Modeling the absolute rate of fossil occurrence

Peter D Smits
Committee on Evolutionary Biology, University of Chicago

February 19, 2015

1 Introduction

The most basic statement about the fossil record is that it is incomplete and that what has been preserved is a (biased) subset of the biodiversity that once existed. This fundamental incomplete incompleteness of the record, along with worker effort biased towards certain intervals, are the two major problems to accurately modeling the fossil record.

Sampling in paleontology has two meanings: geological and statistical. The first is the rate at which an organism is preserved as a fossil, and the second is the rate of observing a fossil occurrence. In this study I focus on the second definition, which I call the occurrence rate.

Some organismal groups are considered to have comparable fossil records meaning that the occurrence patterns can be considered transitive.

This assumes that all members within a group are considered to have identical occurrence rates.

There is known differences in occurrence rate within groups as some members may be much more commonly occurring than others either because of biological abundance, differences in worker effort, or preservational biases.

The Bayesian hierarchical modeling approach used here explicitly models the occurrence rate of a given genus or class in relation to the entire record of fossil occurrences for the entire Phanerozoic.

bayesian hierarchical modeling approach here is new for paleontology

NOTES AND PAPERS

previous approaches to overcoming

Alroy2010c

a foote miller paper using rarefaction along with a million others

Jablonski1991

Marshall

Sadler1981

Wagner2007

Wang2004

Wang2012b

previous approaches to modeling

Alroy2014

Foote1996d

Foote1996e

Foote1997c

Foote1999a

Foote2001

Solow1997

Strauss1989

Wagner2013a

other

Foote2007a

Jernvall2002

Liow2007d

Lloyd2011

Lloyd2012b

Lloyd2012c

Lloyd2013

McGowan/Smith and McGowan

Sepkoski1975

Simpson2009

Wagner2000h

2 Methods

2.1 Fossil occurrence information

Foote and Miller data.

2.2 Hierarchical counting model

First, define y_i as some count of fossil occurrences of genus j in a geologic stage for $i = 1, \ldots, n$ and $j = 1, \ldots, J$.

The Poisson distribution is used as the simplest model of count data, such as the number of observed fossils. The Poisson distribution has one parameter λ which is a "rate" or inverse-scale parameter. λ can be interpreted as the expected count observed $\mathbf{E}[y]$. λ can be reparameterized for use in regression using the log link function $\mathbf{E}[y] = \exp(\alpha)$ where α can be any real number (**?**). This is written as

$$\begin{aligned} y_i &\sim \text{Poisson}(\lambda_i) \\ \lambda_i &= \exp(\alpha_i). \end{aligned} \tag{1}$$

Currently, this model (Eq. 1) does not take into account the generic membership j of the fossil count and assumes that all genera have the same sighting rate. To account for variation in occurrence rate between genera while also modeling mean generic occurrence rate I take a Bayesian hierarchical modeling approach (**??**). First, I redefine α_i as $\alpha_{j[i]}$ to indicate that observation i is a member of genus j. I then assume that genera can be considered exchangeable or that the actual value of j has no meaning. Given this assumption, values of $\alpha_{j[i]}$ are given the following normally distributed prior

$$\alpha_{j[i]} \sim \mathcal{N}(\mu, \sigma_j). \tag{2}$$

The scale hyperparameter σ_j (Eq. 2) is then estimated from the data itself. This approach allows genera with small sample size to pull towards the mean of the prior (μ) while still genera with large sample sizes and strong effects to be modeled. Because σ_j is the standard deviation of the overall genus-level rate of occurrence per collection, values of σ_j close to 0 indicate complete pooling/congruence between all genera while high values of σ_j no pooling or congruence between genera (**?**).

This hierarchical approach can be further extended to account for a genus' class membership. Define k as the class that genus j belongs to, where $k = 1, \ldots, K$. Then, instead of assuming that μ is equal for all classes (Eq. 2), instead the μ is allowed to vary across classes and is written $\mu_{k[j]}$. This is the estimate of the rate of fossil occurrence for classes k. Then, assuming that classes are exchangeable,

values of $\mu_{k[j]}$ are given the same, shared hyperprior. These changes are then written as

$$\begin{aligned} \alpha_{j[i]} &\sim \mathcal{N}(\mu_{k[j]}, \sigma_j) \\ \mu_{k[j]} &\sim \mathcal{N}(\psi, \sigma_k). \end{aligned} \tag{3}$$

ψ here is an estimate of the mean class rate. Similar earlier (Eq. 2), the scale hyperparameter σ_k corresponds to the overall class-level rate of occurrence per collection. Values of σ_k cose 10 indicate completely pooling between all classes while high values correspond to no pooling of classes.

The current model (Eq. 1) does not take into account the number of chances to count an observation. For example, if counting the number of traffic accidents at a street corner it matters if 20 vehicles have passed through the intersection versus 100. To account for this we can define an exposure term u_i for each observation (**?**). In this study, u_i is defined as the number of localities species i occurred in during the given stage. The inclusion of u_i is formulated as

$$\begin{aligned} y_i &= \text{Poisson}(u_i \lambda_i) \\ \lambda_i &= \exp(\log(u_i) + \alpha_{j[i]}). \end{aligned} \tag{4}$$

The inclusion of $\log(u_i)$ in the parameterization of λ_i (Eq. 4) is due to the following relationships

$$\begin{aligned} \frac{\mathbf{E}[y_i]}{u_i} &= \lambda_i \\ \mathbf{E}[y_i] &= u_i \lambda_i \\ \log(\mathbf{E}[y_i]) &= \log(u_i) + \log(\lambda_i) \end{aligned}$$

We can now interpret λ as the expected number of co-occurring species per locality for a given observation. While u_i is called the exposure, $\log(u_i)$ is called the offset (**?**).

One of the major assumptions of the Poisson distribution is that, because there is only one parameter, the variance of the distribution is equal to the mean ($\frac{Var[y]}{E[y]}$). When variance is greater than the mean, this is called overdispersion. We can relax this assumption by assuming that, instead of a Poisson distribution, observations are drawn from a negative bionmial distribution (**?**). Here, I use the following parameterization of the negative binomial

$$\text{Negative binomial}(y|\eta, \phi) = \binom{y + \phi - 1}{y} \left(\frac{\eta}{\eta + \phi}\right)^y \left(\frac{\phi}{\eta + \phi}\right)^\phi \tag{5}$$

where η is the mean and ϕ is the overdispersion. Substituting the negative

binomial for the Poisson, the model as currently defined is written

$$\begin{aligned} y_i &= \text{Negative binomial}(u_i\eta_i, \phi_y) \\ \eta_i &= \exp(\alpha_{j[i]}) \\ \alpha_{j[i]} &\sim \text{Normal}(\mu_{k[j]}, \sigma_j) \\ \mu_{k[j]} &\sim \text{Normal}(\psi, \sigma_k). \end{aligned} \tag{6}$$

Finally, given the Bayesian framework taken here, I have to assign priors to various non-hierarchically modeled parameters. Scale parameters were given weakly informative half-Cauchy (C^+) priors because they have good regulatory priors for constraining hierarchical effects (**??**). For the location parameter ψ, I used a weakly informative prior because it is expected that the most probable values do not have a very high magnitude, while still allowing for that possibility. The priors used here are

$$\begin{aligned} \phi_y &\sim \text{C}^+(2.5) \\ \sigma_j &\sim \text{C}^+(2.5) \\ \psi &\sim \text{Normal}(0, 10) \\ \sigma_k &\sim \text{C}^+(2.5). \end{aligned}$$

The Cauchy distribution is equivalent to the t-distribution with 1 degree of freedom, and the half-Cauchy distribution is the Cauchy folded about 0.

2.3 Model checking

Posterior predictive checks.

y and y^{rep}

count data residuals (**?**)

raw residual: $r_i = \sqrt{y_i} - \sqrt{y_i^{rep}}$

deviance residual: $r_i^D = sign(y_i - y_i^{rep}) \left[y_i \log\left(\frac{y_i}{y_i^{rep}}\right) - (y_i - y_i^{rep}) \right]^{1/2}$ though i'm not sure if i got this right. what is the derivation?