Mehodology on the EMPHASIS framework

1. Introduction

On this document we describe the methodology we use on the EMPHASIS package for parameter estimation on phylogenetic analysis. Aditionally, we include the code to compute every step on the diversity dependance case.

We approach this problem by applying an EM-type algorithm, which defines the EM iteration $\phi^* \to \phi$ as

E-step Compute $Q(\phi|\phi^*) = E_{\phi^*}(\log f(x|\phi)|y)$,

M-step Choose ϕ to be the value of $\phi \in \Omega$ which maximizes $Q(\phi|\phi^*)$

Starting from an initial value for the parameter $\phi \in \Omega \subset \mathbb{R}^m$, we performs this two steps iterativelly until reaching convergence on the parameter space Ω. On the following sections we explain how we implement it on each step on the algorithm.

Notation

T is the total time from crown time to present. We call a *missing species*, those branches on the phylogenetic tree which their tip is not at time T. We assume that the time that a branch j takes to diverge on two new branches has an exponential distribution

$$f(t) = \rho_t e^{-\int \rho_t dt}$$

... We define $\mu_{t,j}$ and $\lambda_{t,j}$ as the extinction and speciation rates of species j at time $t \in (0,T)$. where $\lambda_t = \sum_{j=1}^{n_t} \lambda_{t,j}$[1]

We define n_t as the number of extant species at time t, m_t the number of missing species at time t, and we will be specially interested on the branching points of the phylogenetic tree $\{t_1, t_2, ..., t_d\}$ and their corresponding waiting times $\{\Delta t_1, ..., \Delta t_d\}$. We also define the sum of rates at a specific moment t_i as

$$\sigma_i = \sum_{j=1}^{n_{t_i}} \lambda_{t_i,j} + \mu_{t_i,j}$$

and the the sum of speciation rates at time t_i as

$$s_i = \sum_{j=1}^{n_{t_i}} \lambda_{t_i,j}$$

moreover, we define the topology component of a tree at time t_i as the vector $\tau_i \in \{0,1\}^{2*n_{t_i}}$. So, this vector contains a 1 on the position j if species j has speciated at time t_i, or a 1 on position $(n_{t_i} + j)$ if species j has get extinct on time t_i, and 0´s elsewhere. We finally define

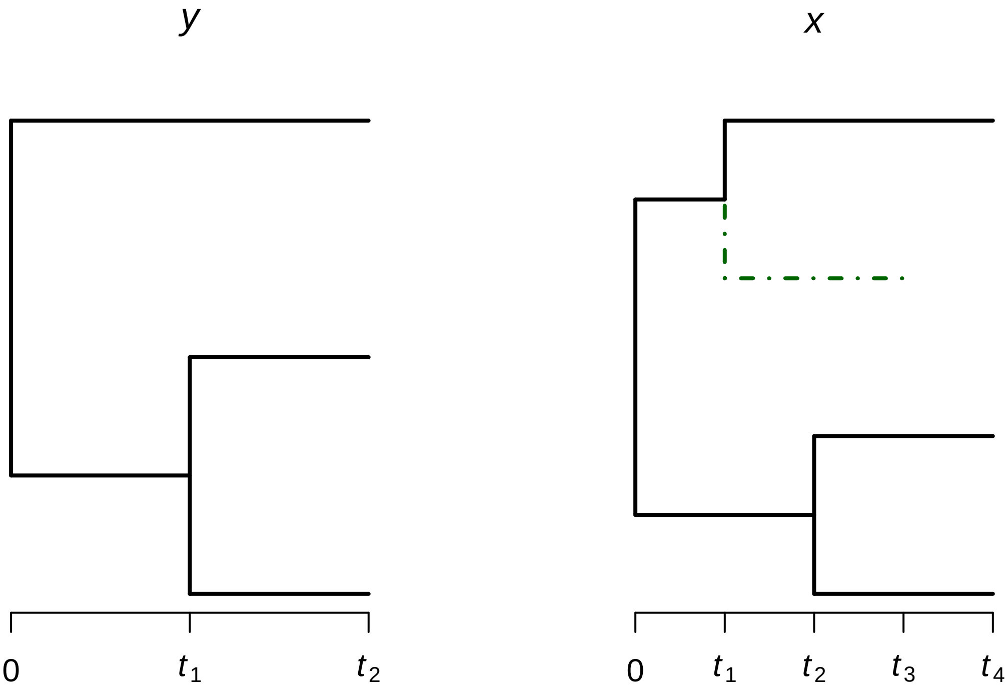
$$\rho_i = \prod_{j=1}^{n_{t_i}} (\lambda_{(t_i,j)})^{\tau_i^{(j)}} \prod_{j=1}^{n_{t_i}} (\mu_{t_i,j})^{\tau_i^{(n_{t_i}+j)}}$$

where $\tau_i^{(j)}$ is the j-component of the vector τ_i. Note that ρ_i is just the speciation or extinction rate of species diversifying on branching time t_i. As we assume that only one speciation or extinction take place at a moment t, the vector τ_i in principle contains only one 1 and 0´s elsewhere and we define it as $\mathbb{1}_j^{(n)}$ for a vector of

[1] Note that $s_i = \lambda_{t_i}$, they are just continuos/discrete caracterizations of the same process.

dimension n with a 1 at position j and zeros elsewhere. The case of multiple speciations is a possible future extension.

Hereafter we refer to $x \in \mathcal{X}$ as a variable representing a complete tree (extinct species included) and $y \in \mathcal{Y}$ as variable representing the observed (ultrametric) tree, moreover, $\mathcal{X}(y)$ is the subset of complete trees that has exactly same extant species at the present than the ultrametric tree y. For a detailed explanation of this time-tree spaces we refer to [**?**]. On the chart bellow we see an example of $y \in \mathcal{Y}$ and $x \in \mathcal{X}(y)$.



2. E-step

On the E-step we want to calculate

$$Q(\phi|\phi^*) = E_{\phi^*}(\log f_X(x;\phi)|y) = \int_{\mathcal{X}(y)} \log f_X(x;\phi) f_{X|Y}(x|y,\phi^*)dx$$

where $f_X(x;\phi)|y)$ is the probability density of a complete tree defined as

$$f_X(x;\phi) = \left(\prod_{i=1}^{d} e^{-\sigma_i \Delta t_i}\right)\left(\prod_{i=1}^{d-1} \rho_i\right) \tag{1}$$

Moreover, $f_{X|Y}(x|y,\phi)$ is the probability density of the complete tree x given that $x \in \mathcal{X}(y)$. That probabily does not have a close form. On the same way, the calculation of $Q(\phi|\phi^*)$ has not a close form neither due to the huge complexity of the space $\mathcal{X}(y)$, then numerical calculations are needed.

2.1 Monte-Carlo aproximation and data arguentation algorithm

One way to perform this task is considering a Monte-Carlo sampling, where, given a set of sampled trees $x_1, ..., x_p$ from $f_{X|Y}(x|y,\phi^*)$, we approximate $Q(\phi|\phi^*)$ by

$$E_{\phi^*}(\log f(x;\phi)|y) \approx \frac{1}{p}\sum_{i=1}^{p} \log f_{X;\phi^*}(x_i;\phi) \tag{2}$$

We can, theoretically, make this approximation as accurate as we want by increasing the sampling size p if we know the exact sampling probability $f_{X|Y}(x|y,\phi^*)$, however that distribution does not hold a close form on our framework. Thus, an approximated data argumentation algorithm, and general approach incorporating importance sampling is needed.

2.1.1 Data argumentation algorithm

In order to simulate the extinct part of a tree when the extant species and their diversification events are observed, we consider the event:

"t_{M_j}: *Time for branch j to speciate into a new species that is going to get extinct before present.*"

which will have a rate of $\lambda_{t,j}(1-e^{-\mu(r-t)})$, where r is the time from the starting point of the process to the present. The probability distribution of this process would be

$$f_{M_j}(\Delta t_{M_j} = t) = \lambda_{t,j}(1-e^{-\mu_{t,j}(r-t.)})e^{-\int_0^t \lambda_{t,j}(1-e^{-\mu_{t,j}(r-z)})dz} \tag{3}$$

Moreover we define $\Delta t_M = \min\{t_{M_1},...,t_{M_n}\}$, then

$$f_M(t_M = t) = \lambda_t(1-e^{-\mu(r-t.)})e^{-\int_0^t \lambda_t(1-e^{-\mu(r-z)})dz} \tag{4}$$

Similarly, if we define m_t as the number of missing species at time t, we can define the processes

"t_{E_k}: *Time for missing species k to get extinct since born.*"

By definition $t_{E_k} < T$, then the probablity distribution would be

$$f_{E_k}(t_{E_k} = t) = f(t_{E_k} = t|t < T) = \frac{\mu_{t,k}e^{-\int_0^t \mu_{t,k}dt}}{(1-e^{-\int_0^t \mu_{t,k}dt})} \tag{5}$$

similarly, we define $t_E = \min\{t_{E_1},...t_{E_m}\}$. Note that if there are not missing species, then $t_E = \min\{\emptyset\} = \infty$. Moreover, the probability density function for the waiting time of a speciation of species j that get extinct at t_{ext} is

$$\begin{aligned} f_D(D = c(t_{spe}, t_{ext}, \mathbb{1}_j^{(n_{t_{spe}})})) =& f_M(t_M = t_{spe})f_{t_{E_j}}(t_{E_j} = t_{ext})P(\tau_j = \mathbb{1}_j^{(n_{t_{spe}})}|\Delta t_{M_j} = t_{spe}) \\ =& \lambda_{t_{spe},j}e^{-\int_0^{t_{spe}} \lambda_t(1-e^{-\mu(r-z)})dz}\mu_{t_{ext},j}e^{-\int_{t_{spe}}^{t_{ext}} \mu_{t,j}dt} \end{aligned} \tag{6}$$

The data argumentation algorithm, given the set of waiting times $\Delta t_1, ..., \Delta t_d$, will be then

1. set $i = 1$
2. $\Delta t = \Delta t_i$. If $i > d$ go to step 8
3. Draw speciation time t_{spe} of next missing species from $f_M(t)$
4. Draw extinction time t_{ext} of next extinction from $f_E(t)$
5. If $\Delta t < t_{ext}$ and $\Delta t < t_{spe}$, set $i \leftarrow i + 1$ and go to step 2

6. If $t_{spe} < t_{ext}$ and $t_{spe} < \Delta t$ add new speciation to the tree, set $\Delta t \leftarrow \Delta t - t_{spe}$, and go to step 3
7. If $t_{spe} > t_{ext}$ and $t_{ext} < \Delta t$ add new extinction to the tree, set $\Delta t \leftarrow \Delta t - t_{ext}$, and go to step 3
8. Return the complete tree.

Once we get the missing part of the tree, we neet to calculate the importance weights by using the sampling probability and the real probability of the whole tree. We do that on next section.

2.1.2 Importance sampling

Due to the complexity of the problem we cannot sample from $f_{X|Y}(x|y,\phi^*)$, but we can still sample from an approximated distribution $g_{X|Y}(x|y,\phi^*)$ (using the data argumentation algorithm described on previous section), and then correct via importance sampling, using a more general approximation instead

$$E_{\phi^*}(\log f(x;\phi)|y) \approx \sum_{i=1}^{p} \log f_{X;\phi^*}(x_i;\phi)\frac{f_{X|Y}(x_i|y,\phi^*)}{g_{X|Y}(x_i|y,\phi^*)}$$

While we can compute $g_{X|Y}(x_i|y,\phi^*)$ readily using the nonhomogeneous Poisson process of equation (4) and (5), we must still find an expression for $f_{X|Y}(x_i|y,\phi^*)$. We can write this as

$$f_{X|Y}(x_i|y,\phi^*) = \frac{f_{X,Y}(x_i,y|\phi^*)}{f_Y(y|\phi^*)}$$

using the law of conditional probabilities. Because the denominator is the same for all x_i and does not depend on ϕ it will not affect our maximization step, and we can simply write

$$Q(\phi|\phi^*) = E_{\phi^*}(\log f(x;\phi)|y) \approx \frac{1}{f_Y(y|\phi^*)} \sum_{i=1}^{m} \log f_{X;\phi^*}(x_i;\phi)\frac{f_{X,Y}(x_i,y|\phi^*)}{g_{X|Y}(x_i|y,\phi^*)}$$

where $\frac{1}{f_Y(y|\phi^*)}$ is just a (unknown) constant value. . Note that the dependence on ϕ (which is important for the maximization step) only occurs in the term $\log f_{X;\phi^*}(x_i;\phi)$.

We call

$$w_i = \frac{f_{X,Y}(x_i,y|\phi^*)}{g_{X|Y}(x_i|y,\phi^*)}$$

the importance weights. So,

$$Q(\phi|\phi^*) \propto \sum_{i=1}^{m} w_i \log f_{X;\phi^*}(x_i;\phi) \tag{7}$$

To calculate $g_{X|Y}(x_i|y,\phi^*)$ we introduce the vector $\xi = (\xi_1, ..., \xi_d)$ where $\xi_i = 1$ if the branching time t_i is drived by a speciation of a missing species and $\xi_i = 0$ else (i.e speciation of observed species or extinction)

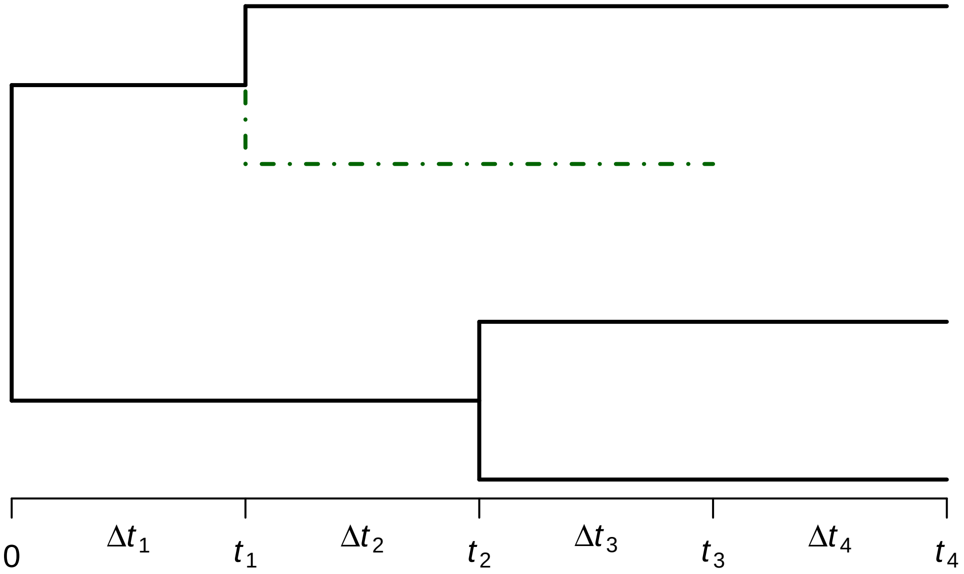
Then

$$g_{X|Y}(x_i|y,\phi^*) = \prod_{i=1}^{d} \left(f_D(D = (\Delta t_i, t_{E_k}, \tau = \tau_i)\right)^{\xi_i} \left(f_M(t_M > \Delta t_i)\right)^{1-\xi_i} \tag{8}$$

Example

On the tree below, we can calculate the sampling probability of the sampled (green) missing species in the following way:

$$g_{X|Y}(x|y,\phi) = f_M(t=\Delta t_1)\frac{f_E(t=\Delta t_2+\Delta t_3)}{f_E(t<\Delta t_1+\Delta t_2+\Delta t_3)}f_M(t>\Delta t_2)f_M(t>\Delta t_3)f_M(t>\Delta t_4)$$



2.2 Application + Code

Case 1: Diversity dependande

The first example we explore is the diversity dependance case. That process define diversification rates as

$$\begin{aligned}\lambda_{t,j} &= \lambda_0-(\lambda_0-\mu_0)\frac{n_t}{K},\\ \mu_{t,j} &= \mu_0\end{aligned}$$

Then $\sigma_i = (\lambda_0-(\lambda_0-\mu_0)\frac{n_t}{K}+\mu_0)n_t$ and $\rho_i = (\lambda_0-(\lambda_0-\mu_0)\frac{n_t}{K})\tau_i+\mu_0(1-\tau_i)$, where in this case the topology value is simplified as a scalar equal to 0 if there was an extinction at time t_i and 1 if there was an speciation at time t_i. That is because all species has same rates at time t.

Below is the code for the negative loglikelihood function

```
#negative logLikelihood of a tree
nllik.tree <- function(pars,tree){
  wt = tree$wt
  to = tree$to
  n = c(2,2+cumsum(to)+cumsum(to-1))
  lambda = (pars[1]-(pars[1]-pars[2])*(n/pars[3]))
  mu = pars[2]
  sigma = (lambda + mu)*n
  rho = pmax(lambda[-length(lambda)]*to+mu*(1-to),0)
  nl = -(sum(-sigma*wt)+sum(log(rho)))
  if(min(pars)<0){nl = Inf}
  return(nl)
}
```

with that, we can define inmediatelly the Q function of equation (7)

```
Q.aprox = function(pars, st){
  m = length(st$rec)
  l = vector(mode = 'numeric',length = m)
  w = vector(mode = 'numeric',length = m)
  for(i in 1:m){
    s = st$rec[[i]] # complete tree
    w[i] = st$w[i] # corresponding weight
    if(w[i]!=0){ # if weight is non-zero, calculate likelihood
      l[i] = nllik.tree(pars,tree=s)
    }else{
      l[i] = 0
    }
  }
  w = w/sum(w) #normalization of weights
  L = sum(l*w)
  return(L)
}
```

The next step is to write down the data argumentation algorithm, so, given an observed (ultrametric) tree, simulate the extincted part:

```
extinction.processes <- function(u,inits,mu0){
  nm = length(u)
  t.ext = vector(mode='numeric',length=nm)
  if(nm > 0){
    for(i in 1:nm){
      t.ext[i] = inits[i] - log(1-u[i])/mu0  #Inverse of the intensity function for constant extinction
    }
  }
  return(t.ext)
}
###  simulation of extincted new version
sim.extinct <- function(brts,pars,model='dd',seed=0){
  if(seed>0) set.seed(seed)
  wt = -diff(c(brts,0))
  ct = sum(wt)
  lambda0 = pars[1]
  mu0 = pars[2]
  K = pars[3]
  dim = length(wt)
  ms = NULL # missing speciations, for now we just add time. When we consider topology we do it with sp
  me = NULL # missing extinctions (in the uniform plane)
  bt = NULL
  to = NULL
  cbt = 0 # current branching time
  N = 2 # number of species (crown time)
  sprob = NULL # sampling probability of Missing|observed
  h = 1 # index to fill probabilities
  for(i in 1:dim){
    cwt = wt[i]
    cbt = sum(wt[0:(i-1)])
    key = 0
    gosttime = 0
    while(key == 0){
      if(model == "dd"){  # diversity-dependence model
        lambda = max(1e-99, lambda0 - (lambda0-mu0)*N/K)
        mu = mu0
        s = N*lambda
      }else{print('Model not implemented yet, try dd')}
      t.spe = rexp(1,s)
      t.ext = extinction.processes(u=me,inits=ms,mu0=mu0)
      #sometimes parameters does not make sense. write a warning when that happens
      t_ext = ifelse(length(t.ext)>0,min(t.ext),Inf)-cbt  # if is not empty gives the waiting time for
      mint = min(t.spe,t_ext)
      if(mint < cwt){
        if(mint == t.spe){#speciation
          u = runif(1)
          if(u < pexp(ct-(cbt+t.spe),mu)){
            ms = c(ms,cbt+t.spe)
            me = c(me,u)
            bt = c(bt,cbt+t.spe)
            to = c(to,1)
            sprob[h] = sampprob(t = t.spe+gosttime, s = s, mu = mu, r = ct-(cbt-gosttime),N=N)
            h = h + 1
            N = N + 1
          }else{gosttime = t.spe + gosttime}
          cwt = cwt - t.spe
          cbt = cbt + t.spe
        }
        else{#extinction
          extinctone = which(t.ext == min(t.ext))
          tspe = ms[extinctone]
          text = t.ext[extinctone]
          bt = c(bt,text)
          to = c(to,0)
          sprob[h] = truncdist::dtrunc(text-tspe,'exp',a=0,b=ct-tspe,rate=mu)*(1-integrate(sampprob,lowe
          ms = ms[-extinctone]
          me = me[-extinctone]
          cwt = cwt - mint
          cbt = cbt + mint
          N = N-1
          h = h+1
          gosttime = 0
        }
      }
      else{
        key = 1
        sprob[h] = (1 - integrate(Vectorize(sampprob),lower = 0, upper = cwt+gosttime,s=s,mu=mu,r=ct-cb
        h = h+1
      }
    }
    N = N+1
  }
  df = data.frame(bt = c(bt,ct-brts),to = c(to,rep(2,length(wt))))
  df = df[order(df$bt),]
  n.tree = list(wt=c(diff(df$bt),ct-df$bt[length(df$bt)]),E=df$to[-length(df$to)])
  if(length(n.tree$E==1) != length(n.tree$E==0)) print('algo mal!!')
  n.tree$xi = vector('numeric',length(n.tree$E))
  n.tree$xi[n.tree$E==2] = 0
  n.tree$xi[n.tree$E==1] = 1
  n.tree$xi[n.tree$E==0] = 0

  n.tree$E[n.tree$E==2] = 1


  lrprob = -nllik.tree(pars,n.tree) #f
  lsprob = sum(log(sprob)) #g
  logweight = lrprob-lsprob
  if(logweight==Inf) logweight = -Inf
  n.tree$weight = exp(logweight)
  n.tree$logweight = logweight
  n.tree$f=lrprob
  n.tree$g=lsprob
  return(n.tree)
}
```

given that on this case speciation rates are piece-wise contant, on every waiting time, we can write equation (4) as

$$P_M(t) = s_\lambda(1-e^{-\mu(r-t.)})e^{-\int_0^t s_\lambda(1-e^{-\mu(r-z)})dz} = s_\lambda(1-e^{-\mu(r-t)})e^{-s_\lambda\left[t+\frac{1}{\mu}(e^{-\mu r}-e^{-\mu(r-t)})\right]}, t<r \qquad (9)$$

then, equation (8) would be equivalent to

$$g_{X|Y}(x|y,\phi) = \prod_{i=1}^{d}\left(s_i\mu_0 e^{-s_i\left[\Delta t+\frac{1}{\mu_0}(e^{-\mu_0 r}-e^{-\mu_0(r-\Delta t)})\right]-\mu_0 t_{ext}}\right)^{\xi_i}\left(e^{-s_i\left[\Delta t+\frac{1}{\mu}(e^{-\mu r}-e^{-\mu(r-\Delta t)})\right]}\right)^{1-\xi_i}$$

and

$$\begin{aligned} log(g_{X|Y}(x|y,\phi)) = \sum_{i=1}^{d}\xi_i\left(log(s_i)+log(\mu_0)-s_i\left[\Delta t+\frac{1}{\mu_0}(e^{-\mu_0 r}-e^{-\mu_0(r-t)})\right]-\mu_0 t_{ext}\right) \\ +(1-\xi_i)\left(-s_i\left[\Delta t+\frac{1}{\mu}(e^{-\mu r}-e^{-\mu(r-t)})\right]\right) \end{aligned} \qquad (10)$$

```
prob.ms <- function(wt,t_ext,s,mu,r){
  s*mu*exp(-s*(wt+(exp(-mu*r)-exp(-mu*(r-wt)))/mu)-mu*t_ext)
}
prob.nospecies <- function(wt,s,mu,r){
  exp(-s*(wt+(exp(-mu*r)-exp(-mu*(r-wt)))/mu))
}

da.prob <- function(xi,wt,t_ext,s,mu,r){
  g = ifelse(tree$xi,prob.ms(wt,t_ext,s,mu,r),prob.nospecies(wt,s,mu,r))
}
```

3. M-step

The M-step consists on the optimization procedure

$$\phi^* = \underset{\phi \in \mathbb{R}^n}{\operatorname{argmax}} Q(\phi|\phi^*)$$

```
mle.st <- function(S,init_par = c(0.5,0.5,100)){
  po = subplex(par = init_par, fn = Q.approx, st=S,hessian = TRUE)
  return(po)
}
```

4. Stopping criteria and sampling size

For the stoping criteria and the sampling size needed to ensure convergence we use the procedure described on Chan et. al.

Summary

The whole framework follows the following steps

The whole MCEM consists on the following steps :

Input: Observed phylogeny y, some ϵ and a preliminary initial value θ_{init}

S1: Run pilot study suggested on [**?**] and described on section A.x and get m, initial values θ, and a tolerance value γ

S2: Sample m$M_{g',i}^{(1)}, ..., M_{g',i}^{(m)}$ and calculate

$$Q(\psi|\widehat{\psi}^{(r)}) \approx \frac{1}{m} \sum_{i=1}^{m} \ln L(D, M_{g',i}^{(r)}|\Psi) \frac{g(M_{g',i}^{(r)}|D, \widehat{\Psi}^{(r)})}{g'(M_{g',i}^{(r)}|D, \widehat{\Psi}^{(r)})}$$

S3: Calculate

$$\Psi^* = \underset{\Psi \in \mathbb{R}^n}{\operatorname{argmax}} Q(\Psi|\widehat{\Psi}^{(r)})$$

and

$$\tilde{\Delta} l_Y(\Psi^*, \Psi) = -log\left(\left(\sum_{k=1}^{m} f_\theta(X)/f(X)\right)/m\right) \tag{11}$$

S4: If $|\tilde{\Delta} l_Y(\Psi^*, \Psi)| \geq \gamma$ update parameter and go to step S.2, else go to step S.5

S.5 Return Ψ^*

Further projects

1. multiple speciation
2. Diversity-dependance on extinction
3. time-dependance (continuous (non step-wise) case) 3.1 check the complete continous version, It might be computationally intractable (?) 3.2 is 3.1 does not work, check what is the discrete aproximation of the continous process.