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/~jpmegaher

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 biosensors 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, such models must be multivariate and understood for evolutionary inference on sets of traits such as echolocation.

Background Methods

Given data $\mathcal{D} \equiv \{Y = (y_1, \ldots, y_n), \mathbf{P}\}$ where $y_n = (x_1, x_2, \ldots, x_n)$ is a set of observed traits for $n$ species and $\mathbf{P}$ 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

$$y(x) \sim \mathcal{GP}(\mu(x), \kappa(x, x'))$$

which has the same distribution as $Y$, where $f(x_i) \sim \mathcal{N}(\mu(x_i), \kappa(x_i, x_i')) \sim \mathcal{GP} (\mu, \kappa)$ is a univariate PGP over $\mathcal{P}$. The kernel used for PGPs is the Ornstein-Uhlenbeck kernel, a Gaussian-Markov process equivalent to a Matern kernel for $\nu = \frac{1}{2}$.

$$k_{O}(p, p') \propto \exp\left(-\frac{d_{p, p'}^2}{\ell^2}\right)$$

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

Phylogenetic Gaussian Process Factor Model

Drop $x$ for ease of notation and let $\phi_n \equiv \phi_n(x)$, independent standard normal random variable of length $n$.

$$y(p) = W\mathbf{x}(p) + \sqrt{\Sigma_p} \phi(p)$$

$$W = \left[\begin{array}{c} w_1 \\ w_2 \\ \vdots \\ w_m \end{array}\right] \in \mathbb{R}^{m \times n}$$

$w_i$ are factor loadings. $\Sigma_p$ is $Q \times Q$ diagonal covariance matrix. $\phi(p)$ is a $Q \times 1$ vector of characteristic length-scales $\ell = (\ell_1, \ell_2, \ldots, \ell_Q)$. We place GP priors on $w_i$, inverse Gamma priors on the diagonal elements of $\Sigma_p$, beta prime priors on the diagonal elements of $R$, and Gamma priors on the elements of $\ell$.

Factor models have well known identifiability problems. Fixing the location and scale of the latent variables $x(\cdot)$ and $W$ 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

$$\text{vec}(W|Y, Z, \Sigma_p) \sim \mathcal{N}(\mu_{\text{vec}(W|Y, Z, \Sigma_p)}, \Sigma_{\text{vec}(W|Y, Z, \Sigma_p)})$$

$$\text{vec}(Z|Y, W, \Sigma_p, \ell) \sim \mathcal{N}(\mu_{\text{vec}(Z|Y, W, \Sigma_p, \ell)}, \Sigma_{\text{vec}(Z|Y, W, \Sigma_p, \ell)})$$

The distribution $p(W, Z, \Sigma_p, \ell)$ can be sampled from using a Metropolis-within-Gibbs scheme.

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

Results

We implement the PGPFM for preprocessed call spectrograms setting $W$ to be the first 3 principal components of the preprocessed call spectrograms and assuming $\Sigma_p$ 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