**Jointly Modeling Discrete and Continuous Traits**

**as a Means of Discovering Hidden Rate Heterogeneity**

James D. Boyko1\*, Brian O’Meara2, Jeremy M. Beaulieu1

*1Department of Biological Sciences, University of Arkansas, Fayetteville, AR, 72701, USA*

*2Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, USA*

*\*To whom correspondence should be addressed (jboyko@uark.edu).*

**Running head**: **Jointly Modeling Discrete and Continuous Traits**

**Abstract**

Rates of evolution have changed throughout the history of life and produced the rich diversity of morphology, behaviour, and ecology that characterizes the biodiversity we see today. As such, models favoring variation in rates and states has fast become the rule, as opposed to the rare exception in empirical applications of comparative methods. The challenge then becomes how to model this heterogeneous process with the limitation that, for most comparative biologists, our data comes exclusively from extant species. Our model, which we call hOUwie, uses hidden Markov models as a statistically rigorous way to learn about variation in the evolutionary process. hOUwie is composed of two processes: one describes the evolution of a discrete character (such as whether a mammal is a carnivore, herbivore, or omnivore) and the other the evolution of a continuous character (such as mammalian body size). To model the evolution of our continuous character we have chosen to use an Ornstein-Uhlenbeck (OU) model. In evolutionary biology, we use it to model the change in traits through time, which is the product of two antagonistic forces continuously pulling trait values towards some optimal value and randomly moving away from the optimum. The dynamics of these forces can also vary by so-called “regimes”. Under a typical OU model, these underlying regimes are taken as a given and are defined a priori leaving no room for inference about the evolution of the regimes themselves. To resolve this problem, we will model the evolution of these regimes as a discrete character via the aforementioned hidden Markov model. One of the most important properties we will utilize is the inclusion of hidden states. Hidden states give us a way to learn about the variation in the evolutionary process without the necessity of it being linked to one of our explanatory variables.

**Introduction**

The two most common approaches for applying OU process with multiple , , and/or have been to either assume selective regimes a priori (e.g. Butler and King 2004; Hansen et al. 2008; Beaulieu et al. 2012) or to infer the underling regimes based on the distribution of the continuous trait often called “shift-detection” (e.g. Ingram and Mahler 2013; Uyeda and Harmon 2014; Khabbazian et al. 2016). With the former approach, regime paintings are often generated through stochastic mapping, a form of ancestral state reconstruction (ASR) (Revell 2013). The advantage of this approach is that there is an explicit model for how regimes change through time. However, the evolution of these regimes is entirely independent of the continuous trait and the probability of these regimes is not explicitly considered. For example, it may be possible that the model which fits the discrete data best generates stochastic maps which do not fit the continuous data. Under this approach, although we have explicitly considered how regimes change, we have not considered whether the changes are appropriate for the continuous trait. The second approach of detecting shifts in underlying regimes can consider both the discrete and continuous character by evaluating the likelihood of the continuous character given a regime painting which matches the discrete data. These methods work quite well at detecting shifts although a modified BIC should be used to avoid overfitting (Khabbazian et al. 2016). The shift detection approach can be taken a step further, as in the case of Uyeda et al. (2018), who modeled a species’ underlying regime as the weighted average of regimes informed by shifts in the discrete character and regimes informed by the continuous data alone. However, this approach will suffer from the opposite problem of the ancestral state reconstruction approach. Although we will have explicitly considered the probability of our continuous data given regimes (which may or may not be informed by discrete character data), the probability of the regimes themselves is not explicitly considered. This could lead to scenarios where regimes are well suited for the continuous trait distribution, but poorly describe the discrete character.

Ornstein-Uhlenbeck models have been conceptualized and interpreted in many ways. Quantitative genetic interpretations are reliant on adaptation towards an optimum and the parameters of the model can have very clear evolutionary interpretations. They have been used in stock prices as models of deterministic stochasticity where a stock price is expected to reach a certain value, but with some uncertainty along the way. However, in macroevolution, perhaps because of our ties to microevolutionary theory, interpretations of the OU process have been focused on stabilizing selection and adaptation towards an optimum. Strictly speaking, this is not a correct interpretation. The work of Thomas Hansen and others has shown that the OU model when parameterized to macroevolutionary data does not match what is expected at the microevolutionary level. Thus, the OU model has begun to take on interpretations of simply phylogenetic signal. But here we may be throwing the baby out with the bath water. Phylogenetic signal exists because of descent with modification. Brownian motion at a macro-level can be seen as microevolutionary adaptation so long as we expect that all lineages phenotypic optima drift randomly on large timescales. OU models do not describe the phenotypic optima of a population evolving at the microevolutionary level. However, they can describe and give insights into the long-term trends of those microevolutionary dynamics. The question then becomes how do we want to describe changes in the dynamics? Typically, OU models are codified with hypotheses informed by intuition of scientists or ancestral state reconstruction. However, we do not know the ancestral state. We can only make a probabilistic statement about what it may be. Most methods until now have not allowed for uncertainty in this reconstruction. Here we describe these dynamics via a hidden Markov model. The microevolutionary interpretation of this model notwithstanding, this model is tractable and widely used. In truth, it may be more appropriate to study macroevolution in ways that are interpretable only in a macroevolutionary setting.

Imagine microevolutionary dynamics are in equilibrium on a macroevolutionary landscape (like a period of stasis as described by Gould and others). The dynamics are loosely described by the OU models in which drift around the phenotypic optima occurs so long as movement is not too far away. From this point how does change occur? There is abundant evidence that a change should occur given the theory of punctuated equilibrium. Most OU modeling is done such that the positions of change are *apriori* decided. This is not a probabilistic statement about where we think change is to occur, it is a statement of fact that a change did occur. However, we do not know where changes occur or what the ancestral state was for certain, and thus a probability statement is far more appropriate. A theory in-line with punctuated equilibrium would be one which allows for a sudden shift to a new microevolutionary landscape. Our model describes such sudden shifts using hidden Markov models.

Additionally, we have allowed for the possibility that regime changes follow 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 et al. 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. If 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 2021). Since the state-dependent process of hOUwie is a joint model of discrete and continuous character we can parse whether differences in continuous or discrete variation are due to the influence of our observed characters ( and ) or hidden states.

Previous work discussing character dependence and independence has focused on whether a focal trait influences diversification rate. Less attention has been placed on the potential biases associated with pure phenotypic evolution.

**Methods**

The hOUwie model

The hOUwie 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 a single 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; Ho and Ané 2014*b*). 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 optimum. In this model, a trait () is pulled towards an optimum at a rate scaled by the parameter , while the optimum itself (which may change through time) is denoted by the parameter . is piecewise constant on intervals and takes values in a finite set. 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 , 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 . has been interpreted as the strength of selection (Simpson 1953, Lande 1976, 1980) and has been referred to as genetic drift (Lande 1976, Hansen 1997). However, both interpretations have been criticized. The former because may be more aptly described as a measure of phylogenetic signal (Ho and Ané 2014*a*) and the latter because the stochasticity modeled by an OU on a macroevolutionary time scale is likely not a consequence of genetic drift (Hansen 2012). Formally, the OU process is an Itô diffusion satisfying:

Here, we use the set of extensions introduced by Beaulieu et al. (2012) and implemented in the R-package OUwie. This allows for multiple primary optima () in which both the pull strength () and the noise intensity () can vary across the phylogeny. The derivation of this model is given in Beaulieu et al. (2012). However, the algorithm used to calculate the likelihood described in Beaulieu et al. (2012) involves matrix inversion - a computationally costly procedure. Therefore, we implement a linear-time computation of the likelihood of Gaussian trait models following (Ho and Ané 2014*b*). To do this we first transform the phylogeny such that its variance covariance matrix,, is 3-point structured. We can write the variance covariance matrix of the untransformed phylogeny as , where following Beaulieu et al. (2012) and Ho and Ané (2014),

where, is the distance from the root to the beginning of the selective regime () for the number of selective regimes along the path from the root to the last common ancestor of and ( or from the root to the terminal tip (). Our transformed phylogeny now has a variance covariance matrix and diagonal matrix . We can then apply the function three.point.compute from the R-package phylolm (Ho et al. 2016) to calculate the quadratic quantities and determinant of . The probability of our continuous trait is given by

where is the number of tips in the phylogeny (, is the continuous trait value of each species, and is the expected value of each species given the continuous trait model – calculated following equation (11) of Beaulieu et al. (2012), *D* is the discrete character data, is a particular regime mapping, and are the parameters of the hOUwie model.

Next we describe the calculation of the probability of the underlying regime structure, , which is the joint probability of discret­e characters () and stochastic mapping () (similar to the pathway likeilhood in Steel and Penny 2000). May and Moore (2020) suggested that the joint probability of a regime structure and the discrete character is the product of the probabilities of exponentially distributed waiting times. By this definition, branch lengths are the sum of waiting times. However, for an equal rates model where the maximum likelihood estimate is zero changes, any path of 0, 2, or 100 changes has the same likelihood. We suspect this is an issue of treating the probability density functions independently when their convolution is required (the time of the second transition depends on the timing of the first and the total branch length available). The convolution of these independent random variables is given by an Erlang distribution. Ultimately, the effect of this inaccuracy is that the number of transitions has no influence on the probability of a branch and that the sum of the joint probabilities will be greater than the marginal probability. For further details and R-code necessary to demonstrate these issues, we refer readers to Appendices A and B.

Instead, to calculate the probability of discret­e characters () and stochastic mapping () we use an approximation. Our approximation relies on a finite number of degree-2 internodes and uses the standard Chapman-Kolmgorov equation to calculate the probabilities of beginning in a particular state and ending in state (Pagel 1994). As the number of internodes increase, the amount of time between nodes decreases and the approximation improves. Formally, we calculate the joint probability of a regime structure and the discrete character as

where the instantaneous rate matrix (, is the phylogeny, is the root probability (Pagel 1994; Yang 2006; Maddison et al. 2007), is the number of external nodes (tips), internal nodes, and internodes (degree-2 nodes) summed, indicates a particular branch, , and is an indicator function which ensures that we only use the probability of states indicated by the specific the regime mapping instead of summing over all possible state combinations (Figure 1).

Our likelihood function for the joint probability of continuous and discrete character is:

The continuous character probability requires a discrete state to be defined along the entire branch, thus we place transitions halfway between nodes (Figure 1d). Finally, there are three important caveats of our model that will be mentioned now. First, our discrete mapping probability ( is merely an approximation. What we calculate here is the probability of starting in a particular state and ending a particular state , summed over all possible paths. However, the continuous model probability is based off a particular pathway history which is defined throughout the entire branch. Our justification for this approximation is that as the length between internodes decreases this approximation improves because the probability of more than a single transition (or no transitions) decrease. The second caveat is that we do not force hOUwie to sum over all possible mappings . This is because the number of mappings will grow exponentially as the number of nodes and internodes increases and the computation will quickly become infeasible. Instead, we simulate node, internodes, and tip states (tip states only in the case of hidden Markov models) using the stochastic mapping procedure described in Bollback (2006) and modified to include the possibility of internodes (Figure 1a-c). We explore this caveat in series of simulation studies below and show that this sampling procedure helps identify high probability mappings which contribute the most to the overall joint probability. Our final caveat is that because we generate node mappings using the stochastic mappings procedure, they would typically be based solely on the discrete character process. This can be particularly problematic for character independent models where the discrete character and continuous character are decoupled. An imperfect solution to this problem is to weight tip probabilities when first calculating the conditional likelihoods (Bollback 2006). Typically, tip values for hidden Markov models are given a weight of 1 for each state (Beaulieu et al. 2013). However, we can modify this value to be proportional to the probability of sampling a particular tip’s continuous trait value given the models theta and sigma value for each discrete state. This is the approach we take; we assume that the conditional probability of a tip being in a particular hidden state is not equal, but rather proportional to ; the probability of sampling the continuous trait value under a normal distribution for each possible state . In cases where there are no differences between parameters, conditional probabilities remain equal.

Exploring the joint model of discrete and continuous phenotypes

For any reasonably large dataset using all possible discrete character mappings is computationally infeasible. However, for trees with fewer than 5 tips and few internodes, we can assess the behavior of our joint model under the ideal circumstance of every possible mapping. Our aim in this section is to determine whether hOUwie can detect signals of character independence and dependence, understand better where that signal comes from in the data, and search for mappings which have incongruent signals for each type of data (i.e., mappings which are highly likely for discrete characters, but poorly fit our continuous trait or vice versa).

We evaluate the performance and importance of the joint estimation of all mappings by simulating a pure birth phylogenetic tree of 4 taxa and coding the exact dataset by hand. We create our datasets such that they are expected to be consistent with different scenarios of character dependence and character independence (Figure 2a). We then evaluate the exact joint likelihood with fixed parameters for each dataset. Finally, we examine our results to determine whether the model detected the signal of character dependence or independence.

Additionally, because the actual implementation of hOUwie generates mappings stochastically and in proportion to the probability of the mapping and the discrete character, we can evaluate whether this leads to bias in our map generation. For example, it is possible that most mappings with high discrete character probabilities have low continuous trait probabilities. If this were the case, generating maps in proportion to the discrete character probabilities would bias our estimates away from possible maximum likelihood solutions. To test the efficiency of each procedure, we sample maps uniformly and sample maps proportionally to their discrete probabilities and compare which sampling procedure more rapidly approaches the true likelihood as we increase the number of maps being sampled.

Possible hOUwie models

We examined 22 possible hOUwie model structures for a binary discrete character. We assumed that transitions between the observed characters were equal although, when hidden states were included, different hidden states were allowed to have different rates. The model structures allowable in hOUwie are a generalized form of those allowed in OUwie and now include models in which only alpha varies (OU1A), only sigma varies (OU1S), and combinations of an OU and BM process (OUBM). The model structures range from completely character dependent to character independent. Character dependent (CD) models are models in which any OU parameter differs between observed state, while character independent models test whether observed states can be described by the same parameters. There are two types of character independent model. First, character independent models include structures where there are no differences between any OU parameters (CID). Under this model the entire evolutionary history of the clade can be described by a single alpha, sigma, and optimum value. To combat this potentially unrealistic assumption we introduce a character independent model which allows for differences in the OU parameters to depend upon an unobserved hidden state (CID+). This addition allows for heterogeneity within the evolutionary process without the necessity of it being linked to a focal trait. In total we examine 22 unique model structures (2 CID, 10 CD, and 10CID+).

Simulation study

We evaluated the performance of the hOUwie model using a simulated phylogenetic dataset. For each dataset, we simulated a pure birth phylogenetic tree with , and rescaled tree height to 1. The parameters used to generate a phenotypic dataset depend on the structure of the generating model. For example, an OUM model and OU1 model can have identical , but they must differ in or else OUM will collapse into OU1. Thus, we set parameter “baselines” at , which were doubled if the model structure allowed the parameter to vary. For example, an OU1 model would be simulated with parameters , but an OUM model would be simulated with parameters . Once a phylogeny and phenotypic dataset were simulated, we fit our models to assess parameter estimation accuracy and model selection power.

The simplest way to assess parameter estimation would be to fit the generating model to a dataset it generated and examine the difference between the simulating values and the maximum likelihood estimate. The main source of error would then come from the likelihood of the model fit to a stochastically generated dataset not necessarily corresponding to the simulating parameters. However, our variables occur at different scales and at different scales (a difference 1 unit is more important to an estimate of than under our simulation protocol because it represents a proportionally larger change). Thus, we assess the difference between generating and estimated parameters in natural log units. Additionally, it is often the case that biological hypotheses do not depend on the exact value of a parameter. In examining a dataset, it would be difficult to make a statement about what value we expect or to take on. Instead, most hypotheses are put in relative terms. For example, do we expect that the rate of evolution will be lower for species with smaller range sizes than species with larger range sizes? Furthermore, if we do find a difference in the relative values of the parameters there is no guarantee that this difference is biologically meaningful. Thus, we focus our evaluation of parameter estimation on assessing type-S (sign) and type-M (magnitude) errors (Gelman and Carlin 2014).

The empirical example

Ever since humans dared look upwards at the night sky to first confront the infinite expanse of darkness punctuated by non-random pricks of light, they have wondered whether the genome size of an herbaceous thing or woody thing changed faster.

**Results**

We find that sampling based on the discrete weighted probabilities is a more efficient way to get the true likelihood (Figure 4). Does this mean we should just use the best ancestral state reconstruction available and get the correct answer? To answer this, we can examine our results for discordance. Is there any evidence of low probability mappings producing major contributions to the likelihood? What proportion of our overall probability comes from the best discrete mapping, and does it have the highest continuous value too? Although discrete probability is a good predictor of the overall likelihood contribution, it is still necessary to include the continuous character as there are mappings with lower discrete probabilities that have higher overall likelihoods (Figure 5; S1).

Parameter estimation in hOUwie did not vary greatly depending on the number of stochastic maps or number of internodes included. Most variation in parameter estimation was instead related to the model complexity and which parameters are allowed to vary. It has been well-established that estimating OU models can be difficult and this is particularly true of variable alpha models. Here we find that when alpha is allowed to vary alongside other parameters it was poorly estimated (as was the case in Beaulieu 2012). However, when modeled with the newly implemented OUA model alpha was estimated well. It is possible that increasing the number of taxa could improve inference of these models, but previous work (Cressler, Beaulieu, Ho) suggests that alpha estimation is often a consequence of the strength of signal relative to the noise in the data.

We find that signs and magnitude are estimated well in the case of character dependent models. However, character independent models which are difficult to estimate. We discuss the possible causes of this below in the context of power analysis.

Preliminary power analysis suggests that hOUwie can detect character independence and character dependence when that is the true model. However, when character independent models are heterogeneous (i.e., there is variation in how a continuous trait evolves unlinked to the focal character) hOUwie favored simpler versions of character independent models that does not have additional variation. This is likely a consequence of the simulation-based approach. Even in cases where the tips are informed by the continuous model, the discrete character maps are generated based solely on discrete trait evolution. This means finding an independent mapping which suits the continuous characters is difficult. Other approaches like lasso and tribble et al., may be better suited for this task although they do not explicitly account for the probability of the discrete mapping. A future extension could combine the discrete character methods developed here with the discrete independent models for the best model.

**Discussion**

Even when data is character dependent the character independent model with multiple hidden states can find a high level of support. This makes sense as hidden states could be assigned at tips to match the observed states (albeit with less certainty) and thus should approach the probability of the character dependent model. This serves as a reminder that although we are focusing on a single focal trait, any trait that matches the distribution at the tips would produce identical results and we should be cautious with over interpretation.

Why should we care about jointly modeling discrete and continuous traits? Firstly, this explicitly allows for the traits to talk with each other – a concept formally referred to as mutual information. The explicit modeling of two or more characters is not only biologically more relevant but can improve our modeling inferences. The quality of our modeling inferences also has a direct consequence on how accurate our hypotheses are. For example, if one is interested in testing the presence of convergent evolution of pollination syndromes. The first step to establishing a syndrome is to test for the convergence of the traits involved. This requires that we examine whether the traits have independently associated multiple times through the history of the clade. Modeling discrete characters independently can alter the certainty and the states of ancestral state reconstruction and can completely alter how we perceive the evolution of convergence. For example, if we are interested in how seed size has evolved alongside climate, reconstructing seed size independently may lead to different conclusions than if we modeled them jointly. This is because the distribution of the discrete trait alone does not utilize the information that is shared between climate and seed size. This relationship allows for a more accurate ancestral state reconstruction and thus a more accurate answer to the question of convergent evolution. The flipside of this is if the traits are perfectly correlated then there is not going to be any difference between independently modeling and jointly modeling. This is because change in one variable is matched perfectly by change in the other variable. Here the advantage of joint modeling is simply in the certainty provided in the reconstruction. If such a strong relationship was detected, the uncertainty that would normally be accumulated by independently modeling the variables will be reduced because we now have multiple sources of information suggesting a particular ancestral state.

If we do not model joint probabilities explicitly, we are still able to test against null models of correlated evolution. The issue with this is that it does not leave room for discovery and null hypothesis testing has been criticized elsewhere extensively. Whether traits are correlated is an important initial step in understanding the evolution of syndromes but could also lead to false positives as we have seen in other comparative methods. This is accounted for in hOUwie because it includes entirely hidden state nulls as well as the more generic trivial null hypotheses present in traditional studies. But more exciting than testing null hypotheses is interpreting the parameters from resulting models. Hybrid models which mix variation due to focal and hidden traits open ways for identification of interesting regions of the phylogeny for future studies to assess.

Often we are biased for looking for adaptation in traits that are easily measured, but that does not mean there are not other important aspects of a species’ natural history which is the true adaptive feature. It is possible and likely that not all of the adaptive features of any particular hypothesis are going to be measured. This is where the utility of hidden Markov models is most apparent. It is possible that in some lineages color is the most important feature and in others things other than color are more important. By identifying regions where the association between color and foraging behaviour do not match we are able to generate testable predictions. And highlight areas where we can readily supplement comparative analyses with experimental and observation evidence. The advantage is that we generate hypotheses and outline where in the phylogeny we can go looking.

Another advantage of modeling joint probabilities explicitly is that it allows us to put our models in the context of likelihood. With an explicit likelihood we are able to do model comparison and model averaging. Rather than testing whether our pet hypothesis is better than a trivial null model we can discover the relative power of each model to describe the variation in our dataset. Then we can use that information to inform our model averaging and examine the parameter estimates from our set of models holistically. The upside of this is that we are able to account for model structure uncertainty in addition to parameter estimation uncertainty. Ultimately the use of modeling in general is to describe our data in ways that we cannot do using summary statistics. In part, this means hypothesis testing and those are important steps. However, parameter estimates are comparable across studies and why support for hypotheses is important, it is reminiscent of vote counting in which we examine blankly whether a hypothesis was supported. Instead, parameter estimates are similar to examining effect sizes and are often directly comparable across studies since they are in the same units.

**Figures**

Figure 1: A diagram illustrating how the branch pathways are calculated.

Figure 2: a) The exact datasets we are coding we are coding. b) Our expectation of whether the datasets are either CD, CID, or CID\*. c) the results of our search.

Figure 3: A conceptual diagram describing the possible hOUwie model structures.

Figure 4: The comparison of sampling procedures.

Figure 5: A graphical breakdown of discrete and continuous probabilities. (x-axis could be discrete probability and the y-axis could be the overall probability).

Figure S1: A pictographic representation of all the possible stochastic maps and their respective discrete and continuous probabilities.

A screenshot of a computer

Description automatically generated with low confidence

![Chart

Description automatically generated]()

*![Chart, box and whisker chart

Description automatically generated]()*