Untitled

J.P. Meagher

2 July 2017

Title Slide (1 min 15 sec) (1 min 15 sec)

Good morning everyone, and thank you very much for that introduction _____ .

My name is Joe Meagher, and I am going to speak to you today about some of the research I have done so far for my PhD at the University of Warwick's Department of Statistics, titled "Phylogenetic Gaussian Processes for the Ancestral Reconstruction of Bat Echolocation Calls".

Over the next 25 minutes or so I will try to:

- Provide some background and motivation for modelling Bat echolocation calls.
- Give a gentle introduction to Ancestral Reconstruction and define function-valued traits in this context.
- Present results reducing models for the evolution of function-valued traits to Gaussian Process Regression.
- Explain how we applied these models to bat echolocation calls and discuss the results.
- Finally, I'll talk a little about where I see this work going in the

Acknowlegdements (1 min 15 sec) (2 min 30 sec)

I have been supported by a few organisations throughout my PhD.

- This work is funded by the EPSRC through the Engage project, which aims to enable non-programming scientists develop systems for detection in audio and visual data.
- The Alan Turing Institute has facilitated interdisciplinary collaboration on this project, particularly last summer.
- The Gaelic Players Association has also supported me, personally, during this research project.

I also want to thank my PhD supervisors, co-authors on this conference paper.

- Prof. Mark Girolami, Chair of Statistics at the Imperial and Director for the Alan Turing Institute-Lloyd's Register Foundation Programme in Data-Centric Engineering.
- Dr. Theo Damoulas, an Assistant Professor in Data Science at Warwick and a Faculty Fellow at the ATI.
- Prof. Kate Jones, Chair of Ecology and Biodiversity at UCL, based at the Centre for Biodiversity and Environment Research

Background and Motivation (4 min) (6 min 30 sec)

So, why might one be interested in studying the evolution of echolocation in bats?

Bat Background

Well, bats, as a whole, comprise the second most speciose order of mammals, after rodents.

There are over 1000 known species of bat, with about 17 species breeding in the UK & Ireland.

Bats are found on every continent, except Antartica, and they have successfully adapted to a massive variety of habitats and foraging strategies. The are also considered to be keystone species in desert and tropical environments. If any of you are unfamiliar with the notion of keystone species I recommend taking a look at a video on the impact of reintroducing wolves in Yellowstone. It presents a pretty incredible narrative.

What sets bats apart from other mammals is that they fly, in the dark most of the time, and even though they are not actually blind,

Ancestral Reconstruction and Function-Valued Traits

Ok, so by now you probably have a fair idea what I mean when I say Ancestral Recostruction, but to be precise:

Definition

Ancestral Reconstruction is the extrapolation back in time from measured characteristics of individuals (or populations) to their common ancestors.

It is an application of phylogenetics. Phylogenetics being the reconstruction and study of evolutionary relationships.

Some examples. Ancestral Reconstruction can be applied to genetic sequences, the amino acid sequence of a protien, or observed traits of an organism, an echolocation call, for example.

The results of an Ancestral Reconstruction study are heavily reliant on the modelling assumptions made. They are the phylogenetic tree of evolutionary relationships and also the model for the evolutionary dynamics, how the object of interest changes through evolutionary time.

Modelling the Evolution of Function Valued Traits

What I have just described, in statistical terms, is a functional data object, and fortunately for me there has been a lot of work done in the area of Functional Data Analysis which informs our analysis of function valued traits.

In fact the machinery for applying Brownian Motion and Ornstein Uhlenbeck models for evolution to function-valued traits has already been developed.

Phylogenetic Gaussian Process Regression

Firstly, I need to introduce Gaussian Processes.

Formally, a Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

A Gaussian Process places a prior distribution over functions, and so for a p dimensional input space, we can define a mean function, which is just the expected value of the functions, and a kernel which captures the covariance between points in the input space. The kernel, and the hyperparameters embeddid within it, defines the

Application to Bat Echolocation Calls

The Data

The actual calls used in this analysis were recorded in north and central Mexico and they relate to 449 individuals from 22 species. Altogether we have just over 1800 individual call recordings and these are not necessarily spread uniformly across the sample with respect to individuals or species.

The dataset aimed for 100 individual calls per species, but some some species only had 4 idividual bats recorded so this means the number of calls per individual bat ranges from 1 to 40. This structure of the data was something we were wary of in the analysis.

We are not trying to infer the phylogenetic tree itself here, and I'm not convinced that using echolocation calls to do that would be a good idea, although maybe it's something we will revisit in the future. For now, we assume a phylogenetic tree for the recorded bats is the same as bat super tree published in 2012. that is what you can see on the slide.

Results

Kernel

We assumed that each phylogenetic Gaussian process was an Ornstein-Uhlenbeck process. We felt this was appropriate because bat echolocation calls seem to exhibit stabilising selection, as they can be thought of as serving a very specific purpose, which does in a sense have an optimal value.

The kernel for such a process is given here, where the hyperparameters arethe phylogenetic noise, the phylogenetic length scale, and the non-phylogenetic noise.

Again, without going into detail, a resampling procedure was set up and Type II maximum likelihood estimation used to estimate the hyperparameters of the phylogenetic Ornstein-Uhlenbeck Process. The results of which I have reported here.

Model Hyperparameters

Now in order to interpret these results let me draw you attention to the third basis function

Conclusions

So what can we conclude from this study.

Well we have identified a phylogenetic signal in the data, as one would hope.

More specifically, we have identified a strong phylogenetic signal at low frequencies, indicating that ancestral bats had lower frequency calls than their descendants.

To be honest, there is more work to do in interpreting the output of this model but it is a preliminary analysis, and I personally am more interested is seeing what the next iteration of this model tells us.

Future Research Directions

Bat calls do in fact have a complex temporal structure and so spectral densities is not a good way of characterising them. A more usual method is to use a spectrogram, shown here.

The first extension to make to this model is to implement it for call spectrograms rather than spectral densities. We hope to complete this over the coming months.

Looking beyond this, we may examine other representations of the echolocation calls, such as Multi-component separation techniques.

Also the model for evolution used could also be improved, using Stable models for evolution seem like a promising route to consider.

That about covers everything I wanted to say. I'd just like thank Winton and Imperial for giving me the chance to present my work, I hope you enjoyed it and I'm happy to try answer any questions you may have.

Thank you.