When do traits matter? A hierarchical Bayesian model of Brachiopod survival

Peter D Smits
Committee on Evolutionary Biology, University of Chicago

April 20, 2015

1 Introduction

2 Methods

2.1 Fossil occurrence information

The dataset analyzed here is derived from the a combination of the occurrence information from Miller and Foote (2009) and the body size data from Payne et al. (2014). The Miller and Foote (2009) dataset is based on the Paleobiology Database (http://www.paleodb.org); see Miller and Foote (2009) for a full description of the inclusion criterion.

Sampled occurrences were restricted to those with latitude and longitude coordinates, assignment to either epicontinental or open-ocean environment, and being of a genus present in the body size dataset. Genus duration was calculated as the number of geologic stages from first appearance to last appearance, inclusive. Genera who's last appearance was in a stage preceding a mass extinction were right censored, and genera with a duration of only one stage and were left censored (see below for explanation of censoring). The covariates used to model genus duration were geographic range size (r), environmental preference (θ), and body size (m).

Geographic range was calculated using an occupancy approach. First, all occurrences were projected onto an equal-area cylindrical map projection. Each occurrence was then assigned to one of the cells of a 70×34 regular raster grid placed on the map. Each grid cell represents approximately 250,000 km^2. Following this, for each stage, the total number of grid cells occupied is calculated. The number of grid cells that each genus present occurs in was then calculated and made relative by dividing by the total number of possible cells. Finally,

mean relative genus occupancy was calculated as the mean of per stage relative occupancy.

The calculation and inclusion of environmental affinity in the subsequent survival model is a statistical procedure that takes into account our uncertainty based on where fossils tend to occur. Because we cannot directly observe if a fossil taxon had occurrences restricted to only a single environmental, instead we can only estimate a probability of occurrence with some amount of uncertainty. 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. open ocean). 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

$$\theta_i \sim \text{Beta}(e_i, o_i). \tag{1}$$

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 E_i as the total number of other fossil occurrences (e.g. excepting for genus i) in epicontinental seas during stages where i occurs and O_i as the number of other fossil occurrences not on epicontinental seas. We can then define a prior for θ as

$$\theta_i \sim \text{Beta}(E_i, O_i). \tag{2}$$

Given the likelihood (Eq. 1) and the prior (Eq. 2), the conjugacy of the prior can be taken advantage of to calculate the full posterior distribution

$$\begin{aligned} p(\theta_i|y_i) &\propto p(y_i|\theta)p(\theta) \\ p(\theta_i|y_i) &= \text{Beta}(e_i + E_i, o_i + O_i). \end{aligned} \tag{3}$$

$p(\theta_i|y_i)$ is the full posterior describing the probability that genus i occurs on an epicontinental sea, where higher values indicate increased affinity for epicontinental seas over open-ocean environments.

Body size data was sourced directly from Payne et al. (2014). Because those measurements are presented with out quantified error, a measurement error model similar to the one for environmental affinity could not be implemented.

Prior to analysis all covariates were transformed to be defined on the entire real line. Geographic range size and environmental preferences, because they are defined as values only between 0 and 1, were logit transformed. Body size, which is defined for all positive real values was natural log transformed. All of these covariates were then standardized by mean centering and dividing by two times their standard deviation following Gelman and Hill (2007).

2.2 Survival model

Genus durations were modeled in a Bayesian parameteric survival analysis framework. Durations were assumed to follow either an exponential or Weibull distribution. Each of these distributions makes strong assumptions about how duration may effect extinction risk. Use of the exponential distribution assumes that extinction risk is independent of duration. In contrast, use of the Weibull distribution allows for age dependent extinction via the shape parameter α, though only as a monotonic function of duration. Importantly, the Weibull distribution is equivalent to the exponential distribution when $\alpha = 1$. In general, the notation used here follows Gelman and Hill (2007), Gelman et al. (2013), and STAN MANUAL.

The simplest model of genus duration includes no covariate or structural information. Define y_i as the duration in stages of genus i, where $i = 1, \ldots, n$ and n is the number of observed genera. These two models are them simply defined as

$$\begin{aligned} y_i &\sim \text{Exponential}(\lambda) \\ y_i &\sim \text{Weibull}(\alpha, \sigma). \end{aligned} \tag{4}$$

Note that λ is a "rate" or inverse-scale while σ is a scale parameter, meaning that $\frac{1}{\lambda} = \sigma$.

These current simple models can then be expanded to include covariate information as predictors by reparameterizing λ and σ as a regression (Klein and Moeschberger, 2003). Each of the covariates of interest is given its own regression coefficient (e.g. β_{range}) along with an intercept term β_0. There are some additional complications to the parameterization of σ associated with the inclusion of α as well as interpretability (Klein and Moeschberger, 2003). Both of these cases are written more fully as

$$\begin{aligned} \lambda_i &= \exp(\beta_0 + \beta_{range} r_i + \beta_{environment} \theta_i + \beta_{size} m_i) \\ \sigma_i &= \exp\left(\frac{-(\beta_0 + \beta_{range} r_i + \beta_{environment} \theta_i + \beta_{size} m_i)}{\alpha} \right). \end{aligned} \tag{5}$$

The regression equations are exponentiated because both λ and σ are only defined for positive reals. Remember that θ_i is the environmental affinity, as defined above, of genus i.

The current model which incorporates both equations 4 and 5 can then be further expanded to allow all of the β coefficients, including β_0, to vary with origination cohort while also modeling their covariance and correlation. This is called a varying-intercepts, varying-slopes model (Gelman and Hill, 2007). It is much easier to represent and explain how this is parameterized using matrix notation. First, define $\mathbf{B}$ as $k \times J$ matrix of the k coefficients including the intercept term ($k = 4$) for each of the J cohorts. Second, define $\mathbf{X}$ as a $n \times k$ matrix where each column is one of the covariates of interest. Importantly, $\mathbf{X}$ includes a columns of all 1s which correspond to the constant term β_0. Third, define $j[i]$ as the origination cohort of genus i, where $j = 1, \ldots, J$ and J is the total number of observed cohorts.

Using the above hierarchical expansion to the model, we then rewrite λ and σ in matrix notation as

$$\begin{aligned} \lambda_i &= \exp(\mathbf{X}_i \mathbf{B}_{j[i]}) \\ \sigma_i &= \exp\left(\frac{-(\mathbf{X}_i \mathbf{B}_{j[i]})}{\alpha}\right). \end{aligned} \tag{6}$$

At face value, the above parameterization (Eq. 6) is opaque as to how the covariance and correlation between elements of $\mathbf{B}$ are estimated. This becomes more apparent after defining the prior distribution of $\mathbf{B}$. Because $\mathbf{B}$ is a matrix, I used a multivariate normal prior with unknown vector of means μ and covariance matrix $\mathbf{\Sigma}$. This is written as

$$\mathbf{B} \sim \text{MVN}(\mu_{\mathbf{B}}, \Sigma_{\mathbf{B}}). \tag{7}$$

$\mu_{\mathbf{B}}$ is length k vector representing the overall mean of the distributions of β coefficients. $\Sigma_{\mathbf{B}}$ is a $k \times k$ covariance matrix of the β coefficients.

What remains is assigning priors the elements of $\mu_{\mathbf{B}}$ and the covariance matrix $\Sigma_{\mathbf{B}}$. Each of the elements of vector $\mu_{\mathbf{B}}$ were given independent, weakly-informative normal priors. The prior for $\Sigma_{\mathbf{B}}$ is a bit more complicated. While the conjugate prior distribution for a covariance matrix is the inverse-Wishart (Gelman et al., 2013), because I am using a variant for Hamiltonian Monte Carlo (HMC) called No U-Turn Sampling (NUTS) for posterior estimation as opposed to Gibbs sampling there is not benefit for using a conjugate prior STAN MANUAL. Additionally, the inverse-Wishart distribution strongly constraints the off-diagonal elements of the covariance matrix. Instead, it is better to model the correlation matrix and separate variance terms for each of the k coefficients. This is possible because of the relationship between a covariance and a correlation matrix, defined as

$$\Sigma_{\mathbf{B}} = \text{Diag}(\tau_B) \Omega_{\mathbf{B}} \text{Diag}(\tau_B) \tag{8}$$

where τ_B is a length k vector of variances and $\text{Diag}(\tau_B)$ is a diagonal matrix.

I used a LKJ prior distribution for $\Omega_{\mathbf{B}}$ as recommended by STAN MANUAL. An LKJ is a single parameter multivariate distribution where values of η greater than 1 concentrate density at the unit correlation matrix, which corresponds to no correlation between the β coefficients. The scale parameter, τ_B, is given a weakly informative half-Cauchy (C^+) prior following Gelman (2006).

Given all the above, the exponential distribution based model is then defined, including priors, as

$$\begin{aligned} y_i &\sim \text{Exponential}(\lambda) \\ \lambda_i &= \exp(\mathbf{X}_i \mathbf{B}_{j[i]}) \\ \mathbf{B} &\sim \text{MVN}(\mu_{\mathbf{B}}, \Sigma_{\mathbf{B}}) \\ \Sigma_{\mathbf{B}} &= \text{Diag}(\tau_B) \Omega_{\mathbf{B}} \text{Diag}(\tau_B) \\ \mu_\kappa &\sim \mathcal{N}(0, 5) \text{ for } \kappa \in 1:k \\ \tau_\kappa &\sim \text{C}^+(1) \text{ for } \kappa \in 1:k \\ \Omega &\sim \text{LKJ}(2). \end{aligned} \tag{9}$$

The Weibull distribution based model is then also defined as

$$\begin{aligned} y_i &\sim \text{Weibull}(\alpha, \sigma) \\ \sigma_i &= \exp\left(\frac{-(\mathbf{X}_i \mathbf{B}_{j[i]})}{\alpha}\right) \\ \mathbf{B} &\sim \text{MVN}(\mu_{\mathbf{B}}, \Sigma_{\mathbf{B}}) \\ \Sigma_{\mathbf{B}} &= \text{Diag}(\tau_B) \Omega_{\mathbf{B}} \text{Diag}(\tau_B) \\ \alpha &\sim \text{C}^+(2) \\ \mu_\kappa &\sim \mathcal{N}(0, 5) \text{ for } \kappa \in 1:k \\ \tau_\kappa &\sim \text{C}^+(1) \text{ for } \kappa \in 1:k \\ \Omega &\sim \text{LKJ}(2). \end{aligned} \tag{10}$$

Note that the above formulations of each model (Eq. 9, 10) does not include how the uncertainty in environmental affinity (Eq. 3) is included nor how censored observations are included. An explination of including censored observations follows.

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 censored genera are those that did not go extinct during the window of observation, or genera 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 the survival function, or $S(t)$. $S(t)$ corresponds to the probability that a genus having existed for t stages will not have gone extinct while $h(t)$ corresponds to the instantaneous extinction rate at taxon age t Klein and Moeschberger (2003). For an exponential model, $S(t)$ is defined as

$$S(t) = \exp(-\lambda t), \tag{11}$$

and for the Weibull distribution $S(t)$ is defined as

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

$S(t)$ 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. 6) 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

$$\begin{aligned}
\mathcal{L} &\propto \prod_{i \in C} \text{Exponential}(y_i|\lambda) \prod_{j \in R} S(y_j|\lambda) \prod_{k \in L} (1 - S(y_k|\lambda)) \\
\mathcal{L} &\propto \prod_{i \in C} \text{Weibull}(y_i|\alpha, \sigma) \prod_{j \in R} S(y_j|\alpha, \sigma) \prod_{k \in L} (1 - S(y_k|\alpha, \sigma))
\end{aligned} \tag{13}$$

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 ($\hat{R} < 1.1$) indicate approximate convergence. Convergence means that the chains are approximately stationary and the samples are well mixed (Gelman et al., 2013).

2.5 Model evaluation

Models were evaluated using both a series of multiple posterior predictive checks and an estimate of out-of-sample predictive accuracy.

The motivation behind posterior predictive checks as tools for determining model adequacy is that replicated data sets using the fitted model should be similar to the original data (Gelman et al., 2013). Systematic differences between the simulations and observed indicate weaknesses of the currently fit model. An example of a technique that is very similar would be inspecting the residuals and Q-Q plots from a linear regression.

The strategy behind posterior predictive checks is to draw simulated values from the joint posterior predictive distribution, $p(y^{rep}|y$, and then compared to the original observed values (Gelman et al., 2013). To accomplish this, for each replicate, a single value is drawn from the marginal posterior distributions of each regression coefficient from the final model (Eq. 9, 10). Then, given the covariate information for each of the observations $\mathbf{X}$, a new set of n genus durations are generated giving a single replicated data set y^{rep}. This is repeated 1000 times in order to provide a distribution of possible values that could have been observed given the model.

In order to compare the fitted model to the observed data, various graphical comparisons or test quantities need to be defined. The principal comparison used here is a comparison between non-parameteric approximation of the survival function $S(t)$ as estimated from both the observed data and each of the replicated data sets. The purpose of this comparison is to determine if the model approximates the same survival/extinction pattern as the original data.

The exponential and Weibull models were compared for out-of-sample predictive accuracy using the widely-applicable information criterion (WAIC) (Watanabe, 2010). Because the Weibull model reduces to the exponential model when $\alpha = 0$, our interest is not in choosing between these models. Instead comparison of WAIC values is useful for better understanding the effect of model complexity on out-of-sample predictive accuracy. The calculation of WAIC used here corresponds to "WAIC 2" formulation recommended by Gelman et al. (2013).

WAIC can be considered fully Bayesian alternative to the Akaike information criterion, where WAIC acts as an approximation of leave-one-out cross-validation which acts as a measure of out-of-sample predictive accuracy. WAIC is calculated

starting with the log pointwise posterior predictive density calculated as

$$\text{lppd} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\Theta^S) \right), \tag{14}$$

where n is sample size, S is the number posterior simulation draws, and Θ represents all of the estimated parameters of the model. This is similar to calculating the likelihood of each observation given the entire posterior. A correction for the effective number of parameters is then added to lppd to adjust for overfitting. The effective number of parameters is calculated, following derivation and recommendations of (Gelman et al., 2013), as

$$p_{\text{WAIC}} = \sum_{i=1}^{n} V_{s=1}^{S}(\log p(y_i|\Theta^S)). \tag{15}$$

where V is the sample posterior variance of the log predictive density for each data point.

Given both equations 14 and 15, WAIC is then calculated

$$\text{WAIC} = \text{lppd} - p_{\text{WAIC}}. \tag{16}$$

When comparing two or more models, lower WAIC values indicate better out-of-sample predictive accuracy.

References

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

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.

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

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

A. I. Miller and M. Foote. Epicontinental seas versus open-ocean settings: the kinetics of mass extinction and origination. *Science*, 326(5956):1106–9, Nov. 2009. ISSN 1095-9203. doi: 10.1126/science.1180061. URL `http://www.ncbi.nlm.nih.gov/pubmed/19965428`.

J. L. Payne, N. A. Heim, M. L. Knope, and C. R. Mcclain. Metabolic dominance of bivalves predates brachiopod diversity decline by more than 150 million years. *Proceedings of the Royal Society B*, 281:20133122, 2014.

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`.

S. Watanabe. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.