Updates on modeling species given their phylogenetic history in Stan

Background & Question: A long line of evidence suggests phenological 'sensitivity' in plants are likely driven in part by shared evolutionary history. As an example, pretend you're a tree who comes from a clade that for some reason in the past evolved to be responsive to variable spring temperatures. To be, effectively highly sensitive (you probably have high forcing, low chilling and low photoperiod cues to make this happen). You might move to a new climate where being less sensitive would be good, but because of your evolutionary history you're still pretty responsive. This would show up in models as a phylogenetic signal in your 'sensitivity' and if you drilled down, it should show up in your forcing. We want to know if there is evidence of this in plants as it would help us predict something about species for which we have no data and it would impose a constraint on how quickly we think these cues can evolve in the future. Lots of people have been interested in this before, but all work has used observational data.

So, given experimental data (hopefully removing confounding effects of geography on phenological dates) how much does phylogeny predict phenology? This question has some interesting embedded problems (or we could call them 'opportunities') as we both the cues which inform the 'sensitivity' and can use these to estimate phenological dates, and then assess the phylogenetic signal in dates, if we wanted—but now we're gotten perhaps very circular.

Common modeling approaches: There are two really common approaches to questions like this: calculate Pagel's λ on a trait, or correct for phylogenetic correlation in residuals in a typical analysis (PGLS) and analyse that correlation along the way (also called λ). These are **not** necessarily identical: one is evaluating the correlation structure of a trait and one is calculating the correlation structure of residuals. Neither one of these is what we want because they are not the model we think is at play (see below) and because both rely on one trait value per species, but I think I should understand them to get anywhere. So let's do the PGLS, which I am copying from the second edition of *Statistical Rethinking*.

$$P \sim MVN(\mu, S) \tag{1}$$

$$\mu_i = \alpha + \beta * x_i \tag{2}$$

μ is a usual linear model. P is a vector of phenological dates (one per species), and S is a covariance matrix with as many rows and columns as species. In ordinary regression this takes the form:

$$S = \sigma^2 I \tag{3}$$

where I is just an identity matrix (all 1s) so we can ignore it. In PGLS we replace S with the phylogenetic covariance matrix (Σ). **You have to make sure of a few things**: the phylogeny must go in as correlation matrix (this makes the diagonals 1s and the off-diagonals the correlation across species due to evolutionary history) and make sure the rows and columns are in the same order as the species will be ordered numerically. One issue with this model is that (I think) It

forces the correlation structure you give it—it does not adjust the correlation structure at all; some iterations of PGLS in certain R packages do this and it is quite an important addition to avoid Type I errors. I believe this just involves estimating a value to multiply the matix by such that:

$$S = \sigma^2(\Sigma * \lambda) \tag{4}$$

So many models set λ to 1, but it's definitely best to ask the model to estimate it (rather than assume the phylogenetic correlation strutcure if how you want to structure the residuals).

PMM (phylogenetic mixed model): Here's my understanding:

$$y = \alpha + \beta x + a + e \tag{5}$$
$$a \sim normal(0, \sigma_P^2 \Sigma) \tag{6}$$
$$e \sim normal(0, \sigma_R^2 I) \tag{7}$$
$$\text{PGLS: } y \sim normal(\alpha + \beta x, \sigma_P^2 \Sigma) \tag{8}$$

... where α and β, respectively, are the intercept and the slope for the co-factor x, a is the phylogenetic random effect, and e is the residual error. Now, the two last terms are assumed to be normally distributed, with Σ as a phylogenetic correlation matrix , I stands for the relevant identity matrix. Our model, thus, assumes that phylogenetic effects are correlated according to the phylogenetic correlation matrix Σ. Note also that our model is estimating two variances: V_P is the variance of the phylogenetic effect and V_R is the residual error (environment effects, intraspecific variance, measurement error, etc.). [This text here is copied from Chapter 11: General Quantitative Genetic Methods for Comparative Biology, by Villemereuil & Nakagawa.]

A big difference often pointed out about PMM versus PGLS is that PGLS does not allow for non-phylogenetically structured error (but I think it sort of does once you scale the phylogenetic effect by λ, no?) and the PMM explicitly models other sources of error through e, and then in the PMM the strength of the phylogenetic effect is measured as:

$$\lambda = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_R^2} \tag{9}$$

which is equivalent to just saying 'what proportion of the variance is due to phylogeny?' (This model has only a and e, while the Joly et al. has a, e and c.)

The multiple-values-per-species version of this just involves within-group centering (I think, need to check), which I don't see as necessary.

One interesting thing I noticed about this model is that it estimates a **separate** intercept for the $y = \alpha + \beta x$ part of the model and I wonder if we need this? (Though the more I think on

this, who is to say what is an intercept versus error in this model? They are both additive terms.)

My underlying model: I think the phylogeny is a better option for partial pooling that treating species as unordered categories. I would ideally like a model that fits mainly the phylogeny, but also allows me to assess the strength of the phylogeny in the analysis. I used to sort of doubt we'd be able to manage this though, so I was (am?) searching for the best alternative option.

But just for fun, what is wrong with this model?

$$y = \alpha_0 + \alpha + \beta x + e \tag{10}$$
$$\alpha \sim MVN(0, \sigma^2_{P\alpha}) \tag{11}$$
$$\beta \sim MVN(0, \sigma^2_{P\beta}) \tag{12}$$
$$e \sim normal(0, \sigma^2_y) \tag{13}$$
$$\sigma^2_{P\alpha} = \alpha_\alpha + \lambda_\alpha * \Sigma \tag{14}$$
$$\sigma^2_{P\beta} = \alpha_\beta + \lambda_\beta * \Sigma \tag{15}$$

Where α_0 is a grand mean and species-level intercepts are partially pooled by phylogeny, scaled by λ_α and slopes are also are partially pooled by phylogeny, scaled by λ_β, and there is some residual error σ_y. Again, I ask what's wrong with this model?

What we have tried: So far, for our phenology project, we have tried the following:

1. Fit the models with traditional partial pooling then estimate Pagel's λ on species-level estimates of forcing, chilling, photoperiod. **What's wrong with this?** We partially pool based on no structure among species (species with noisy and/or low sample size data will pool towards data-rich, low-noise species) so we could obliterate any signal. Plus I am not sure how to feel about applying two different versions of species across analyses.

2. Try to fit PMM with brms. **What's wrong with this?** I have no idea what brms is doing really. I also think it will struggle to fit the model we want (phylo on slope and intercept) and it seems weird to me to partially pool by unordered species on the slopes but not intercept ... so if we like this model maybe we just try hard with regularizing priors or such to make it fit?

3. Wrote my own Stan code that is not yet working, as we don't have good fake data to test it. But maybe I should work more on this?

Update on 1 June 2020

I have started work on some models that try to put the phylogeny on the slopes (which is what I want). Check out `ubermini.stan` and related R code. Here's the main breakthrough from today, which is just me learning math.

The money code is this line I stole from Will Pearse's code:

```
b_force~ multi_normal(rep_vector(0,n_sp),
diag_matrix(rep_vector(null_interceptsb, n_sp)) + lam_interceptsb*Vphy;
```

It says that my vector of slopes is multinormal centered around 0 (why zero? That's how Gaussian processes work, it somehow gets the 'centering' if you will from the $y \sim normal(\hat{y}, \sigma_y)$ bit of the model) and the variance should be the within-species variance on the diagonal, and the between-species variance on the off-diagonals.

To go into painful detail (for me, if no one else) ... `diag_matrix` is taking `null_interceptsb` and making that an n x n matrix (where n is species number), so `null_interceptsb` is on that diagonal (0s on the off-diagonals). Next, the code says to add this to the scaled (by `lam_interceptsb`) n by n `Vphy` matrix. In the end the diagonal will be `null_interceptsb` + `lam_interceptsb` ... and the off-diaganols will be `lam_interceptsb*Vphy`. (Note that given this formula there is a within-species variance that is the same for all species.) Just to show it here ...

Let α be `null_interceptsb` and λ be `lam_interceptsb`.

`b_force` =

$$\begin{bmatrix} \alpha + \lambda * Vphy & \cdots & \lambda * Vphy \\ & \ddots & \\ \lambda * Vphy & \cdots & \alpha + \lambda * Vphy \end{bmatrix} \tag{16}$$

Remember that $Vphy$ is currently set to have 1s down the diagonal (`Vphy=vcv(phylo, corr=TRUE)`) which means we can simplify to this:

`b_force` =

$$\begin{bmatrix} \alpha + \lambda & \cdots & \lambda * Vphy \\ & \ddots & \\ \lambda * Vphy & \cdots & \alpha + \lambda \end{bmatrix} \tag{17}$$

So here's the rub—those ones mean λ gets stuck on the diagonals and wouldn't it be better if the diagonals in the $Vphy$ were 0s so then we get:

`b_force` =

$$\begin{bmatrix} \alpha & \cdots & \lambda * Vphy \\ & \ddots & \\ \lambda * Vphy & \cdots & \alpha \end{bmatrix} \tag{18}$$

Some links:
This is an example of a PMM in Stan code. Varying slope with phylogenetic structure
Someone running a multiple measures model in BRMS with some queries. Not sure how useful this is.

Some random notes:
Cholesky decompostion: From what I understand, this is a trick when you cannot get the model to run with your covariance matrix, *Rethinking* says, it's way to "smuggle the covariance martix out of the prior."