Phylogenetic comparative methods in the lme4-verse

Michael Li and Ben Bolker

The goal of *phylogenetic comparative methods* is to analyze relationships among data where the observations are gathered from nodes (usually tips) of a phylogenetic tree - for example, regression analyses of body temperature as a function of body size for animals within a clade More generally, we can frame this in the usual GLMM way as

$$
\begin{aligned}
y &\sim D(\mu, \phi) \\
\mu &= g^{-1}(\eta) = g^{-1}(X\beta + Zb) \\
b &\sim \mathrm{MVN}(0, \Sigma)
\end{aligned}
$$

where the part that makes it specifically phylogenetic is that Σ captures the *phylogenetic correlation* (PC). The PC is the correlation among observations due to relatedness; recently diverged taxa have higher correlation than more anciently diverged taxa. In the extreme case of a *star phylogeny* (all taxa diverged from each other simultaneously at some point in the past) the phylogenetic correlation collapses to a diagonal matrix and we get back to the simple, uncorrelated regression.

Various P(G)LMM (phylogenetic [generalized] linear mixed model) approaches have been proposed. Many depend on Pagel's lambda transformation, which gives the correlation matrix a particularly simple form (but has been criticized ...)

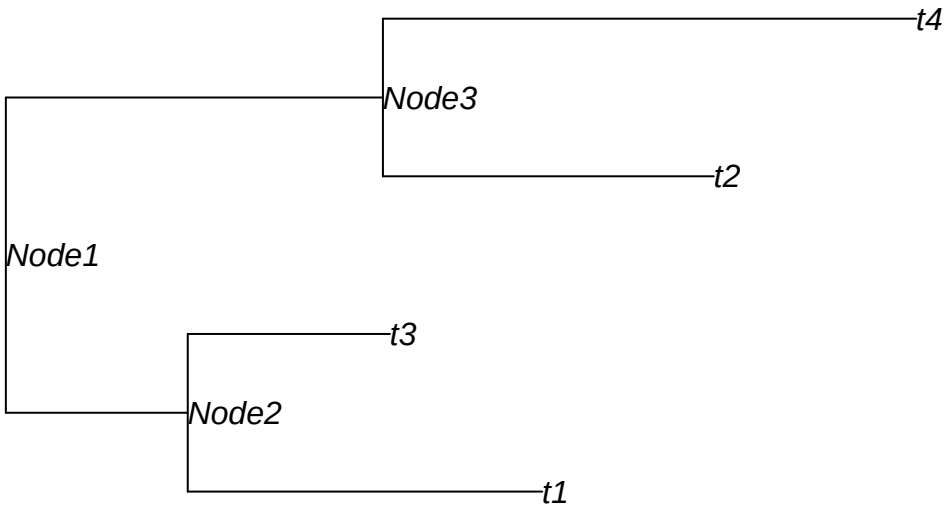
An alternative approach is to model the phylogenetic correlation as a *Gaussian process* (GP). In particular, suppose that the evolutionary process is a Brownian motion process (an almost certainly incorrect/oversimplified model of evolution, but one that many phylogenetic methods are built on). In that case, the phylogenetic variability of a particular observation can be written as the sum of the evolutionary changes that occurred on all of the branches in the phylogeny in its past. If we set up the Z matrix appropriately, we can model everything with a sequence of *independent* errors, rather than having to impose a correlation structure on the random effects.

Nuts and bolts: from a phylogeny to a Z matrix for the GP

```
library(ape)
```

```
## Warning: package 'ape' was built under R version 3.4.4
```

```
library(Matrix)
set.seed(101)
r <- makeNodeLabel(rtree(4))
plot(r,show.node.label=TRUE)
```



Information in a `phylo` object is contained in the *edge matrix*:

> edge: a two-column matrix of mode numeric where each row represents an edge of the tree; the nodes and the tips are symbolized with numbers; the tips are numbered 1, 2, ..., and the nodes are numbered after the tips. For each row, the first column gives the ancestor.

```
t(r$edge)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    5    6    6    5    7    7
## [2,]    6    1    2    7    3    4
```

and a list of *edge lengths*

```
r$edge.length
```

```
## [1] 0.3000548 0.5848666 0.3334671 0.6220120 0.5458286 0.8797957
```

Inspecting this tree, we can figure out (see `$tip.label` and `$node.label` for label-to-number correspondences):

- tips are 1-4, nodes are 5-7
- tip 1 (`t1`) involves branches 2 ($6 \to 1$) and 1 ($5 \to 6$).
- tip 2 (`t3`) involves branches 3 ($6 \to 2$) and 1 ($5 \to 6$)
- tip 3 (`t2`) involves branches 5 ($7 \to 3$) and 4 ($5 \to 7$)
- tip 4 (`t4`) involves branches 6 ($7 \to 4$) and 4 ($5 \to 7$)

So, for example, we can say that the 'error' value corresponding to tip 1 is $\ell_1 b_1 + \ell_2 b_2$, where ℓ_i is the square root of the branch length and the b_i are independent, homoscedastic Normal variates. Alternately, the Z matrix is

$$\begin{pmatrix} \ell_1 & \ell_2 & 0 & 0 & 0 & 0 \\ \ell_1 & 0 & \ell_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ell_4 & \ell_5 & 0 \\ 0 & 0 & 0 & \ell_4 & 0 & \ell_6 \end{pmatrix}$$

where ℓ_i is the length of the i^{th} branch, so that the species effects are Zb.

If we can build the corresponding Z matrix, then we can insert it in the `lme4` modular model-fitting process (see `?modular`).

Here's a (probably not very efficient) way to construct the Z matrix. (There must be a way to not walk the tree multiple times from every tip ...)

```
phylo.to.Z <- function(r,stand=FALSE){
  ntip <- length(r$tip.label)
  Zid <- Matrix(0.0,ncol=length(r$edge.length),nrow=ntip)
  nodes <- (ntip+1):max(r$edge)
  root <- nodes[!(nodes %in% r$edge[,2])]
  for (i in 1:ntip){
    cn <- i  ## current node
    while (cn != root){
      ce <- which(r$edge[,2]==cn)   ## find current edge
      Zid[i,ce] <- 1   ## set Zid to 1
      cn <- r$edge[ce,1]            ## find previous node
    }
  }
  V <- vcv(r)
  sig <- exp(as.numeric(determinant(V)["modulus"])/ntip)
  Z <- t(sqrt(r$edge.length) * t(Zid))
  if(stand){Z <- t(sqrt(r$edge.length/sig) * t(Zid))}
  rownames(Z) <- r$tip.label
  colnames(Z) <- 1:length(r$edge.length)
  return(Z)
}
phylo.to.Z(r)
```

```
## 4 x 6 sparse Matrix of class "dgCMatrix"
##            1         2         3         4         5         6
## t1 0.5477726 0.7647657 .         .         .         .
## t3 0.5477726 .         0.5774661 .         .         .
## t2 .         .         .         0.7886774 0.7388021 .
## t4 .         .         .         0.7886774 .         0.9379743
```

On the other hand, it only takes a few seconds to run for a 200-species phylogeny (see below).

To fit in lme4 and glmmTMB, we need random effect terms in the formula that includes a `(...|phylo)`. Using the lme4 machinery, it will build the appropriate random effect structure multiplied by the phyloZ matrix. `glmmTMB` can be deconstructed in a similar way. In fact, we can re-use a lot of the machinery. Being able to use `glmmTMB` means we can use a broader range of distributions, zero-inflation, etc. (machinery below assumes phylogenetic structure only in the conditional distribution). This is also a little clunky, some adjustment on the `glmmTMB` side might make it a bit easier.