Report

August 2, 2016

Summary

On this report we review the theory behind the diversity-dependence model under the framework described on the introductory essay. In the later section we show results of the algorithm. To materialize this theory we

- Write MLE code for diversity-dependence model
- Paralellize the code
- Create R package

Overview

We consider a phylogenetic tree, mathematicaly expressed as a set $Y = (\mathcal{T}, \Upsilon)$, where $\mathcal{T}$ represent the set of branching times [1], and Υ has the information of the topology of the tree. The markov nature of the process means that the likelihood is exaclty the product of the conditional densities [2], in other words, the likelihood of the tree is then described as a multiplication of an exponenial distribution and a multinomial distribution

$$L(\theta|Y) = \prod_{i}^{N} -\sigma_i(\theta) e^{-\sigma_i(\theta)t_i} \frac{\rho_i(\theta)}{\sigma_i(\theta)}$$

thus, the log-likelihood is

$$l(\theta|Y) = \sum_{i}^{N} -\sigma_i(\theta)t_i + log(\rho_i(\theta))$$

Diversity-dependence model

For the simplest diversity-dependence model

$$\lambda_{i,j} = \lambda_0 - (\lambda_0 - \mu_0)\frac{n_i}{K}, \qquad \mu_n = \mu_0$$

The MLE can be finded partialy analiticaly and partialy numerically. First we consider σ_i and ρ_i

$$\sigma_i = \sum_{j=1}^{N} \lambda_0 - (\lambda_0 - \mu_0)\frac{n_i}{K} + \mu_0 = n_i(\lambda_0 + \mu_0) - n_i^2\beta_0$$

where $\beta_0 = \left(\frac{\lambda_0 - \mu_0}{K}\right)$, and

$$\rho_i = E_i(\lambda_0 - n_i\beta_0) + (1 - E_i)\mu_0$$

[1] that is, t_i is described as the minumun time over all possible times any species could take to speciate/extinct after t_{i-1}

[2] please see the introductory esay for details

Here, n_i is defined as the number of species at time t_i and E_i is a binary vector with 1 if there was an speciation at time t_i or 0 is there was an extinction at time t_i.

Thus, seeking for the MLE values, we analyze the thee equations

$$\begin{cases} \frac{\partial l(\lambda,\beta,\mu|Y)}{\partial \lambda} = 0 \\ \frac{\partial l(\lambda,\beta,\mu|Y)}{\partial \beta} = 0 \\ \frac{\partial l(\lambda,\beta,\mu|Y)}{\partial \mu} = 0 \end{cases}$$

Firstly, after some algebra, we find a very nice analytical solution for the extinction rate parameter

$$\frac{\partial l(\lambda,\beta,\mu|Y)}{\partial \mu} = 0 \Leftrightarrow \hat{u}_0 = \frac{\sum_{i=1}^{N}(1-E)}{\sum_{i=1}^{N}(n_i t_i)} \tag{1}$$

Moreover, with the other two equations, we have the following system

$$\begin{cases} \sum_{i=1}^{N} \frac{E_i}{\lambda - n_i\beta} = \sum_{i=1}^{N} n_i t_i \\ \sum_{i=1}^{N} \frac{E_i n_i}{\lambda - n_i\beta} = \sum_{i=1}^{N} n_i^2 t_i \end{cases}$$

A numerical efficient method to solve this system is described in the appendix.

Algorithm

Under the previous results, we developed an algorithm able to find accurate solution for the MLE. The algorithm is based on the following formulation.

Step 1. get $\hat{\mu_0}$ using eq 1.
Step 2. Consider the function

$$\hat{\beta}(\lambda) = \arg\max_{\beta} L(\lambda, \beta, \hat{\mu}_0)$$

Step 3. Calculate the MLE such that,

$$(\hat{\lambda}, \hat{\beta}, \hat{\mu}) = \arg\max_{\lambda} L(\lambda, \hat{\beta}(\lambda), \hat{\mu})$$

Two important properties of this algorithm ensures convergence (the writting of the proof is pending):

1. $\hat{\beta}(\lambda)$ is a linear function of λ
2. $\arg\max_{\lambda} L(\lambda, \hat{\beta}(\lambda), \hat{\mu})$ is a convex funcion of λ.

This two results ensures the existence of an unique global maximum, which is easily calculated by several classic optimization methods.

As an example, on figure we can see the estimations over 1000 simulations of the diversity-dependence process with true values $\lambda_=0.8, \beta_0 = 0.0175, \mu_0 = 0.1, K = 40$ and crown time $= 15$. We can see accurate estimations to the true value.

Results

On table 1 we can see the bias and precision of the maximum-likelihood estimates, as shown by the median and the 25th and 75th percentiles of the estimated parameters of 100 simulated datasets. There we can see that the results are quite accurate regarding the real values.

Morever, on table 2 we have the estimated parameters of the diversity-dependence model simulated under the DDD package. Here we can see that the algorithm is even able to capture true values simulated under the DDD framework.

Table 1: MLE estimation of 100 simulations. Simulations and estimations are from the algorithm described above.

Simulated Parameters				Estimated parameters (25th, 50th, 75th percentiles)								
				λ_0			μ			K		
λ_0	K	crown age	μ	025th	50th	75th	025th	50th	75th	025th	50th	75th
0.8	40	5	0	0.71	0.87	1.04	0.00	0.00	0.00	31.20	39.09	440.16
			0.1	0.76	0.92	1.11	0.07	0.10	0.13	22.23	32.65	65.39
			0.2	0.80	0.96	1.28	0.13	0.19	0.26	12.74	31.76	83.12
			0.4	1.05	1.23	1.55	0.27	0.36	0.44	7.00	17.61	30.58
		10	0	0.68	0.79	0.87	0.00	0.00	0.00	39.63	40.98	43.92
			0.1	0.71	0.86	0.98	0.09	0.10	0.12	37.11	39.52	42.01
			0.2	0.78	0.91	1.05	0.18	0.20	0.23	34.08	38.13	43.32
			0.4	0.87	1.01	1.20	0.37	0.41	0.46	18.32	30.13	42.13
		15	0	0.69	0.77	0.87	0.00	0.00	0.00	39.58	40.00	41.03
			0.1	0.72	0.80	0.91	0.09	0.10	0.11	38.72	39.89	40.98
			0.2	0.78	0.84	0.96	0.18	0.20	0.22	38.29	40.25	41.93
			0.4	0.79	0.90	1.00	0.38	0.40	0.43	31.40	37.38	43.89

Table 2: MLE estimation of 100 simulations. Simulations are from the 'DDD' package and estimation from p1 algorithm.

Simulated Parameters				Estimated parameters (25th, 50th, 75th percentiles)								
				λ_0			μ			K		
λ_0	K	crown age	μ									
0.8	40	5	0	0.74	0.91	1.11	0.00	0.00	0.00	34.84	41.04	59.34
			0.1	0.94	1.15	1.28	0.08	0.11	0.14	26.57	32.55	39.98
			0.2	1.03	1.22	1.46	0.17	0.21	0.29	17.06	27.85	37.47
			0.4	1.13	1.43	1.67	0.33	0.42	0.52	9.40	17.52	27.29
		10	0	0.75	0.87	0.98	0.00	0.00	0.00	38.55	39.33	40.35
			0.1	0.78	0.90	1.05	0.09	0.10	0.12	37.09	38.55	40.33
			0.2	0.86	0.96	1.08	0.19	0.21	0.23	34.47	37.45	40.38
			0.4	0.96	1.11	1.22	0.38	0.42	0.45	28.11	33.88	40.41
		15	0	0.74	0.83	0.96	0.00	0.00	0.00	38.57	39.05	40.00
			0.1	0.80	0.88	0.99	0.09	0.11	0.12	37.57	38.67	39.57
			0.2	0.88	0.95	1.05	0.20	0.21	0.22	36.98	38.58	40.07
			0.4	0.92	1.03	1.14	0.38	0.41	0.44	34.02	37.45	41.96

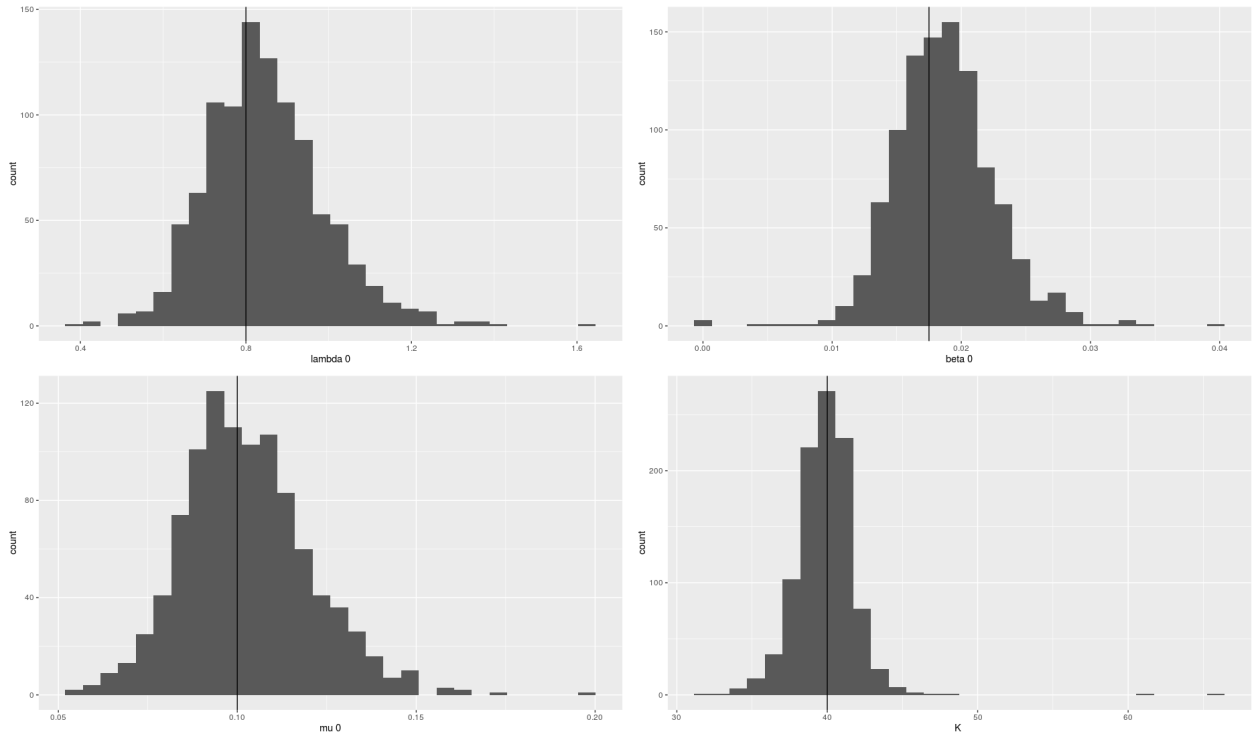

Figure 1: Estimations over 1000 simulations of the diversity-dependence process with true values $\lambda_{=}0.8, \beta_0 = 0.0175, \mu_0 = 0.1, K = 40$ and crown time $= 15$. The black vertical lines shows the real values.

Further steps: First Ideas about MCEM applied to incomplete phylogenies

To complete this first example we would like to estimate trees where extinc species are not observable. Bellow is a draft of the pseudo-code, inspired on a montecarlo EM algorithm approach.

Algorithm 1 Montecarlo EM aproach

```
Set number of iterations, initial parameters
repeat
    for i in 1:nit do
      simulate a reconstructed phylogenetic tree with algorithm 2
      use the simulated tree to estimate parameters
    end for
average parameters and use it as initial parameters

until convergence
```

Algorithm 2 simulation of a reconstructed tree

```
Load the incomplete phylogenetic tree
set t be the vector of branching times of the incomplete tree
for i in 1:length(t) do
    update λ and μ for moment t_i
    simulate a new branching time t_temp from t_i
    while t_temp < t_{i+1} do
        simulate the extinction event from exponential distribution with rate μ_i
        simulate new branching time t_temp from previous t_temp and extinction event
        update λ and μ for moment t_i
    end while
update the reconstructed tree
end for
return Reconstructed tree
```

Appendix

In order to solve the system given by the diversity-dependence set up we consider the following idea,

Given

- $b_1, b_2, \ldots, b_n \in \{0, 1\}$
- $m_1, m_2, \ldots, m_n \in \mathbb{N}$
- $c_1, c_2 \in \mathbb{R}$

define

$$z_i := \frac{1}{\lambda - m_i \beta}$$

we have a system of 2 equaions in $\mathrm{z} \in \mathbb{R}^n$

$$\begin{bmatrix} b_1 & b_2 & \ldots & b_n \\ b_1 m_1 & b_2 m_2 & \ldots & b_n m_n \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

or, in a more succinct form, $\mathrm{Az} = \mathrm{c}$. If $n > 2$, this is an underdetermined system whose **least-norm** solution is

$$\hat{\mathrm{z}} := \mathrm{A}^T (\mathrm{A}\mathrm{A}^T)^{-1} \mathrm{c}$$

If all the entries of $\hat{\mathrm{z}}$ are nonzero, then we have an overdetermined system of equations

$$\begin{aligned} x - m_1 y &= \hat{z}_1^{-1} \\ x - m_2 y &= \hat{z}_2^{-1} \\ &\vdots \\ x - m_n y &= \hat{z}_n^{-1} \end{aligned}$$

Lastly, we compute the **least-squares** solution $(\hat{\lambda}, \hat{\beta})$. If $\hat{z}_i = 0$, then the i-th equation, whose right-hand side is illegal, is simply discarted.