Presentation of Ancestral Reconstruction Results

J.P. Meagher

January 12, 2018

Introduction

We are given a dataset of N bat echolocation call recordings denoted $\{y_n\}_{n=1}^N$. This recording is then processed to produce a set of smooth surfaces over a regular grid denoted $\{S_n\}_{n=1}^N$. This surface, referred to as the call surface hereafter, is produced by smoothing the call spectrogram and mapping it to a regular grid over relevant frequencies and an absolute time scale in the manner as outlined by Pigoli *et al.* [14]. A robust N-D spline smoothing algorithm presented by Garcia [5] was implemented for surface smoothing and a pairwise curve synchronisation algorithm by Tang & Muller [17] was extended to surfaces for time registration.

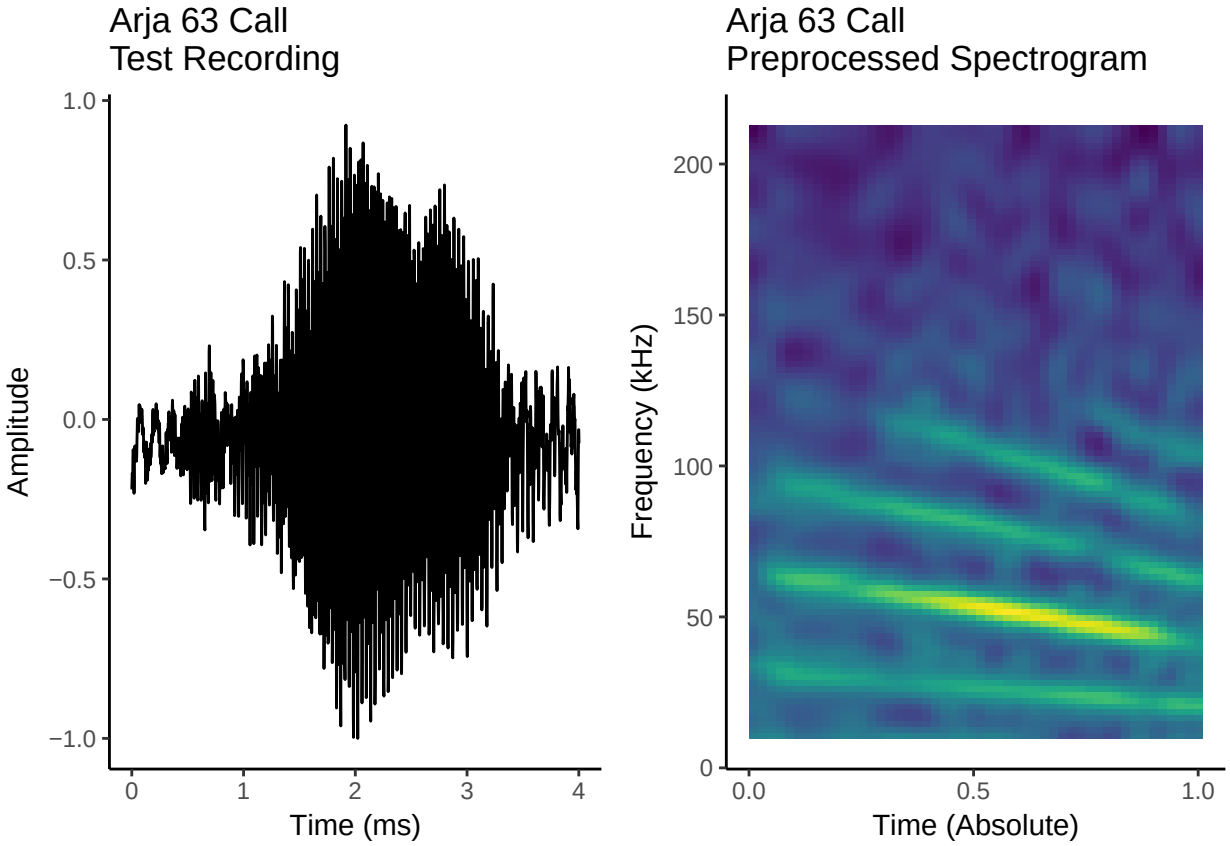

Figure 1: A randomly selected bat call from the species Arja alongside it's corresponding call surface. The call surface is obtained by taking the call spectrogram and treating it as a functional data object. The spectrogram is first smoothed by a robust 2-D spline smoother, then mapped to an absolute time scale and registered in time by a pairwise surface synchronisation, and finally restricted to the 9 - 212 kHz frequency spectrum.

Along with this dataset we are given a phylogeny defining the evolutionary relationships between the species

of bat. The phylogeny was transcribed from a Bat super-tree published by Collen [1].

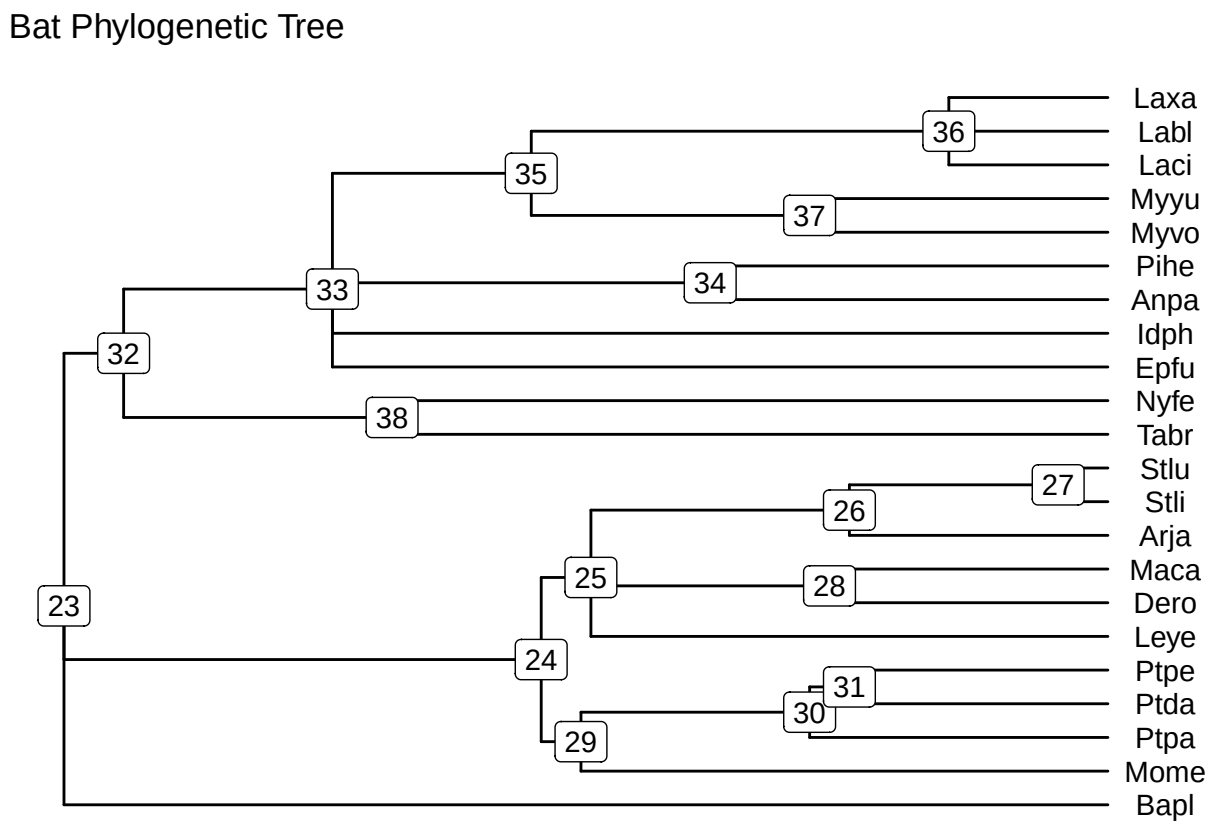

Figure 2: Tree of assumed evolutionary relationships between Bat Species. This phylogeny was transcribed from a recent bat super tree and should represent a 'best guess' for the evolutionary relationships between bat species based on the fossil record alongside morphological and molecular studies of evolutionary relationships.

Implementing the model for evolutionary inference on function valued traits proposed by Jones & Moriarty [8] and further investigated by Hajipantelis [6] allows for reconstruction of ancestral bat call surfaces. By then adapting the RTISI-LA algorithm for spectrogram inversion [19] to these call surfaces, an acoustic signal approximating the corresponding echolocation call is produced.

Thus, for a dataset of Mexican Bat echolocation calls and the given phylogeny, reconstruction of ancestral bat echolocation calls has been performed.

The Current Model

An illustration of the current iteration of a model for the evolution of bat echolocation calls is presented in Figure 4.

In the graphical models presented, a ○ node represents a random variable, a ◉ represents a random variable for which there are observations, and a • represents a deterministic parameter in the model. The plates ▭ indicate how many observations / random variables / parameters of each kind there are.

Figure 4 illustrates the implementation of the model described in the above introduction, from which reconstructions of ancestral echolocation call surfaces can be produced. This representation describes the

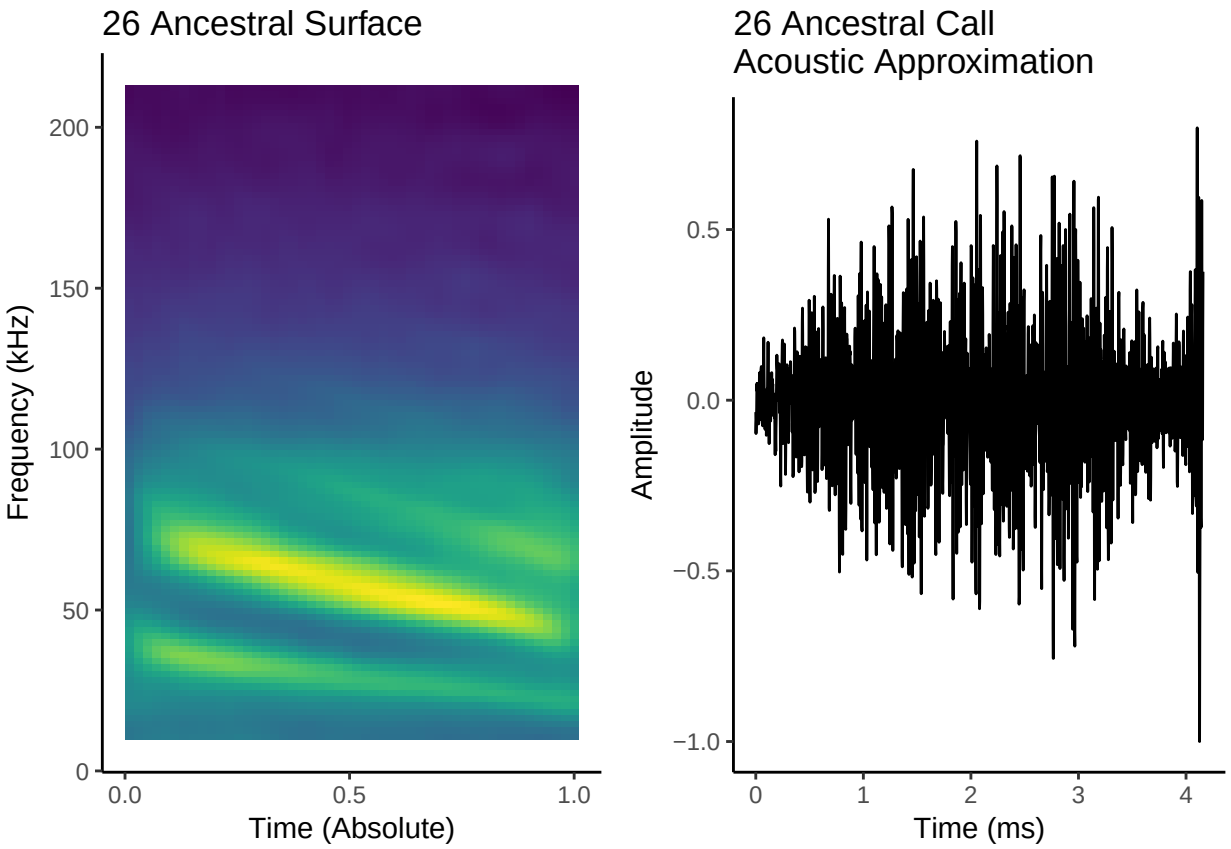

Figure 3: Ancestral call surface and acoustic approximation for the common ancestor of Arja, Stli, and Stlu, which corresponds to node 26 in Figure 2. The reconstruction is given by the MAP estimates for the weight of each evolutionary feature at the node. In this case, evolutionary features were identified by a PCA of the smoothed spectrogram surfaces. The acoustic reconstruction was performed assuming a call duration of approximately 4 ms

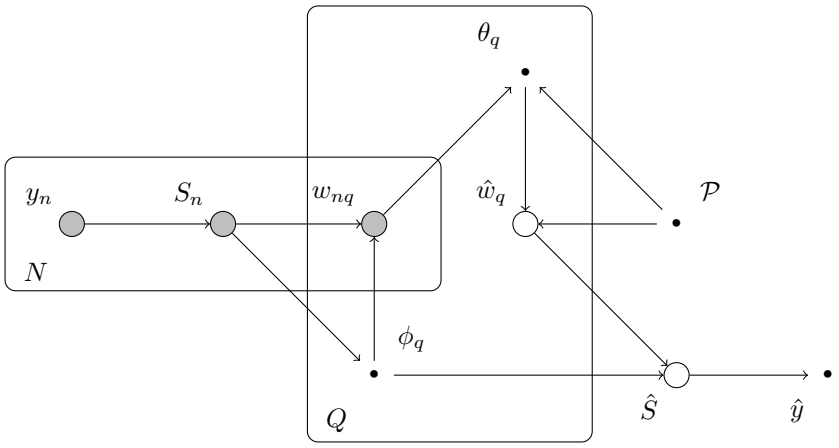

Figure 4: A Graphical model detailing the structure of the model for evolution used to produce reconstructions of ancestral bat echolocation calls. Let y_n be a random variable representing an echolocation call recording. S_n is the random variable representing the smoothed spectrogram surface given by y_n. The Mexican bat call dataset provides $N = 1816$ observations of these random variables. The process of transforming a call recording into a spectrogram surface was covered in my 9 month report. The model assumes that each S_n can be modelled by Q independent deterministic 'evolutionary features' denoted ϕ_q. In this case ϕ_q is inferred by a Principal Components Analysis of $\{S_n\}_{n=1}^N$. The weight of each evolutionary feature in S_n is itself a random variable, where w_{nq} denotes the weight of ϕ_q in S_n. w_{nq} is assumed to behave as an Ornstein-Uhlenbeck Gaussian process for which the input space is the phylogeny $\mathcal{P}$. Each Gaussian process is defined by the deterministic hyperparameters $\theta_q = [\gamma_q, \ell_q, \sigma_q]^\intercal$ which are inferred from the data by Type II maximum likelihood estimation over the observed weights. The phylogeny $\mathcal{P}$ is also assumed to be deterministic in this model and is shown in Figure 2. Ancestral reconstruction is performed by making a prediction for the feature weights, denoted $\hat{w}$, at some point on $\mathcal{P}$. Applying these weights to the evolutionary features produces the ancestral call surface which in turn provides an estimate for the ancestral call.

model for call spectral density curves presented at the Statistical Data Science conference, the conference paper presented alongside this progress update [12], and the current iteration of the model where call spectral density curves have been replaced by call surfaces.

Consider the model than presented in the introduction above with reference to Figure 4. In order to transform the data $\{y_n\}_{n=1}^N$ into $\{S_n\}_{n=1}^N$ some modelling decisions must be made. To find the spectrogram for each call an appropriate window and step size must be chosen. This will affect the time and frequency resolution of the spectrogram. In this case, window size was set to 512 and the step size to 8. The spectrogram is then restricted to the 9 - 212 kHz frequency spectrum as this is the frequency band used by bats for echolocation. The next step involves applying Garcia's robust weighted spline smoothing algorithm [5] to the spectrograms. The algorithm was weighted to ensure that those areas on the spectrogram with the greatest magnitude, corresponding to call harmonics, were not smoothed away in order to reduce noise in unimportant parts of the spectrogram. After smoothing, spectrograms are mapped to an absolute time scale and the grid over which the surface is observed is regularised. In this case, a 50×104 grid over $T \times F$ was used, where $T \in [0, 1]$ and $F \in [9, 212]$. With this smoothing and regularisation done, time registration of the call surfaces was performed by adapting pairwise curve synchronisation [17] to surfaces using a dynamic time warping kernel [2]. In this way $\{y_n\}_{n=1}^N$ was transformed into a new set of observations, $\{S_n\}_{n=1}^N$. Further detail on the preprocessing steps described here was included in my 9 month report and can also be found in [14].

The set of call surfaces $\{S_n\}_{n=1}^N$ can be considered to be a set of function-valued traits [13] each observed at a point $\mathbf{p}$ in the phylogeny $\mathcal{P}$ shown in Figure 2, which is assumed to be a deterministic description of the evolutionary relationships between bat species. Thus $\mathbf{p} \in \mathcal{P}$ can be thought of as a particular species of bat. Jones & Moriarty [8] present results which extend Gaussian processes [15] for evolutionary inference over phylogenies, referred to as phylogenetic Gaussian processes (PGP). Details on PGPs were included in both my 15 month report and in the accompanying conference paper [12]. The implementation of the PGP for echolocation call surfaces follows the simulation study performed by Hajipantelis [6] closely. Firstly, 'evolutionary features' must be defined. These can be thought of as deterministic components, a linear combination of which specifies the call surface exactly. Here, a set of Q evolutionary features were defined by a Principal Components Analysis of $\{S_n\}_{n=1}^N$, denoted $\{\phi_q\}_{q=1}^Q$. For this analysis $Q = 15$. Secondly, once the set of evolutionary features has been defined, the weight of the q^{th} component in the n^{th} call surface, w_{nq}, can be calculated. Finally, it is assumed that $\{w_{nq}\}_{n=1}^N$ comes from a zero-mean phylogenetic Ornstein-Uhlenbeck process (POUP) over the input space $\mathcal{P}$ and is independent of $\{w_{nq'}\}_{n=1}^N$ for any $q' \neq q$. The hyperparameters for the POUP are treated as deterministic parameters and are calculated by type II maximum likelihood estimation given $\{w_{nq}\}_{n=1}^N$. The key aspect of the POUP of the q^{th} evolutionary feature is the kernel defining the covariance matrix,

$$k_q(\mathbf{p}, \mathbf{p}') = \gamma_q \exp\left(\frac{\mathbf{d}(\mathbf{p}, \mathbf{p}')}{\ell_q}\right) + \sigma_q \delta_{\mathbf{p},\mathbf{p}'}, \tag{1}$$

where γ_q is the phylogenetic noise, ℓ_q is the characteristic length-scale, σ_q is the non-phylogenetic noise, $\mathbf{d}(\mathbf{p}, \mathbf{p}')$ is the distance between $\mathbf{p}$ and $\mathbf{p}'$ in $\mathcal{P}$, and $\delta_{\mathbf{p},\mathbf{p}'}$ is the Kronecker delta. We then let $\theta_q = [\gamma_q, \ell_q, \sigma_q]^\mathsf{T}$.

Now, having defined the POUP hyperparameters $\{\theta_q\}_{q=1}^Q$, and assuming $\mathcal{P}$, a Gaussian predictive distribution can be found for each of the random variables $\{\hat{w}_q\}_{q=1}^Q$ observed at some point $\hat{\mathbf{p}} \in \mathcal{P}$. Taking the maximum a posteriori (MAP) estimate for $\{\hat{w}_q\}_{q=1}^Q$ and denoting this as $\{\tilde{w}_q\}_{q=1}^Q$, A reconstruction of the call surface at $\hat{\mathbf{p}}$ is given by

$$\hat{S} = \sum_{q=1}^{Q} \tilde{w}_q \phi_q \tag{2}$$

Which in turn provides an estimate of the acoustic call signal at $\hat{\mathbf{p}}$, $\hat{y}$.

Developing the Current Model

The model presented above allows the reconstruction of ancestral bat echolocation call surfaces, which in turn allows the reconstruction of echolocation call acoustic signals. For my collaborators at the Centre for Biodiversity and Environment Research at UCL, this represents a significant step forward in the study of echolocation in bats. It is expected that over the coming months I will be granted access to the EchoBank database of bat call recordings from over 600 species, and that applying this model will result in an impactful publication. However, there remains much work to do.

The first limitation of the model to be addressed is the assumption that $S_n = \sum_{q=1}^{Q} w_{nq}\phi_q$. That is to say that the model does not allow for observation noise on the call surface. This in turn means that the likelihood of the observed surfaces, $p(\mathbf{S}|\Phi, \Theta, \mathcal{P}, \dots)$, for a given model cannot be calculated. Addressing this issue would represent an extension to the methods presented by Jones & Moriarty [8] and Hajipantelis [6]. With a joint likelihood for the call surfaces in place, issues of model selection can be addressed, particularly for comparing models with evolutionary features identified by PCA against features identified by Independent Components Analysis or VARIMAX, for example.

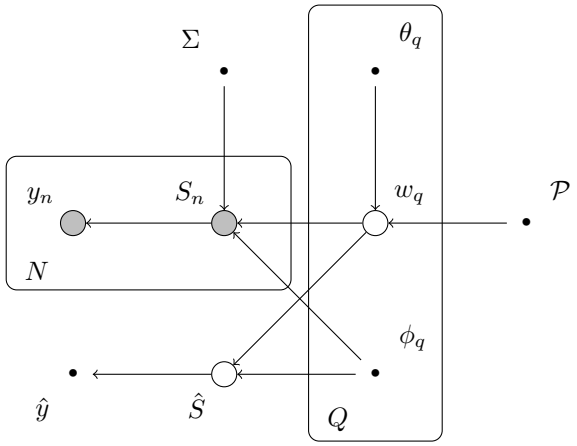

Figure 5: A proposed extension to the model for the evolution a bat echolocation calls presented in Figure 4. This model includes a noise process over the call surface, which would allow the calculation of the model evidence. This in turn would facilitate model selection for various sets of evolutionary features, phylogenies, and Ornstein-Uhlenbeck process hyperparameters.

Further Development

In almost every respect, the model presented in Figure 5 can be refined further.

Consider $\mathcal{P}$, the evolutionary relationships between bat species, which has been treated as a deterministic input space in this model. In fact, the true phylogeny is unknown, and so treating $\mathcal{P}$ as a random variable in the model would be a more accurate representation of our understanding of the problem.

The same argument applies to the evolutionary features identified, $\{\phi_q\}_{q=1}^{Q}$. As such, a probabilistic PCA [18] approach to identifying evolutionary features may be more informative.

The above model assumes that evolution follows an POUP, a popular model for the evolution of continuous character traits introduced by Lande [10] and particularly useful for modelling stabilising selection [7]. However empirical analysis of continuous characteristics suggest that models based on stable distributions may be more appropriate [4] as the heavier tails of these distributions (as compared to a Gaussian distributions) allow for discontinuous jumps in the value of the characteristic between species. One potentially interesting

avenue to explore would be replacing the Gaussian processes in the model with Student-t processes [16], which may go some way to accomodating these heavier tails.

It is also fair to say that a 2-D surface may not be the best approach to the Time-Frequency representation of acoustic signals. It is possible that an alternative representation to the call surfaces, such as that proposed by DiCecco & Gaudette [3], may provide a more revealing perspective on the call structure.

While interesting, I do not necessarily feel that any of these any of these aspects of the model are the most important to address. I am of the opinion that relaxing the assumption of independence between the Q PGPs is the most important next step. An approach to this would be to implement a Gaussian Process Latent Variable Model (GPLVM) [11] [9]. A GPLVM)may even allow the incorporation of other call characteristics into the model, such as duration, maximum frequency, etc. These call characteristics have been the basis of previous attempts at ancestral reconstruction of echolocation calls [1] and incorporating them may lend more weight to any results.

Thesis Outline

As part of this report I am including a brief thesis outline. I envision my thesis consisting of four chapters along with an introduction and a conclusion.

In the first chapter I intend to present an ancestral reconstruction of bat echolocation call spectral density curves based on the model presented in Figure 4, along the lines of the accompanying conference paper [12].

In the second chapter I intend to present a joint model for the evolution of echolocation call surfaces, similar to the model presented in Figure 5.

The third chapter of my thesis will be based around implementing a LVGPM to relax independence assumption between evolutionary features.

I am uncertain as to the contents of my final chapter, however, I have outlined potential avenues to explore above and expect the most promising of these to become apparent over the coming months.

References

[1] AL Collen. *The evolution of echolocation in bats: a comparative approach*. PhD thesis, UCL (University College London), 2012.

[2] Theodoros Damoulas, Samuel Henry, Andrew Farnsworth, Michael Lanzone, and Carla Gomes. Bayesian classification of flight calls with a novel dynamic time warping kernel. In *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, pages 424–429. IEEE, 2010.

[3] John DiCecco, Jason E Gaudette, and James A Simmons. Multi-component separation and analysis of bat echolocation calls. *The Journal of the Acoustical Society of America*, 133(1):538–546, 2013.

[4] Michael G Elliot and Arne Ø Mooers. Inferring ancestral states without assuming neutrality or gradualism using a stable model of continuous character evolution. *BMC evolutionary biology*, 14(1):226, 2014.

[5] Damien Garcia. Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational statistics & data analysis*, 54(4):1167–1178, 2010.

[6] Pantelis Z Hadjipantelis, Nick S Jones, John Moriarty, David A Springate, and Christopher G Knight. Function-valued traits in evolution. *Journal of The Royal Society Interface*, 10(82):20121032, 2013.

[7] Thomas F Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, pages 1341–1351, 1997.

[8] Nick S Jones and John Moriarty. Evolutionary inference for function-valued traits: Gaussian process regression on phylogenies. *Journal of The Royal Society Interface*, 10(78):20120616, 2013.

[9] Karl Krauth, Edwin V Bonilla, Kurt Cutajar, and Maurizio Filippone. Autogp: Exploring the capabilities and limitations of gaussian process models. *arXiv preprint arXiv:1610.05392*, 2016.

[10] Russell Lande. Natural selection and random genetic drift in phenotypic evolution. *Evolution*, pages 314–334, 1976.

[11] Ping Li and Songcan Chen. A review on gaussian process latent variable models. *CAAI Transactions on Intelligence Technology*, 1(4):366–376, 2016.

[12] Joseph Meagher, Theodorus Damoulas, Kate Jones, and Mark Girolami. Phylogenetic gaussian processes for the ancestral reconstruction of bat echolocation calls. 2017.

[13] Karin Meyer and Mark 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.

[14] Davide Pigoli, Pantelis Z Hadjipantelis, John S Coleman, and John AD Aston. The analysis of acoustic phonetic data: exploring differences in the spoken romance languages. *arXiv preprint arXiv:1507.07587*, 2015.

[15] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[16] Amar Shah, Andrew Wilson, and Zoubin Ghahramani. Student-t processes as alternatives to gaussian processes. In *Artificial Intelligence and Statistics*, pages 877–885, 2014.

[17] Rong Tang and Hans-Georg Müller. Pairwise curve synchronization for functional data. *Biometrika*, 95(4):875–889, 2008.

[18] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[19] Xinglei Zhu, Gerald T Beauregard, and Lonce L Wyse. Real-time signal estimation from modified short-time fourier transform magnitude spectra. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1645–1653, 2007.