A brief introduction to character independent HMMs when testing for correlation

James D. Boyko

31/03/2022

Background

This document is a practical guide to creating a general model set suitable to testing for correlation between two (or more) discrete characters and is meant to accompany the work of Boyko and Beaulieu (in prep.). Note: this document will not discuss the collapsed model since what is presented below should be more generally useful to readers.

A character is defined as a general property of the organism that can have alternative states. For example, flower color would be a character but the specific color (e.g., red, white, orange) would be alternative states of that character. States are generally mutually exclusive (an individual cannot have both red and white flowers), but there are ways to model polymorphism within a species. Testing for correlation means we are interested in whether the specific state of a character influences the states of another character and vice versa. For example, whether red or white flowers are more likely to be unscented or scented? Our characters would be flower color and scent and each has two states (flower color can be red or white and scent can be absent or present).

Typically comparative analyses which aimed to test for correlation between discrete characters have used the framework of Pagel (1994). This means fitting two models: one in which change in character X and character Y are dependent on each other and one in which the changes in each character are independent of one another. To determine whether a model says the characters "depend" on one another let us first define our model for two binary characters.

$$F = flower\ color\ where,\ 0 = red\ \&\ 1 = white$$

$$S = scent\ where,\ 0 = unscented\ \&\ 1 = scented$$

$$\begin{bmatrix} - & F_0S_0 \rightarrow F_0S_1 & F_0S_0 \rightarrow F_1S_0 & - \\ - & - & - & - \\ - & - & - & - \\ - & - & - & - \end{bmatrix}$$

Each element of the matrix above represents a potential transition between the combined states of our characters. Shown above is the first row which describes the possible transitions for a red flower (F_0) that is currently unscented (S_0). It can either transition to being scented (S_1) or become a white flower (F_1). Finally, there is also the possibility of a direct transition to being a white and scented flower, but since we are interested in character dependence we need to test whether the most likely pathways between states are in line with character dependence or independence and the presence of dual transitions would make that comparison impossible (see Boyko and Beaulieu, in prep.). The full matrix, with all possible transitions looks like this:

$$F = flower\ color\ where,\ 0 = red\ \&\ 1 = white$$

$$S = scent\ where,\ 0 = \mathit{unscented}\ \&\ 1 = \mathit{scented}$$

$$\begin{bmatrix} - & F_0S_0 \to F_0S_1 & F_0S_0 \to F_1S_0 & - \\ F_0S_1 \to F_0S_0 & - & - & F_0S_1 \to F_1S_1 \\ F_1S_0 \to F_0S_0 & - & - & F_1S_0 \to F_1S_1 \\ - & F_1S_1 \to F_0S_1 & F_1S_1 \to F_1S_0 & - \end{bmatrix}$$

Testing for dependence in the most general sense comes down to determining whether transitions in a focal character depend on the state of the background character. For example, if the transition from red to white flower depends on whether it is scented or unscented, we expect that transitions in our focal character (flower color) will depend on the state of our background character (scent). In the matrix below, we will focus on whether two transition rates are equal or different:

$$F = flower\ color\ where,\ 0 = red\ \&\ 1 = white$$

$$S = scent\ where,\ 0 = \mathit{unscented}\ \&\ 1 = \mathit{scented}$$

$$\begin{bmatrix} - & - & F_0S_0 \to F_1S_0 & - \\ - & - & - & F_0S_1 \to F_1S_1 \\ - & - & - & - \\ - & - & - & - \end{bmatrix}$$

The only state changing in these two transitions is flower color ($F_0 \to F_1$), but the background state of the other character is different (S_0 or S_1). If these two transition rates are equal to one another then the background state has no influence on the focal character, and we can say that the transitions between the states of flower color do not depend on whether the lineage is unscented. However, if they are unequal then there is evidence of dependence (or correlation). Furthermore, we may even suspect that flowers which are unscented are more likely to transition to a red state since they will need to attract pollinators with visual cues.

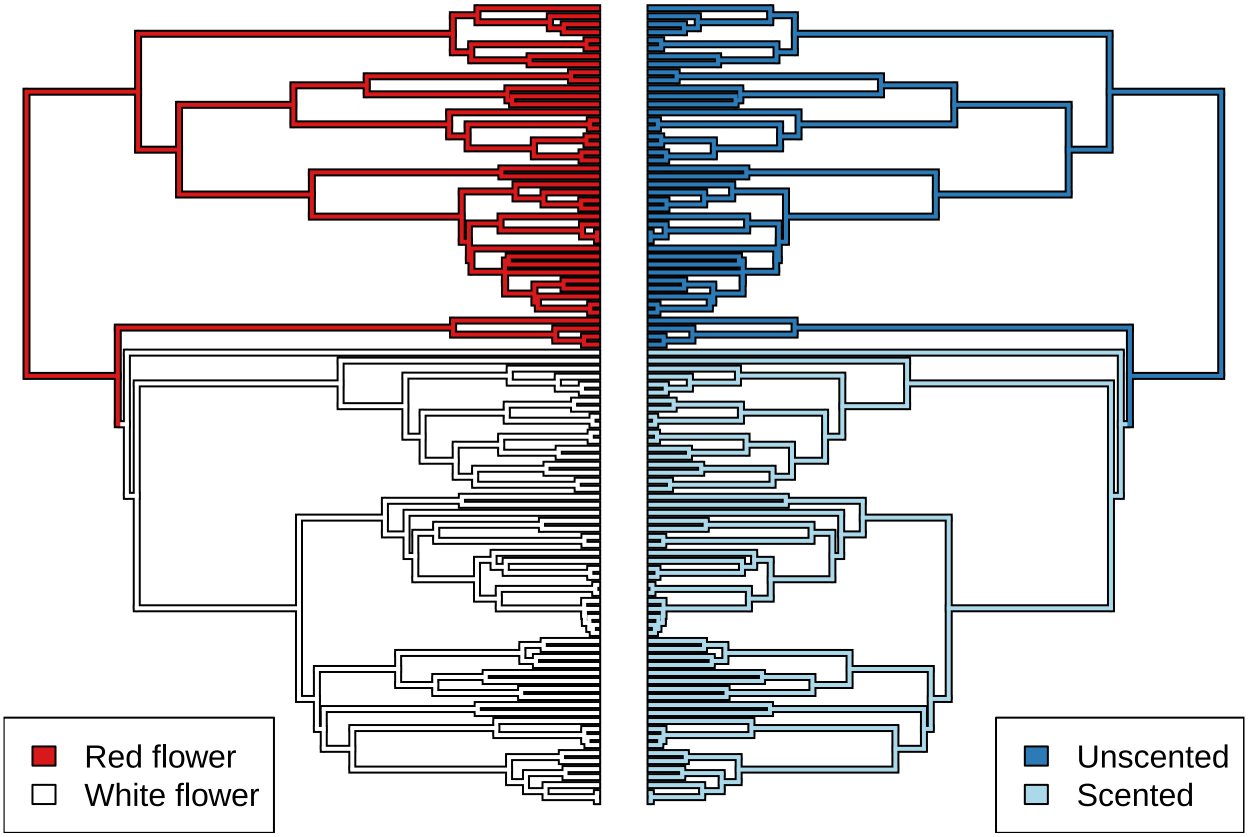
What's the problem?

The problem was described by Maddison and FitzJohn (2015) and can be summarized as PCMs may be biased towards finding correlation. Specifically, they demonstrated that the most widely used phylogenetic method for detecting correlated evolution between categorical characters (Pagel 1994), almost always indicates strong evidence of correlation when singular events deep in time can account for the co-distribution of two characters.

For example, Razoraceae is a fictitious clade of flowering plant originating in the Eocene. The family itself only has about 100 species, but is well-known for a particularly intriguing genus called *Ranolis*. The *Ranolis* genus comprises approximately 40 to 60% of Razoraceae and is well known for their scented white flowers (all other members of Razoraceae have unscented red flowers). The question on everyones mind?

Are flower color and flower scent correlated on a macroevolutionary scale?

The scenario described above is meant to setup a system in which our test of correlation relies on the comparison of two characters that emerged at the same time and are both synapomorphies for the same focal clade. On a phylogeny, a potential distribution of discrete traits that matches this scenario would look like this:



The issue is that, when applied to this scenario (Darwin's scenario ala Maddison and FitzJohn 2015), commonly used comparative methods (Pagel 1994) will almost always indicate strong evidence of correlation despite the dependent relationship arising from little more than a single event deep in time. Although several species have the combination of characters that make it seem like a strong correlation (many species with unscented red flowers or scented white flowers), they are mainly derived from a single clade and thus the observations are not independent. However, because this is in a phylogenetic context, one would hope that comparative methods correct for this obvious bias, but they did not.

Testing for correlation

What we demonstrate in Boyko and Beaulieu (in prep) is that this bias is not a consequence of the methods failing to correctly account for phylogeny, but rather an incomplete set of models being tested. Specifically, we are advocating for the inclusion of a model with character-independent rate heterogeneity.

First, we examine the typical Pagel (1994) framework, which has been the source of the biased correlation. We will be using the functions within the updated version of corHMM (Boyko and Beaulieu 2021) to create our models.

```r
dependent_model_matrix <- getStateMat4Dat(dat, collapse = FALSE)$rate.mat
independent_model_matrix <- getStateMat4Dat(dat, collapse = FALSE,
    indep = TRUE)$rate.mat
# data legend
getStateMat4Dat(dat, collapse = FALSE)$legend
```

```
##     1     2     3     4
## "0_0" "0_1" "1_0" "1_1"
```

```
# the dependent/ correlated model
dependent_model_matrix
```

```
##     (1) (2) (3) (4)
## (1)   0   3   5   0
## (2)   1   0   0   7
## (3)   2   0   0   8
## (4)   0   4   6   0
```

```
# the independent model
independent_model_matrix
```

```
##     (1) (2) (3) (4)
## (1)   0   4   2   0
## (2)   3   0   0   2
## (3)   1   0   0   4
## (4)   0   1   3   0
```

In this case, 1 = 00 is red unscented flowers, 2 = 01 is red scented flowers, 3 = 10 is white unscented flowers, and 4 = 11 is white scented flowers. If you examine the matrices you will see that under a independent model, changes in a focal character do not depend on the background state of the other character. For example, parameter 2 is equal for transitions from (1) → (3) and (2) → (4) which represents the transition for red unscented flowers to white unscented flowers and red scented flowers to white scented flowers. The only character changing in this transition is flower color (from red to white) and by setting these parameters to be equal, we are saying the state of the background character (presence or abscence of scent) does not influence the rate of that transition.

```
dependent_model_fit <- corHMM(phy, dat, rate.cat = 1, rate.mat = dependent_model_matrix,
    node.states = "none", collapse = FALSE)
```

```
## State distribution in data:
## States:  1   4
## Counts:  43  57
## Beginning thorough optimization search -- performing 0 random restarts
```

```
independent_model_fit <- corHMM(phy, dat, rate.cat = 1, rate.mat = independent_model_matrix,
    node.states = "none", collapse = FALSE)
```

```
## State distribution in data:
## States:  1   4
## Counts:  43  57
## Beginning thorough optimization search -- performing 0 random restarts
```

```
c(AIC_dependent_model = dependent_model_fit$AIC, AIC_independent_model = independent_model_fit$AIC)
```

```
##   AIC_dependent_model AIC_independent_model
##              32.43798              40.87499
```

The AIC of the dependent model being 8 lower than the AIC of the independent model represents fairly substantial evidence for correlation when comparing this model set.

Next, we create a hidden Markov independent model (HMIM) and add it to the set. The HMIM is simply the independent model matrix, but with the allowance of rate heterogeneity. We again use the updated corHMM functions to create the HMM.

```
# this is a list of matrices that will represent the
# alternative hidden states (see Boyko and Beaulieu 2021 and
# the corHMM v2.0 vignette for more detail)
state_mats <- list(independent_model_matrix, independent_model_matrix)
```

```
# the matrix which describes transitions between rate classes
rate_cat_mat <- equateStateMatPars(getRateCatMat(2), c(1, 2))
# the HMIM
HMIM_matrix <- getFullMat(state_mats, rate_cat_mat)
HMIM_matrix
```

```
##        (1,R1) (2,R1) (3,R1) (4,R1) (1,R2) (2,R2) (3,R2) (4,R2)
## (1,R1)      0      4      2      0      9      0      0      0
## (2,R1)      3      0      0      2      0      9      0      0
## (3,R1)      1      0      0      4      0      0      9      0
## (4,R1)      0      1      3      0      0      0      0      9
## (1,R2)      9      0      0      0      0      8      6      0
## (2,R2)      0      9      0      0      7      0      0      6
## (3,R2)      0      0      9      0      5      0      0      8
## (4,R2)      0      0      0      9      0      5      7      0
```

We can add this model to the model set, but with two rate classes instead of one (rate.cat = 1).

```
HMIM_fit <- corHMM(phy, dat, rate.cat = 2, rate.mat = HMIM_matrix,
    node.states = "none", collapse = FALSE)
```

```
## State distribution in data:
## States:  1   4
## Counts:  43  57
## Beginning thorough optimization search -- performing 0 random restarts
```

```
c(AIC_dependent_model = dependent_model_fit$AIC, AIC_independent_model = independent_model_fit$AIC,
    AIC_HMIM = HMIM_fit$AIC)
```

```
##   AIC_dependent_model AIC_independent_model              AIC_HMIM
##              32.43798              40.87499              33.98693
```

The HMIM and dependent model have much closer AICs indicating little evidence to distinguish between the two. This lets us know that there is evidence of rate heterogeneity.

Finally, ideally we would not be finding even this much evidence of correlation when examining two synapomorphies. Although the above model is the most general (as most biologists would not examine whether two synapomorphies are correlated on a macroevolutionary scale), our purposed solution isn't satisfying unless it can resolve the problem entirely. Thus, we introduced a simplified version of the independent model and added rate heterogeneity. Note: we also tested several other simplified models (such as a simplified correlated model) and added rate heterogeneity to added hidden states to all of the reported models.

```
simplified_independent_model_matrix <- equateStateMatPars(independent_model_matrix,
    list(c(4, 2), c(1, 3)))
simplified_independent_model_matrix
```

```
##     (1) (2) (3) (4)
## (1)  NA   2   2  NA
## (2)   1  NA  NA   2
## (3)   1  NA  NA   2
## (4)  NA   1   1  NA
```

```
simplified_HMIM_matrix <- getFullMat(list(simplified_independent_model_matrix,
    simplified_independent_model_matrix), rate_cat_mat)
simplified_HMIM_matrix
```

```
##        (1,R1) (2,R1) (3,R1) (4,R1) (1,R2) (2,R2) (3,R2) (4,R2)
## (1,R1)      0      2      2      0      5      0      0      0
## (2,R1)      1      0      0      2      0      5      0      0
## (3,R1)      1      0      0      2      0      0      5      0
## (4,R1)      0      1      1      0      0      0      0      5
## (1,R2)      5      0      0      0      0      4      4      0
## (2,R2)      0      5      0      0      3      0      0      4
## (3,R2)      0      0      5      0      3      0      0      4
## (4,R2)      0      0      0      5      0      3      3      0
```

We can then add each of these models to the set, but we'll just focus on the simplified HMIM.

```
simplified_HMIM_fit <- corHMM(phy, dat, rate.cat = 2, rate.mat = simplified_HMIM_matrix,
    node.states = "none", collapse = FALSE)
```

```
## State distribution in data:
## States:  1   4
## Counts:  43  57
## Beginning thorough optimization search -- performing 0 random restarts
```

```
c(AIC_dependent_model = dependent_model_fit$AIC, AIC_independent_model = independent_model_fit$AIC,
    AIC_HMIM = HMIM_fit$AIC, AIC_simplified_HMIM = simplified_HMIM_fit$AIC)
```

```
##   AIC_dependent_model AIC_independent_model              AIC_HMIM
##              32.43798              40.87499              33.98693
##   AIC_simplified_HMIM
##              25.98693
```

This model has substantial evidence for this kind of dataset. This result is also somewhat intuitive as this model most closely resembles the information available in the data at hand. Specifically, it treats the two characters identically, which since they have identical distributions, makes sense. Next, this model allows for rate heterogeneity, which there is signal of because of the stasis outside of the focal clade and, within the focal clade, a lot of transitions to different state combinations.

There is a great deal more that can be done with these models than simply testing even the set of models presented here. For example, it is straightforward to have different versions of dependent models, independent models, and even mixes between the two. Of course, there is a danger to exhaustively exploring this model space without a priori reason to. Running tests until we find an answer we're looking for isn't good practice. But, clearly there is also danger in underexploring the model space and that our conclusions may be biased by underexploration as well.

I will also mention that I created a function in corHMM which has not yet been exported, but is available as an internal function with the github version. corHMM:::fitCorrelationTest() will run a set of comparisons with these models automatically. ?corHMM:::fitCorrelationTest() for more information.

If you have any questions feel free to email me at jboyko [at] uark [dot] edu

References

Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc. R. Soc. B: Biol. Sci. 255:37–45.

Pagel M., Meade A. 2006. Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. Am. Nat. 167:808–825.

Maddison W.P., FitzJohn R.G. 2015. The Unsolved Challenge to Phylogenetic Correlation Tests for Categorical Characters. Syst Biol. 64:127–136.

Boyko J.D., Beaulieu J.M. 2021. Generalized hidden Markov models for phylogenetic comparative datasets. Methods Ecol Evol. 12:468–478.