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

The reconstruction of ancestral function-valued traits in phylogenetics requires the use computational Statistics for comparative analysis. The reconstruction of ancestral bat echolocation calls is an important part of understanding bats natural history. 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. Further interpretation of these results will be required to deduce the full implication of this model on our understanding of 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.[1,2] 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.

Bats (order *Chiroptera*) have been identified as ideal bioindicators for monitoring climate change and habitat quality,[5] and are of particular interest for monitoring biodiversity acoustically. Typically, a bat broadcasts information about itself in an ultrasonic echolocation call.[6] The development of automatic acoustic monitoring algorithms[3,7] means that large scale,

*Author footnote.
†Affiliation footnote.

non-invasive monitoring of bats is becoming possible.

Monitoring bat populations provides useful information, but an understanding the orders natural history is required to identify the cause and effect of any changes observed. The echolocation call structure, which reflects a bats diet and habitat,[8] is a key aspect of this natural history. Reconstructing ancestral traits[9] relies on a statistical comparative analysis incorporating extant species and fossil records.[10] However, the fossil record is of limited use in inferring ancestral bats echolocation calls. Therefore, statistical data science techniques may shed some light on this topic.

Previous studies of bat echolocation calls for both classification[7] and ancestral reconstruction[11] analysed features extracted from the call spectrogram. These call features relied domain knowledge to ensure they were sensibly selected and applied. More recently, general techniques for the classification of acoustic signals have been developed.[124] General techniques for the ancestral reconstruction of function-valued traits have also been proposed.[13] This piece of research applies some proposed techniques to bat echolocation calls.

A function-valued trait[14] is a characteristic of some organism measured along some continuous scale, usually time, and can be modelled as a continuous mathematical function by functional data analysis.[15] Jones & Moriarty[16] developed a method which extends Gaussian Process Regression[17] to model the evolution of function-valued traits over a phylogeny. A full demonstration of the phylogenetic Gaussian Process method for ancestral reconstruction on a synthetic dataset is presented by Hajipantelis et al.[18]

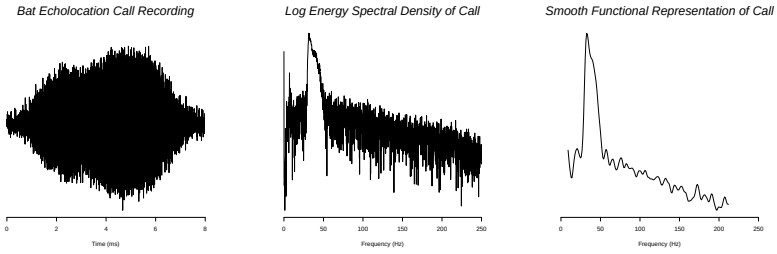
This general approach to evolutionary inference for function-valued traits is implemented here for a set of bat echolocation calls. Our goal in doing so is twofold. These techniques had previously been considered in the context of modelling the evolution of human speech sounds in language.[13] It is hoped that by applying these methods in a simpler context 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 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. 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.[15] 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.[14]



(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 (c).

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 $\{\tilde{x}^{\mathbf{t}}_{lm}(n) : n = 0, \ldots, N^{\mathbf{t}}_{lm} - 1\}$. Thus, $\{x^{\mathbf{t}}_{lm}(\cdot)\}$ is a series of discrete measurements of the function $x^{\mathbf{t}}_{lm}(\cdot)$, the m^{th} call function for the l^{th} individual bat, 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, where $\mathbf{E}[z^{\mathbf{t}}_{lm}(\cdot)] = 0$. Further to this, 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, with $\mathbf{E}[z^{\mathbf{t}}_{l}(\cdot)] = 0$. It is the echolocation call functions at the species level that we re interested in.

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.[19] 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 magnitude of the Fourier transform and the log energy spectral density is given by

$$\mathcal{E}^{\mathbf{t}}_{lm}(\cdot) = 10 \log_{10} \left(|X^{\mathbf{t}}_{lm}(\cdot)|^2 \right). \quad (1)$$

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 ae 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 $F = [0, \frac{f_S}{2}]$.[20]

The discrete Fourier Transform[20] of $\{\tilde{x}^{\mathbf{t}}_{lm}(n)\}$ 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[21] 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 is a collection of random variables, any finite number of which have a joint Gaussian distribution. A Gaussian process prior can be defined as a 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 kernel $k(x, x') = cov(f(x), f(x'))$. Any collection of function values has a joint Gaussian distribution $[f(x_1), f(x_2), \ldots f(x_N)]^{\mathsf{T}} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$ where the $N \times N$ covariance matrix $\mathbf{K}$ has the entries $K_{ij} = k(x_i, x_j)$, and the mean $\boldsymbol{\mu}$ has entries $\mu_i = m(x_i)$. Thus, the properties of the functions are determined by the kernel function. Assuming that the function values $\mathbf{y}$ observed at locations $\{x_n\}_{n=1}^N$ are subject to Gaussian noise, and given the kernel hyperparameters θ, a posterior predictive distribution over Gaussian process functions can be inferred analytically. A marginal likelihood $p(\mathbf{y}|\theta, \{x_n\}_{n=1}^N)$, which can in turn be used to estimate the kernel

hyperparameters, can also be derived analytically. An in depth treatment of Gaussian processes is given by Rasmussen & Williams.[17]

Jones & Moriarty[16] extend Gaussian processes for inference on 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 Gaussian process 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.

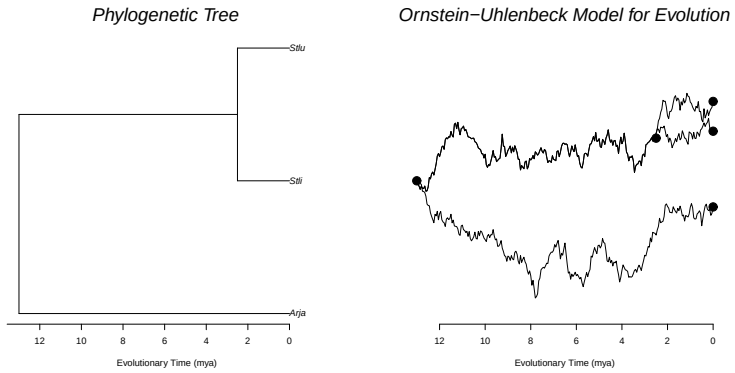


Fig. 2. An example of a Phylogenetic Tree and a simulated univariate phylogenetic Gaussian process of that tree.

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

Under these conditions the phylogenetic covariance function is also separable and so

$$\Sigma_{\mathbf{T}}((f,\mathbf{t}),(f',\mathbf{t}')) = K(f,f')k_{\mathbf{T}}(\mathbf{t},\mathbf{t}'). \quad (2)$$

For a phylogenetic Gaussian Process Y with covariance function given by 2, when K is a degenerate Mercer kernel, 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. The full phylogenetic covariance function of this phylogenetic Gaussian process 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 $\mathbf{T}$.

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

$$\mathcal{E} = X_{\mathbf{T}}\phi^{\mathsf{T}}_F. \quad (3)$$

The problem of obtaining estimators $\hat{\phi}_F$ and $\hat{X}_{\mathbf{T}}$ is dealt with by Hajipantelis et al.[18] $\hat{\phi}_F$ is obtained by Independent Components Analysis, as described by Blaschke & Wiscott[22] after using a resampling procedure to obtain stable principal components for samples of traits balanced by species. Given $\hat{\phi}_F$, the estimated matrix of mixing coefficients is $\hat{X}_{\mathbf{T}} = \mathcal{E}(\hat{\phi}^{\mathsf{T}}_F)^{-1}$.

Each column of $X_{\mathbf{T}}$ is an independent, univariate, phylogenetic Gaussian process, $X^{\mathbf{T}}_i(\sqcup)$, modelled here with phylogenetic Ornsetin-Uhlenbeck process kernel.

The phylogenetic Ornstein-Uhlenbeck process is defined by the kernel

$$k^i_{\mathbf{T}}(\mathbf{t},\mathbf{t}') = (\sigma^i_p)^2 \exp\left(\frac{-d_{\mathbf{T}}(\mathbf{t},\mathbf{t}')}{\ell^i}\right) + (\sigma^i_n)^2\delta_{\mathbf{t},\mathbf{t}'} \quad (4)$$

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

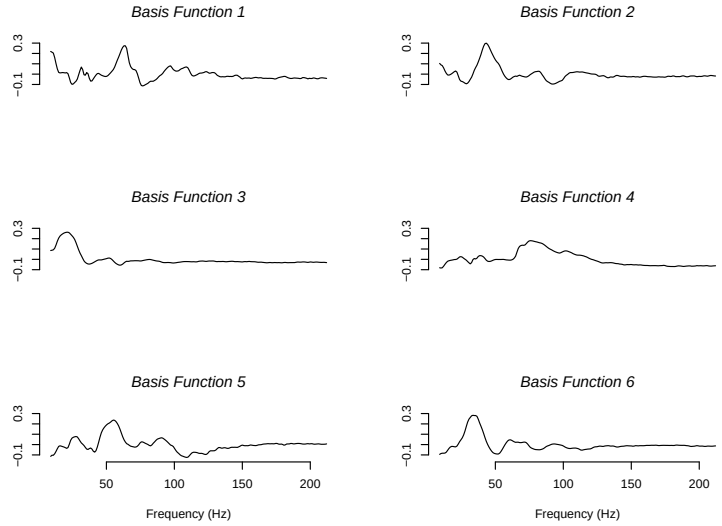


Fig. 3. Set of basis functions identified

$X_i^{\mathbf{T}}(\cdot)$. 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

$$\log(p(\mathcal{E}|\theta)) \propto -\frac{1}{2}\sum_{i=1}^{n}\left(X_i(\sqcup)^{\mathsf{T}}\Sigma^i_{\mathbf{T}}(\sqcup,\sqcup)^{-1}X_i(\sqcup) + \log|\Sigma^i_{\mathbf{T}}(\sqcup,\sqcup)|\right) \quad (5)$$

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

$$A = \sum_{i=1}^{n}\left(\Sigma^i_{\mathbf{T}}(\mathbf{t}^*,\sqcup)\left(\Sigma^i_{\mathbf{T}}(\sqcup,\sqcup)\right)^{-1}X_i^{\mathcal{E}}(\sqcup)\phi_i(\cdot)\right) \quad (6)$$

$$B = \sum_{i=1}^{n}\left(\Sigma^i_{\mathbf{T}}(\mathbf{t}^*,\mathbf{t}^*) - \Sigma^i_{\mathbf{T}}(\mathbf{t}^*,\sqcup)\left(\Sigma^i_{\mathbf{T}}(\sqcup,\sqcup)\right)^{-1}\Sigma^i_{\mathbf{T}}(\mathbf{t}^*,\sqcup)^{\mathsf{T}}\right)\phi_i(\cdot) \quad (7)$$

We note that the elements of θ each 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 variation which can be accounted for by the cophenetic distance between any $\mathbf{t},\mathbf{t}' \in \mathbf{T}$, while σ_n accounts for other sources of variation which may be

observed at $\mathbf{t}$. The length-scale parameter, ℓ then indicates the strength of the correlation along $\mathbf{T}$, where large values of ℓ indicate a slowly decaying correlation.

4. Results

4.1. *Data Description*

The post processed echolocation call data accompanying Stathopoulos et al.[3] was used in this analysis. Echolocation calls were recorded across north and central Mexico with a Pettersson 1000x bat detector (Pettersson Elektronik AB, Uppsala, Sweden). Live trapped bats were measured and identified to species level using field keys.[23,24] Bats were recorded either while released from the hand or while tied to a zip line. The bat detector was set to record calls manually in real time, full spectrum at 500 kHz. In total the dataset consists of 22 species in 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[11] Bat super-tree provided the basis for 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 the echolocation call of the species of bat observed at $\mathbf{t} \in \mathbf{T}$. 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 (4). A short simulation study was performed to investigate the dynamics at play.

Although there are only 22 leaf nodes of the phylogeny $\mathbf{T}$, we are not limited to a single observation at any given $\mathbf{t}$. By sampling at each leaf node of $\mathbf{T}$ multiple times, larger samples can be obtained, improving the quality of the estimators $\hat{\theta}$. With this in mind, 1000 independent, univariate phylogenetic Gaussian processes were simulated for each of $n = \{1, 2, 4, 8\}$ according to the kernel (4) 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 (5) 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

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

stable estimators $\hat{\theta}$. Thus, when implementing the phylogenetic Ornstein-Uhlenbeck process for bat echolocation call traits, resampling methods must be used to provide multiple estimates for $\mathcal{E}^{\mathbf{t}}(\cdot)$

Given the modelling assumptions made in Section 2 the best estimator for $\mathcal{E}^{\mathbf{t}}(\cdot)$ is the sample mean given by

$$\hat{\mathcal{E}}^{\mathbf{t}}(\cdot) = \frac{1}{l_{\mathbf{t}}} \sum_{l=1}^{l_{\mathbf{t}}} \frac{1}{m_l} \sum_{m=1}^{m_l} \hat{\mathcal{E}}^{\mathbf{t}}_{lm}(\cdot) \tag{8}$$

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. Summary of $\hat{\theta}$ for 1000 simulations of independent Ornstein-Uhlenbeck 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)

simply calculating $\hat{\mathcal{E}}^{\mathbf{t}}(\cdot)$ for each species performing the analysis laid out in Section 3 means that we have too few datapoints to obtain a stable model. For this reason we implement a resampling procedure to leverage more datapoints from the dataset in order to produce stable estimates for the model parameters.

A resampled estimator $\hat{\mathcal{E}}_r^{\mathbf{t}}(\cdot)$ 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 (8). 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 $\hat{\mathcal{E}}_r = [\hat{\mathcal{E}}_{r,1}^{\mathbf{t}_1}(\mathbf{f}), \hat{\mathcal{E}}_{r,2}^{\mathbf{t}_1}(\mathbf{f}), \ldots, \hat{\mathcal{E}}_{r,1}^{\mathbf{t}_2}(\mathbf{f}), \ldots]$, where $\mathbf{f}$ is the vector of frequencies over which $\hat{\mathcal{E}}_r^{\mathbf{t}_2}(\cdot)$ is sampled. The methods outlined in Section 3 can then be applied to each resampled $\hat{\mathcal{E}}_r$

Our analysis took $n_r = 4$ and included 4 samples of $\hat{\mathcal{E}}_r^{\mathbf{t}}(\mathbf{f})$ in each $\hat{\mathcal{E}}_r$. 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 averaging the basis identified over all $\hat{\mathcal{E}}_r$ and applying a single set of basis functions to each dataset. Thus $\hat{X}_r$, the matrix of mixing coefficients described by (3), the columns of which are modelled a phylogenetic Ornstein-Uhlenbeck processes, is obtained for each $\hat{\mathcal{E}}_r$. $\hat{\theta}_r$ is then the type II maximum likelihood estimator of (5) given $\hat{\mathcal{E}}_r$. Table 3 presents the results of the hyperparameter estimation procedure.

Ancestral reconstruction by a phylogenetic Gaussian process involves obtaining the posterior predictive distribution of the trait at the ancestral node $\mathbf{t}^* \in \mathbf{T}$ given by (6) and (7).

To perform ancestral trait reconstruction for $\mathcal{E}^{\mathbf{t}^*}(\cdot)$ the species level

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)

traits are estimated by (8) and and the model hyperparameters by the mean values of θ_r reported in Table 3.

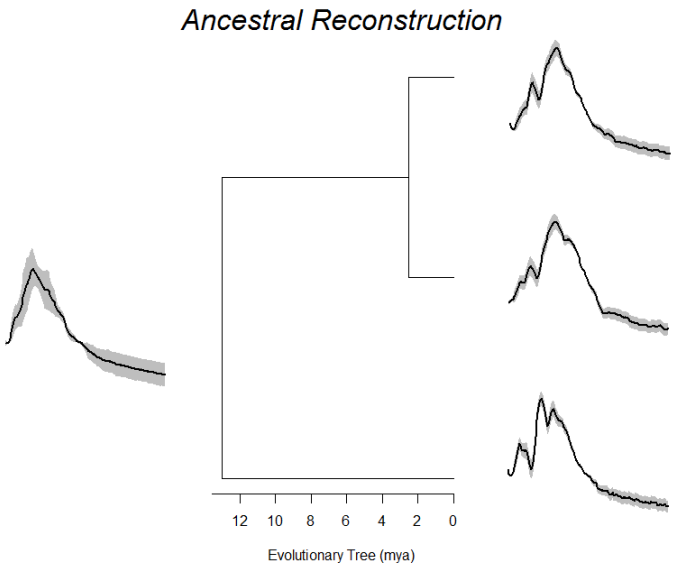


Fig. 4. Ancestral Reconstruction

5. Conclusions and Further Work

This analysis has developed a limited model for the evolution of echolocation in bats and identified a phylogenetic signal which allows the con-

struction 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 evolutionary path 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 attributable solely to the evolutionary path while ℓ provides a measure of how quickly correlation between observations along the evolutionary path decays. We are working towards understanding what the results of this analysis mean for 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 representation, which preserves more of this structure, is the spectrogram. Implementing this model using the spectrogram as a functional representation of the echolocation call is a priority, as the results of such an analysis will hopefully be more meaningful.

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. D. R. Griffin, Echolocation by blind men, bats and radar, *Science.* **100**(2609),

589–590 (1944).
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. B. Joy, R. H. Liang, R. M. McCloskey, T. Nguyen, and A. F. Poon, Ancestral reconstruction, *PLoS Comput Biol*. **12**(7), e1004763 (2016).
10. J. Felsenstein and J. Felenstein, *Inferring phylogenies*. vol. 2, Sinauer associates Sunderland (2004).
11. A. Collen. *The evolution of echolocation in bats: a comparative approach*. PhD thesis, UCL (University College London) (2012).
12. 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).
13. T. F. P. Group, Phylogenetic inference for function-valued traits: speech sound evolution, *Trends in ecology & evolution*. **27**(3), 160–166 (2012).
14. 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).
15. J. O. Ramsay, *Functional data analysis*. Wiley Online Library (2006).
16. 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).
17. C. E. Rasmussen, Gaussian processes for machine learning (2006).
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. 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).
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. 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).
23. G. Ceballos and G. Oliva. Los mamíferos silvestres de méxico. comisión nacional para el conocimiento y uso de la biodiversidad (2005).
24. R. Medellín and H. Arita, O. sánchez h. 2008. identificación de los murciélagos de méxico. claves de campo, *Revista Mexicana de Mastozoología*. **2**, 1–83 .