Chapter 1

Ancestral Reconstruction of Bat Echolocation Calls

J.P. Meagher*

Department of Statistics,
University of Warwick,
J.Meagher@Warwick.ac.uk†

Something something bats something evolution something ancestral reconstruction something magic.

1. Introduction

Advances in technology allowing the precise quantification and storage of information about the world around us continues to drive the emergence of Data Science as a discipline distinct from both Statistics and Computer Science.

Bioacoustics is one area of research generating vast quantities of data which also captures the imagination of the public, as evidenced by successful citizen science initiatives.[1,2] Bioacoustic techniques for biodiversity monitoring[3,4] have the potential to make real policy impacts, particularly with regard to sustainable economic development and nature conservation.

In the acoustic monitoring of biodiversity, bats (order *Chiroptera*) are of particular interest. Bats have been identified as ideal bioindicators for monitoring climate change and habitat quality,[5] largely because bats broadcast information about themselves into their environment in the form of echolocation calls.[6] The development of automatic acoustic monitoring algorithms for classifying species of bats[3,7] means that large scale, non-invasive monitoring is becoming possible.

While monitoring bat populations provides useful information, understanding the root causes and effects of what is observed requires that the

*Author footnote.
†Affiliation footnote.

natural history of extant bat species is also well understood. Traits, such as call structure or body size, exhibited by particular bat species are linked to the bats interactions with its environment.[8] Existing fossil records are of limited use in inferring the traits exhibited by ancestral bats, particularly with respect to echolocation calls. The reconstruction of ancestral traits relies heavily on the comparative analysis[9] of extant bat species. Thus, statistical data science techniques may be particularly useful for inferring the evolutionary dynamics and reconstructing ancestral states of echolocation in bats.

Previous studies of bat echolocation calls for both classification[7] and ancestral reconstruction[10] examined features of the call extracted from the call spectrogram. These, somewhat arbitrary, call features relied on significant domain knowledge to ensure they were sensibly selected and used. More recently however, general techniques for the classification of acoustic signals have been developed.[11,4] These methods do not require, but can be augmented by, domain knowledge. General techniques for ancestral reconstruction of function-valued traits, such as speech sounds or echolocation calls, have been proposed.[12] The study of bat echolocation calls offers an opportunity to examine the efficacy of these techniques.

A function-valued trait is measured along some continuous scale, usually time, and can then be modelled as a continuous mathematical function using techniques for functional data analysis.[13] Jones & Moriarty[14] developed a method which extends Gaussian Process Regression[15] to model the evolution of function-valued traits over a phylogeny. The model facilitates the implementation of two popular models for continuous character state evolution,[16] the Brownian Motion and Ornstein-Uhlenbeck models.[17] A full demonstration of ancestral reconstruction for synthetic data using the method was presented by Hajipantelis et al.[18]

This general approach to evolutionary inference for function-valued traits is implemented here for a set of bat echolocation calls. Our goal in doing so is twofold. These techniques had previously been considered in the context of modelling the evolution of human speech sounds in language.[12] It is hoped that by applying these methods in the simpler context of the evolution of bat echolocation calls that progress can be made towards resolving methodological problems. For example, how do we extend these methods to more realistic models of evolution?

We are also interested in what specifically these models tell us about bats and the evolutionary dynamics driving the development of echolocation. What impact might these results have on our understanding of ancestral

bats and their behaviour?

This paper presents the early stages of our research and some preliminary results.

2. Functional Representation of Bat Echolocation Calls

A functional data object is generated when repeated measurements of some process are taken along a continuous scale, such as time.[13] These measurements can be thought of as representing points on a curve that varies gradually and continuously. In the context of phylogenetics, these functional data objects are called function-valued traits.[19]

Denote the m^{th} call recording of the l^{th} individual bat of species S by $\{\tilde{x}^S_{lm}(n) : n = 0, \ldots, N^S_{lm} - 1\}$. Thus $\{\tilde{x}^S_{lm}(\cdot)\}$ is a noisy realisation of $x^S(\cdot)$, the echolocation call generating process for species S, observed at the time points given by $\frac{n}{f_S}$, where f_S is the process sampling rate in samples per second (Hz).

Call recordings themselves are in fact functional data objects, however modelling the phylogenetic relationships between $\{\tilde{x}^S_{lm}(\cdot)\}$ and $\{\tilde{x}^{S'}_{l'm'}(\cdot)\}$ directly implies that the processes are comparable at point n. This is not the case for acoustic signals, which are sinusoidal and can vary in time without significantly altering the information carried. Thus some alternative functional representation of the signal is required.

The discrete Fourier transform of the signal $\{\tilde{x}^{S'}_{l'm'}(\cdot)\}$ is given by

$$\tilde{X}^S_{lm}(k) = \sum_{n=0}^{N^S_{lm}-1} \tilde{x}^S_{lm}(n) e^{-i2\pi kn/N^S_{lm}}.$$

The energy spectral density of this signal is the magnitude of the Fourier transform and so the log energy spectral density per second (in decibel) is estimated by

$$\tilde{\mathcal{E}}^S_{lm}(k) = 10 \log_{10} \left(\frac{|\tilde{X}^S_{lm}(k)| f_s}{N^S_{lm}} \right).$$

The log energy spectral density estimate for the signal $\tilde{x}^S_{lm}(\cdot)$, $\tilde{\mathcal{E}}^S_{lm}(\cdot)$ is now considered to be a noisy estimate of the log energy spectral density for species S, denoted $\mathcal{E}^S(\cdot)$. The log energy spectral density is a periodic function of frequency which describes the energy of a signal at each frequency on the interval $F = [0, \frac{f_S}{2}]$. $\tilde{\mathcal{E}}^S_{lm}(\cdot)$ has been mapped to the decibel scale and also scaled according to the length, in time, of $\tilde{x}^S_{lm}(\cdot)$.[20] We have now

mapped each echolocation call on the same scale and $\tilde{\mathcal{E}}^S_{lm}(\cdot)$ is now comparable to $\tilde{\mathcal{E}}^{S'}_{l'm'}(f)$ at frequency f. In order for these representations to be considered as function-valued traits however, $\tilde{\mathcal{E}}^S_{lm}(\cdot)$ must be smoothed such the call representation in the frequency domain varies gradually and continuously.

The smoothed log energy spectral density is estimated by smoothing splines where

$$\mathcal{E}^S_{lm}(f) = \arg\min_{\mathcal{E}^S_{lm}(\cdot)} \int_0^{\frac{f_S}{2}} \{\tilde{\mathcal{E}}^S_{lm}(\cdot) - \mathcal{E}^S_{lm}(f)\}^2 df + \lambda \int_0^{\frac{f_S}{2}} \{\mathcal{E}^S_{lm}{}''(f)\}^2 df.$$

The smoothing parameter λ is chosen using a generalised cross-validation procedure.[2122]

We now have a functional representation of each bats echolocation call where the pairs of observations $\{f, \mathcal{E}^S_{lm}(f)\}$ and $\{f, \mathcal{E}^{S'}_{l'm'}(f)\}$ are directly comparable. The function-valued traits can now be modelled for evolutionary inference.

3. Phylogenetic Gaussian Process Regression for Bat Echolocation Calls

3.1. *Gaussian Process Regression on Phylogenies*

The key innovation of Jones & Moriarty[14] in extending Gaussian Process Regression[15] for use in evolutionary inference, was to replace the linear measure of distance between observations with a phylogenetic tree, denoted $\mathbf{T}$. When this condition is imposed each of our observations $\mathcal{E}^S(f)$ correspond to a point $(f, \mathbf{t})$ on the frequency-phylogeny $F \times \mathbf{T}$. It is then by constructing a phylogenetic covariance function $\Sigma_{\mathbf{T}}\left(\mathcal{E}^S(\cdot), \mathcal{E}^{S'}(\cdot)\right)$ that evolutionary inference can be performed.

Deriving a tractable form of the phylogenetic covariance function requires some simplifying assumptions. Firstly, it is assumed that conditional on their common ancestors in the phylogenetic tree $\mathbf{T}$, any two traits are statistically independent.

The second assumption is that the statistical relationship between a trait and any of it's descendants in $\mathbf{T}$ is independent of the topology of $\mathbf{T}$. That is to say that the underlying process driving evolutionary changes is identical along all individual branches of the tree. We call this underlying process along each branch the marginal process. The marginal process depends on the date of $\mathbf{t}$, the distance between a point $\mathbf{t} \in \mathbf{T}$ and the root of $\mathbf{T}$, denoted t.

Finally, we assume that the covariance function of the marginal process is separable over evolutionary time and the function-valued trait space. Thus, by defining the frequency only covariance function $K(f, f')$ and the time only covariance function $k(t, t')$ the covariance function of the marginal process is

$$\Sigma\left((f,t),(f',t')\right) = K(f,f')k(t,t')$$

Under these conditions, Jones & Moriarty[14] show that the phylogenetic covariance function is also separable, that is

$$\Sigma_{\mathbf{T}}\left((f,\mathbf{t}),(f',\mathbf{t}')\right) = K(f,f')k_{\mathbf{T}}(\mathbf{t},\mathbf{t}').$$

It is also shown that for a phylogenetic Gaussian Process with this covariance function, Y, and a degenerate Mercer kernel, $K(\cdot,\cdot)$, there exists a set of n deterministic basis functions $\phi_i : F \to \mathbf{R}$ and univariate Gaussian processes X_i for $i = 1, \ldots, n$ such that

$$g(f,\mathbf{t}) = \sum_{i=1}^{n} \phi_i(f) X_i(\mathbf{t})$$

has the same distribution as Y.

Thus, given an appropriate set of basis functions, $\phi_{\mathcal{E}} = [\phi_1^{\mathcal{E}}(\cdot), \ldots, \phi_n^{\mathcal{E}}(\cdot)]$, and Gaussian Processes, $X_{\mathcal{E}} = [X_1^{\mathcal{E}}(\cdot), \ldots\ X_n^{\mathcal{E}}(\cdot)]$, the set of observations of the echolocation function-valued trait can be expressed in matrix notation as

$$\mathcal{E} = X_{\mathcal{E}} \phi_{\mathcal{E}}^{\mathsf{T}},$$

where $X_{\mathcal{E}}$ is the matrix of mixing coefficients of the fixed basis functions determining the function-valued trait. The values of $X_{\mathcal{E}}$ are modelled as evolving by univariate phylogenetic Gaussian Processes.

3.2. *Deterministic Basis Functions*

Applying this model to observed traits requires that ϕ be estimated somehow. Hajipantelis et al.[18] addressed this problem. The model outlined above implicitly assumes that the rows of $X_{\mathcal{E}}$ are independent. This in turn implies that each of the basis, $\phi_i(\cdot)$, evolved independently of one another. $\hat{\phi}$, the estimate for ϕ, must reflect this.

A Functional Principal Components Analysis[13] of the traits would return a set of orthogonal basis functions. This dimension reduction technique allows the selection of n basis functions which describe some proportion of

the variation in the sample. However, this implicitly assumes that the basis functions are also Gaussian, a strong assumption which may not be realistic.

A less stringent condition is to assumes only that the basis functions are independent. Such a set of components can be found by an Independent Components Analysis. Blasche & Wiskott[23] present a method for deriving Independent Components. This two step procedure first implements a Principal Components Analysis to estimate the effective dimensionality of the dataset, before passing the effective dimensions to the CuBICA algorithm. This algorithm then rotates these effective dimensions until the third and fourth cumulants have also been diagonalised, which produces approximately independent basis functions.

3.3. *The Phylogenetic Ornstein-Uhlenbeck Process for Evolutionary Inference*

The relationships between the mixing coefficients, $X_{\mathcal{E}}$, are modelled by a phylogenetic Gaussian Process, which must be defined. The Ornstein-Uhlenbeck process offers a popular method of modelling stabilising selection in comparative studies.[24,10,18] Here, each independent phylogenetic Gaussian Process, $X_i(\cdot)$, is modelled as an Ornstein-Uhlenbeck process.

The phylogenetic Ornstein-Uhlenbeck process is defined by the kernel

$$k^i_{\mathbf{T}}(\mathbf{t},\mathbf{t}') = (\sigma^i_p)^2 \exp\left(\frac{-d_{\mathbf{T}}(\mathbf{t},\mathbf{t}')}{\ell^i}\right) + (\sigma^i_n)^2 \delta_{\mathbf{t},\mathbf{t}'}$$

where δ is the Kronecker delta, $d_{\mathbf{T}}(\mathbf{t},\mathbf{t}')$ is the cophenetic distance between the points $\mathbf{t}$ and $\mathbf{t}'$ on the phylogeny $\mathbf{T}$, and $\theta^i = [\sigma^i_p, \ell^i, \sigma^i_n]^{\mathsf{T}}$ is the vector of hyperparameters for the process $X_i(\cdot)$.

The full phylogenetic covariance function is then

$$\Sigma_{\mathbf{T}}((f,\mathbf{t}),(f',\mathbf{t}')) = \sum_{i=1}^{n} k^i_{\mathbf{T}}(\mathbf{t},\mathbf{t}')\phi^{\mathcal{E}}_i(f)\phi^{\mathcal{E}}_i(f')$$

and the log likelihood associated with the model is

$$\ell(\mathcal{E}|\theta) = -\frac{1}{2}\sum_{i=1}^{n}\left(X_i(\cdot)^{\mathsf{T}} k^i_{\mathbf{T}}(\cdot,\cdot)^{-1} X_i(\cdot) + \log(\det(k^i_{\mathbf{T}}(\cdot,\cdot))) + |X_i(\cdot)|\log 2\pi\right),$$

where $\theta = [\theta^1,\ldots,\theta^n]$, and $|X_i(\cdot)|$ is the Euclidean length of the vector of realisations of $X_i(\cdot)$.

Model selection can be performed by a type II maximum likelihood estimation procedure which maximises the likelihood of the sample with respect to θ.

The model hyperparameters have intuitive interpretations. The variance of observations in the sample is $\sigma_p + \sigma_n$, where σ_p is the phylogenetic noise, and σ_n is the non-phylogenetic noise. σ_p is the proportion of the variance within the sample which due to phylogenetic relationships, while σ_n accounts for other sources of variation. The length-scale parameter, ℓ then indicates the strength of the correlation between traits at various points on the phylogeny, where strong correlations are given by a large ℓ.

Finally, ancestral reconstruction of the function-valued trait at some unobserved point in the phylogeny, $\mathcal{E}^*$, is given by its posterior distribution

$$p(\mathcal{E}^*|\mathcal{E}) \sim \mathcal{N}(A, B)$$

where

$$A = \sum_{i=1}^{n} k^i_{\mathbf{T}}(\mathbf{t}^*, \cdot) k^i_{\mathbf{T}}(\cdot, \cdot)^{-1} X^{\mathcal{E}}_i(\cdot) \phi^{\mathcal{E}}_i$$

and

$$B = \sum_{i=1}^{n} \left(k^i_{\mathbf{T}}(\mathbf{t}^*, \mathbf{t}^*) - k^i_{\mathbf{T}}(\mathbf{t}^*, \cdot) k^i_{\mathbf{T}}(\cdot, \cdot)^{-1} k^i_{\mathbf{T}}(\mathbf{t}^*, \cdot)^{\mathsf{T}} \right) \phi^{\mathcal{E}}_i$$

Thus, all the tools required to perform an ancestral reconstruction of the function-valued trait have been defined.

References

1. P. E. Allen and C. B. Cooper. Citizen science as a tool for biodiversity monitoring. (2006).
2. N. Pettorelli, J. E. Baillie, and S. M. Durant, Indicator bats program: a system for the global acoustic monitoring of bats (2013).
3. V. Stathopoulos, V. Zamora-Gutierrez, K. E. Jones, and M. Girolami, Bat echolocation call identification for biodiversity monitoring: a probabilistic approach, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* (2017).
4. T. Damoulas, S. Henry, A. Farnsworth, M. Lanzone, and C. Gomes. Bayesian classification of flight calls with a novel dynamic time warping kernel. In *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, pp. 424–429 (2010).
5. G. Jones, D. S. Jacobs, T. H. Kunz, M. R. Willig, and P. A. Racey, Carpe noctem: the importance of bats as bioindicators, *Endangered species research.* **8**(1-2), 93–115 (2009).
6. G. Jones, Echolocation, *Current Biology.* **15**(13), R484–R488 (2005).
7. C. L. Walters, R. Freeman, A. Collen, C. Dietz, M. Brock Fenton, G. Jones, M. K. Obrist, S. J. Puechmaille, T. Sattler, B. M. Siemers, et al., A continental-scale tool for acoustic identification of european bats, *Journal of Applied Ecology.* **49**(5), 1064–1074 (2012).

8. H. Aldridge and I. Rautenbach, Morphology, echolocation and resource partitioning in insectivorous bats, *The Journal of Animal Ecology*. pp. 763–778 (1987).
9. J. Felsenstein and J. Felenstein, *Inferring phylogenies*. vol. 2, Sinauer associates Sunderland (2004).
10. A. Collen. *The evolution of echolocation in bats: a comparative approach*. PhD thesis, UCL (University College London) (2012).
11. V. Stathopoulos, V. Zamora-Gutierrez, K. Jones, and M. Girolami. Bat call identification with gaussian process multinomial probit regression and a dynamic time warping kernel. In *Artificial Intelligence and Statistics*, pp. 913–921 (2014).
12. T. F. P. Group, Phylogenetic inference for function-valued traits: speech sound evolution, *Trends in ecology & evolution*. **27**(3), 160–166 (2012).
13. J. O. Ramsay, *Functional data analysis*. Wiley Online Library (2006).
14. N. S. Jones and J. Moriarty, Evolutionary inference for function-valued traits: Gaussian process regression on phylogenies, *Journal of The Royal Society Interface*. **10**(78), 20120616 (2013).
15. C. E. Rasmussen, Gaussian processes for machine learning (2006).
16. J. B. Joy, R. H. Liang, R. M. McCloskey, T. Nguyen, and A. F. Poon, Ancestral reconstruction, *PLoS Comput Biol*. **12**(7), e1004763 (2016).
17. R. Lande, Natural selection and random genetic drift in phenotypic evolution, *Evolution*. pp. 314–334 (1976).
18. P. Z. Hadjipantelis, N. S. Jones, J. Moriarty, D. A. Springate, and C. G. Knight, Function-valued traits in evolution, *Journal of The Royal Society Interface*. **10**(82), 20121032 (2013).
19. K. Meyer and M. Kirkpatrick, Up hill, down dale: quantitative genetics of curvaceous traits, *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. **360**(1459), 1443–1455 (2005).
20. A. Antoniou, *Digital signal processing*. McGraw-Hill Toronto, Canada: (2006).
21. J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. vol. 1, Springer series in statistics Springer, Berlin (2001).
22. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2017). URL `https://www.R-project.org/`.
23. T. Blaschke and L. Wiskott, Cubica: Independent component analysis by simultaneous third-and fourth-order cumulant diagonalization, *IEEE Transactions on Signal Processing*. **52**(5), 1250–1256 (2004).
24. T. F. Hansen, Stabilizing selection and the comparative analysis of adaptation, *Evolution*. pp. 1341–1351 (1997).