A model of biological and phylogenetic effects on mammalian co-occurrence

Peter D Smits
Committee on Evolutionary Biology, University of Chicago

February 10, 2015

1 Methods

1.1 Species information

1.1.1 Informal phylogeny

1.2 Biogeographic network

A biogeographic network is defined as a bipartite network between localities and taxa (Sidor et al., 2013, Vilhena, 2013, Vilhena et al., 2013). The properties of a bipartite network are such that taxa are connected to localities, but taxa are not connected to taxa and localities are not connected to localities (Fig. 1a). In this study, taxa are defined as species and localities are defined as 2x2 latitude–longitude grid cells from an azimuthal equal-area map projection.

1.3 Co-occurrence model

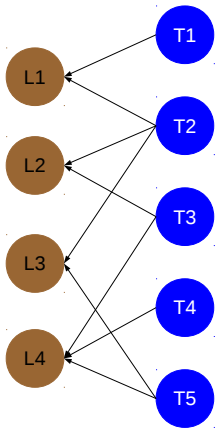
A random graph is one where the probability of any two nodes being connected is equal for all combinations (Erdos and Renyi, 1959). This assumption implies that the distribution of nodal degree should follow a Poisson distribution. What this means is that the degree, or number of connections, of a node in a one mode network is considered drawn from a Poisson distribution with some rate, λ.

The Poisson distribution with rate $\lambda > 0$ is defined

$$\text{Poisson}(y_i|\lambda_i) = \frac{1}{y_i!}\lambda^{y_i}\exp(-\lambda_i) \tag{1}$$

$$\lambda_i = \exp(\beta^T \mathbf{X_i} + h_i + \log(u_i)). \tag{2}$$

(a)



(b)

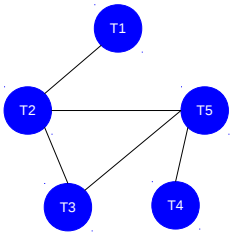


Figure 1: Bipartite and one mode graphs.

The biggest assumption of the Poisson distribution is that mean and variance are equal is very strong and not frequently observed, instead there tends to be excess variance or overdisperssion. The negative binomial distribution can be viewed as an extension of the Poisson distribution allowing for overdispersion.

Negative binomial distribution parameterized with mean $\mu > 0$ and overdispersion $\phi > 0$ is defined

$$\text{Negative binomial}(y_i|\mu_i, \phi) = \binom{y_i + \phi - 1}{y_i} \left(\frac{\mu_i}{\mu_i + \phi}\right)^{y_i} \left(\frac{\phi}{\mu_i + \phi}\right)^{\phi} \tag{3}$$

$$\mu_i = \exp(\beta^T \mathbf{X_i} + h_i + \log(u_i)) \tag{4}$$

$$\phi \sim \text{halfCauchy}(2.5). \tag{5}$$

$\mathbf{X}$ is defined as an $n \times K$ matrix where n is the number of observations and K is the number of covariates of interest. As described above, the covariates of interest are the dietary and locomotor categories of a species and its body mass. While body mass is a continuous covariate, dietary and locomotor categories are index variables that need to be transformed into multiple binary covariates or indicator variables. To do this, for each of these variables were each transformed into $n \times (k-1)$ matrices where $k-1$ is the number of categories of the index variable (3 and 4, respectively). Only $k-1$ indicator variables are necessary as the intercept takes on the remaining category. Finally, a column of 1-s in included in the matrix $\mathbf{X}$ whose corresponding β coefficient is the intercept. In total, K equals 7.

For the parameterizations of the means for both the Poisson (Eq. 2) and negative binomial models (Eq. 4), a unique coefficient β is assigned to each of the covariates and are given a diffuse, weakly informative prior ($\beta_k \sim \mathcal{N}(0, 10)$).

h, or phylogenetic effect, is defined as a random multivariate normally distributed variable whose covariance matrix known up to a constant σ_p (Eq. 6).

$$\begin{aligned} h &\sim \text{Multivariate } \mathcal{N}(0, \mathbf{\Sigma}) \\ \mathbf{\Sigma} &= \sigma_p^2 \mathbf{V}_{phy} \\ \sigma_p &\sim \text{halfCauchy}(2.5). \end{aligned} \tag{6}$$

This parameterization follows Lynch (1991) and Housworth et al. (2004).

u_i is the exposure term for observation i and is defined as the number of localities species i occurred in during the given stage. The inclusion of an exposure is so that the rate is more biologically interpretable as the rate is now in relation to some baseline. In this case, we can 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 in the language generalized linear modeling (Gelman and Hill, 2007). The inclusion of $\log(u_i$ in the parameterization of μ_i is due to the following relationships

$$\begin{aligned} \text{E}(Y) &= u_i \lambda_i \\ \log(\text{E}(Y)) &= \log(u_i) + \log(\lambda_i) \\ \log(\text{E}(Y)) &= \log(u_i) + \beta^T \mathbf{X_i} + h_i \\ \log(\text{E}(Y)) - \log(u_i) &= \beta^T \mathbf{X_i} + h_i \\ \log\left(\frac{\text{E}(Y)}{u_i}\right) &= \beta^T \mathbf{X_i} + h_i. \end{aligned}$$

The overdispersion parameter ϕ of the negative binomial distribution was given weakly informative half-Cauchy prior distribution (Eq. 5).

Graphical summaries of the Poisson and negative binomial models, along with all values for the prior distributions, are presented in Figures 2 and 3, respectively.

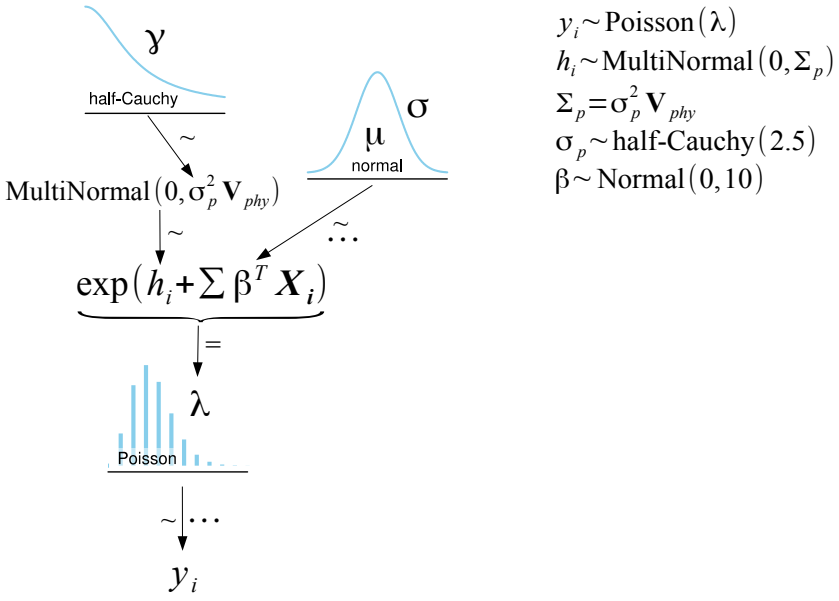


Figure 2: Poisson model

1.4 Posterior predictive checks

A models utility in making predictive and descriptive statements is dependent on its fit to the data and the questions of interest. The most basic assessment of model fit is if, given known covariate information, values simulated from the model (y^{rep}) are similar to the observed (y). This is the idea behind posterior predictive checks (Gelman et al., 2013).

For all estimated parameters, a value is randomly drawn from their respective posteriors. These parameter estimates are then used with covariate information for all observations to make n estimates of the number of co-occurring species for a given species y^{rep}. This process is repeated 1000 times to get a distribution of possible y^{rep}-s, which can then be compared to the original data y to assess the quality of model fit (Gelman et al., 2013).

1.5 Predictive estimates

For each stage,

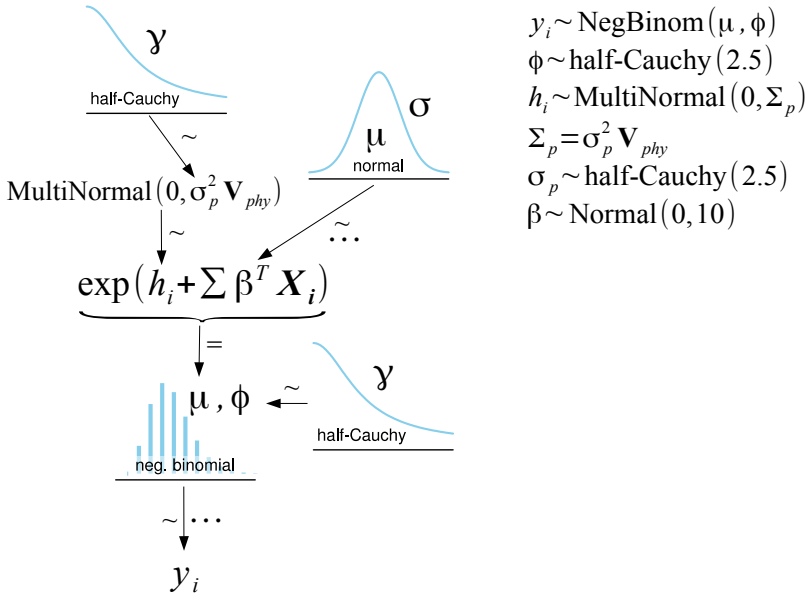


Figure 3: Negative binomial model

References

P. Erdos and A. Renyi. On random graphs I. *Publicationes Mathematicae*, 6: 290–297, 1959.

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.

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

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

C. A. Sidor, D. A. Vilhena, K. D. Angielczyk, A. K. Huttenlocker, S. J. Nesbitt, B. R. Peecook, J. S. Steyer, R. M. H. Smith, and L. A. Tsuji. Provincialization of terrestrial faunas following the end-Permian mass extinction. *Proceedings of the National Academy of Sciences*, 110(20):8129–33, May 2013. ISSN 1091-6490. doi: 10.1073/pnas.1302323110.

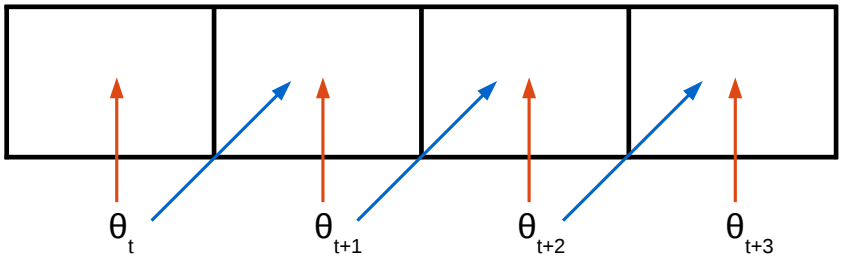


Figure 4: CAPTION

D. A. Vilhena. *Boundaries and dynamics of biomes*. PhD thesis, University of Washington, 2013.

D. A. Vilhena, E. B. Harris, C. T. Bergstrom, M. E. Maliska, P. D. Ward, C. A. Sidor, C. A. E. Strömberg, and G. P. Wilson. Bivalve network reveals latitudinal selectivity gradient at the end-Cretaceous mass extinction. *Scientific reports*, 3:1790, May 2013. ISSN 2045-2322. doi: 10.1038/srep01790.