A hierarchical Bayesian model of Phanerozoic survival patterns

Peter D Smits
Committee on Evolutionary Biology, University of Chicago

April 8, 2015

1 Introduction

2 Methods

2.1 Fossil occurrence information

I analyzed the exact data set used by Foote and Miller (2013) in their analysis of the early survival of marine genera. This data set is a highly pruned and vetted set of all occurrences of important marine groups for the entire Phanerozoic as sampled in the Paleobiology Database (http://www.paleodb.org). Specifically, I analyzed a subset of the Foote and Miller (2013) data set of only genera belonging to a higher-taxonomic level assigned to one of the three Sepkoski Fauna (Table 1).

2.2 Survival model

Define y as a vector of length n where the ith element is the duration in geologic stages of genus i, where $i = 1, \ldots, n$.

The simplest parametric survival model is where species durations are assumed to be distributed exponentially with a single inverse-scale ("rate") parameter λ

$$\begin{aligned} p(y|\lambda) &= \lambda \exp(-\lambda y) \\ y_i &\sim \text{Exponential}(\lambda). \end{aligned}$$

Given a parameteric model of survival, two key functions can be defined: survival $S(t)$ and hazard $h(t)$. $S(t)$ corresponds to the probability that a species having existed for t will not have gone extinct while $h(t)$ corresponds to the instantaneous extinction rate per unit age given taxon age t (Klein and Moeschberger, 2003). For an exponential model, $S(t)$ is defined

$$S(t) = \exp(-\lambda t) \tag{1}$$

and $h(t)$ is defined

$$h(t) = \lambda \tag{2}$$

An exponential model of duration is the parameteric representation of Van Valen (1973) because the right side of Eq. 2 does not include any t terms.

In order to allow species duration to vary with individual covariate information λ can be reparameterized as a regression (Klein and Moeschberger, 2003). Given that λ is only defined for all non-negative reals, I use a log-link function. More specifically, this is written out as

$$\lambda = \exp(\beta). \tag{3}$$

β can then be expanded to included covariates of interest.

2.2.1 Hierarchical modeling

Hierarchical or multilevel models are for when data are structured in known groups and parameters are allowed to vary by groupings (Gelman and Hill, 2007, Gelman et al., 2013). This approach allows for inference to be couched directly in terms of the structure of the data. In a hierarchical model, these groups are considered exchangeable samples from a population distribution (Gelman et al., 2013). By setting up a hierarchical model we can improve predictive accuracy by allowing for more sources of variance.

While a hierarchical modeling approach increases the overall complexity of the model, there are numerous advantages not available to single-level modeling as well as being a more realistic model of the data-generating process (Gelman and Hill, 2007, Gelman et al., 2013).

For example, if we were studying mortality rates across multiple hospitals we might consider each hospital as a sample from a shared “hospital-level” distribution describing the expected “hospital effect.” Additionally, we are able to estimate the hospital-level variance which allows for inference about how consistent effects may be across all hospitals.

Hierarchical modeling can be viewed as a compromise between inference from completely pooled data and from data separated by group or no pooling. Mathematically, complete pooling is when we assume that between group variance is zero while no pooling is when we assume that between group variance is infinite. Compromise is accomplished, for example, by letting the scale hyperparameter of the shared prior be estimated from the data (Gelman et al., 2013). This is effectively a weighted mean of these two extremes where the weighting depends on the sample size of each group.

Groups with small samples sizes or little power are drawn towards the population mean, while groups with larger samples sizes or possibly stronger effects are estimated as necessary. This means that groups with smaller sample sizes borrow power from the larger groups.

2.2.2 Temporal effects

By definition, a genus originates during a single geologic stage. Define an origination cohort, c_j, as the set of all genera that origination during the same geologic stage j where $j = 1, \ldots, J$, J being the total number of origination cohorts (Raup, 1978). Origination cohort is modeled as a hierarchical effect where cohorts are considered exchangeable and drawn from a shared normal distribution. This is written as

$$c_j \sim \mathcal{N}(0, \sigma_c) \tag{4}$$

where the scale parameter σ_c is estimated from the data. This term is incorporated in to Eq. 3 as follows

$$\lambda_i = \exp(\beta + c_{j[i]}) \tag{5}$$

This approach for handling temporal effects can be further expanded, allowing for more temporal structure to be included. For example, each geologic stage can be considered a member of a specific period of "background extinction," or temporal range between two mass extinction events. Effectively, this means modeling the macroevolutionary regime membership of each cohort (Jablonski, 1987). Define regime r_k as the set of all origination cohorts within the temporal span between two mass extinctions k where $k = 1, \ldots, K$, K being the total number of regimes. Regime is then modeled as a hierarchical one "level" above origination cohort, and is also considered exchangeable and drawn from a shared normal distribution. Eq. 4 is then expanded as follows

$$\begin{aligned} c_j &\sim \mathcal{N}(\mu_{k[j]}, \sigma_c) \\ \mu_k &\sim \mathcal{N}(0, \sigma_\mu). \end{aligned} \tag{6}$$

As above, both σ_c and σ_μ are estimated from the data.

2.2.3 Taxonomic effects

By definition, each genus belongs to a single higher-level taxonomic grouping. Define a higher-level taxonomic grouping l_g as the set of all genera belonging to a shared Linnean classification of approximately class level where $g = 1, \ldots, G$, G being the total number of observed groupings. I model taxonomic groupings as a hierarchical effect where groups are conspired exchangeable and drawn from a shared normal distribution. I consider this approach appropriate because there is no comprehensive phylogenetic hypothesis including all genera from major marine groups across the entire Phanerozoic. In cases where there is a more detailed phylogenetic hypothesis, it would be possible to model phylogenetic effect as an individual hierarchical effect modeled as a multivariate normal distribution with covariance matrix known up to a constant (Housworth et al., 2004, Lynch, 1991).

Given the above assumptions, l is modeled as

$$l_g \sim \mathcal{N}(0, \sigma_l) \tag{7}$$

where σ_l is estimated from the data. Just as with temporal effect, this term is incorporated into Eq. 3 as

$$\lambda_i = \exp(\beta + l_{g[i]}). \tag{8}$$

Incorporating both temporal and taxonomic effects into Eq. 3 is simply

$$\lambda_i = \exp(\beta + c_{j[i]} + l_{g[i]}). \tag{9}$$

It is possible to increase the complexity of Eq. 7 by including increasingly more taxonomic levels both above and below the class level. However, I chose to not do this as it adds what is probably an unnecessary amount of complexity and makes interpretation of results extremely difficult.

What is of interest, however, are Sepkoski's three fauna (Sepkoski Jr., 1981) which divide select higher-level classifications into Cambrian, Paleozoic, and Modern fauna (Table 1). To model taxonomic groups as samples from a fauna, first I define a fauna f_h as the set of all taxonomic groups assigned by Sepkoski Jr. (1981) where $h = 1, \ldots, H$, H bing the number of fauna (3). Fauna are considered exchangeable and drawn from a shared normal distribution. Eq. 7 is then rewritten as

$$\begin{aligned} l_g &\sim \mathcal{N}(\omega_{h[g]}, \sigma_l) \\ \omega_h &\sim \mathcal{N}(0, \sigma_\omega) \end{aligned} \tag{10}$$

where both σ_l and σ_ω are estimated from the data.

2.2.4 Age-dependent extinction

We can relax assumption of the exponential model that extinction risk is independent of species age by instead assuming that species durations follow a Weibull distribution. The Weibull distribution has a shape parameter α and a scale parameter σ. Note that, by definition, $\sigma = 1/\lambda$. The Weibull distribution and sampling statement are defined

$$\begin{aligned} p(y|\alpha, \sigma) &= \frac{\alpha}{\sigma}\left(\frac{y}{\sigma}\right)^{\alpha-1} \exp\left(-\left(\frac{y}{\sigma}\right)^{\alpha}\right) \\ y &\sim \text{Weibull}(\alpha, \sigma). \end{aligned} \tag{11}$$

The corresponding $S(t)$ and $h(t)$ functions are defined

$$S(t) = \exp\left(-\left(\frac{t}{\sigma}\right)^{\alpha}\right) \tag{12}$$

$$h(t) = \frac{\alpha}{\sigma}\left(\frac{t}{\sigma}\right)^{\alpha-1}. \tag{13}$$

Similar to the exponential model, to allow for species duration to vary with individual covariate information σ can be reparameterized as a regression (Klein and Moeschberger, 2003) using a log-link function. This reparameterization takes a different form than for the exponential (Eq. 3 and 9) and, incorporating temporal and taxonomic effects, is written

$$\sigma = \exp\left(\frac{-(\beta + c_{j[i]} + l_{g[i]})}{\alpha}\right). \tag{14}$$

Incidentally, this is the formulation for an accelerated time failure model (Klein and Moeschberger, 2003). α can also be allowed to vary by, for example, hierarchical effects though it is normally assumed constant for all observations (Klein and Moeschberger, 2003).

2.2.5 Biological covariates

We cannot directly observe if a fossil taxon had occurrence restricted to any single environmental type, instead we estimate a probability with uncertainty based on proxies. One advantage of using a Bayesian analytical context is that both parameters and data are considered random samples from some underlying distribution, which means it is possible to model the uncertainty in our covariates of interest (Gelman et al., 2013). In this case, this is the probability that a genus will occur in an epicontiental sea or not. A genus' probability of occurring in an epicontinental environment, θ, was calculated using a fully Bayesian extension of (Simpson and Harnik, 2009) where occurrence probability (e.g. affinity) is defined as a distribution and not a point estimate. The reasoning behind this approach is that it allows for our uncertainty to properly propagate through our model.

Define e_i as the number of occurrences of genus i in an epicontiental sea and o_i as the number of occurrences of genus i not in an epicontinental sea (e.g. offshore). Because the value of inters is the probability of occurring in an epicontinental environment, given the observed fossil record, I assume that probability follows a beta distribution. We can then define our sampling statement as

$$\theta \sim \text{Beta}(e_i, o_i). \tag{15}$$

It is extremely important, however, to take into account the overall environmental occurrence probability of all other genera present at the same time as genus

i. This is incorporated as an informative beta prior, which is conveniently the conjugate prior for the beta distribution. Define η_i as the total number of other fossil occurrences (e.g. excepting for genus i) in epicontinental seas during stages where i occurs and π_i as the number of other fossil occurrences not on epicontinental seas. We can then define a prior for θ as

$$\theta \sim \mathrm{Beta}(\eta_i, \pi_i). \tag{16}$$

Given the likelihood (Eq. 15) and the prior (Eq. 16), the conjugacy of the prior can be taken advantage of to calculate the full posterior distribution

$$\begin{aligned} p(\theta|y) &= p(y|\theta)p(\theta) \\ p(\theta|y) &= \mathrm{Beta}(e_i + \eta_i, o_i + \pi_i). \end{aligned} \tag{17}$$

Now given the full definition of θ (Eq. 17), it can then be included in the definition of σ (Eq. 14) as a covariate with its own slope,

$$\sigma = \exp\left(\frac{-(\beta_0 + \beta_1\theta + c_{j[i]} + l_{g[i]})}{\alpha}\right). \tag{18}$$

2.2.6 Priors

Given the Bayesian framework used here, what remains is the important step of assigning priors probability statements to all estimated parameters. While many of the (hyper)parameters have already been given (hyper)priors, there currently remain a few improper priors.

The intercept term β_0 is given a weakly informative prior, $\beta_0 \sim \mathcal{N}(0, 10)$, reflecting that while little information is directly known, the value is most likely not extremely large or small. The environmental slope term β_1 is given a slightly more informative prior ($\beta_1 \mathcal{N}(0, 5)$) for similar reasons.

The prior terms for the various scale parameters of the hierarchical effects (Eq. 6 and 10) I use a weakly informative, regularizing half-Cauchy (C^+) prior distribution in order to better constrain all parameter estimates (Gelman et al., 2013). This is especially appropriate in the case of Eq. 10 as there are only three factors at the fauna level.

$$\begin{aligned} \sigma_\mu &\sim \mathrm{C}^+(1) \\ \sigma_\omega &\sim \mathrm{C}^+(1) \end{aligned} \tag{19}$$

Note that the half-Cauchy distribution is equivalent to a folded t-distribution with 1 degree-of-freedom (Gelman et al., 2013).

Finally, the Weibull shape parameter α is also given a weakly informative half-Cauchy prior, $\alpha \sim \mathrm{C}^+(2.5)$.

2.3 Censored observations

A key aspect of survival analysis is the inclusion of censored, or incompletely observed, data points (Ibrahim et al., 2001, Klein and Moeschberger, 2003). The two classes of censored observations encountered in this study were right and left censored observations. Right censoring is when a species does not go extinct during the window of observation, or species that are still extant. Left censored observations are those species that it is only known when a species was extinct by. To put another way, this is a species that went extinct but the observed duration is an over estimate of the actual duration.

In the context of this study, I considered all genera that had a duration of only one geologic stage to be left censored as we do not have a finer degree of resolution. Conceptually, this is similar to if I was studying, say, survival patterns in rats and an individual had died between the start of the experiment and next time the rats were observed. We know the rat lived no more than day.

The key function for modeling censored observations is $S(t)$ (Eq. 1 and 12). $S(t)$ is equivalent to the complementary cumulative distribution function, $1 - F(t)$ (Klein and Moeschberger, 2003). For right censored observations, instead of calculating the likelihood as normal (Eq. 1 or 11) the likelihood of an observation is evaluated using $S(t)$. Conceptually, this approach calculates the likelihood of observing a species that existed for at least that long. For left censored data, instead the likelihood is calculated using $1 - S(t)$ which corresponds to the likelihood of observing a species that existed no longer than t.

The full likelihood statements incorporating fully observed, right censored, and left censored observations are then

$$\begin{aligned}
\mathcal{L} &\propto \prod_{i \in C} \text{Exponential}(y_i|\lambda) \prod_{j \in R} S(y_j|\lambda) \prod_{k \in L} (1 - S(y_k|\lambda)) \\
\mathcal{L} &\propto \prod_{i \in C} \text{Weibull}(y_i|\alpha, \sigma) \prod_{j \in R} S(y_j|\alpha, \sigma) \prod_{k \in L} (1 - S(y_k|\alpha, \sigma))
\end{aligned} \tag{20}$$

where C is the set of all fully observed species, R the set of all right censored species, and L the set of all left-censored species.

2.4 Parameter estimation

Given the above likelihood and prior statements, the posterior probabilities of all parameters was approximated using a Markov-chain Monte Carlo routine using a variant of Hamiltonian Monte Carlo called the No-U-Turn Sampler (Hoffman and Gelman, 2014) as implemented in the probabilistic programming language Stan (**?**). The estimate of the posterior distribution were approximated from four parallel chains run for XXX draws split half warm-up and half sampling thinned to every XXX sample for a total of XXX samples. Chain convergence was assessed via the scale reduction factor $\hat{R}$ where values close to 1, or less than

Fauna	Taxa
Cambrian	Trilobita, Polychaeta, Monoplacophora (Tergomya), Inarticulata (Lingulata)
Paleozoic	Articulata (Rhynchonellata), Crinodea, Ostracoda, Cephalopoda, Anthozoa, Stenolaemata (Cyclocystoidea), Stelleroidea (Asteroidea, Ophiuridea)
Modern	Gastropoda, Bivalvia, Osteichthyes, Malacostraca, Echinoidea, Gymnolaemata, Demospongea, Chondrichthyes

Table 1: Sepkoski's three evolutionary fauna. In parentheses are the taxonomic names used in this study when there was a conflict between Sepkoski's designations and mine.

or equal to 1.1, indicate approximate convergence. Convergence means that the chains are approximately stationary and the samples are well mixed (Gelman et al., 2013).

References

M. Foote and A. I. Miller. Determinants of early survival in marine animal genera. *Paleobiology*, 39(2):171–192, Mar. 2013. ISSN 0094-8373. doi: 10.1666/12028. URL `http://www.bioone.org/doi/abs/10.1666/12028`.

A. Gelman and J. Hill. *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY, 2007.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall, Boca Raton, FL, 3 edition, 2013.

M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1351–1381, 2014.

E. A. Housworth, P. Martins, and M. Lynch. The Phylogenetic Mixed Model. *The American Naturalist*, 163(1):84–96, 2004.

J. G. Ibrahim, M.-H. Chen, and D. Sinha. *Bayesian Survival Analysis*. Springer, New York, 2001.

D. Jablonski. Heritability at the species level: analysis of geographic ranges of cretaceous mollusks. *Science*, 238(4825):360–363, Oct. 1987. ISSN 0036-8075. doi: 10.1126/science.238.4825.360. URL `http://www.ncbi.nlm.nih.gov/pubmed/17837117`.

J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 2nd edition, 2003.

M. Lynch. Methods for the analysis of comparative data in evolutionary biology. *Evolution*, 45(5):1065–1080, 1991.

D. M. Raup. Cohort Analysis of generic survivorship. *Paleobiology*, 4(1):1–15, 1978.

J. J. Sepkoski Jr. A kinetic model of Phanerozoic taxonomic diversity. I. Analysis of marine orders. *Paleobiology*, 4(3):223–251, 1978.

J. J. Sepkoski Jr. A kinetic model of Phanerozoic taxonomic diversity. II. Early Phanerozoic families and multiple equilibria. *Paleobiology*, 5(3):222–251, 1979.

J. J. Sepkoski Jr. A factor analytic description of the Phanerozoic marine fossil record. *Paleobiology*, 7(1):36–53, 1981.

J. J. Sepkoski Jr. A kinetic model of Phanerozoic taxonomic diversity. III. Post-Paleozoic families and mass extinctions. *Paleobiology*, 10(2):246–267, 1984.

C. Simpson and P. G. Harnik. Assessing the role of abundance in marine bivalve extinction over the post-Paleozoic. *Paleobiology*, 35(4):631–647, Dec. 2009. ISSN 0094-8373. doi: 10.1666/0094-8373-35.4.631. URL `http://www.bioone.org/doi/abs/10.1666/0094-8373-35.4.631`.

L. Van Valen. A new evolutionary law. *Evolutionary Theory*, 1:1–30, 1973. URL `http://ci.nii.ac.jp/naid/10011264287/`.