# Automatic Discovery of Optimal Discrete Character Models through Regularization

**James D. Boyko**[1][2]

**Affiliations**

[1] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA

[2] Michigan Institute of Data Science, University of Michigan, Ann Arbor, Michigan 48109, USA

## Abstract

Modeling discrete character evolution in a Markovian framework has become the standard in phylogenetic comparative methods. The increasing size and complexity of these models reflects a trend of analyses to include more taxa and more discrete characters. However, as complexity the models increase, so do the number of potential model structures and number of estimable parameters, making it nearly impossible to consider all modeling options for a given dataset. To overcome this issue, I apply a combination of regularization and parameter sharing optimization to models of discrete character evolution. This allows for the automatic searching and optimization across different model structures without user specification. By adding a penalty term to the likelihood function, regularization helps balance model complexity and goodness-of-fit, while also mitigating the risk of over-fitting. I test this framework under several simulation scenarios including hidden rates and multiple discrete characters. The results demonstrate that regularized models outperform traditional unregularized methods, achieving a XX% increase in accuracy for parameter estimation and a XX% reduction in error rates for ancestral state reconstruction. I illustrate the power of regularized models by revisiting several classic examples of ancestral state reconstruction of discrete characters including XX, YY, and ZZ. In all cases I find that the root state and rates of evolution differ markedly from the original study. In general, regularizing comparative methods may provide a useful alternative to standard model selection practices, especially when dealing with increasingly complex macro-evolutionary models. It may be worth considering the perspective that, although a regularized method will never be exactly correct, it is less likely to produce severely misleading results.

**Running Head.** REGULARIZATION IN DISCRETE CHARACTER EVOLUTION MODELS

**Key Words.** Discrete character evolution; Regularization; Ancestral state reconstruction; Machine learning; Phylogenetic comparative methods; hidden Markov model; Bias-variance trade-off

Complex discrete character models are becoming increasingly common in phylogenetic comparative methods (PCMs). These models, which were initially limited to relatively few characters and simple processes, have been expanded in an attempt to incorporate various biological processes and sources of variation. Complexity has been introduced to these models through correlated character evolution (Pagel, 1994), hidden rate variation (Beaulieu et al., 2013, Boyko and Beaulieu 2021), state-dependent speciation and extinction (SSE) (Maddison et al., 2007; FitzJohn, 2012, Beaulieu et al. 2016), and the incorporation of continuous character information (May and Moore, 2020; Boyko et al., 2023).

This increasing complexity has also led to higher generality. For example, hidden Markov models which were initially introduced to model two binary characters (Beaulieu et al. 2013), have been expanded to allow for any number of characters, observed states, and hidden states (Boyko and Beaulieu 2021). While this additional flexibility allows for customization of the phylogenetic comparative method to the system at hand, the large model space can make it challenging for biologists to find the most appropriate model for their particular dataset. The typical approach, multi-model inference framework (Burnham & Anderson, 2002), has a biologist decide on a set of potentially realistic models. The relative support for each model is then evaluated based on their fit to the data using information criteria such as Akaike Information Criterion (AIC). This approach is powerful because it allows for model averaging, where inferences are made based on a weighted average of the models' included in the set (Burnham & Anderson, 2002). However, the effectiveness of multi-model inference relies heavily on the appropriateness of the model set. Several PCMs have been criticized for high false positive rates due to the exclusion of the "correct" null hypothesis (Rabosky & Goldberg, 2015; Davis et al., 2016). Although these criticisms have been addressed by introducing new model structures to serve as better null hypotheses (Beaulieu & O'Meara, 2016; Boyko & Beaulieu, 2021), the problem, when recast as a failure to include a complete model set (Boyko and Beaulieu 2023), suggests that there is still a vast space of unexplored model structures.

Often the model set chosen for discrete character evolution are nested, with the difference being which parameters are variable and which are fixed. For example, for a two character binary state dataset, the difference between a correlated model (k=8) and the independent model (k=4) is whether changes in the focal character depend on the state of the background character (Pagel and Meade 2016). In the independent model parameters representing this dependency process are fixed to be equal, while in the correlated model they are freely estimated (Pagel and Meade 2016, Boyko and Beaulieu 2023). In this way, one can think of discrete models adding complexity by adding parameters to increase "biological realism" and represent new processes or relationships between variables. This increasing model complexity then leads to another, more technical, challenge. As the number of estimable parameters grows it reaches a point where the number of parameters (k) approaches the number of taxa (N). However, the rate at which the number of parameters increase often surpasses the rate at which information can be gained through data (Felsenstein 2012). In the case of correlated discrete character models, this is because each additional character requires considering its relationship to all other traits. For instance, the most complex discrete model for a single binary character has 2 parameters, while the most complex model with two binary characters has 12 parameters, and the most complex model with three binary characters has 30 parameters. In each instance we have added a single character, but because we must consider the new character's relationship to all other existing characters, the model complexity (as measured by the number of parameters) outpaces the potential information gained from the new data. This is problematic because although likelihood-based methods are consistent estimators when N >> k, their performance deteriorates as models become more parameter-rich, potentially leading to unreliable and biased parameter estimates (Huelsenbeck et al., 2001). However, for complex models with a finite  data, it is unlikely that all parameters will be essential or necessary to best model the data (Gelman et al., 2013; Lemey et al., 2009).

To address the challenges associated with increasing discrete model complexity, I introduce regularized discrete character models and a method to optimize parameter sharing structures. Regularization techniques, such as L1 (lasso) and L2 (ridge), constrain the magnitude of parameter estimates and encourage simpler, more generalizable models (Tibshirani, 1996; Hoerl & Kennard, 1970). By incorporating regularization, we can balance model complexity and goodness-of-fit, reducing the risk of overfitting and improving the stability of parameter estimates. Furthermore, the automatic evaluation of different parameter sharing structures can help mitigate model complexity and improve interpretability. I implement this framework within corHMM (Boyko and Beaulieu 2021) as corHMMDredge. To test the dredge framework, I conduct an extensive set of simulations to explore the bias-variance trade-off associated with regularization. Under regularization it is expected that models will have increased generality and lower variance. I test these expectations by examining regularized model's accuracy of parameter estimates and predictions for ancestral states. Additionally, I perform a more detailed simulation test on a subset of historically important discrete models, such as the precursor model (Mazarri et al. 2012) and correlated character evolution model (Pagel 1999, Pagel and Meade 2016; Boyko and Beaulieu 2021), to ensure that the dredge framework has acceptable false positive and negative rates. Finally, I revisit several empirical case studies. The empirical case studies are not chosen to overtly challenge existing results. Rather, they are chosen such that the necessary complexity for modeling increases. To that end I begin with examining a simple binary character in the form of oviparity and viviparity in reptiles. I then examine how limb loss in squamates. Finally I revisit the ancestral angiosperm flower and reconstruct the phyllotaxy of the first flower.

## Methods

*Regularization and Parameter sharing*

The likelihood of a discrete character model with its underlying framework as a continuous time Markov chain (CTMC) is calculated as $L=P(D|Q,\Phi)$, which is the probability of observing the data (D) given an instantaneous rate matrix (Q) and a phylogeny with a fixed topology and set of branch lengths (Φ). The data (D) consist of the observed character states (S) at the tips of the phylogeny, while the rate matrix (Q) contains the rates of character state transitions ($q_{ij}$). The likelihood function is then computed by integrating the product of transition probabilities along the branches of the phylogeny (more detailed descriptions can be found in Pagel 1994; Lewis 2001; Felsenstein 2004; Yang 2006). Though not mathematically necessary, it is also useful to consider a mapping matrix (M) which gives the structure of the discrete model by indicating which transition rates are estimated and/or fixed to be equal. For example, if we consider a simple binary character with states 1 and 2 the instantaneous rate matrix is given by $Q=\begin{bmatrix} -q_{12} & q_{12} \\ q_{21} & -q_{21} \end{bmatrix}$. However, the mapping matrix can specify several alternative model structures. If $M_{er}=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, then $q_{12}=q_{21}$ and we have what is commonly known as the "equal rates" model. An "all rates different" model can be indicated by $M_{ard}=\begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix}$, where $q_{12}\neq q_{21}$ and a unidirectional model can be constructed via $M_{dir}=\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, where $q_{21}=0$. To determine which model structure is optimal (best balances goodness-of-fit and complexity) for the dataset, we can compute the maximum likelihood estimate for each model structure and compare them using a information criterion such as AIC (Akaike 1974; Burnham and Anderson 2002).

This framework has been successfully applied for many years within PCMs (O'Meara 2006), but growing model complexity is making it untenable for users to define all possible relevant model structures or for method developers to construct a complete set of default models. As such, regularization techniques may be a necessary step in discovering optimal model structures for large and complex datasets. I incorporate a regularization approach analogous to *l1* regularization (Hastie et al. 2015) into the likelihood estimation. Specifically the regularized likelihood is defined as:

$$L_{reg}(D|Q,\Phi;\lambda)=L(D|Q,\Phi)-\lambda\, tr(Q),$$

where $L(D|Q,\Phi)$ denotes the standard likelihood, the term $tr(Q)$ is the trace of the instantaneous rate matrix Q, and $\lambda$ is a hyper-parameter that adjusts the severity of the penalty, ranging from 0 (no regularization) to 1 (full penalization). This penalization scheme is particularly effective at removing unnecessary parameters such that the dredge algorithm can efficiently move from more complex to simpler models (for example, from $M_{ard}$ to $M_{dir}$ by dropping parameter 1 of $M_{ard}$ ). It is important to note that $\lambda$ cannot be jointly estimated alongside the parameters via the maximized likelihood (Clavel et al. 2019). Instead a form of cross-validation is necessary to tune $\lambda$ (see *Phylogenetically informed k-fold cross validation*).

The second component of the dredge algorithm involves a method to explore and assess alternative model structures that utilize parameter sharing (e.g., is $M_{ard}$ or $M_{er}$ optimal). To implement parameter sharing, I adopt a strategy where the most complex possible model for $C$ characters and $S$ states is first fit to the data by maximizing $L_{reg}$ . Next, any parameters which were estimated to be zero are omitted from the model and the most similar pair of parameters within regularized model are equated. If equating these parameters results in a $\Delta AIC$ of less than 2 units (or another specified threshold) from the current best model, the next most similar pair of parameters are equated and the process continues. Note that information criterion are calculated using $L_{reg}$ rather than $L$ . Importantly, this algorithm also readily applies to hidden Markov models (Boyko and Beaulieu 2021). To evaluate whether additional rate variation is necessary through the presence of a hidden state, corHMMDredge will first find the optimal model for a single rate category model. Upon increasing the number of rate categories, the most complex possible two rate class HMM is evaluated by maximizing $L_{reg}$ . At least one round of parameter sharing is tested before the HMM is compared to the current best model (including single rate class models). If this results in a $\Delta AIC$ of less than 2 units, the search will continue and may increase rate categories until the AIC no longer improves. In practice, another stopping can be useful for datasets that contain little to no evidence of hidden rate variation. In these cases, the first HMM evaluated will drop all parameters associated with an additional rate class. Of course, that means it is no longer necessary to evaluate alternative parameter sharing structures as the HMM will have collapsed to the initial most complex model with one rate class.

When defining the initial model it becomes important to distinguish between characters and states. Mathematically, it is unnecessary to distinguish between the two since the model is entirely defined by the $S$ states since $S=\prod_{m=1}^{C} s_m$ where $s_m$ is the number of states for each character $m$ . Nonetheless, biologically it is often useful to think of characters as independently evolving units that can take on some number of states. This distinction is particular relevant when specifying $M$ to avoid "dual-transitions" (Pagel 1994). These are transitions in which the states of more than one character change in an instant in time (Pagel and Meade 2006). As such, the most complex model is not defined by $k=S\times(S-1)$ parameters (which would be the case if all possible transitions were allowed). Rather, it is defined by $k=(\sum_{m=1}^{C}(s_m-1))\times S$ for a single rate class.

*Phylogenetically informed k-fold cross validation*

To optimize $\lambda$ I choose to employ a phylogenetically informed k-fold cross-validation procedure. K-fold cross-validation is a common way to optimize $\lambda$ when using regularization (Hastie et al. 2009). It involves dividing the dataset into k separate subsets (or folds). The model is trained on k-1 of these folds with the remaining fold used as the test set to evaluate the models performance. This process is repeated k times, with each of the k folds used exactly once as the test set. The results of these k tests is then averaged to produce a single estimate (Kohavi, 1995). In the context of a discrete character model the metric I choose for evaluation is reconstruction of hidden tip state values. Each species within a k-fold is coded as having an unknown state and a reconstruction based on the fitted model is used. The score for a fold is then determined by the average Jensen-Shannon divergence between the predicted and actual tip states. Specifically, $score_k=\frac{1}{N_k}\sum_{i=1}^{N_k} D(L_i,\hat{L}_i)$ where $N_k$ is the number of tips in the k-th fold, $L_i$ is the observed likelihoods at the tips in the k-th fold, $\hat{L}_i$ are the predicted states for the tips in the k-th fold and $D$ is the Jensen-Shannon (JS) divergence (or any form of divergence). The k scores are then averaged over all folds to obtain a single score for the model. The JS divergence is chosen because it a symmetric and bounded measure of difference between predicted and observed probability distributions while being robust to zero probabilities. An advantage of using tip values to measure the model fit is that the phylogenetic structure remains in tact throughout the entire cross-validation procedure because no tips need to be explicitly dropped when fitting the model. They only need to be set to an unknown value. The samples within a given k-fold are chosen without replacement and with probabilities equal to $w=(I^T C^{-1})\vec{1}$ (Rohlf 2001) to ensure phylogenetically even sampling between the k-folds.

*Simulation*

*test one*

Relatively simple simultion set up to assess the magnitude of the bias as well as examine the variation of the standard model. I will use RMSE which can break into bias and variance.

*test two*

A more focused simulation setup focusing on correlated character evolution. Hidden rate models. And ordered state models. Here I am more interested in whether the chosen model strucutre is in line with the simulating strucutre. To assess wheter we know, I will also fit the standard model.

*Empirical case studies*

Oviparity and vivparity. Limb loss in lizards. Floral morphology.

**Discussion**

*Revisiting classic empirical studies*

We find differences? Sometimes? All the time?

*The value of more traits*

Correlated character evolution allows for information from one character to influence the rate of change of another. This means that if there are areas of the phylogeny or the ancestral state reconstruction where one would be uncertain about a trait if it were modeled independently, when accounting for correlated character evolution we are given access to the shared information (formally mutual information) of the two characters which improves inference (Boyko and Beaulieu 2021). Of course, correlation between discrete characters is a hypothesis rather than a given, and as such it must first be tested. These tests need to be conducted in such a way that the null hypothesis includes character independent rate variation (Boyko and Beaulieu 2023) and thus there are more parameters than the standard independent models (Pagel 1994).

*The cost of regularization...*

They introduce a bias, but we get a much lower variance. How much do low rates really matter? Like we’re not talking about a major bias in most cases, in fact it’s often low and doesn’t seem to have much of a negative impact. But, we get the benefits of the lower varianace.

*Model comparison versus dredge...*

Our prior knowledge of macroevolutionary events is rather limited even with fossils. It’s often not clear which model structures should or should not be included. Though there are certainly cases where we are interested in testing particular hypotheses and that can still be done. It may be worth considering what the optimal model structure could be. What are the interpretations of that model?

## Conclusion

Determining which models to test in an empirical setting has the potential to be an incredibly valuable process in which biologists structure their PCM to meet the needs of their system. However, if only default models are considered, there is a risk of overlooking potentially reasonable and important model structures. This issue is exacerbated by the fact that the knowledge necessary to manipulate PCMs is often difficult to acquire and may be hidden in highly technical texts, making it challenging for biologists to explore and customize model sets effectively (Cooper et al., 2016). It is evident that a comprehensive model set is important for trustworthy inferences in comparative biology (Rabosky and Goldberg, Beaulieu and O’Meara, Maddfitz, Boyko and Beaulieu). Nonetheless, the growing complexity of discrete character models makes it challenging for users to determine which models are potentially realistic and important to consider. The dredge framework may help alleviate this burden, enabling biologists to focus on model interpretations and potentially discover new model structures that may have otherwise never been considered.