Phylogeny model updates

7 October 2020

What I will cover...

- Review: The way phylogeny is generally handled in ecological statistical models
  - PGLS
  - PMM
- Review: The way I want to handle it for OSPREE (but I could be wrong)
- Review: My progress on that goal (when last we met in June)
- New! My progress on that goal since last met
- New! How you can help!

A couple quick notes ...

- I am skipping over why we care about phylogeny, check out phylogenyfoundering.pdf for some of that.
- Reminder: Phylogenetic structure in a value (say, a 'trait') is measured based on a value called λ. When $\lambda = 1$ the trait is perfectly predicted by the phylogeny; when it's 0, the phylogeny does not matter.

Common modeling approaches

Part 1: PGLS (adapted from second edition of *Statistical Rethinking*)

Consider...

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

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

μ is a usual linear model. y 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).

Common modeling approaches

Part 1: PGLS

You have to make sure of a few things:

- 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.
- This model forces the correlation structure you give it—it does not adjust the correlation structure at all.

If you don't want to assume $\lambda = 1$, then you need to estimate a value to multiply the matix by such that:

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

Common modeling approaches

Part 1: PGLS

Actually, to be exact, I think you want to end up like this:

$$y \sim MVN(\mu, S)$$

$$\mu_i = \alpha + \beta * x_i$$

$$S = \begin{bmatrix} \sigma_\beta & \lambda\sigma_\beta & \lambda\sigma_\beta \\ \lambda\sigma_\beta & \sigma_\beta & \lambda\sigma_\beta \\ \lambda\sigma_\beta & \lambda\sigma_\beta & \sigma_\beta \end{bmatrix}$$

Common modeling approaches

Part 2: PMM (phylogenetic mixed model)

Now PMM...

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

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

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

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

... where α is the intercept β is the slope for the co-factor x, a is the phylogenetic random effect, and e is the residual error. This model estimated 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.).

Common modeling approaches

Part 3: PMM vs PGLS

- People often stay 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?)
- PMM explicitly models other sources of error through e.

In PMM the strength of the phylogenetic effect is measured as:

$$\lambda = \frac{\sigma^2_P}{\sigma^2_P + \sigma^2_R} \tag{9}$$

which is equivalent to just saying 'what proportion of the variance is due to phylogeny?'

What's wrong with this model?

What I want for OSPREE

Here's the dream

$$y = \alpha_0 + \alpha + \beta x + e \tag{10}$$

$$\alpha \sim MVN(0, \sigma^2_{P\alpha}) \tag{11}$$

$$\beta \sim MVN(0, \sigma^2_{P\beta}) \tag{12}$$

$$e \sim normal(0, \sigma^2_y) \tag{13}$$

$$\sigma^2_{P\alpha} = \alpha_\alpha + \lambda_\alpha * \Sigma \tag{14}$$

$$\sigma^2_{P\beta} = \alpha_\beta + \lambda_\beta * \Sigma \tag{15}$$

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.

Where am I at? (through Sept 2020)

What I have done

- I have stolen code from Will Pearse, and gotten him to write me more code.
- I have contacted Simone Blomberg (31 Aug) and Tony Ives (4 Sep) with my issue.
- I have one version of the code running for forcing, chilling, photo slopes... with the real OSPREE data, but it does not want to have phylogeny on the intercepts also (I tried it with non-partially pooled intercepts also, my notes on this say "null.interceptsbf, null.interceptsbp struggling some, but close...").
- I have done fake data

My progress

Here's the code last we saw it (June 2020)

bforce $\sim$ MVN(rep.vector(0,n.sp),
diag.matrix(repector(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.

My progress

Here's the code last we saw it (June 2020)

Let α be null.interceptsb and λ be lam.interceptsb.

bforce $=$

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

My progress

Here's the code last we saw it (June 2020)

If *Vphy* is set to have 1s down the diagonal, we can simplify to this:

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

My progress

Here's the code last we saw it (June 2020)

If *Vphy* is set to have 0s down the diagonal, we can simplify to this:

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

This seems good—now the within-species estimate is α and the between-species is $\lambda * Vphy$.

My progress

Here's the code last we saw it (June 2020)

- I can return the slopes
- λ is wrong and I don't know why.
- I have a tried a few things ...
  - Set nind to 1 (1 value per species) .... no dice.
  - Tried taking it off slopes and putting on intercepts ... no dice.
  - I made two versions of test data.
  - I asked Will for more help.

My progress

Test data

- In ubertoy.R I use geiger to adjust the slopes (slopes are correct, lam.interceptsb is 0.7 when λ set to 0.9)
- In ubertoy_nogeiger.R I set up the matrix myself (returns null.interceptsb and lam.interceptsb to always be similar, even when set to different numbers)

So, on a side note, setting up the test data is not straight-forward.

My progress

Will wrote back

Will decided he did not like his old formulation and gave me a new one (see ubermini_2.R)

His new code has:
bforce $\sim$ MVN(rep$_v$*ector*(*bz*, *nsp*),
lamvcv(Vphy, lam.interceptsb, sigma.interceptsb)),

meaning there is now ONE value in the first spot of the MVN and the second spot relies on ...

My progress

Will wrote back

... this function:

```
matrix lambda_vcv(matrix vcv, real lambda, real sigma){
    matrix[rows(vcv),cols(vcv)] local_vcv;
    local_vcv = vcv * lambda;
    for(i in 1:rows(local_vcv))
       local_vcv[i, i] = vcv[i, i];
    return(local_vcv * sigma);}
```

My progress

Will wrote back

Which I think (altogether) does this ...

$$\beta \sim MVN(\mu, \Phi\Sigma)$$

$$\Phi\Sigma = \begin{bmatrix} \sigma_\beta & \lambda\sigma_\beta & \lambda\sigma_\beta \\ \lambda\sigma_\beta & \sigma_\beta & \lambda\sigma_\beta \\ \lambda\sigma_\beta & \lambda\sigma_\beta & \sigma_\beta \end{bmatrix}$$

...which is similar to what PGLS does.

My progress

I run Will's new code.

I wrote to Will when he sent this code the following: I tried with nspecies=300 and my results were still off:

- model said 0.3172732 when it was 0.2985636
- model said 0.3253274 when it was 0.1229
- model said 0.3985724 when it was 0.7413879

But when I tried it last night, it seemed better!

Next steps ...

Maybe you can help

- Try to run OSPREE data using phyr, then I can pester Tony for help
- Have someone else test Will's code, maybe it works?
- If so, I need help getting Willi's new code to run on OSPREE (it did not run when I tried it)
- If not, post Will's code or some version of it to Stan Discouse.
- Someone could look at how I wrote the matrix out, versus geiger, versus Will's new code (I'd like to understand this better)

Next steps ...

Maybe you can help

- Try to run OSPREE data using phyr, then I can pester Tony for help
- ...

Tony wrote, "... just from your description, I don't see a particular reason for non-identifiability. I'd try the model using the package phyr (new version just released) that does PGLMMs for either frequentists or Bayesians. It should take (phylogenetically) random slopes pretty easily. Having multiple observations for the same species shouldn't be a problem; you can just include a species-level random effect. (Okay, to do this in phyr you have to code the matrices yourself, but you've already done that for stan)"