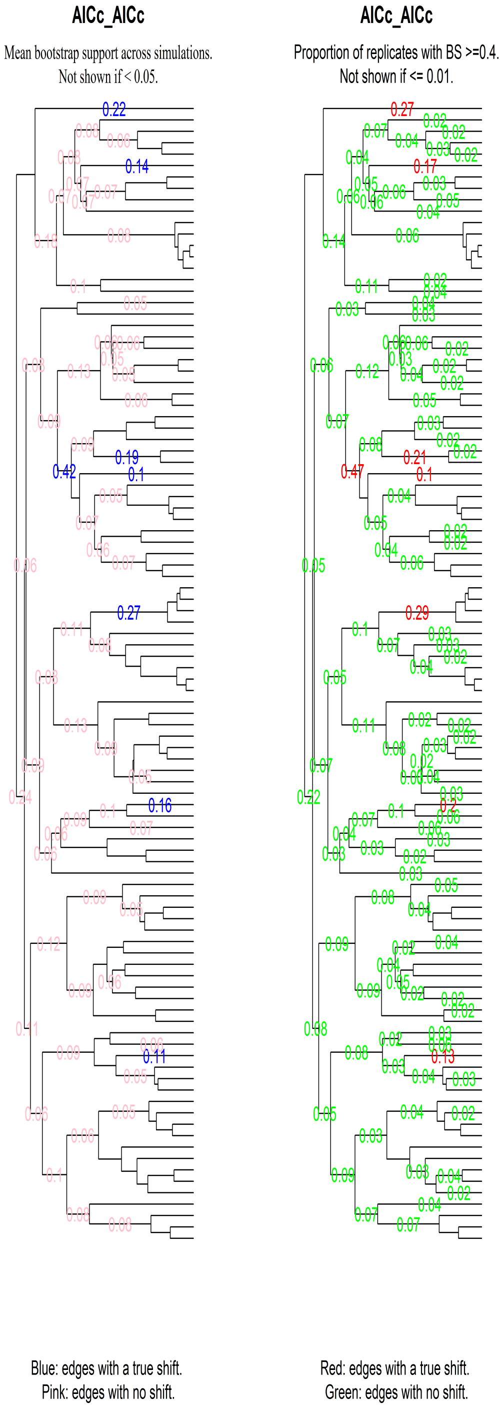
Conclusion: For the combination AICc+AICc, there are many false detection of shifts close to true shifts. Those wrong detections are hard for scientists to differentiate and can be misleading. On the contrary, shifts detected by AICc+pBIC are not all shifts (8 shifts in total). However, all of the detected edges are true shifts.

Conclusion

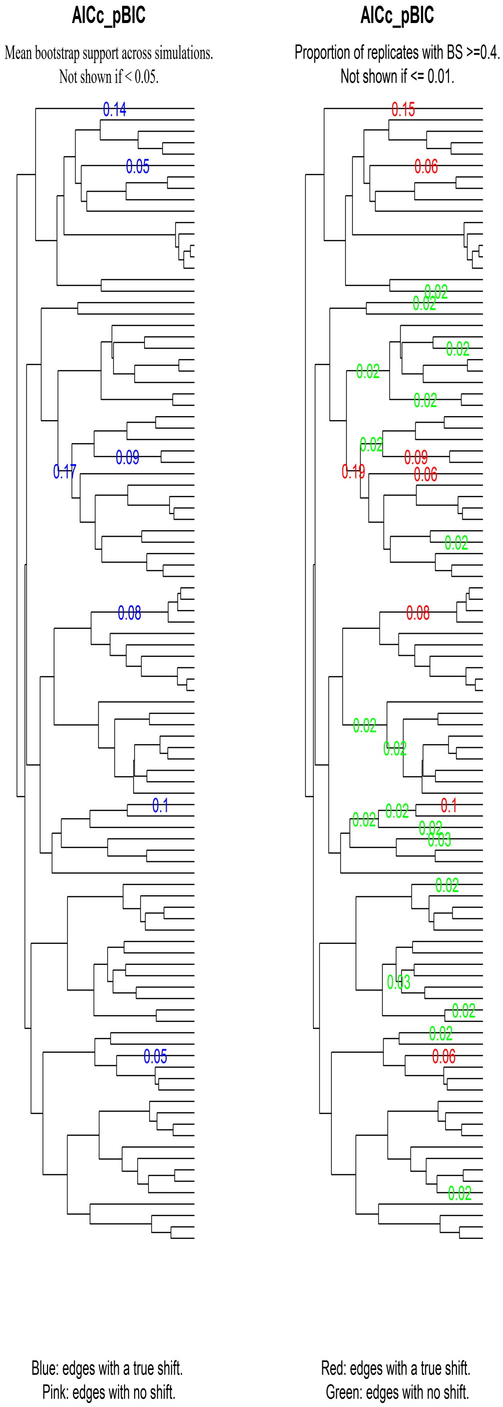
I have spent one year and a half on this research project. While investigating statistical modeling and analyzing results, I discovered and helped fix bugs in ℓ1ou R package and I learned about robust software package development which is beneficial for my future career.

I mentioned quite a few things on the patterns I recognized for contrasts under different scenarios but I am still looking for mathematics formulas to prove they are correct. For example, there should be some mathematics work to explain why the root contrasts are zero from scenario 2 to 4 and why contrasts get over or under estimated if they are close to the root or close to the tips.

My work informs necessary changes to the semi-parametric bootstrap resampling, to get bootstrap support for shifts, we need to exclude the root contrast, which are zeros. Then, we need to exclude or remove the bias from nodes before shifts. Lastly, it is better to use AICc on the original data to detect all true shifts, even if that causes false shifts, to reduce bias in the mean.


AICc_AICc
Mean bootstrap support across simulations.
Not shown if < 0.05.
0.22
0.14
0.42
0.19
0.1
0.27
0.16
0.11
Blue: edges with a true shift.
Pink: edges with no shift.
AICc_AICc
Proportion of replicates with BS >=0.4.
Not shown if <= 0.01.
0.27
0.17
0.47
0.21
0.1
0.29
0.2
0.13
Red: edges with a true shift.
Green: edges with no shift.

(k)


AICc_pBIC
Mean bootstrap support across simulations.
Not shown if < 0.05.
0.14
0.05
0.17
0.09
0.08
0.1
0.05
Blue: edges with a true shift.
Pink: edges with no shift.
AICc_pBIC
Proportion of replicates with BS >=0.4.
Not shown if <= 0.01.
0.15
0.06
0.19
0.09
0.06
0.08
0.1
0.06
Red: edges with a true shift.
Green: edges with no shift.

(l)

Acknowledgement

I want to express my great thanks to Professor Cécile Ané and Mohammad Khabbazian who spent huge amount of time guiding me on this research project from the very beginning. Especially when I found out a bug related to incorrect contrast mean and variance but had no experience looking for the bugs, Professor Cécile Ané helped me analyze all potential problems that could be wrong and instructed me how to write convincing reports to be able to efficiently contact with people remotely. Also, I want to thank for Trewartha Family who financially support me to work on this research. This research experience is both challenging and rewarding. Again, I need to thanks for all the people mentioned above for their great contribution to this project.

Reference Page

Khabbazian et al. ”Fast And Accurate Detection Of Evolutionary Shifts In Ornstein-Uhlenbeck Models”. Methods in Ecology and Evolution 7.7 (2016): 811-824.

Ho and Ané 2014. Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. Methods in Ecology and Evolution 5:11331-146.

Rabosky et al. 2014. BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. Methods in Ecology and Evolution 5:701-707.

Stone, E.A. (2011) Why the phylogenetic regression appears robust to tree misspecification. Systematic Biology, 60, 245?260.

Tibshirani and Taylor 2011. The solution path of the generalized lasso. Ann. Statist. 39:1335-1371.

Uyeda et al. 2015. Comparative analysis of principal components can be misleading. Systematic biology 64:677-689.

Appendix

Timeline of bugs I have helped fix: Please refer to my Github page for all the details

2016-07-28 Bug: There were many false-positives close to root and shift edges. Also, the mean of some contrasts were not zero.

Reason for the bug: the bug was due to mean μ that helped to centralize the contrasts was wrongly calculated in the function ”estimate-shift-configuration”. Instead, μ should be the optimal value at the tips instead of the expected value. In our situation, we also considered that the α we used in this model was not or even close to the true alpha.

The two non-shift edges that had significantly high bootstrap support were not detected as shift edges, meaning the bug was due to the incorrect alpha value.

2016-08-25 Fixed the bug. From version 1.24 to 1.25

2016-09-01 Bug: we set θ (the optimal value) to be one-fourth of the original value. The error happened when l1ou chose a model with only one shift as the best model. With original θ, it never happened to choose a model with one shift. But it became possible when we reduced the θ values.

Reason for the bug: if we selected the S columns in which contained many selected columns we needed from the matrix M, we could not guarantee S was not empty in this case. So, we replaced M[, S] by as.matrix(M[, S]) to avoid the error thrown because of the empty set S.

2016-09-02 I helped find a bug in package 'l1ou' version 1.27. There was a bug due to the function 'normalized-tree'. For the tree has a root edge, the root edge was not used in the function. In addition, the distance from the root to all tips did not include the length of the root.edge. Due to this bug, the square root of the covariance matrix generated by the two models ('OUrandomRoot' and 'OUfixedRoot') are the same except for the last column which is the covariance of the root with other edges.

2016-09-22 Fixed the bug. From version 1.27 to 1.28. Now, the bug has been fixed and the square root of the covariance matrix generated by the two models are different. See the full report as a webpage, please go to report

2016-11 Fixed the bug. From version 1.28. to 1.40: Reason for the bug: intercept correctly handled after noise-whitening (results may change for variables with a mean far from 0) bug fix in the function calculating the square-root (and inverse) of the phylogenetic covariance matrix.