Conference Talk

J.P. Meagher

1 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 future.

Before I go any further however, there are some organisations and people I want to thank.

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, most bats use echolocation to percieve the world around them.

Given all this, I think it's fair to say that bats are pretty remarkable creatures, and worthy of our attention, although its not difficult to convince people that this is the case.

Public Engagement

Bats capture peoples imagination.

The appear consistently in folklore and mythology, from being symbols of good fortune in ancient China, to Bram Stoker's more sinister Count Dracula.

Research into bats generates public interest, Mark did a BBC television interview discussing some of the groups research and Kate supervised work, gathering the data used in this analysis actually, that featured in this BBC article.

This level of public engagement means that citizen science initiatives for gathering and labelling bat call data have been successful.

This brings me on to the specific motivation for wanting to reconstruct ancestral bat calls.

Acoustic Biodiversity Monitoring

Bats have been identified as ideal Bioindicatiors for biodiversity monitoring.

Because they leak information about themselves into their environment through their echolocation calls, they can be monitored non-invasively using acoustic monitoring equipment. This picture shows the sort of monitoring device that can be used.

Ideally, these recordings will provide us with information about bats activity which can then be used as a proxy for the state of the environment in general. For this process to be really effective, a good understanding of bats natural history is required.

Historically, how have bats interacted with the flora and fauna in a given environment. Understanding the evolution of echolocation is a part of this.

General Application of Techniques

The final aspect of motivation for this project that I am going to discuss, is the general applicability of these techniques.

The techniques used is this analysis have a lot in common with to those used for bat call and bird song classification.

There is also work being done to reconstruct latin phonetically from modern romance languages.

Phyologenetics is a big, active area of research.

I don't thing that it is too much of a stretch to think that insights uncovered in this research project could eventually feed into these other areas.

Summary of Research Project

So, to summarise what we are hoping to achieve with this research project, let me refer you to this diagram.

We have audio recordings of the echolocation calls of various species of bat.

Assuming that these bats are related according to some phylogenetic tree, can we then compare and model the calls of existing bats in such a way that we can make some sensible statments about the calls of ancestral bats.

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.

Evolutionary Models for Continuous Character Traits

There are a few approaches that can be taken to modelling continuous character traits, that is traits that vary along some continuous scale.

The most popular of these methods take the change in the trait over some time period to have a Gaussian distribution, which gives us Brownian Motion and Ornstein-Uhlenbeck models for the evolution of continuous character traits.

Now, Brownian Motion allows the trait to vary in an unrestricted way, evolution is essentially modelled as a random walk over the scale the trait is measured on. This is not necessarily the case in reality, as traits usually serve some purpose and so may have some optimal value which they vary around.

This is where The Ornstein-Uhlenbeck process for modelling stabilising selection comes into play. An Ornstein-Uhlenbeck process essentially behaves a Brownian Motion process with a tendency to some central value. This is more appropriate for modelling a trait with some optimal value which the trait then varies around, which is called stabilising selection.

And so what you see on this slide here is a simulation of an Ornstein Uhlenbeck Evolutionary Process over a given phylogenetic tree. The value at the root of the tree is 0 which is the value the trait varies around.

So these are the processes typically used to model the evolution of continuous character traits. However echolocation calls are not continuous character traits. So now I want to introduce the idea of a function-valued trait.

Function-Valued Traits

A function-valued trait is a trait that is repeatedly measured, along some continuous scale, say time, where measurements can represent points on a curve.

Typical examples are growth curves, where the size of an individual is measured repeatedly over their lifetime.

For a function-valued trait, the mean for the trait over the space in which it is measured can vary, but this variation is both gradual and continuous. The same is true for the covariance of the trait.

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 structure we place on the observed functions.

Nick Jones and John Moriarty, of Imperial and Manchester respectively extended Gaussian Processes to Phylogenies, which allows Evolutionary Inference on function-valued traits.

Assumptions

For a phylogenetic Gaussian Process, the input space, $\mathbf{x}$, becomes the Phylogenetic tree, and so the distance between observations becomes the distance between points in evolutionary time.

So, now we have observations which exist in the trait space and evolutionary time.

In order to make work in this context we need to make some simplifying assumptions.

Firstly, traits are conditionally independent given their common ancestors, which seems reasonable.

The second assumption states that the process governing the evolutionary dynamics are the same along every individual branch of the phylogenetic tree. This assumption is less reasonable and may be something you would look at relaxing as the analysis develops.

The third assumption just says that kernel of the marginal Gaussian process is space time separable. The marginal process being the process from an ancestor to its direct descendant.

With these assumptions in place Jones and Moriarty prove a few results, the most important of those for this analysis being this one here.

Theoretical Results

What this result states is that a phylogenetic Gaussian process which is time-space separable is equivalent to the sum of univariate Gaussian Processes times their corresponding deterministic basis function, which we could also call a component.

Therefore, if we can define a set of basis functions our evolutionary inference over a function-valued trait reduces to the sum of n independent univariate Gaussian processes, which are the same processes that are used for modelling continuous character state evolution.

So now we can go ahead and model the evolution of our function-valued trait.

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.

Echolocation Calls as Function-Valued Traits

Now, as regards the function-valued traits, the echolocation calls are repeated measurements of a trait the varies over a continuous scale, one of which is plotted here. However if someone has any ideas on how to compare acoustic waveforms directly for this sort of analysis I would be delighted to hear from them.

In our analysis we transformed the echolocation calls into the frequency domain with a Fourier transform. This gave us the energy spectral density of the call, which I will probably refer to as the spectral density from now on.

This gives us the energy intensity of the call along the frequency spectrum.

Finally, we smoothed each spectral density with smoothing splines and restricted each one to the 9-212 kHz frequency band, as this is the range of frequencies bats use for their echolocation calls.

These smoothed restricted spectral densities are the function valued traits we modelled.

Independent Basis Functions

So now that we have a sample of function-valued traits, we needed to identify the deterministic basis functions that the independent Gaussian processes would act upon.

Now I dont have time to deal with the specifics of what we did but the details are covered in the conference paper.

Broadly speaking, we set up a resampling procedure to aviod bias in the basis functions identified. We then found Independent Principal Components from the resampled datasets. The basis functions identified can be seen here.

Our interpretation of these basis functions is that they reflect concentrations of energy at various bandwiths and identify the the frequency band used by different species for their echolocation calls.

Once the set of basis functions was identified we could then move on to inferring each of the independent univariate Gaussian processes.

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.

If we refer back to our set of basis we can see that this represents a high concentration of energy at lower frequencies.

The next point I want to draw your attention to is the length scale hyperparameter ℓ. This can be interpreted as millions of years in evolutionary time over which the phylogenetic signal decays. Given the phylogenetic tree only goes back 50 million year, a lenght-scale of 70 indicates quite a strong phylogenetic signal.

Now I'd like to consider the noise hyperparameters, those for phylogenetic and non-phylogenetic noise. I thing what is important here is the ratio of these hyperparameters, which indicates what proportion of the variance in the sample is due to phylogenetic factors.

In this case more than 3/4 of the variation is attributable to phylogenetic factors.

So having inferred the model hyperparameters, we can now perform our Ancestral Reconstruction.

Posterior Predictive Distributions

Ancestral resconstruction in this context reduces to finding the posterior predictive distribution for internal nodes of the tree.

What I'm showing you here is an ancestral reconstruction for the spectral density of the echolocation call for the most recent common ancestor of 3 extant species of mexican bat.

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.