Phylogenetic Gaussian Process Factor Models

The ancestral reconstruction of bat echolocation

J.P Meagher*, **T. Damoulas, K. Jones, & M. A. Girolami**

The University of Warwick, Department of Statistics

* https://warwick.ac.uk/jpmeagher

THE UNIVERSITY OF WARWICK

EPSRC Engineering and Physical Sciences Research Council

Introduction

Bats are found in nearly every habitat on Earth. They have evolved into nocturnal, flying echolocators. Ultrasonic echolocation calls mean bats leak information as they forage making them ideal bioindicators for biodiversity monitoring and the topic of much research.

Ancestral reconstruction extrapolates back in time from existing populations to their common ancestors. It can be thought of as a prediction problem without a validation set. Thus, statistical models must be both interpretable and informed by theory. We develop a model based on phylogenetic Gaussian processes allowing for evolutionary inference on sets of traits such as echolocation.

Background Methods

Given data $\mathcal{D} \equiv \{\mathbf{Y} = (\mathbf{y}_1, \dots \mathbf{y}_N), \mathcal{P}\}$ where $\mathbf{y}_n = (y(x_{n1}, \mathbf{p}_n), \dots, y(x_{nD}, \mathbf{p}_n)) \in \Re^D$ is a set of observed traits for $x \in \mathcal{X}$. $\mathbf{p}_n$ denotes the position of n^{th} individual on the phylogenetic tree $\mathcal{P}$.

Assume that the data is drawn from a phylogenetic Gaussian process (PGP) such that

$$y(x, \mathbf{p}) \sim \mathcal{GP}(0, k(x, \mathbf{p}, x', \mathbf{p}')).$$

It is shown by Jones & Moriarty [1] that, if the kernel is assumed to be trait-phylogeny separable, subject to regularity conditions there exists

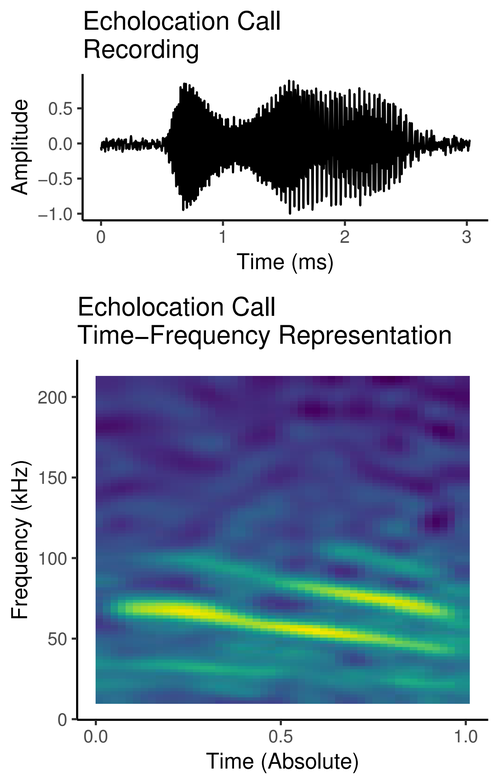
$$g(x, \mathbf{p}) = \sum_{i=1}^{Q} \phi_i(x) F_i(\mathbf{p})$$

which has the same distribution as $\mathbf{Y}$, where $\int \phi_i(x)\phi_j(x)dx = \delta_{ij}$ and $F_i(\mathbf{p}) \sim \mathcal{GP}\left(0, k^i_{\mathcal{P}}(\mathbf{p}, \mathbf{p}')\right)$ is a univariate PGP over $\mathcal{P}$.

The kernel used for PGPs is the Ornstein-Uhlenbeck kernel, a Gauss-Markov process equivalent to a Matern kernel for $\nu = \frac{1}{2}$.

$$k^q_{\mathcal{P}}(\mathbf{p}, \mathbf{p}') \propto \exp\left(-\frac{d_{\mathcal{P}}(\mathbf{p}, \mathbf{p}')}{\ell_q}\right)$$

where $d_{\mathcal{P}}(\mathbf{p}, \mathbf{p}')$ is the distance between $\mathbf{p}, \mathbf{p}' \in \mathcal{P}$.



Following the methods laid out by Pigoli *et al* [3] echolocation call recordings are transformed into smooth spectrogram surfaces for further analysis.

Phylogenetic Gaussian Process Factor Model

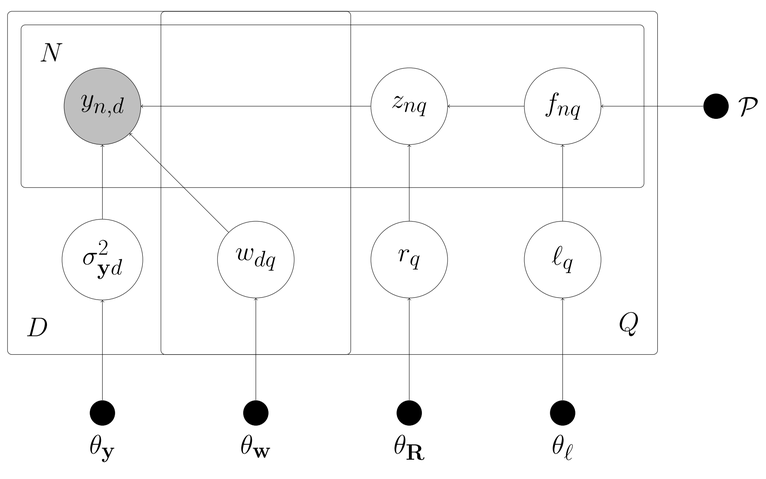
Drop x for ease of notation and let $\boldsymbol{\varphi}_m$ represent an independent standard normal random variable of length m.

$$\begin{aligned}\mathbf{y}(\mathbf{p}) &= \mathbf{W}\mathbf{z}(\mathbf{p}) + \sqrt{\Sigma_{\mathbf{y}}}\boldsymbol{\varphi}_D \\ &= \mathbf{W}(\sqrt{\Sigma_{\mathbf{f}}}\mathbf{f}(\mathbf{p}) + \sqrt{\Sigma_{\mathbf{z}}}\boldsymbol{\varphi}_Q) + \sqrt{\Sigma_{\mathbf{y}}}\boldsymbol{\varphi}_D \\ &= \mathbf{W}(\sqrt{(\mathbf{R}^{-1} + \mathbf{I}_Q)^{-1}}\mathbf{f}(\mathbf{p}) + \sqrt{(\mathbf{R} + \mathbf{I}_Q)^{-1}}\boldsymbol{\varphi}_Q) + \sqrt{\Sigma_{\mathbf{y}}}\boldsymbol{\varphi}_D\end{aligned}$$

where $\mathbf{W} = (\mathbf{w}_1, \dots \mathbf{w}_Q)^{\mathsf{T}} \in \Re^{D \times Q}$ for $\mathbf{w}_q = (w_{1q}, \dots, w_{Dq})^{\mathsf{T}}$ are factor loadings. $\mathbf{f}(\mathbf{p}) = (f_1(\mathbf{p}), \dots, f_Q(\mathbf{p}))^{\mathsf{T}} \in \Re^Q$ for $f_q(\mathbf{p}) \sim \mathcal{GP}\left(0, k^q_{\mathcal{P}}(\mathbf{p}, \mathbf{p}')\right)$ with unit variance. $\Sigma_{\mathbf{y}}$ is a $D \times D$ diagonal covariance matrix. $\Sigma_{\mathbf{f}}$ and $\Sigma_{\mathbf{z}}$ are $Q \times Q$ diagonal covariance matrices with $\Sigma_{\mathbf{f}} + \Sigma_{\mathbf{z}} = \mathbf{I}_Q$. Define the phylogenetic signal to noise ratio $\mathbf{R} = \Sigma_{\mathbf{f}}\Sigma_{\mathbf{z}}^{-1}$. We also have the vector of characteristic length-scales $\boldsymbol{\ell} = (\ell_1, \dots, \ell_Q)^{\mathsf{T}}$.

We place GP priors on $\mathbf{w}_q$, inverse Gamma priors on the diagonal elements of $\Sigma_{\mathbf{y}}$, beta prime priors on the diagonal elements of $\mathbf{R}$, and Gamma priors on the elements of $\boldsymbol{\ell}$.

Factor models have well known identifiability problems. Fixing the location and scale of the latent variables $\mathbf{z}(\cdot)$ and $\mathbf{W}$ to be lower triangular strips out rotational invariance. If necessary, reflection invariance can be dealt with by postprocessing samples from the posterior.



Inference

We have the closed form conditional distributions

$$vec(\mathbf{W})|\mathbf{Y}, \mathbf{Z}, \Sigma_{\mathbf{y}} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}|\mathbf{Y},\mathbf{Z},\Sigma_{\mathbf{y}}}, \Sigma_{\mathbf{w}|\mathbf{Y},\mathbf{Z},\Sigma_{\mathbf{y}}})$$

$$vec(\mathbf{Z})|\mathbf{Y}, \mathbf{W}, \Sigma_{\mathbf{y}}, \mathbf{R}, \boldsymbol{\ell} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{Y},\mathbf{W},\Sigma_{\mathbf{y}},\mathbf{R},\boldsymbol{\ell}}, \Sigma_{\mathbf{z}|\mathbf{Y},\mathbf{W},\Sigma_{\mathbf{y}},\mathbf{R},\boldsymbol{\ell}})$$

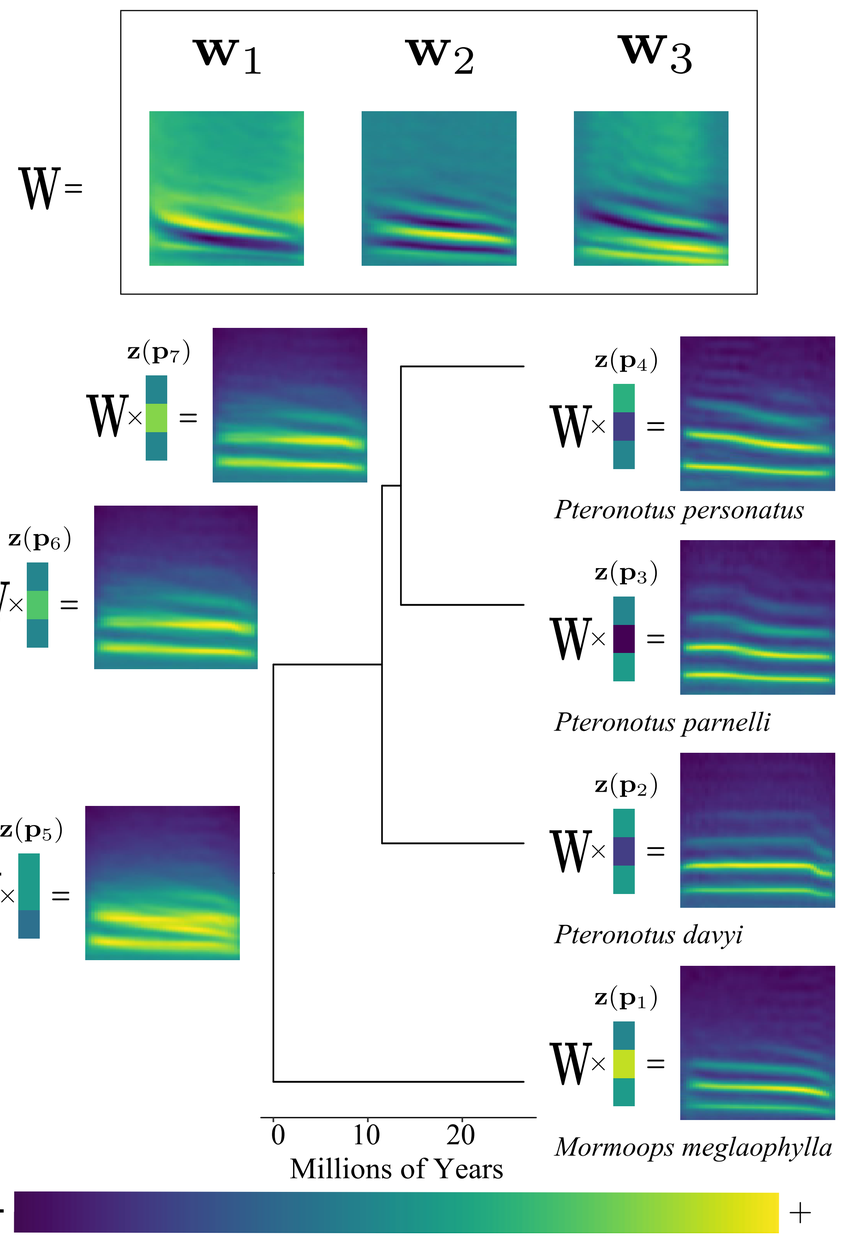
$$(\Sigma_{\mathbf{y}})_{ii}|\mathbf{Y}, \mathbf{W}, \mathbf{Z} \sim \mathcal{IG}(\alpha_{(\Sigma_{\mathbf{y}})_{ii}|\mathbf{Y},\mathbf{W},\mathbf{Z}}, \beta_{(\Sigma_{\mathbf{y}})_{ii}|\mathbf{Y},\mathbf{W},\mathbf{Z}})$$

The distribution $p(\mathbf{W}, \mathbf{Z}, \Sigma_{\mathbf{y}}, \mathbf{R}, \boldsymbol{\ell})$ can be sampled from using a Metropolis-within-Gibbs scheme.

Ancestral reconstruction is performed by considering the posterior distribution of $\mathbf{y}(\mathbf{p}^*)$ where $\mathbf{p}^* \in \mathcal{P}$ is some internal node on the phylogeny.

Results

We implement the PGPFM for preprocessed call spectrograms setting $\mathbf{W}$ to be the first 3 principal components of the preprocessed call spectrograms and assuming $\Sigma_{\mathbf{y}}$ to be negligible. Data presented by Meagher *et al* [2] was used in this analysis.



Discussion

- We have developed a Gaussian process model which builds on existing theory and practise in evolutionary inference to provide a Bayesian approach to the ancestral reconstruction of sets of continuous valued traits.
- Model parameters and hyperparameters can be interpreted intuitively, informing any conclusions drawn from the analysis.

References

[1] 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.

[2] JP Meagher, T Damoulas, KE Jones, and M Girolami. Phylogenetic gaussian processes for bat echolocation. *Statistical Data Science*, page 111, 2018.

[3] Davide Pigoli, Pantelis Z Hadjipantelis, John S Coleman, and John AD. The analysis of acoustic phonetic data: exploring differences in the spoken romance languages. *arXiv preprint arXiv:1507.07587*, 985, 2015.