Chapter 1

Phylogenetic Gaussian Processes for the Ancestral Reconstruction of Bat Echolocation Calls

J.P. Meagher*, T. Damoulas†, K.E. Jones‡, and M. Girolami§

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

The reconstruction of ancestral echolocation calls is an important part of understanding the evolutionary history of bats. General techniques for the ancestral reconstruction of function-valued traits have recently been proposed. A full implementation of phylogenetic Gaussian processes for the ancestral reconstruction of function-valued traits representing bat echolocation calls is presented here. A phylogenetic signal was found in the data and ancestral reconstruction performed. This promising preliminary analysis paves the way for more realistic models for the evolution of echolocation in bats.

1. Introduction

The emerging field of Data Science is driven by research which lies at the nexus of Statistics and Computer Science. Bioacoustics is one such area generating vast quantities of data, often through citizen science initiatives [1]. Bioacoustic techniques for biodiversity monitoring [2] [3] have the potential to make real policy impacts, particularly with regard to sustainable economic development and nature conservation.

Bats (order *Chiroptera*) have been identified as ideal bioindicators for monitoring climate change and habitat quality [4], and are of particular interest for monitoring biodiversity acoustically. Typically, a bat broadcasts information about itself in an ultrasonic echolocation call [5]. The develop-

*J.P Meagher would like to thank the EPSRC for funding this work, and also acknowledge the support of the Alan Turing Institute and Gaelic Players Association.
†The University of Warwick and The Alan Turing Institute
‡Centre for Biodiversity and Environment Research, University College London
§Imperial College London and The Alan Turing Institute

ment of automatic acoustic monitoring algorithms [2] [6] means that large scale, non-invasive monitoring of bats is becoming possible.

Monitoring bat populations provides useful information, but an understanding of the evolutionary history is required to identify the cause and effect of any changes observed. The echolocation call structure, which reflects a bats diet and habitat [7], is a key aspect of this evolutionary history. Reconstructing ancestral traits [8] relies on a statistical comparative analysis incorporating extant species and fossil records [9]. However, the fossil record is of limited use in inferring ancestral echolocation calls in bats. Therefore, statistical data science techniques may shed some light on this topic.

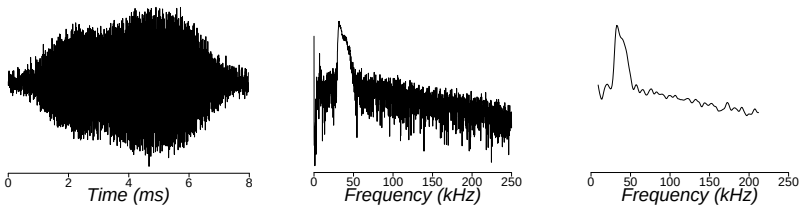
Previous studies of bat echolocation calls for both classification [6] and ancestral reconstruction [10] analysed features extracted from the call spectrogram. These call features relied upon domain knowledge to ensure they were sensibly selected and applied. More recently, general techniques for the classification of acoustic signals have been developed [11] [3]. General techniques for the ancestral reconstruction of function-valued traits have also been proposed [12]. Jones & Moriarty [13] extend Gaussian Process Regression [14] to model the evolution of function-valued traits [15] over a phylogeny, a method which was demonstrated for synthetic data by Hajipantelis *et al.* [16]. This current research investigates these techniques in the context of the bat echolocation calls.

The structure of this paper is as follows, section 2 presents details on representing echolocation call recordings as function-valued traits. Section 3 develops the model for evolution given by a phylogenetic Gaussian Process. The results of the analysis of bat echolocation calls are then presented and discussed in sections 4 and 5.

2. Echolocation Calls as Function-Valued Traits

A functional data object is generated when repeated measurements of some process are taken along a continuous scale, such as time [17]. 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 function-valued traits [15].

Given a phylogenetic tree $\mathbf{T}$, representing the evolutionary relationships between the recorded bat species, we denote the m^{th} call recording of the l^{th} individual bat of the species observed at point $\mathbf{t} \in \mathbf{T}$ by $\{\hat{x}^{\mathbf{t}}_{lm}(n)\}_{n=0}^{N^{\mathbf{t}}_{lm}-1}$. Thus, $\{\hat{x}^{\mathbf{t}}_{lm}(n)\}$ is a series of discrete measurements of the function $x^{\mathbf{t}}_{lm}(\cdot)$,



(a) (b) (c)

Fig. 1. A recording of a bat echolocation call (a) along with the log energy spectral density of the call (b) and the smooth functional representation of that spectral density restricted to the range $[9 - 212]$kHz (c).

observed at the time points given by $\frac{n}{f_S}$, where f_S is the sampling rate, in samples per second (Hz), of the recording. Assume then that $x^{\mathbf{t}}_{lm}(\cdot) = x^{\mathbf{t}}_{l}(\cdot) + z^{\mathbf{t}}_{lm}(\cdot)$, where $x^{\mathbf{t}}_{l}(\cdot)$ is the representative call function for the l^{th} individual and $z^{\mathbf{t}}_{lm}(\cdot)$ is the noise process for the m^{th} call. Further, assume that $x^{\mathbf{t}}_{l}(\cdot) = x^{\mathbf{t}}(\cdot) + z^{\mathbf{t}}_{l}(\cdot)$ where $x^{\mathbf{t}}(\cdot)$ is the representative call function for the bat species at $\mathbf{t}$ and $z^{\mathbf{t}}_{l}(\cdot)$ is the noise process for the l^{th} individual. It is the phylogenetic relationship between the species level echolocation call functions that we are interested in modelling.

The call recordings themselves are functional data objects, however modelling the phylogenetic relationships between $\{x^{\mathbf{t}}_{lm}(t)\}$ and $\{x^{\mathbf{t}'}_{l'm'}(t)\}$ directly implies that the processes are comparable at time t. This is not the case for acoustic signals, a phenomenon which is often addressed by dynamic time warping [18]. Another approach to this issue is to consider an alternative functional representation of the signal.

The Fourier transform of $x^{\mathbf{t}}_{lm}(\cdot)$ is given by

$$X^{\mathbf{t}}_{lm}(f) = \int_{-\infty}^{\infty} x^{\mathbf{t}}_{lm}(t) e^{-i2\pi ft} dt.$$

The energy spectral density of $x^{\mathbf{t}}_{lm}(\cdot)$ is the squared magnitude of the Fourier transform and the log energy spectral density is given by

$$\mathcal{E}^{\mathbf{t}}_{lm}(\cdot) = 10 \log_{10} \left(|X^{\mathbf{t}}_{lm}(\cdot)|^2 \right).$$

Similarly to the call functions, $\mathcal{E}^{\mathbf{t}}_{lm}(\cdot)$ is the log energy spectral density of the m^{th} call of the l^{th} individual from the species at $\mathbf{t}$ where $\mathcal{E}^{\mathbf{t}}_{lm}(\cdot) = \mathcal{E}^{\mathbf{t}}_{l}(\cdot) + \mathcal{Z}^{\mathbf{t}}_{lm}(\cdot)$ and $\mathcal{E}^{\mathbf{t}}_{l}(\cdot) = \mathcal{E}^{\mathbf{t}}(\cdot) + \mathcal{Z}^{\mathbf{t}}_{l}(\cdot)$ where $\mathcal{Z}^{\mathbf{t}}_{lm}(\cdot)$ and $\mathcal{Z}^{\mathbf{t}}_{l}(\cdot)$ are noise processes, each with an expected value of zero. The log energy spectral density is a periodic function of frequency which describes the energy of a signal at each frequency on the interval $F = [0, \frac{f_S}{2}]$. [19]

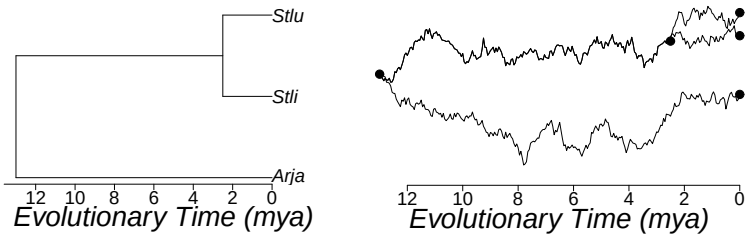
The discrete Fourier Transform [19] of $\{\hat{x}^{\mathbf{t}}_{lm}(n)\}$ provides an estimate for the log energy spectral density, the positive frequencies of which are denoted $\{\mathcal{E}^{\mathbf{t}}_{lm}(k) : k = 0, \ldots, \frac{N^{\mathbf{t}}_{lm}}{2} + 1\}$. Smoothing splines [20] are applied to this series to obtain $\hat{\mathcal{E}}^{\mathbf{t}}_{lm}(\cdot)$, a smooth function estimating $\mathcal{E}^{\mathbf{t}}_{lm}(\cdot)$.

We now have a functional representation of each bats echolocation call where the pairs of observations $\{f, \hat{\mathcal{E}}^{\mathbf{t}}_{lm}(f)\}$ and $\{f, \hat{\mathcal{E}}^{\mathbf{t}'}_{l'm'}(f)\}$ are directly comparable. These function-valued traits can now be modelled for evolutionary inference.

3. Phylogenetic Gaussian Processes

A Gaussian process places a prior distribution over functions, $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$, where $x \in \mathbf{R}^P$ is some input variable, the mean function $m(x) = \mathbf{E}[f(x)]$, and the covariance function $k(x, x') = \mathrm{cov}(f(x), f(x'))$. Given observations $\mathbf{y}$ at locations $\{x_n\}_{n=1}^N$, Gaussian noise, and kernel hyperparameters θ, a posterior predictive distribution over functions can be inferred analytically. See Rasmussen & Williams [14] for an in depth treatment.

Jones & Moriarty [13] extend GPs for the inference of function-valued traits over a phylogeny. Consider $\mathcal{E}^{\mathbf{t}}(\cdot)$, a functional representation of the echolocation call of the species observed at the point $\mathbf{t}$ on the phylogenetic tree $\mathbf{T}$ with respect to frequency. Modelling this as GP function, where $\mathcal{E}^{\mathbf{t}}(f)$ corresponds to a point $(f, \mathbf{t})$ on the frequency-phylogeny $F \times \mathbf{T}$, requires that a suitable phylogenetic covariance function, $\Sigma_{\mathbf{T}}\left((f, \mathbf{t}), (f', \mathbf{t}')\right)$, is defined.



(a) (b)

Fig. 2. A sub tree from the full phylogeny $\mathbf{T}$ (a) and a simulated univariate phylogenetic OU GP over that sub tree (b).

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 $\mathbf{t}$ and the root of $\mathbf{T}$, denoted t.

Finally, it is assumed 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 the phylogenetic covariance function is also separable and so

$$\Sigma_{\mathbf{T}}\left((f, \mathbf{t}), (f', \mathbf{t}')\right) = K(f, f')k_{\mathbf{T}}(\mathbf{t}, \mathbf{t}'). \tag{1}$$

For a phylogenetic Gaussian Process Y with covariance function given by (1), when K is a degenerate Mercer kernel, there exists a set of n deterministic basis functions $\phi_i : F \to \mathbf{R}$ and univariate GPs 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. The full phylogenetic covariance function of this phylogenetic GP is

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

where $\int \phi_i(f)\phi_j(f)df = \delta_{ij}$, δ being the Kronecker delta, and so the phylogenetic covariance function depends only on $\mathbf{t}, \mathbf{t}' \in \mathbf{T}$.

Thus, given function-valued traits observed at $\mathbf{f} \times \sqcup$ on the frequency-phylogeny, where $\mathbf{f} = [f_1, \ldots, f_q]^{\mathsf{T}}$ and $\sqcup = [\mathbf{t}_1, \ldots, \mathbf{t}_Q]^{\mathbf{T}}$, an appropriate set of basis functions $\phi_F = [\phi^F_1(\mathbf{f}), \ldots, \phi^F_n(\mathbf{f})]$ for the traits $\mathcal{E} = [\mathcal{E}^{\mathbf{t}}(\mathbf{f}), \ldots, \mathcal{E}^{\mathbf{t}'}(\mathbf{f})]$, and Gaussian Processes, $X_{\mathbf{T}} = [X^{\mathbf{T}}_1(\sqcup), \ldots\ X^{\mathbf{T}}_n(\sqcup)]$, the set of observations of the echolocation function-valued trait are then

$$\mathcal{E} = X_{\mathbf{T}}\phi^{\mathsf{T}}_F. \tag{2}$$

The problem of obtaining estimators $\hat{\phi}_F$ and $\hat{X}_{\mathbf{T}}$ is dealt with by Hajipantelis *et al.* [16]. $\hat{\phi}_F$ is obtained by Independent Components Analysis, as described by Blaschke & Wiscott [21] after using a resampling procedure to obtain stable principal components for the observed traits. Given $\hat{\phi}_F$, the estimated matrix of mixing coefficients is $\hat{X}_{\mathbf{T}} = \mathcal{E}(\hat{\phi}_F^{\mathsf{T}})^{-1}$.

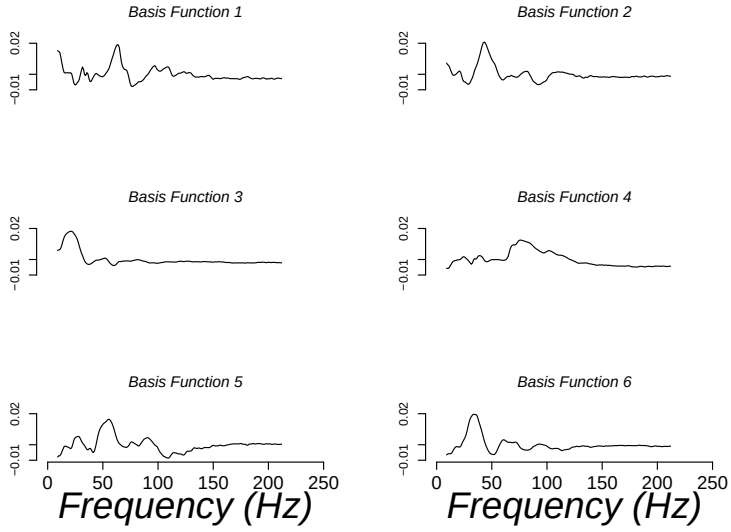


Fig. 3. Set of independently evolving basis functions identified for bat echolocation calls.

Each column of $X_{\mathbf{T}}$ is an independent, univariate, phylogenetic GP, $X_i^{\mathbf{T}}(\sqcup)$, modelled here with phylogenetic Ornstein-Uhlenbeck (OU) process kernel.

The phylogenetic OU process is defined by the kernel

$$k^i_{\mathbf{T}}(\mathbf{t}, \mathbf{t}') = (\sigma_p^i)^2 \exp\left(\frac{-d_{\mathbf{T}}(\mathbf{t}, \mathbf{t}')}{\ell^i}\right) + (\sigma_n^i)^2 \delta_{\mathbf{t},\mathbf{t}'} \tag{3}$$

where δ is the Kronecker delta, $d_{\mathbf{T}}(\mathbf{t}, \mathbf{t}')$ is the distance along $\mathbf{T}$ between $\mathbf{t}$ and $\mathbf{t}' \in \mathbf{T}$, and $\theta^i = [\sigma_p^i, \ell^i, \sigma_n^i]^{\mathsf{T}}$ is the vector of hyperparameters for $X_i^{\mathbf{T}}(\cdot)$. The phylogenetic covariance matrix for $X_i^{\mathbf{T}}(\sqcup)$ is denoted $\Sigma^i_{\mathbf{T}}(\sqcup, \sqcup)$ and the marginal likelihood of the observed data given θ is

$$\log(p(\mathcal{E}|\theta)) \propto -\frac{1}{2}\sum_{i=1}^{n}\left(X_i(\sqcup)^{\mathsf{T}}\Sigma^i_{\mathbf{T}}(\sqcup,\sqcup)^{-1}X_i(\sqcup) + \log|\Sigma^i_{\mathbf{T}}(\sqcup,\sqcup)|\right) \tag{4}$$

and so θ can be estimated by type II maximum likelihood estimation.

Ancestral Reconstruction of the function valued trait for the species at $\mathbf{t}^*$ then amounts to inferring the posterior predictive distribution $p(\mathcal{E}^{\mathbf{t}^*}(\cdot)|\mathcal{E}) \sim \mathcal{N}(A, B)$ where

$$A = \sum_{i=1}^{n} \left(\Sigma^i_{\mathbf{T}}(\mathbf{t}^*, \sqcup) \left(\Sigma^i_{\mathbf{T}}(\sqcup, \sqcup) \right)^{-1} X_i^{\mathcal{E}}(\sqcup) \phi_i(\cdot) \right) \tag{5}$$

$$B = \sum_{i=1}^{n} \left(\Sigma^i_{\mathbf{T}}(\mathbf{t}^*, \mathbf{t}^*) - \Sigma^i_{\mathbf{T}}(\mathbf{t}^*, \sqcup) \left(\Sigma^i_{\mathbf{T}}(\sqcup, \sqcup) \right)^{-1} \Sigma^i_{\mathbf{T}}(\mathbf{t}^*, \sqcup)^{\mathsf{T}} \right) \phi_i(\cdot) \tag{6}$$

We note that the elements of θ each have intuitive interpretations. The total variation of observed points is $\sigma_p + \sigma_n$, where σ_p is the phylogenetic noise, and σ_n is the non-phylogenetic noise. σ_p is the variation depending on the evolutionary time between any $\mathbf{t}, \mathbf{t}' \in \mathbf{T}$, while σ_n accounts for variation that does not depend on the phylogeny. The length-scale parameter, ℓ, indicates the strength of the correlation between points on $\mathbf{T}$, where large values of ℓ indicate a correlation that decays slowly as $d_{\mathbf{T}}(\mathbf{t}, \mathbf{t}')$ increases.

4. Results

4.1. *Data Description*

Post processed echolocation call data accompanying Stathopoulos *et al.* [2] was used in this analysis. Live bats were caught, identified, and recorded at a sampling frequency of 500 kHz. In total the dataset consists of 22 species from five families, 449 individual bats and 1816 individual echolocation call recordings. The distribution of these call recordings across species is summarised in Table 1.

Collen's [10] Bat super-tree provided the phylogenetic tree of the recorded bat species, $\mathbf{T}$.

4.2. *Hyperparameter Estimation and Ancestral Trait Reconstruction with Phylogenetic Gaussian Processes*

We are interested in modelling the evolution of $\mathcal{E}^{\mathbf{t}}(\cdot)$, the function valued trait representing $x^{\mathbf{t}}_{lm}(\cdot)$, with a phylogenetic GP. However, only 22 species of bat are represented in $\mathbf{T}$. The relatively small size of this dataset presents challenges for the estimation of the kernel hyperparameters in (3). A short simulation study was performed to investigate the the accuracy of estimated hyperparameters for a phylogenetic GP over $\mathbf{T}$.

Table 1. Echolocation Call Dataset

Species	Key	Individuals	Calls
Family: Emballonuridae			
1 *Balantiopteryx plicata*	Bapl	16	100
Family: Molossidae			
2 *Nyctinomops femorosaccus*	Nyfe	16	100
3 *Tadarida brasiliensis*	Tabr	49	100
Family: Vespertilionidae			
4 *Antrozous pallidus*	Anpa	58	100
5 *Eptesicus fuscus*	Epfu	74	100
6 *Idionycteris phyllotis*	Idph	6	100
7 *Lasiurus blossevillii*	Labl	10	90
8 *Lasiurus cinereus*	Laci	5	42
9 *Lasiurus xanthinus*	Laxa	8	100
10 *Myotis volans*	Myvo	8	100
11 *Myotis yumanensis*	Myyu	5	89
12 *Pipistrellus hesperus*	Pihe	85	100
Family: Mormoopidae			
13 *Mormoops megalophylla*	Mome	10	100
14 *Pteronotus davyi*	Ptda	8	100
15 *Pteronotus parnellii*	Ptpa	23	100
16 *Pteronotus personatus*	Ptpe	7	51
Family: Phyllostomidae			
17 *Artibeus jamaicensis*	Arja	11	82
18 *Desmodus rotundus*	Dero	6	38
19 *Leptonycteris yerbabuenae*	Leye	26	100
20 *Macrotus californicus*	Maca	6	53
21 *Sturnira ludovici*	Stlu	8	51
22 *Sturnira lilium*	Stli	4	20

We are not limited to a single observation at any given $\mathbf{t}$. By repeatedly sampling at each observed $\mathbf{t}$, larger samples can be obtained, improving the quality of the estimators $\hat{\theta}$. With this in mind, 1000 independent, univariate phylogenetic GPs were simulated for each of $n = \{1, 2, 4, 8\}$ according to the kernel (3) with $\theta = [1, 50, 1]^{\mathsf{T}}$, where n is the number of samples generated at each leaf node. The likelihood of each of these samples (4) is then maximised to give a type II maximum likelihood estimator $\hat{\theta}$ and the results summarised in Table 2. This simulation study indicates that at least $n = 4$ observations are needed at each leaf node to provide stable estimators $\hat{\theta}$.

Given the modelling assumptions made in Section 2 an unbiased esti-

Table 2. Summary of $\hat{\theta}$ for 1000 simulations of independent OU processes with $\theta = [1, 50, 1]^{\mathsf{T}}$ reporting: sample mean (standard error)

n	$\hat{\sigma}_p$	$\hat{\ell}$	$\hat{\sigma}_n$
1	1.09 (0.47)	10^{14} (10^{15})	0.57 (0.54)
2	0.97 (0.29)	10^{13} (10^{14})	0.99 (0.15)
4	0.97 (0.25)	63.66 (136.96)	1.00 (0.09)
8	0.99 (0.24)	56.21 (48.24)	1.00 (0.06)

mator for $\mathcal{E}^{\mathbf{t}}(\cdot)$ is the sample mean given by

$$\hat{\mathcal{E}}^{\mathbf{t}}(\cdot) = \frac{1}{l_{\mathbf{t}}} \sum_{l=1}^{l_{\mathbf{t}}} \frac{1}{m_l} \sum_{m=1}^{m_l} \hat{\mathcal{E}}^{\mathbf{t}}_{lm}(\cdot) \tag{7}$$

where m_l is the total number of recordings for the l^{th} individual and $l_{\mathbf{t}}$ is the number of individuals recorded from the species at $\mathbf{t} \in \mathbf{T}$. However, Table 2 indicates that 22 samples is not enough to obtain a stable $\hat{\theta}$ by type II maximum likelihood estimation. We implement a resampling procedure to leverage multiple estimates for each $\mathcal{E}^{\mathbf{t}}(\cdot)$ from the dataset. This will produce a stable estimator, $\hat{\theta}$.

A resampled estimator $\hat{\mathcal{E}}^{\mathbf{t}}_r(\cdot)$ is obtained by sampling at random one call from n_r individuals of the species at $\mathbf{t}$ and calculating the arithmetic mean of the sample, similarly to (7). This can be repeated to create an arbitrary number of estimates for $\mathcal{E}^{\mathbf{t}}$. Resampling across all the species in the dataset we create a resampled dataset $\hat{\mathcal{E}}_r = [\hat{\mathcal{E}}^{\mathbf{t}_1}_{r,1}(\mathbf{f}), \hat{\mathcal{E}}^{\mathbf{t}_1}_{r,2}(\mathbf{f}), \ldots, \hat{\mathcal{E}}^{\mathbf{t}_2}_{r,1}(\mathbf{f}), \ldots]$, where $\mathbf{f}$ is the vector of frequencies over which $\hat{\mathcal{E}}^{\mathbf{t}_2}_r(\cdot)$ is sampled. The methods outlined in Section 3 can then be applied to each resampled $\hat{\mathcal{E}}_r$

Our analysis set $n_r = 4$ and included 4 samples of $\hat{\mathcal{E}}^{\mathbf{t}}_r(\mathbf{f})$ in each $\hat{\mathcal{E}}_r$ for $r = 1, \ldots, 1000$. This reflected the structure of the dataset, for which the minimum number of individuals per species was 4, and the results of the simulations study which showed that 4 observations per species provided reasonably stable estimates for θ. Note also that $\mathbf{f} = [9, 10, \ldots, 212]^{\mathsf{T}}$, which reflects the spectrum of frequencies over which bats emit echolocation calls. $\hat{\phi}_F$ was obtained by identifying the first six principal components, which accounted for approximately 85% of the variation, in each $\hat{\mathcal{E}}_r$. By averaging over each sample, a single set of six stable, approximately orthogonal, basis functions were identified. These basis functions were then passed through Blaschke & Wiscott's [21] algorithm to produce a set of six independent basis functions for $\mathcal{E}^{\mathbf{t}}(\cdot)$. Thus $\hat{X}_r$, the matrix of mixing coefficients described

by (2), the columns of which are modelled a phylogenetic OU processes, is obtained for each $\hat{\mathcal{E}}_r$. $\hat{\theta}_r$ is then the type II maximum likelihood estimator of (4) given $\hat{\mathcal{E}}_r$. Table 3 presents the results of the hyperparameter estimation procedure.

Table 3. Summary of $\hat{\theta}_r$ over 1000 $\hat{\mathcal{E}}_r$ samples reporting: sample mean (standard error)

Basis	$\hat{\sigma}_p$	$\hat{\ell}$	$\hat{\sigma}_n$
1	2.30 (0.11)	12.27 (4.18)	1.18 (0.11)
2	3.17 (0.11)	27.63 (3.70)	1.26 (0.13)
3	4.05 (0.32)	70.50 (20.31)	1.19 (0.12)
4	3.32 (0.17)	22.86 (8.95)	1.96 (0.19)
5	3.00 (0.13)	26.93 (2.85)	1.21 (0.11)
6	3.70 (0.14)	12.82 (4.52)	1.28 (0.15)

Ancestral reconstruction by a phylogenetic GP involves obtaining the posterior predictive distribution of the trait at the ancestral node $\mathbf{t}^* \in \mathbf{T}$ given by (5) and (6).

To perform ancestral trait reconstruction for $\mathcal{E}^{\mathbf{t}^*}(\cdot)$ the species level traits are estimated by (7) and and the model hyperparameters are set to be the mean values of θ_r reported in Table 3.

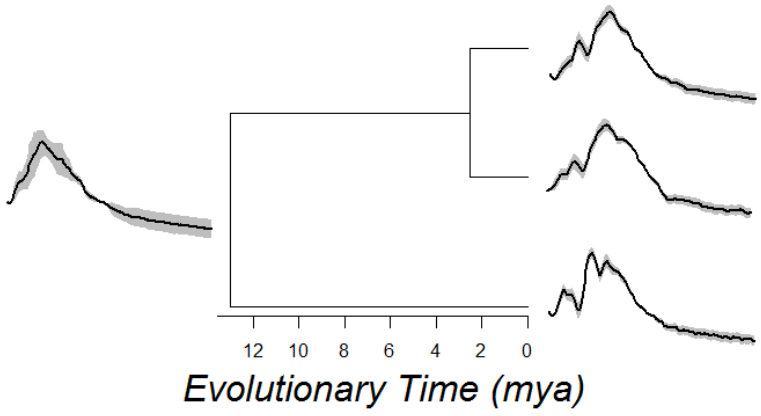


Fig. 4. Ancestral Reconstruction of the function valued trait representing the echolocation calls of the bat species included in the sub tree shown in Figure 2 (a). Grey shaded region represents one standard deviation of variation around $\hat{\mathcal{E}}^{\mathbf{t}}(\cdot)$.

5. Conclusions and Further Work

This preliminary analysis has developed a model for the evolution of echolocation in bats and identified a phylogenetic signal which allows the construction of a posterior predictive distribution for ancestral traits. The log energy spectral density has been identified as a trait representative of the echolocation call in bats. This trait, representing the energy intensity of the call across the frequency spectrum, is modelled as a series of independent components, combinations of energy intensities across the spectrum, each of which evolves according to a phylogenetic Ornstein-Uhlenbeck process. Estimating the hyperparameters governing these Ornstein-Uhlenbeck processes from observed traits provides an insight into the evolution of these traits. Each of the hyperparameters has an intuitive interpretation where $\frac{\sigma_p}{\sigma_p+\sigma_n}$ indicates the proportion of variation in the sample accounted for by the phylogenetic distance between species, while ℓ provides a measure of how quickly correlation along the phylogeny decays. We are working towards understanding what the results of this analysis could mean with respect to the evolution of echolocation in bats.

One particular limitation of the model is the representation of the echolocation call by a log energy spectral density. Echolocation calls have complex spectral and temporal structures, much of which is lost in the log energy spectral density representation. An alternative time-frequency representation, which preserves more of this structure, is the spectrogram. Modelling the evolution of bat echolocation calls with spectrograms, and implementing this model for a larger dataset of bat echolocation calls, is to be the subject of future research.

The interested reader can access the datasets and code used to produce these results through the R package 'sdsBAT' which is still under development and can be found at https://github.com/jpmeagher/sdsBAT.

References

[1] N. Pettorelli, J. E. Baillie, and S. M. Durant, Indicator bats program: a system for the global acoustic monitoring of bats (2013).

[2] 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).

[3] T. Damoulas, S. Henry, A. Farnsworth, M. Lanzone, and C. Gomes. Bayesian classification of flight calls with a novel dynamic time warping kernel. In *Ma-*

chine Learning and Applications (ICMLA), 2010 Ninth International Conference on, pp. 424–429 (2010).

[4] 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).

[5] D. R. Griffin, Echolocation by blind men, bats and radar, *Science.* **100** (2609), 589–590 (1944).

[6] 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).

[7] H. Aldridge and I. Rautenbach, Morphology, echolocation and resource partitioning in insectivorous bats, *The Journal of Animal Ecology.* pp. 763–778 (1987).

[8] J. B. Joy, R. H. Liang, R. M. McCloskey, T. Nguyen, and A. F. Poon, Ancestral reconstruction, *PLoS Comput Biol.* **12**(7), e1004763 (2016).

[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] 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).

[14] C. E. Rasmussen, Gaussian processes for machine learning (2006).

[15] 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).

[16] 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).

[17] J. O. Ramsay, *Functional data analysis.* Wiley Online Library (2006).

[18] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, vol. 10, pp. 359–370 (1994).

[19] A. Antoniou, *Digital signal processing.* McGraw-Hill Toronto, Canada: (2006).

[20] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning.* vol. 1, Springer series in statistics Springer, Berlin (2001).

[21] 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).