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 reply 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*.

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

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

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. **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

force 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:

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

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:

$$y = \alpha + \beta x + a + e \tag{5}$$

$$a \sim normal(0, \sigma_P^2 \Sigma) \tag{6}$$

$$e \sim normal(0, \sigma_R^2 I) \tag{7}$$

$$\text{PGLS: } y \sim normal(\mu + \beta x, \sigma_P^2 \Sigma) \tag{8}$$

... 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.]

The multiple-values-per-species version of this just involves within-group centering, 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?

My underlying model: I think the phylogeny is a better option for partial pooling that treating species as under-ordered 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 sort of doubt we'll be able to manage this though. So I am searching for the best alternative option.

What we have tried: So far, 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 it seems poor form to apply 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.

3. Wrote my own Stan code that is not yet working (runs great, but does not estimate the phylogenetic effect correctly).

```
> summary(m1)
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: resp ~ force.z + chill.z + photo.z + (1 | spps)
   Data: bb.stan (Number of observations: 3856)
Samples: 2 chains, each with iter = 2000; warmup = 500; thin = 1;
         total post-warmup samples = 3000

Group-Level Effects:
~spps (Number of levels: 219)
              Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)    16.20      0.83    14.64    17.89        409 1.00

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept    32.00      1.12    29.78    34.33        290 1.00
force.z      -5.36      0.36    -6.07    -4.67       2213 1.00
chill.z      -7.01      0.31    -7.63    -6.43       2262 1.00
photo.z      -1.30      0.28    -1.85    -0.74       2746 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma    15.04      0.18    14.69    15.39       3853 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
> summary(m2)
```

Figure 1: No phylogenetic structure, just species on the intercept.

Some links:
Varying slope with phylogenetic structure

```
> summary(m2)
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: resp ~ force.z + chill.z + photo.z + (1 | spps)
   Data: bb.stan (Number of observations: 3856)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Group-Level Effects:
~spps (Number of levels: 219)
             Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)    16.18      0.91    14.45    18.01        615 1.01

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept    32.06      1.14    29.95    34.39        387 1.02
force.z      -5.36      0.36    -6.05    -4.65       3403 1.00
chill.z      -7.01      0.30    -7.59    -6.43       4774 1.00
photo.z      -1.30      0.28    -1.84    -0.75       5943 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma    15.04      0.17    14.70    15.39       6061 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
>
```

Figure 2: Adding phylogenetic structure (and species separately?) on the intercept.

```
> summary(m3)
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: resp ~ force.z + chill.z + photo.z + (force.z + chill.z + photo.z | spps)
   Data: bb.stan (Number of observations: 3856)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Group-Level Effects:
~spps (Number of levels: 219)
                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sd(Intercept)           17.17      0.98    15.35    19.20        982 1.01
sd(force.z)              6.71      0.87     5.16     8.56       1205 1.00
sd(chill.z)             14.13      1.01    12.26    16.19       1078 1.00
sd(photo.z)              2.78      0.42     2.05     3.68       1852 1.00
cor(Intercept,force.z)  -0.50      0.14    -0.73    -0.20        816 1.00
cor(Intercept,chill.z)  -0.15      0.09    -0.32     0.04        961 1.01
cor(force.z,chill.z)    -0.08      0.14    -0.35     0.19        313 1.01
cor(Intercept,photo.z)  -0.10      0.19    -0.46     0.27       3307 1.00
cor(force.z,photo.z)    -0.40      0.16    -0.67    -0.07       1635 1.00
cor(chill.z,photo.z)     0.35      0.21    -0.09     0.69       1386 1.00

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept    31.04      1.25    28.52    33.44        401 1.00
force.z      -7.55      0.99    -9.48    -5.59        829 1.00
chill.z      -9.29      1.17   -11.55    -6.99       1036 1.01
photo.z      -1.33      0.47    -2.24    -0.39       2494 1.00

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma    12.69      0.16    12.38    13.00       4038 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Figure 3: Adding phylogenetic structure and species on the intercept and slope? Or not ... not sure!