Chapter 1

Ancestral Reconstruction of Bat Echolocation Calls

J.P. Meagher*

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

The stated aim of the Statistical Data Science workshop jointly organised by the Department of Mathematics and Data Science Institute at Imperial College London, and Winton Global investment management, is “exploring the nature of the relationship between statistics and data science”.

1. Introduction

1.1. *Motivation*

Data science, the “ study of the generalisable extraction of knowledge from data”[1] is motivated by availability of interesting datasets. One field producing such datasets is Bioacoustics, where data is collected, often through citizen science[2] initiatives, for monitoring and conservation purposes.[3] Bats (order *Chiroptera*) have been of particular interest.[4]

Bats form the second most speciose order of mammals, behind rodents, with over 1200 species.[5] Bats have been identified as ideal bioindicators for monitoring climate change and habitat quality.[6] Good bioindicators are easy to identify and sample, well distributed geographically, respond to changes in habitat in a manner correlated with other taxa, and have a well understood natural history.[7] Bats have the potential to satisfy these characteristics excellently.

Bats are, for the most part, nocturnal, flying echolocators.[8] Echolocation refers to the bats use of, typically ultrasonic,[9] calls and their echoes to

*Author footnote.
†Affiliation footnote.

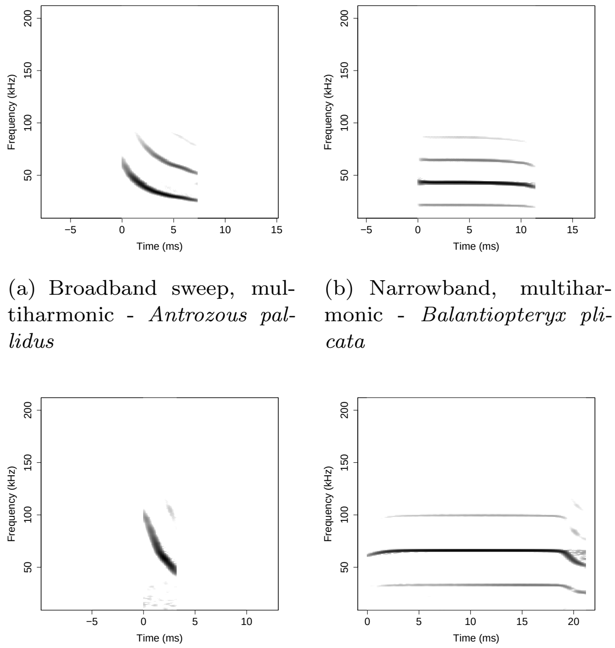
forage and navigate in the night sky.[10] Thus, bats leak information about themselves into the environment, allowing acoustic monitoring.[4] Work towards the development of automatic acoustic monitoring algorithms for bats[11][12] is ongoing. Similar algorithms have been developed for insects[13] and birds,[14] reflecting the level of interest in bioacoustic monitoring.

The growing database of bat call recordings[15] also opens up other avenues for research. Bat echolocation calls are diverse and can generally be sorted into categories according to the duration, bandwidth and use of harmonics in the call.[16] It has been observed that closely related species have similar call structures indicating that some of the variation is due to a shared evolutionary history.[17] Comparative analysis of bat echolocation call parameters for ancestral reconstruction[18] have been performed.[15] However, an approach based Statistical Data Science principles may shed further light on the evolution of echolocation in bats, without relying heavily on domain specific knowledge.

1.2. *Literature review*

Echolocation, or biosonar, is a process whereby sound is produced and the echoes that return from objects are used to perceive the environment.[19] Animals that use echolocation include bats, toothed whales, some birds, and some shrews and rats.[19] Bat echolocation calls, ranging from 9 to 212 kHz, differ from those of other echolocators in that they are often laryngeal, complex, and of a relatively long duration.[20] The evolutionary dynamics that resulted in bats being nocturnal, flying echolocators have resulted in a wide variety of call structures across species.[17] Echolocation in bats demonstrates convergent evolution, and bat species have been arranged in 'guilds' based on habitat and foraging strategy, with similar call structures observed within a 'guild'.[21][22][23] However, it has been acknowledged that "echolocation signals reflect a phylogenetically determined basic call structure shaped by specific ecological conditions."[24] Ancestral Reconstruction is a key aspect of understanding the variation is call structure that is due to phylogenetic relationships between species.

Ancestral Reconstruction involves the extrapolation back in time from measured characteristics of current populations to the ancestral state, the estimate of the same characteristic in common ancestors.[18] Effective ancestral reconstruction relies on an accurate model for evolution, that is, both the dynamic evolutionary process and the phylogenetic relationships between species.[18] There have been maximum parsimony,[25] maximum like-



(a) Broadband sweep, multiharmonic - *Antrozous pallidus*

(b) Narrowband, multiharmonic - *Balantiopteryx plicata*

(c) Broadband sweep, single harmonic - *Myotis yumanensis*

(d) Constant frequency with broadband sweep - *Pteronotus parnellii*

Fig. 1. Representative call spectrograms for some bat species, demonstrating echolocation call structures.

lihood,[26] and Bayesian[27] approaches taken to ancestral reconstruction.

Other studies have considered bat echolocation calls for ancestral reconstruction.[2815] Fenton[28] hypothesised that early bats used short, broadband clicks for echolocation based on the observation that this is the method of echolocating employed by all animals except bats. Collen[15] proposed instead that early bats used short, multi-harmonic, narrowband laryngeal calls for echolocation, based on a comparative analysis of echolocation call parameters. An approach to the reconstruction of early bat echolocation calls which is not so selective in the information it employs may shed further light on this topic. Gaussian Process Regression on Phylogenies for function valued traits provide a promising path to explore.[29] Jones & Moriarty proposed an extension to Gaussian Process Regression[30] which allowed the generalisation of Ornstein-Uhlenbeck[31] models of continuous-time character evolution for traits considered as functional data objects,[32] through a phylogeny. This study represents an effort to apply these methods to the ancestral reconstruction of bat echolocation calls.

1.3. *The evolution of function valued traits*

Bat echolocation calls can be described as 'function valued' traits.[33] In this context, 'function valued' refers to traits measured along a continuous scale, which can then be represented by a continuous mathematical function. Data of this nature can be viewed as functional data objects.[32] The Functional Phylogenies Group[34] argue the case for performing evolutionary inference, including ancestral reconstruction, on function valued traits, with a particular focus on the evolution of linguistic speech sounds. They propose performing inference on a functional representation of the speech sound itself, by extending Gaussian Process Regression[30] to take advantage of the tree structure of phylogenetic relationships.[29]

Phylogenetic Gaussian Process Regression (PGPR), as developed by Jones & Moriarty,[29] extends Ornstein-Uhlenbeck[31] models for continuous-time character evolution to function valued traits. In comparative studies of genetic data the O-U model is used to describe stabilising selection with random genetic drift, the notion that traits have some optimal value, around which they vary.[35] The PGPR is based on a kernel with 3 hyperparameters. Phylogenetic noise and length scale hyperparameters govern the correlation between nodes on a tree due to the phylogeny, while non-phylogenetic noise accounts for other sources of variation. Estimating these hyperparameters allows posterior distributions for traits at internal nodes to be constructed. In this way estimates for the ancestral state can be made. This model does not require that variance of the trait is homogenous across the whole function space. By defining a set of deterministic basis functions, combinations of which can be used to describe the observed trait functions, the variance of traits can be modelled to vary over the function space. An implementation of this method on simulated data is given by Hajipantelis.[36]

1.4. *Paper Description*

In a similar manner to the representation of speech sounds as spectrograms,[37] bat echolocation calls will be represented by their Energy Spectral Density[38] (ESD). The ESD, the energy of a signal as a function of a frequency, provides a straightforward characterisation from which a PGPR can be implemented.

Using a synthetic dataset based on the bat phylogeny the conditions for an effective implementation of PGPR will be demonstrated.

PGPR will then be implemented for the set of bat calls and posterior distributions of ESD estimated.

Spectral representations of the calls will be described, the implementation of PGPR. Independent Component Analysis. Simulations study to demonstrate the PGPR can work for this dataset. Application to Bats.

2. Methods

2.1. *Data Description*

The post processed echolocation call data accompanying Stathopoulos et al.[11] 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.[3940] 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.

Phylogenetic relationships between species are described and dated according to the Bat super-tree produced by Collen.[15]

2.2. *From Echolocation Call recordings to Spectral Curves*

Echolocation calls can be considered to be continuous, non-periodic, pulse signals which are a function of time $x(t)$.

The Fourier Transform of a signal is an expression of the signal in the frequency domain, that is, as a function of frequency. It is given by

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-2\pi ft}dt.$$

The energy of the signal is given by

$$E = \int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} |X(f)|^2 df,$$

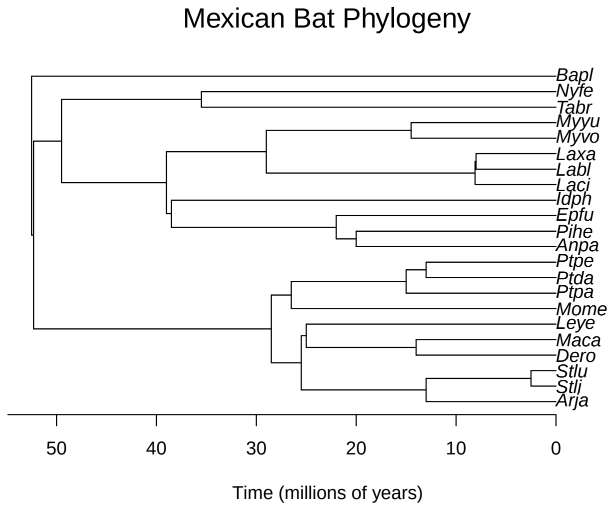
by Parseval's Theorem.

Table 1. Echolocation Call Dataset Statistics

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

The ESD is then $|X(f)|^2$. This is a representation of a sound and so it is appropriate to scale this to a decibel scale. Not also that $|X(f)|^2$ is a periodic function and symmetric around 0. Given that bat echolocation calls range from $9 - 212$kHz we are interested in $S(f) = log_{10}|X(f)|^2, f \in [9, 212]$. This can then be treated as a functional data object.

By treating $S(f)$ as a functional data object it is treated as a set of noisy realisations taken of a smooth underlying curve. Estimating this underlying curve is performed by smoothing. Another aspect of functional data analysis is registration, in this case it would consist of separating out phase and amplitude variation. As the energy at specific frequencies is



q

of interest here, registration along the frequency axis is not appropriate. However the level of the amplitude depends on factors such as the distance of the bat from the microphone while being recorded. This variation is not interesting. Thus registration on the amplitude axis is appropriate. thus all curves are registered such that the amplitude lies on the interval $[0, 1]$.

Once smoothed and registered the curves are sampled from at every kHz. These Spectral curves are the functions to be analysed in the comparative analysis.

2.3. *Phylogenetic Gaussian Processes*

Given a dataset, X, of N observations of an input vector $\mathbf{x}$ associated with a vector of response variables $\mathbf{y}$, a Gaussian Process model assumes that the relationship between $\mathbf{x}$ and y is governed by a multivariate Gaussian distribution. Formally, a Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution.[30]

In a function-space view of Gaussian Processes, a Gaussian Process $f(\cdot)$ is completely characterised by its mean and covariance functions. For simplicity, the mean function can be set such that $m(\mathbf{x}) = \mathbf{E}[f(\mathbf{x})] = 0$. The covariance function, defined by a kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$, describes the relationship between coordinates within the input space. Thus, setting $y_i = f(\mathbf{x}_i)$, models the vector of response variables as a multivariate Gaussian distribution, $\mathbf{y} \sim \mathcal{N}(0, K)$, where $K = K(X, X)$ is the covariance matrix given by $K_{ij} = K(X, X)_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

Once a Gaussian Process has been fitted for the observed data, a conditional distribution for unobserved points, X_*, on the input space can be estimated. The key predictive equations for Gaussian Process Regression are

$$\mathbf{f}_*|X, \mathbf{y}, X_* \sim \mathcal{N}(\bar{\mathbf{f}_*}, cov(\mathbf{f}_*))$$

where

$$\bar{\mathbf{f}_*} \equiv \mathbf{E}[\mathbf{f}_*|X, \mathbf{y}, X_*] = K(X_*, X)(K)^{-1}\mathbf{y}$$

and

$$cov(\mathbf{f}_*) = K(X_*, K_*) - K(X_*, X)(K)^{-1}K(X, X_*).$$

Thus the key aspect of Gaussian Process Regression is the definition of an appropriate kernel, which is context dependant.

The Gaussian Process framework provides a flexible general approach to regression problems, a framework which Jones & Moriarty extended to phylogenteic inference for function valued traits. Suppose the phylogenetic tree, $\mathbf{T}$, denotes the relationships between the observed data in evolutionary time. The trait itself exists in some function space $\mathcal{X}$ and so each observation is indexed by $(\mathbf{x}, \mathbf{t})$, where $\mathbf{x}$ describes the observed functional trait, and $\mathbf{t}$ the position of the observation in $\mathbf{T}$, evolutionary time. A Phylogenetic Gaussian process can then be defined by a suitable kernel. The form of this kernel is given by Jones & Moriarty.[29]

The results are presented for a Phylogenetic Gaussian Process subject to a number of simplifying assumptions. It is assumed that conditional on common ancestors in $\mathbf{T}$ any two traits are independent. It is also assumed that marginal process of evolution, that is the path on $\mathbf{T}$ from the root to any tip, is identical along all paths. Finally, it is also assumed that the marginal process is separable by evolutionary time and the function valued trait space.

In order to fit a Phylogenetic Gaussian Process model with spatially inhomogenous variance, a set of basis functions must be defined. Intuitively, these basis functions should define 'components' of the trait, each of which are subject to an independent univariate Phylogenetic Gaussian Process, which can be additively combined to model the trait at various points in evolutionary time. The cubICA[41] implementation of Independent Components Analysis offers a method of estimating such a set of basis functions, referred to as components hereafter.

The implementation here models the Phylogenetic Gaussian process with an Ornstein-Uhlenbeck kernel, a special case of the Matern class of kernels, with additive gaussian noise.

$$k(x,x) = phylogenetic + nonphylogenetic$$

In this way the PGP can account for both phylogenetic and non-phylogenetic aspects of the evolutionary process.

3. Results

3.1. *Simulation Study*

a simulation study examining PGP for a random tree was performed for a random tree with 128 tips.[36] However the phylogenetic tree presented here is very different from that, having only 22 tips. In order to understand how effectively hyperparameters can be estimated for this OU process a simulation study of my own was performed.

Each observation is associated with some tip on the phylogenetic tree, however we are not limited to a single observation at each tip. we are interested in the variation between species and so we can place multiple observations at each tip. Care was taken to ensure the dataset was balanced in some sense and so the same number of observations were taken from each tip.

The object of this simulation study is to demonstrate the conditions required for type II maximum likelihood estimation to successfully estimate hyperparameters for a Phylogenetic Ornstein-Uhlenbeck Process.

The Phylogenetic Ornsetin-Uhlenbeck Process kernel is

$$k(x,x') = \sigma_p \exp\left(\frac{|x-x'|}{\ell}\right) + \sigma_n$$

where σ_p is the phylogenetic noise, ℓ the phylogenetic length-scale, and $sigma_n$ the non-phylogenetic noise.

The key parameter here is ℓ. It can be easily shown that when $\ell >> |x-x'|$ there is a high degree of correlation between process observations at x and x'. When $\ell << |x-x'|$ there is little correlation and the process becomes indistinguishable from white noise. With this in mind fix $\sigma_p = 1$ and $\sigma_n = 1$.

Investigate the accuracy for the process over the tree for $\ell = \{10, 25, 50, 100, 200\}$ considering $n = \{1, 2, 4, 8\}$ observations at each tip of the tree.

Create a dataframe storing the results of 1000 simulations for the vlaues of ℓ, n of interest.

In order to demonstrate the accuracy of type 2 MLE estimation the mean estimatior(sample variance) of each hyperparameter is given below.

Table 2. Echolocation Call Dataset Statistics

			ℓ		
	10	20	50	100	200
$n = 1$					
2					
4					
8					

More observations better than less.

Exponential distribution of length scale.

Some pictures, pairs plot maybe. Compare different sample sizes, distributions. Note plots cannot use colour.

given hypers, what happens along the given tree

3.2. *Mexican Bat Echolocation call Dataset*

What happens when the method is applied to the Mexican bat calls?

4. Discussion

4.1. *What do we learn from this?*

Phylogenetic / Non-phylogenetic Signal. Not just noise. Although speculative and probably limited going back to root may be useful for more recent ancestors.

4.2. *Future Work in this Area*

Implement for Spectrograms, produce some recordings.

References

1. V. Dhar, Data science and prediction, *Communications of the ACM*. **56**(12), 64–73 (2013).

2. C. Kullenberg and D. Kasperowski, What is citizen science?–a scientometric meta-analysis, *PloS one.* **11**(1), e0147152 (2016).
3. P. E. Allen and C. B. Cooper. Citizen science as a tool for biodiversity monitoring. (2006).
4. N. Pettorelli, J. E. Baillie, and S. M. Durant, Indicator bats program: a system for the global acoustic monitoring of bats (2013).
5. N. B. Simmons, Order chiroptera, *Mammal species of the world: a taxonomic and geographic reference.* **1**, 312–529 (2005).
6. 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).
7. C. E. Moreno, G. Sánchez-Rojas, E. Pineda, and F. Escobar, Shortcuts for biodiversity evaluation: a review of terminology and recommendations for the use of target groups, bioindicators and surrogates, *International Journal of Environment and Health.* **1**(1), 71–86 (2007).
8. T. Kunz and E. Pierson, Bats of the world: an introduction, *Walkers Bats of the World.* pp. 1–46 (1994).
9. J. F. Corso, Bone-conduction thresholds for sonic and ultrasonic frequencies, *The Journal of the Acoustical Society of America.* **35**(11), 1738–1743 (1963).
10. D. R. Griffin, Echolocation by blind men, bats and radar, *Science.* **100**(2609), 589–590 (1944).
11. 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).
12. 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).
13. D. Chesmore, Automated bioacoustic identification of species, *Anais da Academia Brasileira de Ciências.* **76**(2), 436–440 (2004).
14. F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley, A. S. Hadley, and M. G. Betts, Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach, *The Journal of the Acoustical Society of America.* **131**(6), 4640–4650 (2012).
15. A. Collen. *The evolution of echolocation in bats: a comparative approach.* PhD thesis, UCL (University College London) (2012).
16. A. Maltby, K. E. Jones, and G. Jones, .4-understanding the evolutionary origin and diversification of bat echolocation calls, *Handbook of Behavioral Neuroscience.* **19**, 37–47 (2010).
17. G. Jones and E. C. Teeling, The evolution of echolocation in bats, *Trends in Ecology & Evolution.* **21**(3), 149–156 (2006).
18. J. B. Joy, R. H. Liang, R. M. McCloskey, T. Nguyen, and A. F. Poon, Ancestral reconstruction, *PLoS Comput Biol.* **12**(7), e1004763 (2016).
19. G. Jones, Echolocation, *Current Biology.* **15**(13), R484–R488 (2005).
20. J. A. Thomas, *Echolocation in bats and dolphins.* University of Chicago Press

(2004).
21. H. Aldridge and I. Rautenbach, Morphology, echolocation and resource partitioning in insectivorous bats, *The Journal of Animal Ecology*. pp. 763–778 (1987).
22. G. Neuweiler, Auditory adaptations for prey capture in echolocating bats, *Physiological reviews*. **70**(3), 615–641 (1990).
23. H.-U. Schnitzler and E. K. Kalko, Echolocation by insect-eating bats: We define four distinct functional groups of bats and find differences in signal structure that correlate with the typical echolocation tasks faced by each group, *Bioscience*. **51**(7), 557–569 (2001).
24. H. Schnitzler, E. Kalko, and A. Denzinger, Evolution of echolocation and foraging behavior in bats, *Echolocation in bats and dolphins*. pp. 331–339 (2004).
25. W. M. Fitch, Toward defining the course of evolution: minimum change for a specific tree topology, *Systematic Biology*. **20**(4), 406–416 (1971).
26. T. Pupko, I. Pe, R. Shamir, and D. Graur, A fast algorithm for joint reconstruction of ancestral amino acid sequences, *Molecular Biology and Evolution*. **17**(6), 890–896 (2000).
27. M. Pagel, A. Meade, and D. Barker, Bayesian estimation of ancestral character states on phylogenies, *Systematic biology*. **53**(5), 673–684 (2004).
28. M. Fenton, D. Audet, M. K. Obrist, and J. Rydell, Signal strength, timing, and self-deafening: the evolution of echolocation in bats, *Paleobiology*. **21** (02), 229–242 (1995).
29. 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).
30. C. E. Rasmussen, Gaussian processes for machine learning (2006).
31. G. E. Uhlenbeck and L. S. Ornstein, On the theory of the brownian motion, *Physical review*. **36**(5), 823 (1930).
32. J. O. Ramsay, *Functional data analysis*. Wiley Online Library (2006).
33. 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).
34. T. F. P. Group, Phylogenetic inference for function-valued traits: speech sound evolution, *Trends in ecology & evolution*. **27**(3), 160–166 (2012).
35. T. F. Hansen, Stabilizing selection and the comparative analysis of adaptation, *Evolution*. pp. 1341–1351 (1997).
36. 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).
37. D. Pigoli, P. Z. Hadjipantelis, J. S. Coleman, and J. A. Aston, The analysis of acoustic phonetic data: exploring differences in the spoken romance languages, *arXiv preprint arXiv:1507.07587* (2015).
38. A. Antoniou, *Digital signal processing*. McGraw-Hill Toronto, Canada: (2006).
39. G. Ceballos and G. Oliva, Los mamíferos silvestres de méxico, comisión na-

cional para el conocimiento y uso de la biodiversidad (2005).

40. 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 .
41. 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).