# hOUwie Notebook

James Boyko

2021-04-15

## Section 1: Introducing hOUwie

### 1.1: Motivation

Many evolutionary questions involve addressing how rates of character evolution change through time either agnostically (Venditti et al. 2006; Harmon et al. 2010; Eastman et al. 2011) or dependent on a particular explanatory variable (O'Meara et al. 2006; May and Moore 2020). Typically, these questions follow a structure which posits that the rate of a continuous character is dependent on state of a discrete character. For example, it could be that the beak length of a Galapagos island finch is dependent on the presence or absence of a drought (Grant and Grant 2006), or that the genome size of a particular plant lineage depends on whether they are a short-lived herbaceous organism or a long-lived woody organism (Beaulieu et al. 2012). In these examples and many others, the evolution of discrete and continuous traits are not independent from each other and are therefore subject to the biases of existing methodology. hOUwie seeks to model the joint evolution of a continuous character with evolution of the discrete variable. In this way, we are able to pose questions that are more specific to the particular biology of the system and are able to provide users a means to apply their biological expertise in a statistically rigorous framework.

### 1.2 Goals

Our goal is to identify the principal determinants of model-selection power and parameter-estimation accuracy in hOUwie. In the following, we first describe and explain the model. Next, we outline the design of our simulation study. We then present results of model-selection power (probability of choosing the correct model) and parameter-estimation accuracy (expected error in estimated parameters). Finally, the inclusion of hidden states adds utility and complexity. With hidden states we gain access to a type of null model, where rate variation in the continuous character can be independent of observed state and is instead dependent on the hidden state. Hidden states also allow for a more complicated model structure.

## Section 2: hOUwie

### 2.1 The Model

**The model of continuous character evolution** Our model is composed of two processes: one describes the evolution of a discrete character and the other the evolution of a continuous character. To model the evolution of our continuous character we have chosen to use an Ornstein-Uhlenbeck (OU) model (Hansen 1997; Butler and King 2004, Hansen et al. 2008, Beaulieu et al. 2012). This model combines the stochastic evolution of a trait through time with a deterministic component which models the tendency for a trait to evolve towards an adaptive optima. In this model, a trait ($X(t)$) is is pulled towards an optimum at a rate that is scaled by the parameter $\alpha$, while the optimum itself (which may change through time) is denoted by the parameter $\theta(t)$. $\theta(t)$ is piecewise constant on intervals and takes values in a finite set $\{\theta_i\}$. This represents the set of "selective regimes", "regimes", or Simpson's "adaptive zones" based on narrational preference (Cressler et al. 2015). Additionally, random deviations are introduced by Gaussian white noise

$dB(t)$, which is distributed as a normal random variable with mean zero and variance one. The magnitude of these deviations is scaled by the noise intensity $\sigma$. $\alpha$ has been interpreted as the strength of selection (Simpson 1953, Lande 1976, 1980) and $\sigma$ has been referred to as genetic drift (Lande 1976, Hansen 1997). This latter interpretation has been criticized on the grounds that stochasticity may arise from factors other than genetic drift. The OU process is an Itô difussion satisfying:

$$X(t) = \alpha(\theta(t) - X(t)) + \sigma dB(t)$$

**The model of discrete character evolution** Most previous phylogenetic comparative models of an OU process have assumed that the intervals and regimes are known, though the optimum trait values associated with each regime are not (but see, Revell 2013 and May and Moore 2020). This leaves no room for inference about the regimes themselves and how they change through time, nor the possibility that the evolution of the continuous character could influence change in the regimes or vice versa. To resolve this problem, we will model the evolution of these regimes as a discrete character. For this, we have assumed that regime change follows a hidden Markov model (Felsenstein and Churchill 1996; Yang 1994; Beaulieu et al. 2013).

Hidden Markov models have a hierarchical structure that can be broken down into two components: a "state-dependent process" and an unobserved "parameter process" (Zucchini, MacDonald, & Langrock, 2017). Under an HMM, observations are generated by a given state-dependent process, which in turn depends on the state of the parameter process. In other words, the observed data are the product of several processes occurring in different parts of a phylogeny and the parameter process is way of linking them. It is initially unknown what the parameter process corresponds to biologically, hence the moniker "hidden" state. Nevertheless, the information for detecting hidden states comes from the differences in how the observed states change. As long as the transitions between observed states of different lineages are more adequately described by several Markov processes rather than a single process, there will be information to detect hidden states (Boyko and Beaulieu 2020). In comparative biology, for characters that take on discrete states the standard "state-dependent process" is a continuous-time Markov chain with finite state-space (CTMC-FS). The observed states could be any discretized trait such as presence or absence of extrafloral nectaries (Marazzi et al., 2012), woody or herbaceous growth habit (Beaulieu et al., 2013), or diet state across all animals (Román-Palacios et al. 2019). However, a simple Markov process that assumes homogeneity through time and across taxa is often not adequate to capture the variation of real datasets (e.g. Beaulieu et al., 2013), and thus one way to use HMMs is to link multiple evolutionary models (such as a standard Mk) together.

Under a standard Mk model, transitions between discrete states occurs at a particular rate $q_{ij}$, where a state changes from state $i$ to state $j$. These transition rates are often presented in the form of a matrix (transition rate matrix). This matrix describes the possible transitions between all $k$ states of the system. For example, an Mk model with $k = 3$ states has the following form as it's most general form:

$$Q_{Mk} = \begin{bmatrix} - & q_{12} & q_{13} \\ q_{21} & - & q_{23} \\ q_{31} & q_{32} & - \end{bmatrix}$$

The $-$ on the diagonal is a quantity required to make each row sum to zero. The Mk modeling framework can be modified and expanded in several ways. These modifications can lead to hypotheses of, for example, ordered transitions (e.g. it may not be possible to directly go from $1 \rightarrow 3$, but instead requires $1 \rightarrow 2 \rightarrow 3$), or models of correlated evolution where the transition rate of character $X$ is dependent the state of character $Y$. Regardless of the state-dependent process, we can link these together in a way that allows for rate variation. For example, say rate matrix $Q_{ord}$ described the evolution of a particular as being ordered and $Q_{er}$ modeled the possibility that all transitions between the states of a particular character are equal. We can combine these different state-dependent structures in a hidden Markov model generally as a block matrix.

$$Q_{hmm} = \begin{bmatrix} Q_{ord} & q_{Q_{ord} \rightarrow Q_{er}} \\ q_{Q_{er} \rightarrow Q_{ord}} & Q_{er} \end{bmatrix}$$

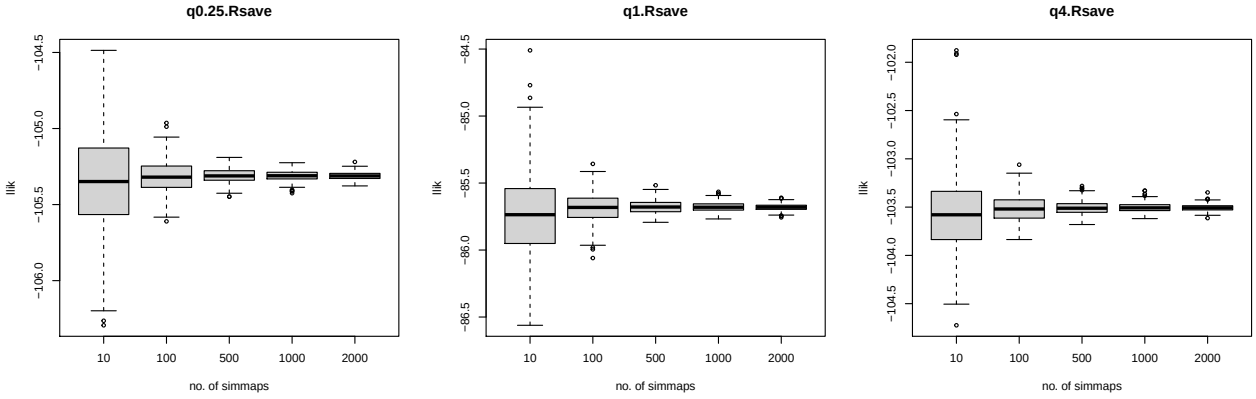
**Combining discrete and continuous character evolution** Here we get to the main goal of this project, which is to design a framework for simultaneously optimizing a discrete character model while also fitting a continuous character model with OUwie. The idea behind this is that we assume a correlation between the two trait types, but we reconstruct the regimes first, then fit OUwie separately. However, this is not what we want, because rates of continuous character evolution may not be optimal without information from how the regimes evolved and regime change may not be optimal without information from how rates change. Here we use a hierarchical maximum likelihood approach using the stochastic mapping now implemented in `corHMM()`. The likelihood function was derived by Jeremy Beaulieu and is computed by summing over all possible maps.

$$L(X, D|\Theta, \Phi) = P(D|\Phi) \sum_{z} P(z|D, \Phi) g(X|\Theta, z),$$

where $P(D|\Phi)$ is the probability of observing the vector states, $D$, at the tips of a phylogeny given parameters $\Phi$, and $P(z|D, \Phi)$ provides the probability of a character map $z$ given a vector of character state D and parameters $\Phi$ (note we use $\Phi$ to generically define any model that will estimate a character history). The mixture probability, $g(X|\Theta, z)$, is the BM or OU likelihood given a set of parameters $\Theta$ and character map $z$.

It is unclear how one would go about calculating $P(z|D, \Phi)$, especially with a larger tree and many character states. Instead, it is easier to sample a large number of realizations of the character transitions given $D$ and $\Phi$. This is where the stochastic maps come in. The summation above would then be over a set of unique character maps that appear in the sample. The frequency of each unique map, $f_z$, would then be used as an estimate of $P(z|D, \Phi)$. If each character map is unique, then the summation is simple an average of likelihoods across the set of maps. This approach is costly from a computational perspective because for every new set of $\Phi$ proposed a new set of character maps would have to be generated and reevaluated.

It is unclear how many maps would suffice to get a good approximation of $P(z|D, \Phi)$. To test this I simulated a birth-death tree (b=1, d=0.5) of 250 tips and scaled tree height to 1. I then simulated a discrete dataset and simmap following the an equal-rates Mk model at rates of $q = 0.25, q = 1, q = 4$. An OUM-type OU model was simulated with $\alpha = 0.5, \sigma^2 = 0.5, \Delta\theta = 1$. The true parameter values were then refit 500 times for each Mk rate value to get a sense of how many simmaps is necessary before the same set of parameters produce the same likelihood.




100 simmaps reduced the variance a great deal, but we will use 250 simmaps going forward. It remains to be tested whether certain parameter values lead to higher variance and need more simmaps to estimate well. One possible solution to utilize a faster computation of ~50 simmaps to locate a reasonable global optimum and 250 simmaps to find the local optimum. Another possibility is to sample simmaps until we acheive a certain variance (after a burn-in period). However, this approach would be computationally costly as it is most efficient to generate many simmaps simultaneously.

### 2.2 Evaluating hOUwie Through Simulation

#### 2.2.1 A brief aside discussing the parameters of the hOUwie model

**2.2.1.1 The number of parameters in hOUwie** The maximum number of parameters ($k$) for the most complex hOUwie model is going to depend on the number of observed discrete states and the number of desired hidden rate categories (for the global model, the OU portion of hOUwie will always be OUMVA and therefore the number of parameters will be determined by the number of regimes).

$$k_{hOUwie} = 3 * k_{obs} * k_{hid} + k_{hmm}$$

Where $k_{obs}$ is the number of observed states, $k_{hid}$ is the number of hidden rate categories, $k_{hmm} = k_{Mk} * k_{hid} + k_{hid}^2 - k_{hid}$ is the number of parameters in a hidden Markov model, and $k_{Mk} = k_{obs}^2 - k_{obs}$ is the number of parameters in an Mk model. For example, the global model if we had $k_{obs} = 3$ and $k_{hid} = 2$ would be:

```
cat("The HMM rate matrix\n")
```

```
## The HMM rate matrix
```

```
rate.mat <- getFullMat(list(getRateCatMat(3), getRateCatMat(3)), getRateCatMat(2))
print(rate.mat)
```

```
##        (1,R1) (2,R1) (3,R1) (1,R2) (2,R2) (3,R2)
## (1,R1)      0      3      5     14      0      0
## (2,R1)      1      0      6      0     14      0
## (3,R1)      2      4      0      0      0     14
## (1,R2)     13      0      0      0      9     11
## (2,R2)      0     13      0      7      0     12
## (3,R2)      0      0     13      8     10      0
```

```
cat("\nThe OU index matrix\n")
```

```
##
## The OU index matrix
```

```
ou.mat <- getOUParamStructure("OUMVA", "three.point", FALSE, FALSE, dim(rate.mat)[1])
print(ou.mat)
```

```
##        [,1] [,2] [,3] [,4] [,5] [,6]
## alpha     1    2    3    4    5    6
## sigma2    7    8    9   10   11   12
## theta    13   14   15   16   17   18
```

```
cat("\n", max(rate.mat, na.rm = TRUE) + max(ou.mat, na.rm = TRUE), "total params.\n")
```

```
##
##  32 total params.
```

With such complexity possible for just 3 observed states and 2 hidden states we cannot reasonably explore all parameter space and all of the potential model structures nested within the global model. Instead, we will split model evaluation into two groups. Group one will focus on Markov dynamics and the joint probability of observed discrete and continuous characters. In group one, we will evaluate the performance of hOUwie by simulating data under various Mk + BM|OU dynamics and then estimate the variance and bias of inferred rates. We will compare hOUwie's performance to the performance of other models which are currently available. Group two will evaluate hOUwie's hidden dynamics, including the character independent model (CID). The CID model is proposed to detect rate variation that is not associated with the observed character. Here we will explore model selection in the context of hOUwie. For e.g., we could simulate a hOUwie model where the discrete character has no influence on the continuous character and determine if we would find a spurious correlation between discrete and continuous character.

**2.2.1.1 Building intuition as to the meaning of parameters** How can we determine when a state transition occurs? Suppose at time $t$, we are in state $i$, and we are interested in the distribution of $\tau$, the time until transition to a different state. A key property of $\tau$ is that it is independent of how much time we have already spent in $t_i$. That is, if were to observe an extant species in a particular state, the distribution of $\tau$ is exactly the unconditioned distribution of $T + \tau$:

$$f_{\tau|\tau>T}(t+T) = f_{\tau}(t)$$

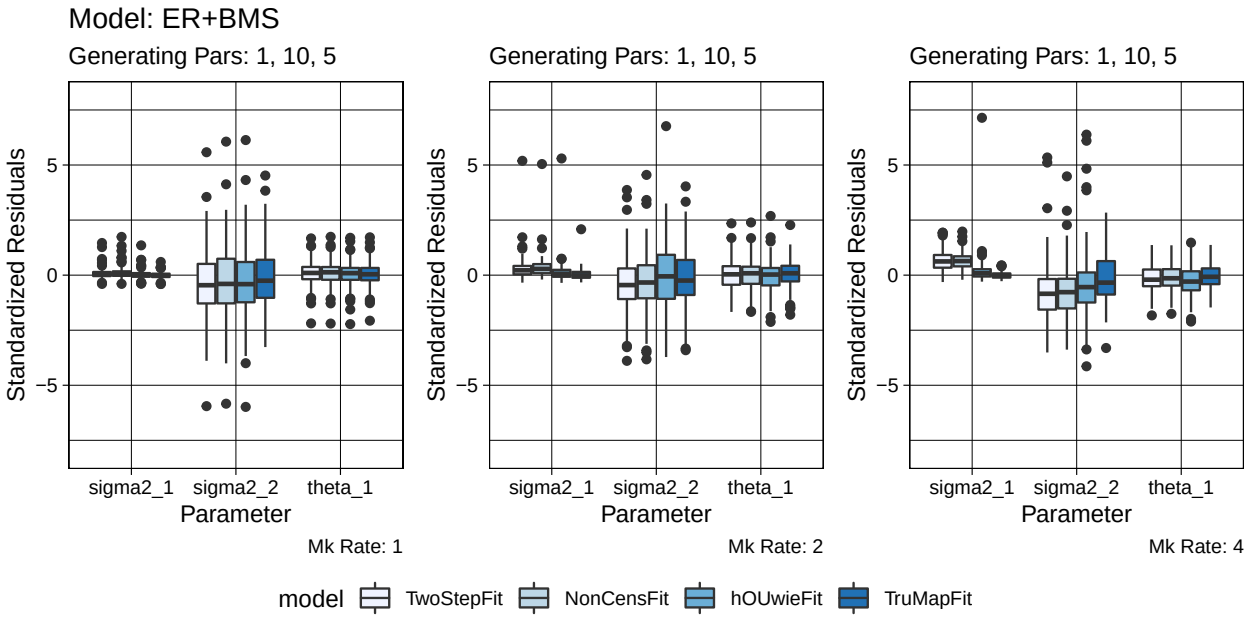
This is the statement that $\tau$ is memoryless (and time-homogenous) and therefore is described by the exponential. This means that the transition times at each state are exponentially distributed and transitions between specific pairs of states are also described by the rate of an exponential random variable. This means that the waiting time for each regime in our hOUwie model will be exponentially distributed with a mean $1/q_{ii}$. This will give us an idea of the expected rates of regime change (something that was not possible in other models where regimes were assumed to be static). How specifically regimes change (e.g. whether we move from a marine habitat to freshwater habitat, or marine to terrestrial depends on the model and is given by the relative rates of moving to the states $j$). The advantage of using $1/q_{ii}$ instead of all $1/q_{ij}$ is that it simplifies cases with hidden states where there are multiple pathways to move between observed states. Furthermore, knowing how long a species remains in a regime can tell us how long it has to reach its optimum phenotypic value, and with knowledge of half-life we can get reasonably complete idea of a particular lineages long-term evolutionary trajectory.

**2.2.2 Parameter Estimation** Our goal in this section is to examine error in hOUwie's parameter estimation compared to other potential frameworks. We will fit hOUwie and two alternative models based on approaches an empiricist may take: (1) TwoStep - maximize corHMM and OUwie separately (data analyzed completely separately). (2) NonCensored - maximize corHMM, simulate $n_{map}$ simmaps based on corHMM's $\hat{\theta}$, fit OUwie to the maps, summarize OUwie. (3) hOUwie - jointly maximize OUwie and corHMM with $n_{map}$ iterations.

**2.2.2.1 An Mk BMS model**

```
## The Mk rate matrix

##    R1 R2
## R1  0  1
## R2  1  0

##
## The OU index matrix

##        [,1] [,2]
## alpha    NA   NA
## sigma2    1    2
## theta     3    3

##
##  4 total params.
```

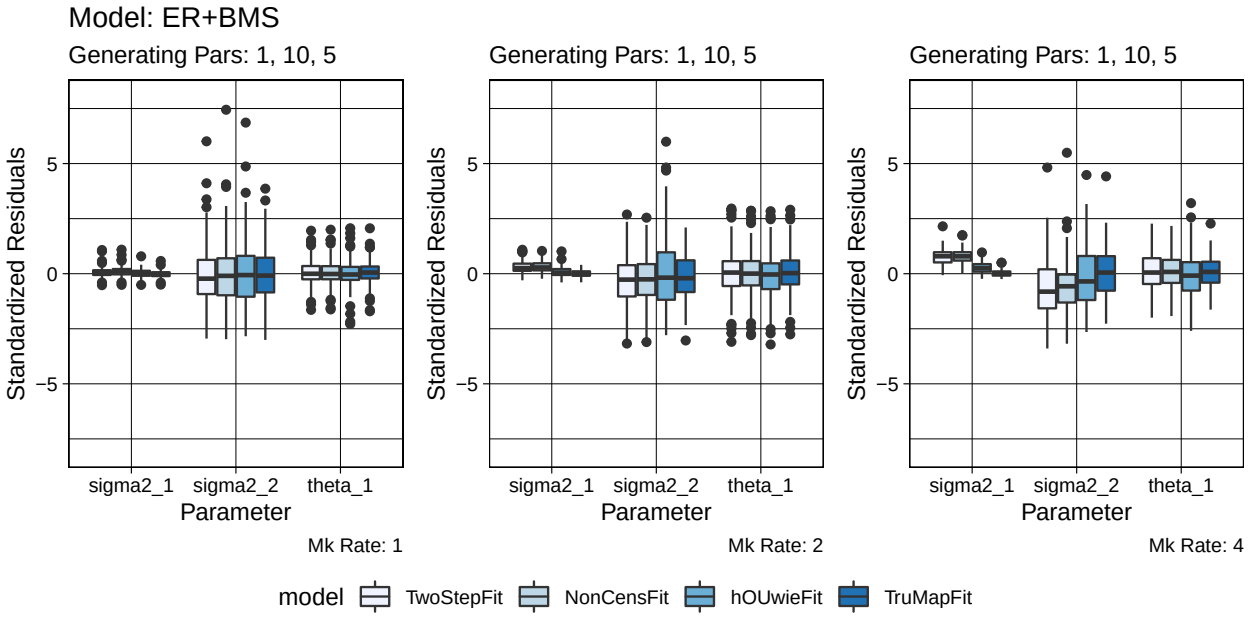
This model framework was used by Revell (2013) to demonstrate that using stochastic maps to estimate variable rate $\sigma_i^2$ (where $i$ is a regime specific rate parameter) was biased to be more similar towards one another than the true simulating values. I replicated this study simulating 500 datasets per single phylogeny ($b = 1, d = 0.5, nTip = 100, H = 1$). The simulations followed Revell (2013) where $\sigma_1^2 = 1$ and $\sigma_2^2 = 10$. We evaluated only the higher Mk rates $q = 1, q = 2, q = 4, q = 8$ since Revell (2013) found no bias at lower rates.




For the 100 tip tree examined here, an Mk Rate of 1 corresponds to about 10 discrete transitions, Mk Rate of 2 is about 35 discrete transitions, and an Mk Rate of 4 is about 70 discrete transitions. Here, standardized residuals are the estimated value for a particular model subtracted from the simulating value then divided by the standard deviation of all residuals. We can see that at lower transition rates $\sigma_1^2$ is estimated with low variance and low bias, but $\sigma_2^2$ is estimated with high variance. What's particularly odd about this result is that the high variance also applies to the the TruMapFit which was given the exact stochastic map used to simulate the data. As we move to Mk higher rates we begin to see bias in the estimates with $\sigma_1^1$ being overestimated and $\sigma_2^1$ being underestimated. This is what Revel (2013) found and although the bias is present in hOUwie estimates, it is less severe.

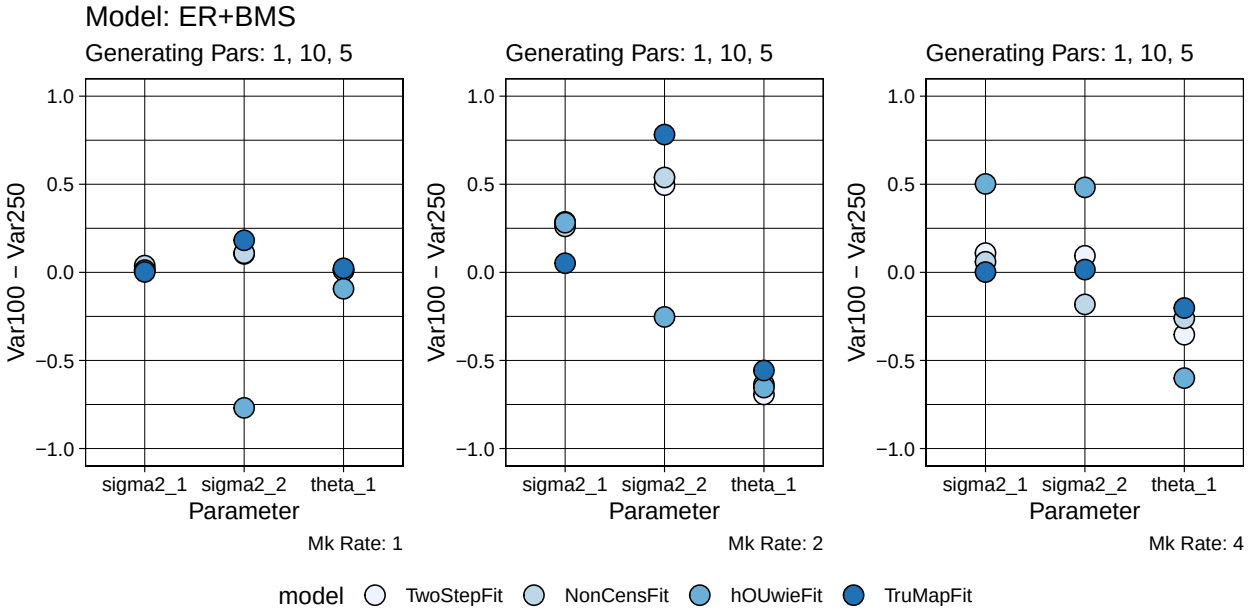
It is possible that the variance around the parameter estimates is a consequence of sample size (we only used a 100 tip tree). Thus we can increase $nTaxa$ to 250.

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```




We find the same general patterns of bias and variance exist even with more taxa: especially at higher rates,

$\sigma_1^2$ is generally overestimated and $\sigma_2^2$ is underestimated. Let's take a closer look at what increasing the number of taxa does to the variance around the parameter estimates.




In these plots, negative values correspond to decreasing variance of paramater estimates when we increase tip sampling. If focus on $\sigma_2^2$, we can see as we increase sampling the variance around hOUwie's parameter estimates decrease, but this effect is dependent on the rate of Mk evolution. However, given hOUwie preforms well at an Mk rate of 1, where most empirical estimates are likely to be found, I think we are doing well. The next step here would be to increase the number of simmaps hOUwie evaluates to determine if the reason for increased bias is the likelihood of any given simmap will increase as we increase the Mk rate. Incidentally, this simmap effect would also be effected by the number of tips in the tree, since the number of viable regime paintings will increase as we increase the total amount of time in the phylogeny (which is proportional to number of tips when birth and death parameters are the same).

```
## Total branch length in the phylogeny.

## 100 Taxa 250 Taxa
## 21.35679 43.82696
```

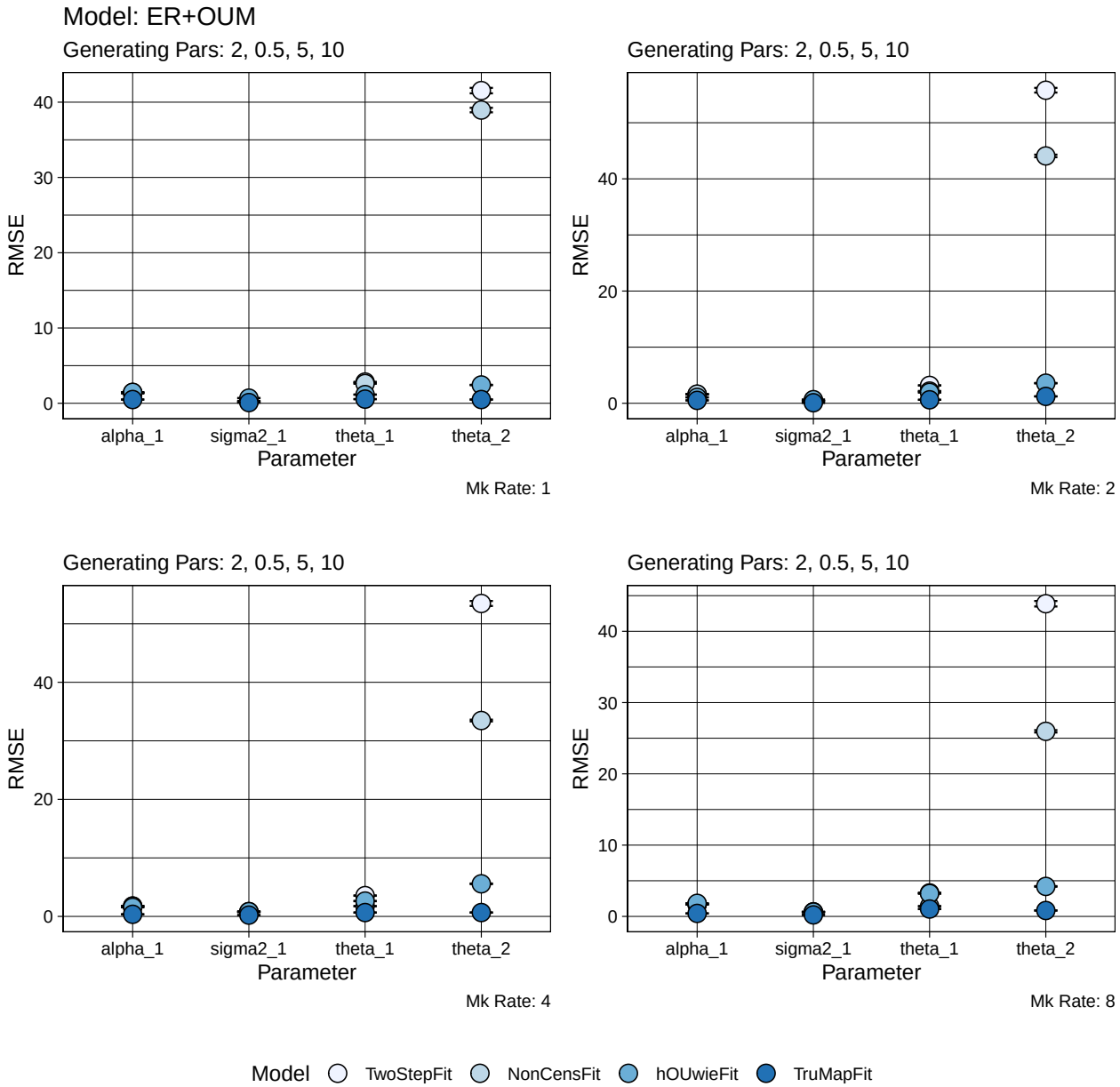
In a sense, both the number of tips and rate of Markov evolution influence time. As we increase the number of tips we allow more time for transitions to occur and as we increase the rate of evolution, the amount of time needed to get the same number of transitions will decrease. Since we are attempting to integrate over all possible simmaps, it is possible that as these parameters increase the effective time in the phylogeny we are not generating a suitably large distribution.

#### 2.2.2.2 An Mk OUM model

```
## The Mk rate matrix

##    R1 R2
## R1  0  1
## R2  1  0

##
## The OU index matrix

##        [,1] [,2]
## alpha     1    1
## sigma2    2    2
```

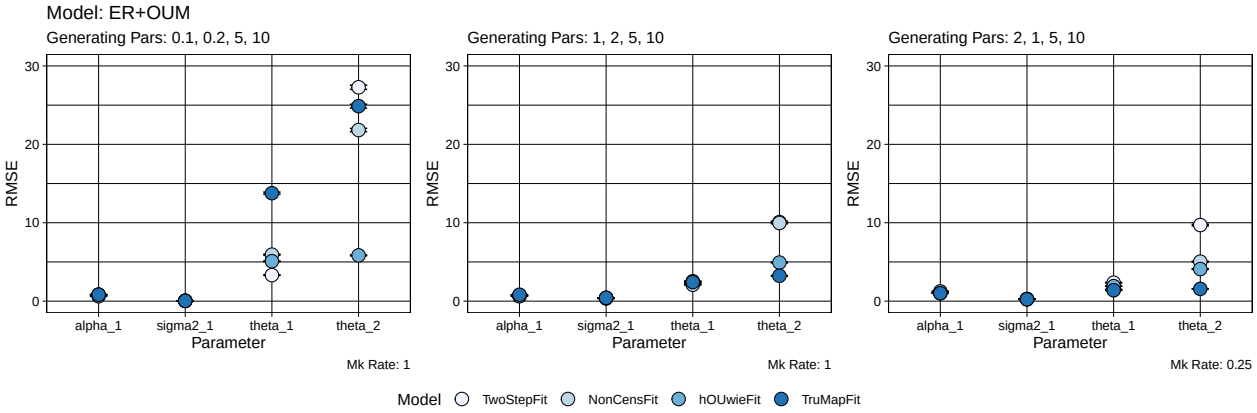
```
## theta     3    4

##
##  5 total params.
```

Just like the Mk BMS models, I was interested in whether the rate of discrete evolution would bias our estimates of continuous evolution. Therefore, I simulated under varting Mk rates of $q = 1, 2, 4, 8$, $\alpha = 2$, $\sigma^2 = 0.5$, $\theta_1 = 5$, and $\theta_1 = 10$. For each Mk rate I simulated 100 datasets per single phylogeny ($b = 1, d = 0.5, nTip = 100, H = 1$). Below I present results in terms of reduced mean square error (RMSE). I am still working out how best to examine differences in parameter estimates and RMSE is useful in that it simaltaneously gives an idea of bias (how far from the original parameter you are) and variance (how consistently the estimate is good). Per the suggestion of Brian O'Meara, I'd like to change this to sign and magnitude error (https://statmodeling.stat.columbia.edu/2004/12/29/type_1_type_2_t/). Sign and magnitude error are more inline with what an empiricist would be interested (was the rate of evolution for herbaceous things higher than woody? and was the estimate reasonable within biological variation?).

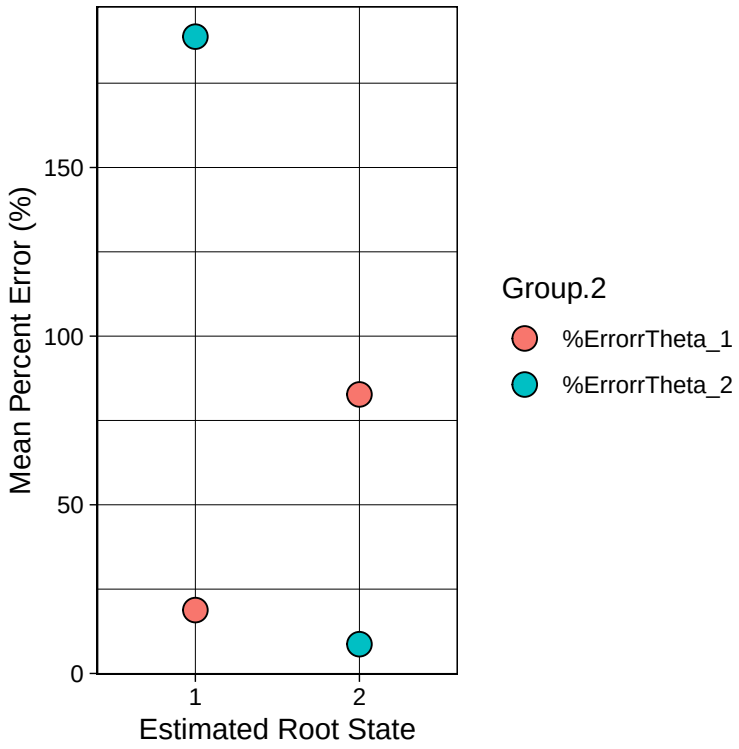


There seems to be little effect of altering the Mk rate with the overwhelming pattern being the poor estimation of $\theta_2$ of the NonCens and TwoStep approaches. Cressler et al. (2015) outlined regions of lower power for OU models defined by selection opportunity ($\eta = \alpha T$) and noise intensity ($\gamma = \sigma T / \Delta\theta$). The higher the selection opportunity and the lower the noise intensity the greater the power of an OU model. In our case, simulating parameters should have been reasonably distinguishable since $\Delta\theta = 5$, but we can increase the power by increasing the values of $\alpha$ and decreasing values of $\sigma^2$ (increasing selection opportunity and decreasing noise around optima).

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```

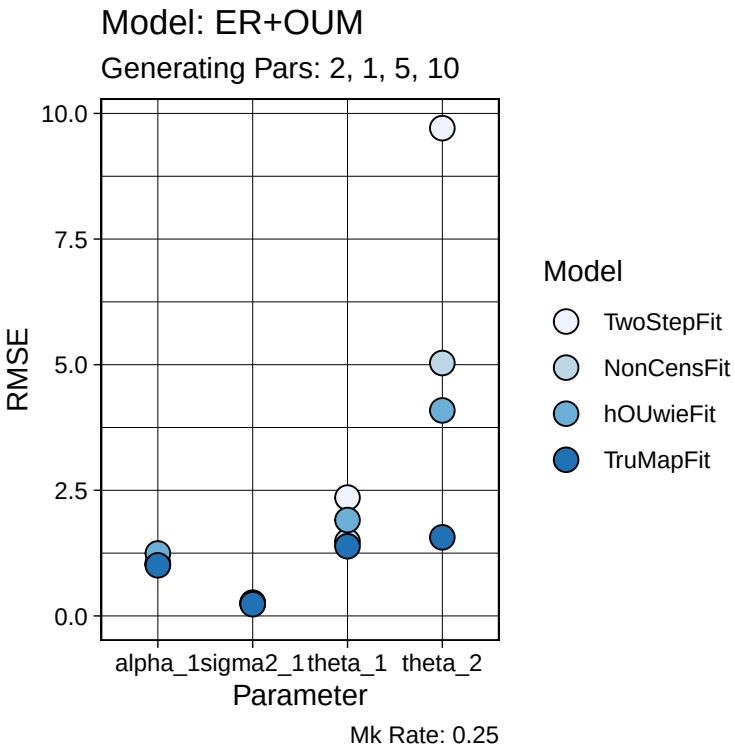



The first plot repeats what we had found earlier with a high RMSE for $\theta_2$, and as we increase the power of the OU model there is less error in $\theta_2$, but it comes at the cost of error in other parameters. I'd be remiss if I did not point out that hOUwie preforms well across all scenarios, but that does not explain why we see the errors. In fact, the actual cause of these large errors is not lower power of the OU models, but incorrect estimation of the root. Additionally, this error applies to both $\theta_1$ and $\theta_2$, but because $\theta_2$ estimates tend to be very large when misestimated they appear more obviously in our RMSE plots. To make the illustration more clear I will be focusing on the TwoStep fit since that only has a single root state to consider. It should be noted, that the model structure of discrete character evolution is not misspecified. Data were simulated under an equal rates Mk model and the ancestral state reconstructions are based on that model structure.




From this plot we see that $\theta_2$ is estimated well when the root state is in that particular regime, but poorly when it is not and this also applies to $\theta_1$ (although less extreme). We can examine the relationship between ancestral state reconstruction and estimation of theta by simulating a set of data where the rate of discrete evolution is low and therefore the root state is more likely to be correctly estimated.

```
## [[1]]
```



Clearly an improvement over higher rates of discrete character evolution for NonCens approach (which also uses a simmap approach), but the error is still high for the TwoStep approach. What gives? Well it seems that the OU model is trying to approximate a trend-type model where a particular optimum is very far away from the current position and a lineage very slowly moves towards it (low alpha and sigma). Thus, we end up with a linear increase (or decrease) through time.

A weakness of these simulations is that most of our complexity comes from the OU process (more free parameters). Boyko and Beaulieu (2020) demonstrated that the power for inferring an HMM comes predominantly from the difference between state-dependent processes. One interesting part of hOUwie is that it gives an additional axis to distinguish between state-dependent processes. Standard HMMs distinguish between state-dependent processes based on differences in how observed states change (eg. rate class A changes quickly between woody and herbacious while rate class B changes slowly). With hOUwie, we add information about how the continuous trait changes in distinguishing between the state-dependent processes. This point forms the basis of our null model.

**2.2.2 Model Selection** These models are currently running on the server. The set of models being run are chosen to answer some basic questions about hOUwie that I think an empiricist would like to see answered:

1) Can we detect hidden states when they're present? From previous work we know that we can detect hidden states when there are differences in the state dependent process with discrete, but what when the differences are entirely due to the OU process? E.g. The rate of limb length evolution is greater on an island than it is on the mainland, but we don't expect that transitions between mainland and island will differ by lineage.

2) Can we avoid detecting hidden states when they're absent? There are often problems in PCMs where overly complex models are selected. One of the advantages of hOUwie is that it explicitly models the process of regimes and is time-heterogeneous. This may assist in avoiding the large-p-small-n problem, but we should make sure it doesn't always select overly complex models.

3) Can we detect character dependence when it's present? A common question when using hOUwie could be does my focal trait associate with my ecological variable? In this case we should be able to recover character dependence.

4) Can we avoid detecting character dependence when it's absent? If there is no association between the OU parameters and the discrete regimes we should not find evidence for an association. This is potential problem reminiscent of the SSE controversy where name length was correlated with diversification rates. Fortunately, hOUwie allows for the construction of character independent models.

5) How does hOUwie perform when the model is outside the original set? In an empirical situation, the true model is outside the set of comprehensible models and is therefore unlikely to be in the set of our biologically informed models. However, with model averaging and tip averaging maybe there are certain aspects of this true models we can begin to understand. Here we will simulate an incredibly complex model (that will not be included within our analysis set) on a tree that is 100 times the size of the tree we will be pruned down for analysis. When we go to fit our set of models maybe we will see that we partially captured the dynamics of this complex model.

## Section 3: A brief case study

The field of evolutionary biology would have quickly fallen into disrepute without the elegant grace of Anolis lizards. They are the single most important genus on this planet and, should life exist outside of our solar system, Anoles would be the most important genus in the galaxy. Therefore, their continued scrutiny is crucial to the success of evolutionary study. Here I present a modest contribution to the study of Anolis by examining the Lesser Antillean anoles.

There are two main competing hypotheses to explain the body size differences in lesser Antillean anoles: the taxon cycle and character displacement. Under the taxon cycle hypothesis, if an anole were to disperse to an

island that already had congener on it, the establishment of the invader would depend on whether it or the native species were larger. If the invader is smaller, it is competitively excluded from the island. If it is larger, then its invasion is successful and the two anoles would co-evolve with the native species becoming smaller until extinction as the invader maintains its intermediate optimum. The character displacement hypothesis proposes that two anoles on the same island will compete and each will diverge from the intermediate optimum that is predicted for allopatric species. Thus, under character displacement we expect that allopatric anoles have an intermediate optima, and sympatric anoles have two optimum (small and large body size).

Taking a look at the data we see the general format of a hOUwie model. It follows OUwie and corHMM in that the first column is the species, the second column is the discrete trait, and the final column is the continuous character. But this data.frame can be extended to include multi-variable discrete traits (e.g. column 1 = sympatry or allopatry, column 2 = presence or absence of avian predators). And although we don't allow for multivaraite continuous characters, there is an option to include measurement error as part of the continuous character. In this case, 2 represents sympatry (2 species on a given island) and 1 corresponds to allopatry.

```
print(head(dat))
```

```
##           sp reg     size
## 1     aeneus   2 2.139624
## 2     roquet   1 2.240792
## 3   extremus   1 2.208070
## 4  richardii   2 2.604961
## 5    griseus   2 2.726213
## 6 trinitatis   2 2.149603
```

Now, we will construct hOUwie models that match the character displacement and taxon cycle hypotheses as well as some alternatives (defaults). The matrices below are going to the basic building blocks for which we will define regime changes. An ARD model would suggest that the rates of movement between allopatry and sympatry differ depending on whether a species is sympatric or allopatric. An ER model says the rate of change will be the same regardless of the state. The a2s model stands for allopatric to sympatric and is an explicit constraint that allopatric species move into sympatry and never out of it.

```
ARD <- getStateMat4Dat(data = dat[,c(1,2)])$rate.mat
ER <- equateStateMatPars(ARD, c(1,2))
a2s <- dropStateMatPars(ARD, 1)
```

Below, we construct the character displacement hypothesis of regime change. In both rate class 1 (R1) and rate class 2 (R2), we allow transitions between sympatry and allopatry and expect they occur at the same rate. However, we only allow transitions between rate classes to occur if a species is allopatric. This is codifies the assumption that once an anole becomes sympatric in a particular rate class it cannot transition to the optimum of the other sympatric optimum. That is to say, if our hypothesis is correct and sympatric species have two distinct optima (large and small body size) we don't have a sympatric anole move from large body size to small body size without first becoming allopatric again first.

```
CD.Cor <- getFullMat(list(a2s, a2s), ER)
CD.Cor <- equateStateMatPars(CD.Cor, c(1,2,3)) # not really any case for one evolving faster than anoth
CD.Cor[2,] <- c(1,0,0,0)
CD.Cor[4,] <- c(0,0,1,0)
print(CD.Cor)
```

```
##        (1,R1) (2,R1) (1,R2) (2,R2)
## (1,R1)      0      1      1      0
## (2,R1)      1      0      0      0
## (1,R2)      1      0      0      1
## (2,R2)      0      0      1      0
```

The OU model structure below codifies our assumption that there will be three optima total: allopatric species will have the same optima (parameter 3) regardless of rate class, and sympatric species will have two

optima (parameter 4 and 5) depending on the rate class.

```
CD.OU <- getOUParamStructure("OUM", "three.point", FALSE, FALSE, dim(CD.Cor)[1])
CD.OU[3,] <- c(3,4,3,5)
print(CD.OU)
```

```
##        [,1] [,2] [,3] [,4]
## alpha     1    1    1    1
## sigma2    2    2    2    2
## theta     3    4    3    5
```

The model structure corresponding to the taxon cycle follows the same principles as above. The essential differences is that there are only two phenotypic optima: an intermediate optima (which corresponds to invaders and allopatric anoles) and a small optima which corresponds to the native anoles increasingly small size. The model structure of regime change is more constrained. Here, allopatric anoles (1R1) can transition to sympatry (2R1) only when in R1. This is because not all allopatric anoles can succesfully invade, only ones that are slightly larger than the rest should be able to. From there, those anoles will remain at their intermediate optima but are allowed transition back to an allopatric state (1R2). From there they can move back into their initial state (1R1). It is also possible for a species in an allopatric state to be the native species and in this case they move from their initial allpatric state into a differeny sympatric state (2R2). This sympatric state has a different optima that is associated with the pressure of the larger congener invading their island.

```
TC.cor <- getFullMat(list(a2s, ER), ER)
TC.cor[1,] <- c(0,1,0,1)
TC.cor[2,] <- c(0,0,1,0)
TC.cor[3,] <- c(1,0,0,0)
TC.cor[4,] <- c(0,0,0,0)
TC.OU <- getOUParamStructure("OUM", "three.point", FALSE, FALSE, dim(TC.cor)[1])
TC.OU[3,] <- c(3,3,3,4)
print(TC.cor)
```

```
##        (1,R1) (2,R1) (1,R2) (2,R2)
## (1,R1)      0      1      0      1
## (2,R1)      0      0      1      0
## (1,R2)      1      0      0      0
## (2,R2)      0      0      0      0
```

```
print(TC.OU)
```

```
##        [,1] [,2] [,3] [,4]
## alpha     1    1    1    1
## sigma2    2    2    2    2
## theta     3    3    3    4
```

Place our models into lists which define our model set.

```
index.cor.list <- list(
  CD = CD.Cor,
  TC = TC.cor
)
index.ou.list <- list(
  CD = CD.OU,
  TC = TC.OU
)
rate.cats <- c(2,2)
```

Because these models can take a long time to run I will load in our results, but below you can see the code

that was used to run the models. I use a function called fit.hOUwie.set which will fit all the models in a set, but also prunes redundant models (models that are identical and lead to the same likelihood), creates a table which contains the relative support for each model, and gets model averaged tip parameters.

```
#fit.hOUwie.set(phy = phy, data = dat, rate.cats = rate.cats,
#index.cor.list = index.cor.list, index.ou.list = index.ou.list,
#root.p = "maddfitz", nSim = 10, lb.cor = 0.002, ub.cor = 0.2)
load("~/2020_hOUwie/data/LA_Anolis_hOUwieSetFit.Rsave")
LA_Anolis_hOUwieSetFit
```

```
## Model Fit Table
##    Model np   lnLik   AICc dAICc AICcwt
## TC    TC  5  -9.139 21.404  0.00  0.861
## CD    CD  6 -10.127 25.054  3.65  0.139
```

Because we used fit.hOUwie.set our print object is a model table and we can see the taxon cycle hypothesis is best supported, but there is also evidence of the character displacement hypothesis.

```
tail(round(LA_Anolis_hOUwieSetFit$model.pars, 3))
```

```
##            alpha sigma optim half.life station.var dist.2.opt wait.time
## gingivinus 0.768  0.06 2.123     0.902       0.023      0.004    71.293
## leachii    0.768  0.06 2.347     0.902       0.023     -0.170    38.129
## pogus      0.768  0.06 1.909     0.902       0.023      0.111   271.345
## schwartzi  0.768  0.06 1.915     0.902       0.023      0.092   242.419
## wattsii    0.768  0.06 1.908     0.902       0.023      0.026   271.236
## acutus     0.768  0.06 2.337     0.902       0.023      0.218    23.286
```

Above are the model averaged parameter estimates for each tip state and we can see some interesting results. A. gingivinus, A. leachii, A. pogus, A. schwartzi, and A. wattsii are all sympatric species, but each species has a distinct optimal value according to its marginal probability of being in a particular hidden state across models. The half-life is the time it would take to get half way from the current phenotypic value to the optimal phenotype and we can see that it is several orders of magnitude lower than the wait.time (which is the amount of time a species is expected to stay in a particular state). Because we did not vary alpha and sigma in our results, the half-life and stationary variance are the same for each species. Dist.2.opt gives us an idea of how far from the optima each species (in the input units (log(mm) in this case)) - e.g. leachii is 1.18mm larger than it's optimum would suggest. This is inline with the taxon cycle hypothesis which proposes that large anoles move towards their energetic optimum (an intermediate size). These were the parameter transformations that I thought could be insightful, but others are possible. This is also not a full set of models that could be explored. Here we've only varied the model structure of the regimes, but it is also possible to vary the rates of change and allow for different paths. It's also possible to allow for variable alpha and sigma. A variable alpha because in the case of sympatric competition there is predicted to be a strong selective force. A variable sigma because allopatric anoles are predicted to experience density dependent effects or, if you don't believe that the population process would scale to interspecific variance, because allopatric anoles are more likely to be dictated by their specific environments instead of the more consistent effect of direct competition.

## References


Beaulieu, J. M., B. C. O'Meara, and M. J. Donoghue. 2013. Identifying Hidden Rate Changes in the Evolution of a Binary Morphological Character: The Evolution of Plant Habit in Campanulid Angiosperms. Systematic Biology 62:725–737.

Beaulieu, J. M., D.-C. Jhwueng, C. Boettiger, and B. C. O'Meara. 2012. Modeling stabilizing selection: expanding the Ornstein–Uhlenbeck model of adaptive evolution. Evolution: International Journal of Organic Evolution 66:2369–2383.

Boyko, J. D., and J. M. Beaulieu. 2021. Generalized hidden Markov models for phylogenetic comparative datasets. (N. Cooper, ed.)Methods in Ecology and Evolution 12:468–478.

Butler, M. A., and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. The American Naturalist 164:683–695.

Cressler, C. E., M. A. Butler, and A. A. King. 2015. Detecting Adaptive Evolution in Phylogenetic Comparative Analysis Using the Ornstein–Uhlenbeck Model. Systematic Biology 64:953–968.

Eastman, J. M., M. E. Alfaro, P. Joyce, A. L. Hipp, and L. J. Harmon. 2011. A Novel Comparative Method for Identifying Shifts in the Rate of Character Evolution on Trees. Evolution 65:3578–3589.

Felsenstein, J., and G. A. Churchill. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Molecular Biology and Evolution 13:93–104.

Grant, P. R., and B. R. Grant. 2006. Evolution of Character Displacement in Darwin's Finches. Science 313:224–226.

Hansen, T. F. 1997. Stabilizing Selection and the Comparative Analysis of Adaptation. Evolution 51:1341–1351.

Hansen, T. F., J. Pienaar, and S. H. Orzack. 2008. A Comparative Method for Studying Adaptation to a Randomly Evolving Environment. Evolution 62:1965–1977.

Harmon, L. J., J. B. Losos, T. J. Davies, R. G. Gillespie, J. L. Gittleman, W. B. Jennings, K. H. Kozak, et al. 2010. Early Bursts of Body Size and Shape Evolution Are Rare in Comparative Data. Evolution 64:2385–2396.

May, M. R., and B. R. Moore. 2020. A Bayesian Approach for Inferring the Impact of a Discrete Character on Rates of Continuous-Character Evolution in the Presence of Background-Rate Variation. Systematic Biology 69:530–544.

O'Meara, B. C., C. Ané, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for Different Rates of Continuous Trait Evolution Using Likelihood. Evolution 60:922–933.

Revell, L. J. 2013. A Comment on the Use of Stochastic Character Maps to Estimate Evolutionary Rate Variation in a Continuously Valued Trait. Systematic Biology 62:339–345.

Venditti, C., A. Meade, and M. Pagel. 2006. Detecting the Node-Density Artifact in Phylogeny Reconstruction. Systematic Biology 55:637–643.

Zucchini, W., I. L. MacDonald, and R. Langrock. 2017. Hidden Markov models for time series: an introduction using R. Chapman and Hall/CRC.