university of groningen / faculty of science and engineering / mathematics and applied mathematics

A statistical approach on a diversification model of species

Bachelor's Project Mathematics

August 2017

Student: C.A.A. Schoenmaker

First supervisor: prof. dr. E.C. Wit

Second assessor: dr. W.P. Krijnen

In cooperation with: F. Richter Mendoza

Abstract

This thesis describes a general statistical model for the problem of diversification of species. Diversification is a complex problem that has an enormous amount of variables influencing it. Therefore, it is necessary to carefully investigate what variables are important in the diversification process. The model described in this thesis uses GLM functions for the speciation and extinction rates of species, which ensures the model to be general since these functions may contain every possible variable. To simulate the model, we make use of the Gillespie algorithm. And for inference, the likelihood function of the model is maximized to be able to estimate the parameters of the GLM function.

1 Introduction

1.1 Diversification of species

There is a wide variety of species living on earth. There are many species alive at this moment, but taking into account all the species that have ever lived, there have been many more. All these species have a past that describes how they have evolved in what they have become. The process of evolving, the act of becoming more diverse, we call diversification. There are many possible factors that play a role in evolution of species. Physical characteristics, environment and neighboring species are a few examples of variables that could influence the process of evolving. To understand on what variables diversification depends, means that we are able to model it and use those models to predict the future. Modelling helps us to display a difficult process in a more ordered way, which gives us a nice overview of what's happening. A representation of our best hypothesis of how species have evolved from their ancestors, is called the phylogenetic tree. This hypothesis is based on information we've gathered from DNA or fossil records. The use of DNA information allows us to compare more than one trait at the same time, as genes contain information for hundreds of traits. Information from DNA is compared using computer programs and statistical algorithms. The less differences there are between DNA sequences of species, the more likely it is that different species share a common ancestor. This way we construct the most likely past course of evolution, starting at the time we live in now, going back in time.

In the phylogenetic tree below (Figure 1), you see how species evolve. From a common ancestor, new species branch into two new species, we will call this "speciation". Their DNA differs so much here, that we call them different names, different species. Different species cannot mate together. Later in time, they either speciate again, or they go extinct. We want to know which factors are important in causing speciation and extinction in a species life.

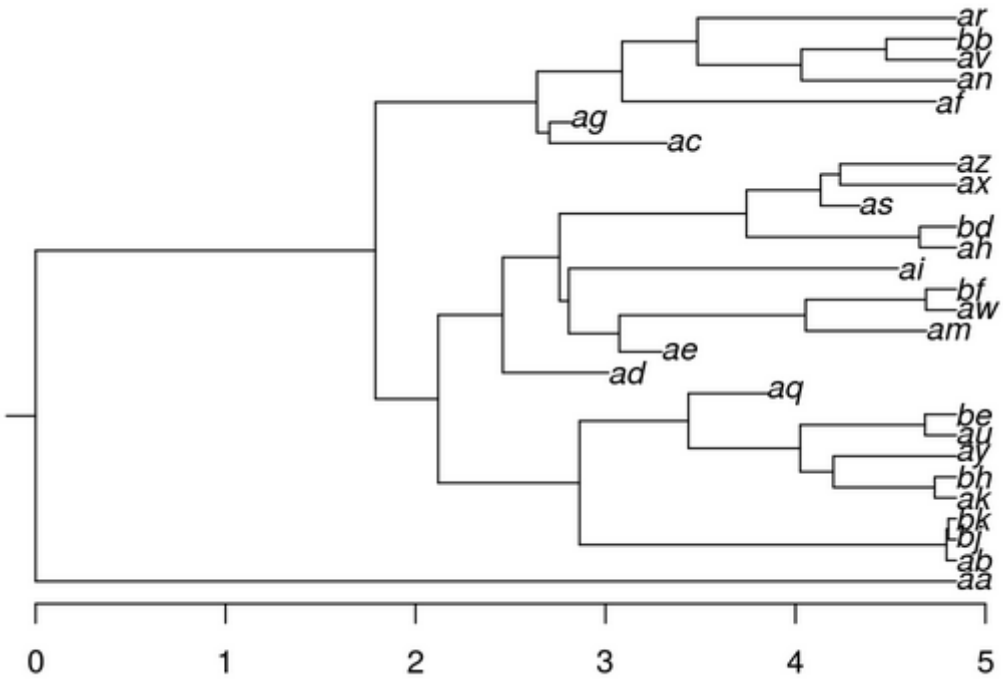


Figure 1: An example of a phylogenetic tree. On the x-axis the time from 0 until recent time. At the far right of the tree, the extant species (species that are still alive) are listed. All the other species, displayed more to the left, have gone extinct.

1.2 Problem

Every species has a big set of factors that determine their lives and the way they evolve. These factors are variables such as traits, which is information on their DNA. For example, body size could define the evolution speed. Bacteria evolve much faster than elephants. But is their body size the cause for this? Maybe it is, but are there more reasons? Also environmental factors can be important, such as temperature, availability of food, humidity etc. Every species has a different set of factors that can influence their way of evolving, these factors we will call diversification variables from now on, and diversification variables all have a different impact. The question of this thesis is whether we can build a general model to infer diversification variables as drivers of the process of speciation and extinction. We do this, to be able find out which variables are the main influencer in the diversification process of different species, and how influential they actually are.

1.3 State of the art

Because of the complexity of diversification of species it is difficult to make a general model. Early models for the tree of life consisted of data based on physical characteristics and metabolic features. [7] This made these models less various and it was difficult to extend to additional covariates. More recent models are based on DNA sequences, which contain several characteristics on just one gene, which gives a different phylogenetic tree. This allows for a more efficient way of modelling.
An example of this is the study to species interactions [2] which is an important

factor that causes complexity of diversification rates.

Since there is little data available from extinct species, models have to be made from data obtained only from extant species, species that are alive in the present. These models can be compared to the little fossil data we have, to check how realistic they are.

In the case of constant diversification rates, a lot of research has been done. Constant rates means that speciation and extinction are constant over time, so the probability for speciation of a certain species remains the same, and so does the probability for extinction. The rates are not dynamical. Since these models are often oversimplified, not taking all factors into account, the results are not very realistic. Though, this research is important for illustrative purposes and understanding of the theory. [10]

Sean Nee (2006) presented the first model based on data from only extant species, an important starting point for the research to a general model for diversification. This model produced an extinction rate equal to zero, so an assumption that extinction doesn't exist, which is obviously false. [9]

After Nee, several studies have been done that allow diversification rates to change over time, making for more realistic models. However, they are all still too simple, mostly changing only under the influence of environmental factors and not species traits.[11]

Two new methods are BiSSE and QuaSSE.

The BiSSE (binary state speciation and extinction) method uses binary traits. [1] This method assumes that the whole phylogenetic tree is known and complete, there is no possibility of undiscovered animals or undiscovered traits or data. Also using only binary traits is limiting for a general model.

Richard Fitzjohn developed the phylogenetic likelihood-based method QuaSSE (quantitative state speciation and extinction) that can be used to test hypotheses based on quantitative variables, such as body size.[5] The downside to these methods, however, is that they use limited kinds and amounts of variables. Since not all variables can be concluded in the model, it either can not be applied to every sort of situation, or it gives an unrealistic outcome. An even more general model is needed.

1.4 My proposal

The goal of this project is to propose a feasible evolutionary diversification model that is easy to estimate from biological data. The model needs to be general, such that it can be applied in any situation. It has to be easily extendable to additional covariates, to include all possible variables. This is also useful for inference of parameters, such that this can be done in a general way. The proposed model will be based on a quasi reaction model, and a Generalized Linear Model, which both will be explained in the next chapter.

2 Diversification through reaction modelling

The model that we propose for the diversification process uses quasi reaction functions, considering speciation and extinction as if they are chemical reactions. With speciation, the reaction goes from one species to two species and with extinction, the reaction goes from one species to an empty set. These reactions happen over time and although they seem to appear randomly, they have a certain probability that they will happen. The probabilities for speciation and extinction are different. The time it takes for an event (either speciation or extinction) to happen, is exponentially distributed, as will be explained.

2.1 SDE model

A stochastic process is a mathematical object of a process that appears to change randomly over time. The starting point might be known, but the process can evolve in many directions. Diversification is stochastic since it's ultimately caused by random mutations of DNA. In a stochastic process, the next step does not depend on the previous states, it only depends on the current state and on time. So the steps in our diversification process can be given by Stochastic Differental Equations (SDE's). [4] The Gillespie algorithm can be used to simulate such a stochastic process, this algorithm will be explained later on.

Our state space is given by

$$\{\mathcal{T}(t_j)|t \geq 0\}$$

Where $\mathcal{T}(t_j)$ is a left continuous rooted tree with binary splits. $t \in [0, CT]$ where CT is the current time.
The process can be described as follows:
We start with 1 species

$$\mathcal{T}(t_0) = \mathcal{T}(0) = 1$$

Given $\mathcal{T}(t_j) = s_1, ..., s_{n_t}$ is the set of species at time t.
We define the following process,

quasi reactions:

$$T_{S,i}^{(t_j)}|\mathcal{T}(t_j) \sim Exp(\lambda_{t_j})$$

$$T_{E,i}^{(t_j)}|\mathcal{T}(t_j) \sim Exp(\mu_{t_j})$$

Where $i = 1, ..., n_t$ is the number of different kinds of species, S,i is the case of a speciation of species i, E,i is an extinction of species i. t_j is the time, and $j = 1, ..., s$ is the number of time intervals. Then T is the time it takes for either a speciation or an extinction to happen. We call this time the waiting time. Why the waiting time is exponentially distributed will be explained in the next section on the Gillespie algorithm.

λ_{t_j} and μ_{t_j} are the diversification rates, λ_{t_j} speciation rate and μ_{t_j} extinction rate. These rates will be modelled using a generalized linear model:

$$\lambda_{t_j} = g^{-1}(x_{t_j}^{'}\alpha)$$

$$\mu_{t_j} = g^{-1}(x_{t_j}^{'}\gamma)$$

Where g is the link function, x_{t_j} is a vector of biological and environmental variables at time t_j and α and β are vectors of unknown parameters. α and β determine the importance of the components of x_{t_j}.

At $T_{t_j} = min\{T_{S,i}^{(t_j)}, T_{E,i}^{(t_j)}\ i = 1, ..., s_{t_j}\}$ the event with the smallest waiting time happens, such that the tree at $\mathcal{T}(t_j + T_{t_j})$ is the tree with 1 additional species or 1 fewer species. If $T_{S,i}^{(t_j)} < T_{E,i}^{(t_j)}$, a speciation takes place, so 1 species is added, however if $T_{S,i}^{(t_j)} < T_{E,i}^{(t_j)}$, an extinction takes place, then there will be 1 species less.

2.2 Gillespie algorithm

The simulation model we'll make, uses λ_{t_j} and μ_{t_j} in a sort of chemical reaction kind of way, where there are two possibilities, speciation and extinction.

$\mathrm{A}_i \xrightarrow{\lambda_{\mathrm{t_j}}} \mathrm{A}_i + \mathrm{B}_i$ in case of speciation

$\mathrm{A}_i \xrightarrow{\mu_{\mathrm{t_j}}} \emptyset$ in case of extinction

Where A_i and B_i are species that are present during the time Δt_j that the simulation is running. $i = 1, ..., n_t$ where n_t is the number of species at time t.

After each branch splitting in the tree, the species will be waiting to either speciate or go extinct. The Gillespie algorithm simulates the trajectory of these events.

Gillespie algorithm [6]

0. Pick a starting situation, a time t and associated numbers of species n_t. At the start of a tree (or a clade of a tree), you start with one species and set the time to 0.

1. Calculate $\sigma_{t_j} = \sum_{i=1}^{n_{t_j}} k_{t_j,i}(\lambda_{t_j,i} + \mu_{t_j,i})$ where $\lambda_{t_j,i}$ are the speciation rates of species i, $\mu_{t_j,i}$ are the extinction rates of species i and $k_{t_j,i}$ is the number of species i that is alive at time t_j.

2. Decide when the event is going to happen by sampling Δt (waiting time) from $exp(\sigma_{t_j})$

3. Sample an event from a multinomial distribution $(\frac{\lambda_{t_j,1}}{\sigma_{t_j}}, ..., \frac{\lambda_{t_j,n_{t_j}}}{\sigma_{t_j}}, \frac{\mu_{t_j,1}}{\sigma_{t_j}}, ..., \frac{\mu_{t_j,n_{t_j}}}{\sigma_{t_j}})$

4. Execute the event.

5. Update time $t_{j+1} = t_j + \Delta t$, return to step 1.

6. Repeat until j reaches s.

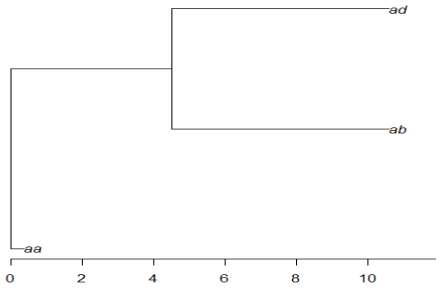
The algorithm is applicable in this model because of the memorylessness property of the exponential distribution. This means that the probability that an event will happen, does not depend on how much time has past.

Proof. Let a random variable $X \sim exp(\lambda)$. Then, for independent times t and s,

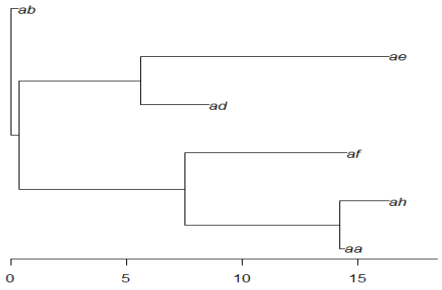
$$\begin{aligned} P(X > t+s | X > s) &= \frac{P(X > t+s; X > s)}{P(X > s)} \\ &= \frac{P(X > t+s)}{P(X > s)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} \\ &= P(X > t) \end{aligned}$$

Concluding, $P(X > t+s|X > s) = P(X > t)$ so the probability that a certain event X happens after time t, is independent of the probability that it happened after time s. □

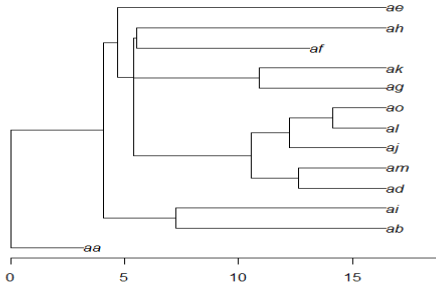
Some examples of simulations can be seen in Figure 2, using a basic model with constant diversification parameters, $\Delta t = 15$, varying the parameters λ_{t_j} and μ_{t_j}. (see code in Appendix 6.1). As can be seen, for low diversification rates, ($\lambda = 0.1$ and $\mu = 0.1$), the tree contains far less events than the trees that have high diversification rates ($\lambda = 0.9$ and $\mu = 0.9$) This can easily be explained if you look at the formula that describes the tree. The time it takes for an event to happen, we call the waiting time t_j, is distributed exponentially with parameters λ and μ. If $\lambda + \mu$ has a relatively low value, the value of $(\lambda + \mu)e^{-(\lambda+\mu)t_j}$ will be relatively high, resulting in longer waiting times. You can also see in the tree that, if the parameter λ (the speciation rate) is higher than μ (the extinction rate), there are significantly more speciations than extinctions, and the reverse holds if μ is higher than λ.



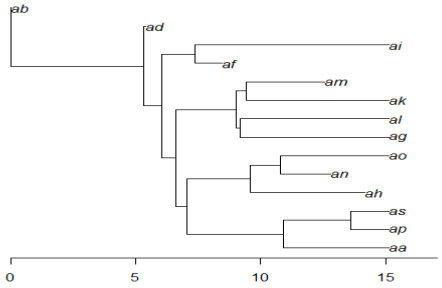
(a) $\lambda_{t_j} = 0.1, \mu_{t_j} = 0.1$



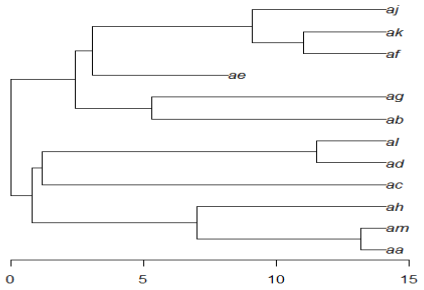
(b) $\lambda_{t_j} = 0.1, \mu_{t_j} = 0.9$



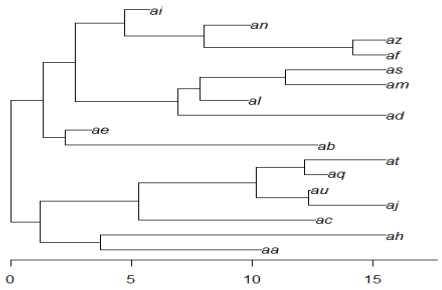
(c) $\lambda_{t_j} = 0.5, \mu_{t_j} = 0.1$



(d) $\lambda_{t_j} = 0.5, \mu_{t_j} = 0.9$



(e) $\lambda_{t_j} = 0.9, \mu_{t_j} = 0.1$



(f) $\lambda_{t_j} = 0.9, \mu_{t_j} = 0.9$

Figure 2: Phylogenetic trees of contant rate models with different rates

The code for plotting the phyologenetic trees can be found in Appendix 6.3.

2.3 Specific examples

In this project, a binary model is used for the simulation of the phylogenetic tree. A binary model contains traits of animals that they either do have, or don't have, such as an animal being a carnivore, or vegetarian. Another example is of a binary variable is Whether animals are live-bred, or born out of an egg. In the case of binary characters, there are always two possibilities, so we assume a set $X = \{0, 1\}$ where 0 is one possibility and 1 is the other one.
An example of the diversification rates of a binary model may look like this

$$\lambda = \lambda_C * \mathbb{1}_0 + \lambda_V * \mathbb{1}_1$$
$$\mu = \mu_0$$

Where λ_C is the speciation rate in case the species is a carnivore (X = 0), and λ_V is the speciation rate in case the species is a vegetarian (X = 1).

2.3.1 My model; predators and prey

The choice of binary model for this thesis has become the model of predator and prey. We chose this model to do further calculations with.

$$\lambda_{t_j}^{predator} = x_{t_j-1}^{prey}\phi$$
$$\mu_{t_j}^{predator} = \mu_0$$
$$\lambda_{t_j}^{prey} = \lambda_0$$
$$\mu_{t_j}^{prey} = x_{t_j-1}^{predator}\omega$$

Speciation and extinction rates of predator and prey at time t. x_t is the biological variable that describes how many individual predators or prey are alive at time t. The speciation rate at time t of the predators depends on the #prey that are alive at time t-1, since the rate at which the predators are able to grow, depends on how much they can hunt and eat. The extinction rate at time t of the prey depends on how many predators are alive at time t-1, this is because when there are more predators around.
ϕ, μ_0, λ_0 and ω are the parameters that we want to estimate in order to check the importance of the biological variables.
The code for the predator and prey model is added in the Appendix of this thesis (6.2).

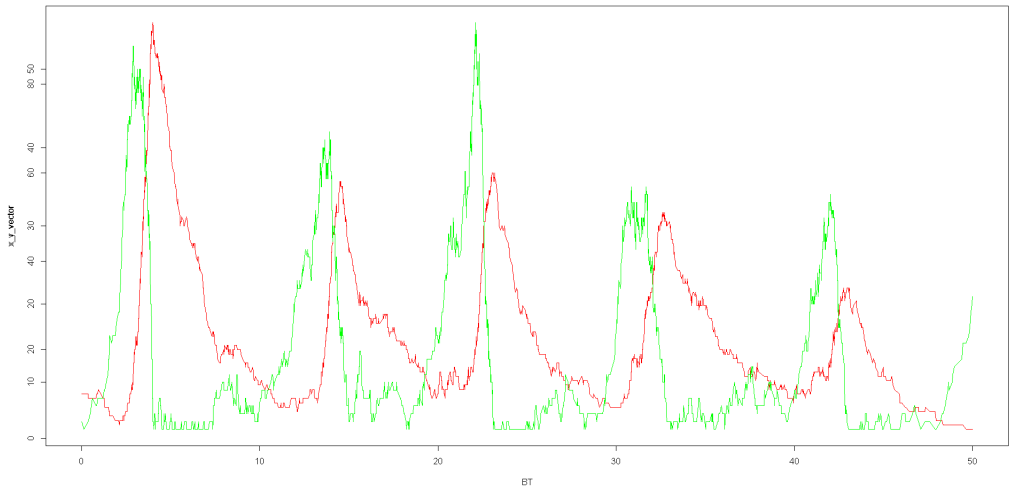


Figure 3: Graph containing #species, red = predators, green = prey

You see in figure 3 the correlation between the number of individual predators and the number of individual prey. As soon as the community of predators begins to rise, the number of prey sinks. And then, the amount of predators goes down too, which makes the amount of prey begin to rise again.

Simulating this phylogenetic tree produces as output vectors of the waiting times, the branching times (times at which a new branch occurs), a vector containing the speciations and extinctions in the case of carnivores and in the case of vegetarian animals.

3 Inference of a diversification model

3.1 Data

We assume that the data consists of a single complete tree $\{\mathcal{T}(t_j)|t \in [0,T]\}$ where T is the present time. The data that we use is derived from using DNA sampling of extant species. The tree is constructed by comparing DNA. When more communal genetic information is observed, we assume that species have more common ancestors than would be the case with species that have less overlap in genetic information. This way, the past connections of species are constructed. The reconstruction is uncertain, since we can never be completely sure of what happened in the past, because we don't have enough data to back us up. The biggest problem with this method is that we don't have much information on extinct species. From fossil records, we can get some information, but when using the method described above, the tree will not be complete. This can be solved using an algorithm called the EM algorithm[12], but the application of this algorithm is beyond the scope of this thesis.

3.2 Likelihood

The phylogenetic tree space is described using two sets: waiting times (T_{t_j}) and events (Y_{t_j}). As described in chapter 2.1, the waiting times T_{t_j} are exponentially

distributed with rate parameters λ_{t_j} and μ_{t_j}. For convenience of the likelihood function, we combine the two separate exponential distributions into one, with a new rate parameter $\sigma_{t_j} = \sum_{i=1}^{n_{t_j}} k_{t_j,i}(\lambda_{t_j,i} + \mu_{t_j,i})$

$$T_{t_j} \sim exp(\sigma_{t_j})$$

The probability that a certain event Y_{t_j} will happen is multinomially distributed with probability rate sampled from $(\frac{\lambda_{t_j,1}}{\sigma_{t_j}}, ..., \frac{\lambda_{t_j,n_{t_j}}}{\sigma_t}, \frac{\mu_{t_j,1}}{\sigma_{t_j}}, ..., \frac{\mu_{t_j,n_{t_j}}}{\sigma_t})$

Where $\frac{\lambda_{t_j,i}}{\sigma_{t_j}}$ is the probability that the i^{th} species speciates, and $\frac{\mu_{t_j,i}}{\sigma_{t_j}}$ is the probability that the i^{th} species goes extinct. We call this sample $\frac{\eta_{t_j}}{\sigma_{t_j}}$.

$$Y_{t_j} \sim \frac{\eta_{t_j}}{\sigma_{t_j}}$$

Recall that diversification rates λ_{t_j} and μ_{t_j} were described by a GLM with parameters α and γ.

$$\lambda_{t_j} = g^{-1}(x'_{t_j}\alpha)$$
$$\mu_{t_j} = g^{-1}(x'_{t_j}\gamma)$$

The probability density function of the phylogenetic tree space is given by

$$f(T_{t_j}, Y_{t_j}; \alpha, \gamma) = \sigma_{t_j} e^{-\sigma_{t_j} T_{t_j}} \eta_{t_j} / \sigma_{t_j}$$

The likelihood function is then

$$\begin{aligned} L(T_{t_j}, Y_{t_j}; \alpha, \gamma) &= \Pi_{j=1}^{s} \eta_{t_j} e^{-\sigma_t T_{t_j}} \\ &= \Pi_{j=1}^{s} \eta_{t_j} e^{-\sum_{i=1}^{n_t} x_{t_j,i}(g^{-1}(x'_{t_j}\alpha) + g^{-1}(x'_{t_j}\gamma))T_{t_j}} \end{aligned}$$

The likelihood function is simplified in the case of the predator/prey model, which has a simpler function for λ_{t_j} and μ_{t_j} since g = identity.

$$\begin{aligned} \sigma_{t_j} &= x_{t_j}^{predator}(\lambda_{t_j}^{predator} + \mu_{t_j}^{predator}) + x_{t_j}^{prey}(\lambda_{t_j}^{prey} + \mu_{t_j}^{prey}) \\ &= x_{t_j-1}^{predator}(x_{t_j-1}^{prey}\phi + \mu_0) + x_{t_j-1}^{prey}(\lambda_0 + x_{t_j-1}^{predator}\omega) \end{aligned}$$

So the likelihood function looks like this

$$L(T_{t_j}, Y_{t_j}; \alpha, \gamma) = \Pi_{j=1}^{s} \eta_{t_j} e^{-(x_{t_j-1}^{predator}(x_{t_j-1}^{prey}\phi + \mu_0) + x_{t_j-1}^{prey}(\lambda_0 + x_{t_j-1}^{predator}\omega)T_{t_j}}$$

For inference of the parameters α and γ, we want to maximize the likelihood function.

3.3 Maximum likelihood estimation

For maximum likelihood estimation, we consider the likelihood function

$$L(T_{t_j}, Y_{t_j}; \alpha, \gamma) = \Pi_{j=1}^{s} \sigma_{t_j} e^{-\sigma_t T_{t_j}} \eta_{t_j} / \sigma_{t_j}$$

We want to find the log-likelihood function $l(\beta) = l_{exp}(\beta) + l_{mutlin}(\beta)$, which is divided into the exponential part and the multinomial part of the function. To find the parameter $\beta = \begin{bmatrix} \alpha \\ \gamma \end{bmatrix}$ we maximize this function.
Recall, the link functions of the GLM are

$$g(\lambda_{t_j}) = x_{t_j}^{'}\alpha$$

$$g(\mu_{t_j}) = x_{t_j}^{'}\gamma$$

We choose a different link function g' such that

$$g'(E(T_{t_j})) = x_{t_j}^{'}\beta = \eta_{t_j}$$

In general the estimating equation has the form

$$\beta^{(m)} = \beta^{(m-1)} + [\mathfrak{J}^{(m-1)}]^{-1}U^{(m-1)}$$

Which can also be written as

$$\mathfrak{J}^{(m-1)}\beta^{(m)} = \mathfrak{J}^{(m-1)}\beta^{(m-1)} + U^{(m-1)} \tag{1}$$

Where $\beta^{(m)}$ is the vector of estimates for α and γ at the m^{th} iteration. $[\mathfrak{J}^{(m-1)}]$ is the information matrix and $U^{(m-1)}$ is the score function. $\mathfrak{J}$ can be written as $\mathfrak{J} = \mathbf{X^T W X}$ Where

$$w_{ii} = \frac{1}{var(T_{t_j})}\frac{\delta E(T_{t_j})}{\delta \eta_{t_j}}$$

And $\mathbf{X}$ contains the entries of the vectors x_{t_j}.
The right hand side of equation (1) can be written as $\mathbf{X^T W z}$ where $\mathbf{z}$ has elements

$$z_i = x_{t_j,i}^{'}\beta + (T_{t_j} - E(T_{t_j}))(\delta\eta_{t_j}/E(T_{t_j}))$$

Hence, equation (1) can be written as $\mathbf{X^T W X}\beta^{(m)} = \mathbf{X^T W z}$ which can be solved by the Iteratively Reweighed Least Squares method (IRWLS) [3]

We divide the likelihood function into the exponential part and the multinomial part. Then

$$\mathbf{X^T W X} = \mathbf{X_{exp}^T W_{exp} X_{exp}} + \mathbf{X_{multin}^T W_{multin} X_{multin}}$$

and

$$\mathbf{X^T W z} = \mathbf{X_{exp}^T W_{exp} z_{exp}} + \mathbf{X_{multin}^T W_{multin} z_{multin}}$$

Where

$$X = \begin{bmatrix} X_{exp} \\ X_{multin} \end{bmatrix}$$

$$W = \begin{bmatrix} W_{exp} & 0 \\ 0 & W_{multin} \end{bmatrix}$$

$$z = \begin{bmatrix} z_{exp} \\ z_{multin} \end{bmatrix}$$

Thus, we need to formulate expressions for $X_{exp}, X_{multin}, W_{exp}, W_{multin}, z_{exp}, z_{multin}$.

3.3.1 IRWLS for the exponential part

Since

$$T_{t_j} \sim exp(\sigma_{t_j})$$

Mean and variance are as follows: $E(T_{t_j}) = \frac{1}{\sigma_{t_j}}$ and $Var(T_{t_j}) = \frac{1}{\sigma_{t_j}^2}$

with

$$\begin{aligned}
\sigma_{t_j} &= \sum_{i=1}^{n_{t_j}} k_{t_j,i}(\lambda_{t_j,i} + \mu_{t_j,i}) \\
&= \sum_{i=1}^{n_{t_j}} k_{t_j,i}(x'_{t_j,i}\alpha + x'_{t_j,i}\gamma) \\
&= \sum_{i=1}^{n_{t_j}} k_{t_j,i}x'_{t_j,i}(\alpha + \gamma) \\
&= \sum_{i=1}^{n_{t_j}} k_{t_j,i}x'_{t_j,i}\beta \\
&= x'_{t_j}\beta = \eta_{t_j}
\end{aligned}$$

Thus, we get the link function $g'(x) = \frac{1}{x}$

Then we calculate

$$\begin{aligned}
w_{jj}^{exp} &= \frac{1}{var(T_{t_j})}(\frac{\delta E(T_{t_j})}{\delta \eta_{t_j}})^2 \\
&= \sigma_{t_j}^2(-\frac{1}{\sigma_{t_j}^2})^2 = 1/\sigma_{t_j}^2
\end{aligned}$$

$$\begin{aligned}
z_j^{exp} &= x'_{t_j,i}\beta + (T_{t_j} - E(T_{t_j}))(\delta\eta_{t_j}/E(T_{t_j})) \\
&= \sigma_{t_j} + (T_{t_j} - \frac{1}{\sigma_{t_j}})(-\sigma_{t_j}^2) \\
&= \sigma_{t_j}(2 - T_{t_j}\sigma_{t_j})
\end{aligned}$$

Thus,

$$W_{exp} = \begin{bmatrix} \frac{1}{\sigma_{t_1}^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_{t_2}^2} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_{t_s}^2} \end{bmatrix}$$

$$z_{exp} = \begin{bmatrix} \sigma_{t_1}(2 - T_{t_1}\sigma_{t_1}) \\ \sigma_{t_2}(2 - T_{t_2}\sigma_{t_2}) \\ \vdots \\ \sigma_{t_s}(2 - T_{t_s}\sigma_{t_s}) \end{bmatrix}$$

and

$$X_{exp} = \begin{bmatrix} \sum_p x_{t_1,1,p} & \sum_p x_{t_1,2,p} & \dots & \sum_p x_{t_1,n_{t_j},p} \\ \sum_p x_{t_2,1,p} & \sum_p x_{t_2,2,p} & \dots & \sum_p x_{t_2,n_{t_j},p} \\ \vdots & & \ddots & \vdots \\ \sum_p x_{t_s,1,p} & \sum_p x_{t_s,2,p} & \dots & \sum_p x_{t_s,n_{t_j},p} \end{bmatrix}$$

Remember, s is the number of time intervals, n_{t_j} the number of species at time t_j and here p is the index of the p^{th} element of the vector of biological variables $x_{t_j,i}$.

Now we can solve $\mathbf{X_{exp}^T W_{exp} X_{exp}} \beta^{(m)} = \mathbf{X_{exp}^T W_{exp} z_{exp}}$

3.3.2 IRWLS for the multinomial part

For the maximum likelihood estimation of the multinomial part we use a similar process. Since

$$Y_{t_j} \sim multinom(\frac{\eta_{t_j}}{\sigma_{t_j}})$$

Where $\frac{\eta_{t_j}}{\sigma_{t_j}}$ is a random sample from $(\frac{\lambda_{t_j,1}}{\sigma_{t_j}}, ..., \frac{\lambda_{t_j,n_{t_j}}}{\sigma_t}, \frac{\mu_{t_j,1}}{\sigma_{t_j}}, ..., \frac{\mu_{t_j,n_{t_j}}}{\sigma_t})$.
To simplify, we call

$$\frac{\eta_{t_j}}{\sigma_{t_j}} = \pi_{t_j} = \frac{\eta_{t_j}}{\eta_{t_j} + c_{t_j}}$$

Where $\eta_{t_j} = x'_{t_j,i_{t_j}} \beta$ and $c_{t_j} = \sum_{i \neq i_{t_j}}^{n_{t_j}} x'_{t_j,i} \beta$

Mean and variance: $E(Y_{t_j}) = \pi_{t_j}$ and $Var(Y_{t_j}) = \pi_{t_j}(1 - \pi_{t_j})$

$$\begin{aligned} w_{jj}^{multin} &= \frac{1}{var(Y_{t_j})} (\frac{\delta E(Y_{t_j})}{\delta \eta_{t_j}})^2 \\ &= \frac{1}{\pi_{t_j}(1 - \pi_{t_j})} \frac{c_{t_j}^2}{(\eta + c_{t_j})^4} \\ &= c_{t_j} \pi^2 \end{aligned}$$

$$\begin{aligned} z_j^{multin} &= x'_{t_j,i} \beta + (Y_{t_j} - E(Y_{t_j}))(\delta \eta_{t_j} / E(Y_{t_j})) \\ &= x'_{t_j,i} \beta + \frac{\pi_{t_j} c_{t_j}^2}{(1 - \pi_{t_j})^3} \end{aligned}$$

$$W_{multin} = \begin{bmatrix} c_{t_1} \pi_{t_1}^2 & 0 & \dots & 0 \\ 0 & c_{t_2} \pi_{t_2} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & c_{t_s} \pi_{t_s} \end{bmatrix}$$

$$z_{multin} = \begin{bmatrix} z_1^{multin} \\ z_2^{multin} \\ \vdots \\ z_s^{multin} \end{bmatrix}$$

and

$$X_{multin} = \begin{bmatrix} x_{t_1,1,p} & x_{t_1,2,p} & \dots & x_{t_1,n_{t_j},p} \\ x_{t_2,1,p} & x_{t_2,2,p} & \dots & x_{t_2,n_{t_j},p} \\ \vdots & & \ddots & \vdots \\ x_{t_s,1,p} & x_{t_s,2,p} & \dots & x_{t_s,n_{t_j},p} \end{bmatrix}$$

3.3.3 IRWLS combined

By combining the two parts, we found the complete iterative equation

$$\mathbf{X^T W X}\beta^{(m)} = \mathbf{X^T W z}$$
$$(\mathbf{X^T_{exp} W_{exp} X_{exp}} + \mathbf{X^T_{multin} W_{multin} X_{multin}})\beta^{(m)} = \mathbf{X^T_{exp} W_{exp} z_{exp}} + \mathbf{X^T_{multin} W_{multin} z_{multin}}$$

For the m^{th} iteration.

β can be solved until the difference between $\beta^{(m-1)}$ and $\beta^{(m)}$ is sufficiently small. Then $\beta^{(m)}$ may be taken as the maximum likelihood estimator.

3.4 Analytical solution for the predator/prey model

For inference of the parameters ϕ, μ_0, λ_0 and ω an analytical solution can be obtained, since the predator/prey model is relatively simple and uses a link function g which is identity. However, using a model that has a more complex link function, the analytical derivation of the likelihood function will be hard or even impossible. In that case the numerical IRWLS approach has to be used.

$$\delta l/\delta\phi = p/\phi - \sum_{i=1} x_{t_j}^{predator} x_{t_j}^{prey} T_{t_j}$$

for $Y_i = 1$ (speciation of predators), $\eta_i = x_{i,y}\phi$ and p is the number of times that $Y_i = 1$. Solving $\delta l/\delta\phi = 0$ yields $\hat{\phi} = p/\sum_{j=1}^{s} x_{t_j}^{prey} T_{t_j}$

Similarly

$$\hat{\mu_0} = q/\sum_{j=1}^{s} x_{t_j}^{predator} T_{t_j}$$

$$\hat{\lambda_0} = r/\sum_{j=1}^{s} x_{t_j}^{prey} T_{t_j}$$

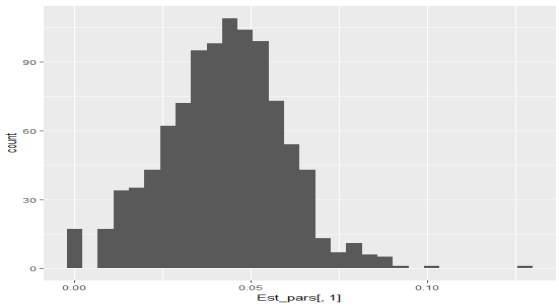
$$\hat{\omega} = s/\sum_{j=1}^{s} x_{t_j}^{predator} x_{t_j}^{prey} T_{t_j}$$

Where q is the number of times that $Y_i = 2$, r is the number of times that $Y_i = 3$ and s is the number of times that $Y_i = 4$.

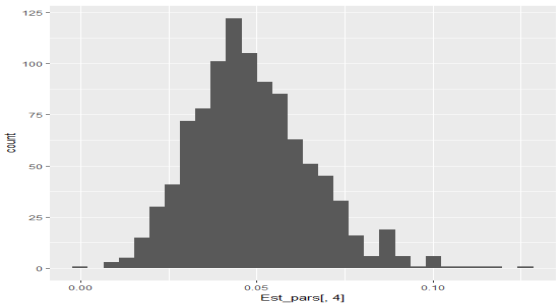
When simulating the analytical solution for the predator/prey model in R for a 1000 different phylogenetic trees, these histograms show the result of the parameters ϕ, μ_0, λ_0 and ω. $\Delta t = 2$ and $\Delta t = 5$.

In the histograms of the analytical solutions, you see a nice binomial distribution that starts to approach a bell curve. In the case of $\Delta t = 2$ you can see that the paramters are slightly underestimated. When $\Delta t = 5$, the estimations are much better. This could be explained by the fact that, with a longer time interval, more events happen so we have more data points to estimate the parameters.

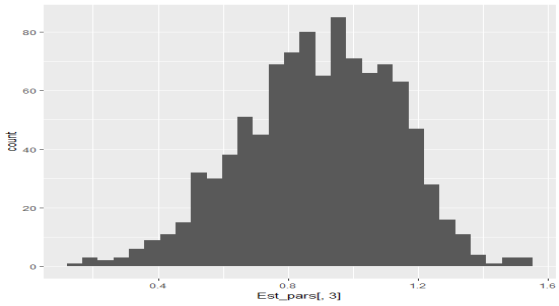
$\Delta t = 2$



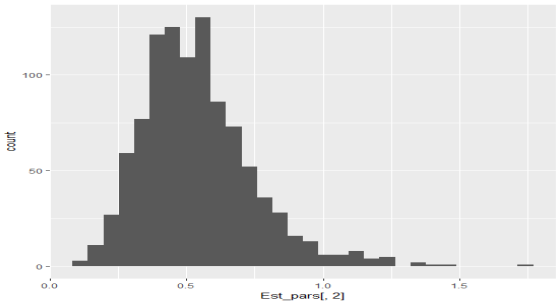
$\phi = 0.05$



$\mu_0 = 0.5$

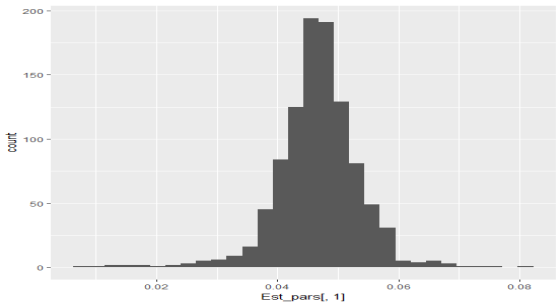


$\lambda_0 = 1.0$

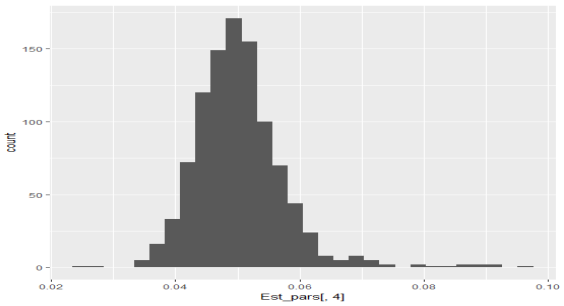


$\omega = 0.05$

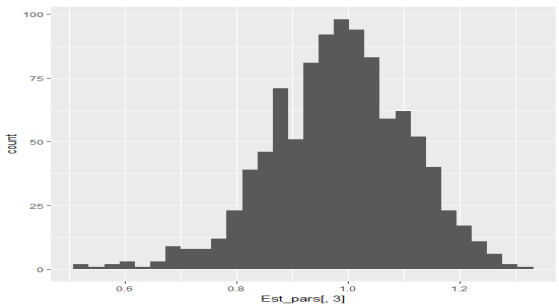
$\Delta t = 5$



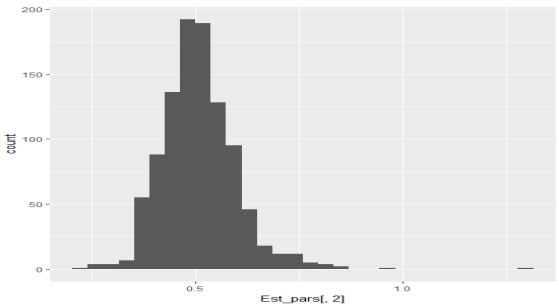
$\phi = 0.05$



$\mu_0 = 0.5$



$\lambda_0 = 1.0$



$\omega = 0.05$

3.5 Uncertainty of estimates

We determine the uncertainty of estimates $\hat{\phi}, \hat{\mu_0}, \hat{\lambda_0}, \hat{\omega}$ of the predator/prey model by evaluating the variance. This way we can check if the estimated parameters can be accepted as useful. Because of the asymptotic efficiency of a maximum likelihood estimator, the variance equals the inverse Fisher information matrix. [8]

$$var(T_{t_j}) = \begin{bmatrix} \frac{\delta^2 l}{\delta\phi^2} & \frac{\delta^2 l}{\delta\phi\delta\mu_0} & \frac{\delta^2 l}{\delta\phi\delta\lambda_0} & \frac{\delta^2 l}{\delta\phi\delta\omega} \\ \frac{\delta^2 l}{\delta\mu_0\delta\phi} & \frac{\delta^2 l}{\delta\mu_0^2} & \frac{\delta^2 l}{\delta\mu_0\delta\lambda_0} & \frac{\delta^2 l}{\delta\mu_0\delta\omega} \\ \frac{\delta^2 l}{\delta\lambda_0\delta\phi} & \frac{\delta^2 l}{\delta\lambda_0\delta\mu_0} & \frac{\delta^2 l}{\delta\lambda_0^2} & \frac{\delta^2 l}{\delta\omega\delta\lambda_0} \\ \frac{\delta^2 l}{\delta\omega\delta\phi} & \frac{\delta^2 l}{\delta\omega\delta\mu_0} & \frac{\delta^2 l}{\delta\omega\delta\lambda_0} & \frac{\delta^2 l}{\delta\omega^2} \end{bmatrix}^{-1}$$

We know then that $\hat{\phi} \sim N(\phi, \frac{\frac{\delta^2 l}{\delta\phi^2}^{-1}(\hat{\phi})}{n})$

Where $\frac{\frac{\delta^2 l}{\delta\phi^2}^{-1}(\hat{\phi})}{n}$ is the Cramer-Rao uncertainty. n is the number of simulations that we did. And similarly for the parameters μ_0, λ_0, ω.

$$\frac{\delta^2 l}{\delta\phi^2} = \frac{-p}{\phi^2}$$

Similar for μ_0, λ_0, ω. All the non-diagonal elements of the Fisher information matrix are zero.

Calculating the uncertainty in R gives us the following values:

The true parameters we used for the simulation

```
> parameters
[1] 0.05000000 0.50000000 1.00000000 0.05000000
```

The average corresponding standard deviation ($\sqrt{var}$) after 1000 simulations

```
> sigma_av
[1] 0.01594777 0.17578733 0.21287544 0.01773934
```

The average corresponding z-value after 1000 simulations

```
> z
[1] 0.6446774  0.1007690  0.3375665  0.2945482
```

(code for this simulation is to be found in Appendix 6.5)

These values of z are well below the critical value of 1.96 which is the normal critical value for a confidence interval of 95%. Which means that our estimates of parameters $\phi, \mu_0, \lambda_0, \omega$ are very good.

4 Conclusion

The general model for diversification of species uses a state space that consists of two parts: the time it takes for an event to happen, and the type of event that happens. They are distributed respectively exponentially and multinomially. Using the Gillespie algorithm, the events can be modelled using quasi reaction functions. The model is based on speciation and extinction parameters λ and μ that are described

using a generalized linear model, which contains biological traits and all other variables that may be of influence on the diversification of a species. This ensures that it is possible to use any factor that you can think of in relation to evolution. These factors are the parameters of the model.
The likelihood function of the model looks like this

$$L(T_{t_j}, Y_{t_j}; \alpha, \gamma) = \Pi_{j=1}^{s} \eta_{t_j} e^{-\sum_{i=1}^{n_t} x_{t_j,i}(g^{-1}(x'_{t_j}\alpha)+g^{-1}(x'_{t_j}\gamma))T_{t_j}}$$

The Iteratively Reweighed Least Squares method can be used to maximize the parameter sets α and γ.

For this thesis, an example of a model was made using a predator/prey relation. The parameters were solved analytically and by an uncertainty of estimates analysis it was checked that the parameters that we estimated can be accepted as good estimates.

5 Discussion

For a general model, the approach that is described in this thesis is a good start. However, there are two big issues that may make the model unreliable. The first is the uncertainty of the past. In this model we use random multinomial and exponential samples to model the tree, and we can never be sure if this is a good estimation, as we have no data to back us up. In the case of this thesis, we neglected this uncertainty. To minimize the uncertainty, it is necessary to keep comparing model and data and keep adjusting the model to make a better fit.

The second issue is the problem of missing data of extinct species. Since we use only data of extant species, the tree is not complete. We don't have an exact idea of species that have gone extinct, but there are probably many more extinct species than species that are alive. To keep extensive data can help us to see over time what factors are important in the extinction of species. There also exists an algorithm, the Expectation-Maximization algorithm, to estimate the underlying presence–absence logistic model for presence-only data. This is an important next step in the general diversification model. [12]

For the model described in this thesis, we did not succeed yet to apply the IRWLS method. This would be a very interesting goal for the future, since with this method, parameters of any model can be maximized. With the predator/prey model, as used in this thesis, it was fortunate coincidence that the problem could also be solved analytically. If a model has a link function g that is not equal to identity, or the model uses more complex parameter relations, it is often not possible to solve the maximization problem analytically. In such a case, the IRWLS can provide an important solution.
The code for the IRWLS method is added in the appendix (6.4) of this thesis, although it is not yet working.

A general diversification model of species can also be used for other purposes, anything that evolves under the influence of different factors. For example languages, which can be compared to species, they change, new languages originate and languages go extinct.

References

[1] W Ayne P M Addison, P Eter E M Idford, and S Arah P O Tto. Estimating a Binary Character ' s Effect on Speciation and Extinction. 56(5):701–710, 2007.

[2] Timothy G Barraclough. How Do Species Interactions Affect Evolutionary Dynamics Across Whole Communities ? 2015.

[3] Annette J. Dobson and Adrian Barnett. *An introduction to generalized linear models*. 2008.

[4] Joseph L. Doob and Joseph L. Doob. *Stochastic processes*. Wiley, New York, vol. 7. no. 2. edition, 1953.

[5] Richard G. FitzJohn. Quantitative Traits and Diversification. 59(6):619–633, 2010.

[6] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, 1976.

[7] Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butter, Alex W Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A Relman, Kari M Finstad, Ronald Amundson, Brian C Thomas, and Jillian F Ban. OPEN A new view of the tree of life. 1(April):1–6, 2016.

[8] Guy Lebanon. Asymptotic Efficiency of the Maximum Likelihood Estimator. Technical Report 2, 2009.

[9] Sean Nee. Birth-Death Models in Macroevolution. *Annu. Rev. Ecol. Evol. Syst*, 37:1–17, 2006.

[10] Emmanuel Paradis. Letter to the Editor. (3):476–480, 1997.

[11] Tanja Stadler. Inferring speciation and extinction processes from extant species data. *Proceedings of the National Academy of Sciences*, 108.39 16145-16146., 2011.

[12] Gill Ward, Trevor Hastie, Simon Barry, Jane Elith, and John R. Leathwick. Presence-Only Data and the EM Algorithm. *Biometrics*, 65(2):554–563, 6 2009.

6 Appendix

6.1 Code in R for constant rate model

```
install.packages("devtools")
devtools::install_github("franciscorichter/dmea")

library(dmea)

rm(list = ls())

   ct=5
sumt=0
model = 'cr'
lambda0 = 0.6
mu0 = 0.05
N = 1
reboot = 0

WT = c()
SE = c()
BT = c()



while(sumt < ct){
  #update parameter
  if(model == 'cr'){
    lambda = lambda0
    mu = mu0
  }
  s = lambda + mu
  wt = rexp(n = 1,rate = s)

  BD = rmultinom(1,1,c(rep(lambda,N),rep(mu,N))/s)
  L = which(BD == 1)

  # if speciation
  # find out where the 1 in rmultinom occurs, so where does the event occ
  if(L<=N){

 # update tree  T, tau...
    N= N+1
  SE = append(SE,1)
  }
  #else (if extinction)
  if (L>N) {
```

```
  # update
  N = N-1              # let op! what if N=0??
  SE = append(SE,0)
  }
  WT = append(WT,wt)
  sumt = sum(WT)
  BT =  append(BT, sumt)
  population = sum(SE)

  if (N == 0) {
    reboot = reboot + 1
    N=1
    WT = c()
    SE = c()
    BT = c()
  }

}

p = list(wt=WT,E=SE)
p2 = p2phylo(p)
plot(p2)
axisPhylo(backward = F)
```

6.2 Code in R for simulating phylogenetic tree of predator and prey model

```
library(dmea)

rm(list = ls())

ct=1
sumt = 0
model = 'bt'

phi = 0.05     #dependency parameter of predators
mu_r0 = 0.5    #constant
lambda_y0 = 1 #constant
omega = 0.05   #dependency parameter of prey

x_r = 10 # nr. of predators
x_y = 5  # nr. of prey

WT = c(0) # waiting time
Sigma = c()
SE = c()  # speciation = 1, extinction = 0
BT = c(0) # branching time, sum of waiting times
L_vector = c() # vector that shows what kind of event takes place (sp/ex
```

```
x_r_vector = c(x_r) # nr. of predators at every wt
x_y_vector = c(x_y) # nr. of prey at every wt
ind_vector = c()


while(sumt < ct){
  #update parameter
  if(model == 'bt'){
    lambda_r = phi*(x_y-1)
    mu_r = mu_r0
    lambda_y = lambda_y0
    mu_y = omega*x_r
  }

  if (x_y<5){mu_y=0}

   vector_rates = c(rep(lambda_r,x_r),rep(mu_r,x_r),rep(lambda_y,x_y),rep
   s   = sum(vector_rates)
   wt = rexp(n = 1,rate = s) # waiting time is exponentially distributed
   BD = rmultinom(1,1,vector_rates/s) # event: birth/death, multinomially
   L = which(BD == 1)
   L_vector = append(L_vector, L)
   Sigma = append(Sigma,s)

  # find out where the 1 in rmultinom occurs, so where does the event occ
  # speciation of predators (x_r)
  ind = 0
  if(L <= x_r){
    # update tree
    x_r= x_r+1
    SE = append(SE,1)
    x_r_vector = append(x_r_vector,x_r)
    x_y_vector = append(x_y_vector,x_y)
  ind=1
  }
  # extinction of predators (x_r)
  if (x_r< L & L <= (2*x_r) & ind == 0) {

    x_r = x_r-1
    SE = append(SE,0)
    x_r_vector = append(x_r_vector,x_r)
    x_y_vector = append(x_y_vector,x_y)
  ind=2
  }
  #speciation of prey (x_y)
  if ((2*x_r)< L & L <= (2*x_r+x_y) & ind == 0) {
    x_y= x_y+1
    SE = append(SE,1)
```

```
      x_r_vector = append(x_r_vector,x_r)
      x_y_vector = append(x_y_vector,x_y)
    ind=3
  }
  #extinction of prey (x_y)
  if ((2*x_r+x_y) < L & ind == 0) {
    x_y = x_y-1
    SE = append(SE,0)
    x_r_vector = append(x_r_vector,x_r)
    x_y_vector = append(x_y_vector,x_y)
  ind=4
    }

  WT = append(WT, wt)
  sumt = sum(WT)
  BT = append(BT, sumt)
  ind_vector = append(ind_vector, ind)


  # if the nr. of predators becomes 0, the tree should reboot
if (x_r == 0) {
 break()
}
}
vector_rates = c(rep(lambda_r,x_r),rep(mu_r,x_r),rep(lambda_y,x_y),rep(mu
  s  = sum(vector_rates)
  Sigma = append(Sigma, s)              # This part is needed for IRWLS, to
  ind_vector = append(ind_vector, 1)
```

6.3 Code in R for plotting the tree

```
  # plot of phylogenetic tree

p = list(wt=WT,E=SE)
p2 = p2phylo(p)
plot(p2)
axisPhylo(backward = F)

  # plot of relationship between predators and prey

plot(BT, x_r_vector, col='red',type = 'l')
par(new=TRUE)
plot(BT, x_y_vector, col='green',type = 'l')
```

6.4 Code in R for IRWLS - not fully working

```
# input: sigma=sum(lambda's+mu's), x_r_vector=nr. of predators at times
# x_y_vector=nr. of predators at times t_i, T=times t_i ---- (all vector
# ind_vector

#iwrls <- function (Sigma, x_r_vector, x_y_vector, BT, ind_vector){

  l = length(BT)
  W_exp = matrix(0,ncol=l, nrow=l)
  Z_exp = matrix(0,l,1)
  X_exp = matrix(0,l, 4)
  W_multin = matrix(0,ncol=l, nrow=l)
  Z_multin = matrix(0,l,1)
  X_multin = matrix(0,l, 4)
  phi = 0.5
  mu = 0.05
  lambda = 1
  omega = 0.05


  eta = rep(0,l)                  #multinomial rate of specific event
  pi_vector = rep(0,l)            #multinomial rate
  c_vector = rep(0,l)             #multinomial rate c = sigma - eta

  i = 1

  distance = 1                    # distance = nr of iterations
  while(distance <= 2) {

    # IRWLS of exponential part

    while (i <= l) {
      W_exp[i,i] = 1/Sigma[i]^2
      Z_exp[i] = Sigma[i]*(2-Sigma[i]*BT[i])
      X_exp[i,1] = sum(x_r_vector[1:i]*x_y_vector[1:i])
      X_exp[i,2] = sum(x_r_vector[1:i])
      X_exp[i,3] = sum(x_y_vector[1:i])
      X_exp[i,4] = sum(x_r_vector[1:i]*x_y_vector[1:i])
      i = i+1
    }


    # IRWLS of multinomial part

    i=1
    while (i <= l) {

      ind2 = 0
      if (ind_vector[i]==1) {
```

```
      eta[i] = x_y_vector[i]*phi
      ind2 = 1
    }
    if (ind_vector[i]==2 & ind2 ==0) {
      eta[i]  = mu
      ind2 = 2
    }
    if (ind_vector[i]==3 & ind2 ==0) {
      eta[i]  = lambda
      ind2 = 3
    }
    if (ind_vector[i]==4 & ind2 ==0) {
      eta[i]  = x_r_vector[i]*omega
      ind2 = 4
    }

    pi_vector = eta/Sigma
    c_vector[i] = x_y_vector[i]*phi + mu + lambda + x_r_vector[i]*ome

   W_multin[i,i] = c_vector[i]*pi_vector[i]^2
    Z_multin[i] = eta[i] + pi_vector[i]*c_vector[i]^2/(1-pi_vector[i]

    X_multin[i,1] = x_y_vector[i]
    X_multin[i,2] = 1
    X_multin[i,3] = 1
    X_multin[i,4] = x_r_vector[i]

    i = i+1
  }

  # combining exponential and multinomial parts

  X = rbind(X_exp, X_multin)
  W = cbind(rbind(W_exp,matrix(0,1,1)),rbind(matrix(0,1,1),W_multin))
  Z = rbind(Z_exp, Z_multin)


  beta =  solve( t(X) %*% W %*% X) %*% t(X) %*% W %*% Z
#maximum likelihood estimate



  phi = beta[1]
  mu = beta[2]
  lambda = beta[3]
  omega = beta[4]

  distance = distance + 1
```

```
    }

# Error in solve.default(t(X) %*% W %*% X) :
# system is computationally singular: reciprocal condition number = 6.715
# ???????????
# beta goes further from the true value when more iterations are done
# ???????????


  #}
```

6.5 Code in R for Analytical inference of ϕ, μ_0, λ_0 and ω and plotting histograms

```
# Analytical inference of parameters phi, mu0, lambda0 and omega


rm(list = ls())
simulation <- function(ct=2,model = 'bt',phi = 0.05,mu_r0 = 0.5,lambda_y0

  sumt = 0
  WT = c(0) # waiting time
  SE = c()  # speciation = 1, extinction = 0
  BT = c(0) # branching time, sum of waiting times
  L_vector = c() # vector that shows what kind of event takes place (sp/e
  x_r_vector = c(x_r) # nr. of predators at every wt
  x_y_vector = c(x_y) # nr. of prey at every wt
  ind_vector = c()


  while(sumt < ct){
    #update parameter
    if(model == 'bt'){
      lambda_r = phi*(x_y-1)
      mu_r = mu_r0
      lambda_y = lambda_y0
      mu_y = omega*x_r
    }

    if (x_y<5){mu_y=0}

    vector_rates = c(rep(lambda_r,x_r),rep(mu_r,x_r),rep(lambda_y,x_y),re
    s   = sum(vector_rates)
    wt = rexp(n = 1,rate = s) # waiting time is exponentially distributed
    BD = rmultinom(1,1,vector_rates/s) # event: birth/death, multinomiall
    L = which(BD == 1)
    L_vector = append(L_vector, L)

    # find out where the 1 in rmultinom occurs, so where does the event o
```

```
  # speciation of predators (x_r)
  ind = 0
  if(L <= x_r){
    # update tree
    x_r= x_r+1
    SE = append(SE,1)
    x_r_vector = append(x_r_vector,x_r)
    x_y_vector = append(x_y_vector,x_y)
    ind=1
  }
  # extinction of predators (x_r)
  if (x_r< L & L <= (2*x_r) & ind == 0) {

    x_r = x_r-1
    SE = append(SE,0)
    x_r_vector = append(x_r_vector,x_r)
    x_y_vector = append(x_y_vector,x_y)
    ind=2
  }
  #speciation of prey (x_y)
  if ((2*x_r)< L & L <= (2*x_r+x_y) & ind == 0) {
    x_y= x_y+1
    SE = append(SE,1)
    x_r_vector = append(x_r_vector,x_r)
    x_y_vector = append(x_y_vector,x_y)
    ind=3
  }
  #extinction of prey (x_y)
  if ((2*x_r+x_y) < L & ind == 0) {
    x_y = x_y-1
    SE = append(SE,0)
    x_r_vector = append(x_r_vector,x_r)
    x_y_vector = append(x_y_vector,x_y)
    ind=4
  }

  WT = append(WT, wt)
  sumt = sum(WT)
  ind_vector = append(ind_vector, ind)


  # if the nr. of predators becomes 0, the tree should reboot
  if (x_r == 0) {
    break()
  }
}
```

```
  p = list(wt=WT,E=SE)    # instead of using SE, use ind_vector as eventfu
  return(list(p=p,WT=WT,ind_vector=ind_vector,x_r_vector=x_r_vector,x_y_v
}

# Inference

estimation <- function(p,WT,x_y_vector, ind_vector,x_r_vector){

  p = sum(ind_vector==1)
  q = sum(ind_vector==2)
  r = sum(ind_vector==3)
  s = sum(ind_vector==4)

  phi_hat = p / sum(x_r_vector*x_y_vector*WT)
  mu_r0_hat = q / sum(x_r_vector*WT)
  lambda_y0_hat = r / sum(x_y_vector*WT)
  omega_hat = s / sum(x_r_vector*x_y_vector*WT)

  var = matrix(0, 4, 4)
  var[1,1] = phi_hat^2/p
  var[2,2] = mu_r0_hat^2/q
  var[3,3] = lambda_y0_hat^2/r
  var[4,4] = omega_hat^2/s
  sigma = diag(sqrt(var))

  return(list(phi_hat=phi_hat,mu_r0_hat=mu_r0_hat,lambda_y0_hat=lambda_y0

  }

sim = simulation()
est = estimation(p=sim$p,WT = sim$WT,x_y_vector=sim$x_y_vector, ind_vecto


## Experiment

phi = 0.05     #dependency parameter of predators
mu_r0 = 0.5    #constant
lambda_y0 = 1 #constant
omega = 0.05 #dependency parameter of prey

n_sim = 1000

Est_pars = matrix(nrow=n_sim,ncol=4)
z_tot = c(0,0,0,0)
sigma_tot = c(0,0,0,0)

for(i in 1:n_sim) {
  sim = simulation(phi = phi, mu_r0 = mu_r0, lambda_y0 = lambda_y0, omega
```

```
  est = estimation(p=sim$p,WT = sim$WT,x_y_vector=sim$x_y_vector, ind_vec
  Est_pars[i,1] = est$phi_hat
  Est_pars[i,2] = est$mu_r0_hat
  Est_pars[i,3] = est$lambda_y0_hat
  Est_pars[i,4] = est$omega_hat

  est_difference = c(phi-Est_pars[i,1], mu_r0 - Est_pars[i,2], lambda_y0
  z = est_difference / est$sigma
  if (any(is.na(z))==FALSE) {z = c(0,0,0,0)}
  if (any(is.na(sigma))==FALSE) {est$sigma = c(0,0,0,0)}

  z_tot = z_tot + z
  sigma_tot = sigma_tot + est$sigma

}

z_av = z_tot/n_sim
sigma_av = sigma_tot/n_sim

qplot(Est_pars[,1],geom='histogram')
qplot(Est_pars[,2],geom='histogram')
qplot(Est_pars[,3],geom='histogram')
qplot(Est_pars[,4],geom='histogram')
```