Evolutionary inference for function-valued traits: Gaussian process regression on phylogenies

Nick S. Jones and John Moriarty

1Department of Mathematics, Imperial College London, London SW7 2AZ, UK
2School of Mathematics, University of Manchester, Oxford Road, Manchester M13 9PL, UK

Biological data objects often have both of the following features: (i) they are functions rather than single numbers or vectors, and (ii) they are correlated owing to phylogenetic relationships. In this paper, we give a flexible statistical model for such data, by combining assumptions from phylogenetics with Gaussian processes. We describe its use as a non-parametric Bayesian prior distribution, both for prediction (placing posterior distributions on ancestral functions) and model selection (comparing rates of evolution across a phylogeny, or identifying the most likely phylogenies consistent with the observed data). Our work is integrative, extending the popular phylogenetic Brownian motion and Ornstein–Uhlenbeck models to functional data and Bayesian inference, and extending Gaussian process regression to phylogenies. We provide a brief illustration of the application of our method.

1. Introduction

In this paper, we consider statistical inference for function-valued data that are correlated owing to phylogenetic relationships. A schematic is given in figure 1a: in this case, given functional data observed at the tips of a phylogeny, the task is to perform inference on the (unobserved) functional data at the root of the phylogeny. Alternatively, if the phylogeny is uncertain we may wish to perform phylogenetic inference, or our interest may be inferring the dynamics of the evolutionary process that produced the data. The term ‘function-valued’ is meant in the sense of Ramsay [2], where a datum is a continuous function \( f(x) \) of a variable \( x \), such as time or temperature: examples are therefore curves for ambient temperature versus growth rate for caterpillars, a heart rhythm time series [3], or a spectrogram of audio data. Our approach is to combine the theory of Gaussian processes with assumptions from phylogenetics, to obtain a flexible non-parametric model for such data. Because this model effectively specifies the evolutionary dynamics of the data through the phylogeny, we note that (i) our approach generalizes the Brownian motion and Ornstein–Uhlenbeck models of continuous-time character evolution from quantitative genetics [4], and (ii) the phylogenetic tree will play the role of evolutionary time: to avoid confusion, we therefore refer to the indexing variable \( x \) above as ‘space’. The model may be used as a prior for Bayesian inference, which opens up phylogenetically aware approaches to the prediction of ancestral functions and to model selection.

Because of their generality, flexibility and mathematical simplicity, there has been substantial recent interest in the use of Gaussian process priors for Bayesian non-parametric regression [5–7]. This report may be viewed as an extension of Gaussian process regression to take account of functional data and a tree topology. Our work relates to the field of spatial statistics, in that it involves multidimensional index sets with both distance and topology; however, it is the tree topology and the conditional independence of siblings given their common ancestors that makes our approach particularly suited to the study of phylogenetically correlated data.
Functional representations of data have been in use for at least 30 years [2, 8]. Their use in (non-phylogenetic) evolutionary studies was proposed for quantitative genetics in Kirkpatrick & Heckman [9]. While the wider debate concerning classical versus functional data types is outside the scope of this report; in Ramsay [2], it is argued that statistical challenges for functional data, so we may ask: in the application of our models to statistical inference, which classical approaches are dual to our functional approach? Classical data types investigated in the phylogenetic context include sequences of discrete symbols (e.g. encoding the presence or absence of certain characters [10], or alternatively in models for the evolution of genetic sequences), the spatial locations of a number of fixed landmark points in geometric morphometrics [11], and multivariate vectors of continuous characters or summary statistics [12]. Within the evolutionary study of multivariate vectors, the relative effects of phylogenetic and (species-)specific variation have been studied [13], which relates closely to our equation (2.14) below, and the challenges of ancestor prediction and model selection have been addressed and explored statistically [14–16]. We address model selection in the supplementary material to this report and more fully in a companion paper [1], which also contains a discussion of statistical issues and performance. Beyond this duality, however, our model is suitable for use as a prior in Bayesian analysis, and in the remainder of this report, we derive our models and provide details for their use in Bayesian regression.

2. Phylogenetic Gaussian processes

2.1. Gaussian processes and regression

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution. Examples are the Wiener (Brownian motion) and Ornstein–Uhlenbeck processes that have received considerable recent attention in the study of evolution [12,14,15,17]. Gaussian processes are characterized by their first two (cross-)moments (although, unless otherwise stated, the mean will be assumed to be zero—a common assumption for Gaussian process priors, see Rasmussen & Williams [5]). The covariance function \(\sigma\), which specifies how the sample values (co-)vary, typically depends on a vector \(\theta\) of parameters.

Suppose that a Gaussian process \(f\) is observed at a vector of coordinates \(L\). Then, the resulting vector of sample values \(f(L)\) has a multivariate Gaussian distribution of dimension equal to \(|L|\), the number of points of measurement: 

\[
\mathbf{f}(L) \sim \mathcal{N}(\mathbf{0}, \Sigma(L, L, \theta)).
\]

Here, \(\mathcal{N}\) represents the Gaussian distribution, its two arguments being the mean vector and the covariance matrix \(\Sigma(L, L, \theta)\), which is the matrix of the covariances between all pairs \((\ell_i, \ell_j)\) of observation coordinates in \(L\) (where \(\ell_i \in L\)) so that

\[
[\sigma(L, L, \theta)]_{ij} = \sigma(\ell_i, \ell_j) = E[f(\ell_i)f(\ell_j)]|\theta|. \tag{2.1}
\]

In §2.3, we derive the structure of some particular covariance functions, including the phylogenetic Ornstein–Uhlenbeck process.

The log-likelihood of the sample \(f(L)\) is then [5]

\[
\log p(f(L)|\theta) = -\frac{1}{2} f(L)^T \Sigma(L, L, \theta)^{-1} f(L) \\
- \frac{1}{2} \log(\det(\Sigma(L, L, \theta))) - \frac{|L|}{2} \log 2\pi. \tag{2.2}
\]

We might be interested in making inferences about the unobserved values of our random function \(f\) at a vector \(M\) of coordinates, given samples at the coordinates \(L\). The posterior distribution of the vector \(f(M)\) given \(f(L)\) is also Gaussian and of the form [5]:

\[
f(M)|f(L) \sim \mathcal{N}(A B) \tag{2.3}
\]

where

\[
A = \sigma(M, L, \theta)\Sigma(L, L, \theta)^{-1} f(L), \tag{2.4}
\]

and

\[
B = \sigma(M, M, \theta) - \sigma(M, L, \theta)\sigma(L, L, \theta)^{-1} \sigma(M, L, \theta)^T \tag{2.5}
\]

and \(\sigma(M, L, \theta)\) denotes the \(|M| \times |L|\) matrix of the covariance function \(\sigma(m_i, l_j)\) evaluated at all pairs \(m_i \in M, l_j \in L\).
From equation (2.4), the posterior mean vector $A$ consists of linear combinations of the observations, while the posterior covariance matrix $B$, given by (2.5), is independent of the observations. Gaussian process regression is non-parametric in the sense that no assumption is made about the structure of the model: the more data gathered, the longer the vector $f(L)$, and the more intricate the posterior model for $f(M)$.

We are able to combine evolutionary dynamics with functional data, because the index variable $\ell$ introduced above can be a point in a space of arbitrary dimension. Therefore, if we wish to model the time evolution of a function-valued trait $f(x)$ with indexing variable $x$, we could consider each point of observation $\ell$ as corresponding to a point $(\ell,t)$ in both space and evolutionary time. Then, $f(L)$ would represent the values of a random space–time surface at various space–time coordinates, $L$ (analogously, the values $f(L)$ are like a set of altimeter recordings recorded at different locations $(L)$ on a map). A cross section through this space–time surface at a fixed value of time $t$ yields a random curve that can be viewed as a single function-valued trait at the fixed time $t$ in its evolution. In §2.2, we will extend this view to processes indexed by phylogenies.

2.2. Phylogenetic covariance function

Our aim in this section is to build a Gaussian process model for the evolution of a function-valued trait along a phylogenetic tree $T$ by allowing $T$ to play the role of evolutionary time, rather than the linear time variable $t$ used above. Suppose, therefore, that each observation $\ell$ corresponds to a point $(\ell,t)$ in the space–phylogeny $S \times T$: that is, $x \in S$ is the value under consideration of the spatial (indexing) variable, and $t \in T$ is the point under consideration on the phylogeny ($t$ is not just a time coordinate but also indicates a branch of the phylogeny). We will do this by constructing a covariance function $\Sigma_T(l,l)$ when the $l,l$ are points in $S \times T$, calling it the phylogenetic covariance function. In order to obtain a unique phylogenetic covariance function $\Sigma_T$, we will make two assumptions that are natural in the context of evolution [12]:

**Assumption 2.1.** Conditional on their common ancestors in the phylogenetic tree $T$, any two function-valued traits are statistically independent.

**Assumption 2.2.** The statistical relationship between a function-valued trait and any of its descendants in $T$ is independent of the topology of $T$.

Assumption 2.2 means that our statistical model of the evolutionary process is identical along paths through $T$ from the root to any tip, and we call this the marginal process (this assumption could be generalized, for example to model unequal rates of evolution along different branches of $T$). As in Felsenstein [12] and related work, we need the assumption that $T$ is a chronogram, having both a tree topology and a distance metric. We will call the distance between a point $t \in T$ and the root of $T$ the ‘depth’ of $t$, and denote it by the plain typeface symbol $t$.

The marginal process is of course Gaussian, and so it is sufficient to specify its covariance function $\Sigma$ on $S \times T$, where $T$ is the set of all dates in $T$. In fact, each different marginal covariance function specifies a different phylogenetic covariance function, and so a wide class of phylogenetic covariance functions may be constructed. In turn, these models offer an equally wide choice of priors for Bayesian evolutionary inference with function-valued traits.

In proposition 2.3, we present our mathematical results in the simplest case when the marginal covariance function is space–time separable: that is, when there exist a space-only covariance function $K(x_1,x_2)$ and a time-only covariance function $k(t_1,t_2)$ such that

$$\Sigma((x_1,t_1),(x_2,t_2)) = K(x_1,x_2)k(t_1,t_2). \quad (2.6)$$

**Proposition 2.3.**

1. If the marginal covariance function $\Sigma$ is space–time separable, then the phylogenetic covariance function $\Sigma_T$ is also space–time separable, i.e.

$$\Sigma_T((x_1,t_1),(x_2,t_2)) = K(x_1,x_2)\Sigma_T(t_1,t_2). \quad (2.7)$$

where $k_{\Sigma_T}(t_1,t_2)$ is the phylogenetic covariance function constructed from the time-dependent component $k$ in (2.6) and $K(x_1,x_2)$ is the space-dependent component.

2. When the time-dependent component $k$ of (2.6) specifies a process that is Markovian in time, we have the simple expression

$$k_{\Sigma_T}(t_1,t_2) = k(t_1,t_{12})k(t_{12},t_2)^{-1}k(t_2,t_{12}) \quad (2.8)$$

where $t_{12}$ is the most recent common ancestor of $t_1$ and $t_2$ (and $t_{12}$ is its depth in $T$). In particular, we have the following corollaries:

(a) if $k(t_1,t_{12}) = \min(t_1,t_{12})$ so that we have a Wiener process in evolutionary time as in Felsenstein [12], then $k_{\Sigma_T}(t_1,t_2) = t_{12}$, which is the variance of the evolutionary time component of variation evaluated at the most recent common ancestor $t_{12}$.

(b) if $k$ is isotropic so that $k(t_1,t_2)$ is a function of $|t_1 - t_2|$ only, it does not necessarily follow that $k_{\Sigma_T}(t_1,t_2)$ is isotropic (meaning a function of the patristic distance between $t_1$ and $t_2$ only). In fact, $k_{\Sigma_T}$ is only isotropic when $k$ is the Ornstein–Uhlenbeck covariance.

3. Let $Y$ be a phylogenetic Gaussian process with a space–time separable covariance function $\Sigma_T$ which factorizes as in (2.7). If $K$ is a continuous degenerate Mercer kernel, then there exist an integer $n$ and deterministic functions $\phi_i : S \to R$ and univariate Gaussian processes $X_i$ for $i = 1 \ldots n$ such that the Gaussian process given by

$$f(x,t) = \sum_{i=1}^n \phi_i(x)X_i(t) \quad (2.9)$$

has the same distribution as $Y$.

Further mathematical detail can be found in the electronic supplementary material. Part 2(b) of proposition 2.3 helps clarify the relationship between phylogenetic Gaussian process models and studies such as [18] based on autocorrelation functions of patristic distance, namely that the two are compatible only when the phylogenetic Ornstein–Uhlenbeck process is assumed. Part 3 of the proposition establishes that a convenient expansion using basis functions can be used for a wide range of phylogenetic Gaussian processes with space–time separable covariance functions. This expansion is useful for statistical inference, as it justifies the use of dimension-reduction techniques (see the paper [1]).

2.3. Examples

We now illustrate two ways in which proposition 2.3 may be used to make priors for Bayesian inference on phylogenetically related function-valued traits (henceforth referred to simply as ‘traits’). They differ in their approach to modelling spatial
function, and may be called the spatially inhomogeneous (figure 1a) and spatially homogeneous (figure 1b) models, respectively. In both cases, we assume the simplest possible structure for the marginal covariance function $\Sigma$: that it is space–time separable as in (2.6). We also assume that, conditional on any given trait, its ancestor and progenitor traits are statistically independent: this corresponds to choosing a temporal component which is Markovian. The only Markovian Gaussian processes are the class of Ornstein–Uhlenbeck processes, and the stationary examples have the covariance function $k(t_1, t_2) = \exp(-|t_1 - t_2|/\theta^2)$, where the hyperparameter $\theta$ specifies the characteristic length scale for the evolutionary dynamics. From proposition 2.3, we obtain

$$k_T(t_1, t_2) = \exp\left(-\frac{d_T(t_1, t_2)}{\theta^2}\right),$$

(2.10)

where $d_T(t_1, t_2)$ denotes the patristic distance between $t_1$ and $t_2$ (note that the phylogenetic Ornstein–Uhlenbeck model is isotropic, see end of proposition 2.3).

2.3.1. Spatially homogeneous model

If our prior belief is that variation in the trait is homogeneous over all values of $x \in S$, then we should choose the spatial covariance $K$ to be stationary, and if the traits are typically smooth we should choose $K$ to generate smooth random functions. An example is the squared exponential covariance function:

$$K(x_1, x_2) = \exp\left(-\frac{(x_1 - x_2)^2}{2\theta^2}\right).$$

(2.11)

Here, the hyperparameter $\theta$ fixes the characteristic length scale of the random functions. This choice of $K$ gives

$$\Sigma_K((x_1, t_1), (x_2, t_2)) = \Sigma_T((x_1, t_1), (x_2, t_2))$$

(2.12)

$$= \exp\left(-\frac{(x_1 - x_2)^2}{2\theta^2} - \frac{d_T(t_1, t_2)}{\theta^2}\right)$$

(2.13)

This simple example may be combined by summation to construct other spatially homogeneous phylogenetic covariance functions, although the separability property is typically then lost. The following phylogenetic covariance function, $\Sigma_T$, contains an uncorrelated noise term whose influence is controlled by the choice of the parameter $\sigma_n^2$:

$$\Sigma_T((x_1, t_1), (x_2, t_2)) = (1 - \sigma_n^2)\exp\left(-\frac{(x_1 - x_2)^2}{2\theta^2} - \frac{d_T(t_1, t_2)}{\theta^2}\right) + \sigma_n^2 \delta_{t_1 t_2} \delta_{x_1 x_2}$$

(2.14)

where $\delta$ is the Kronecker delta. When functional data are sampled at a finite set of space–time points $L$, the phylogenetic covariance function (2.14) may be used to specify a prior for Bayesian regression as described in §2.1. Figure 1b gives a schematic of functional data and the posterior distribution for the root function, when all functions are discretely sampled on a regular lattice. An illustrative example developing figure 1b is provided in the electronic supplementary material.

2.3.2. Spatially inhomogeneous model

We may alternatively construct a prior using the representation (2.9). This involves choosing deterministic spatial basis functions $\phi_1, \ldots, \phi_n: S \to \mathbb{R}$ and univariate phylogenetic Gaussian processes $X_1, \ldots, X_n$. The spatial basis may be specified a priori, or alternatively obtained by functional decomposition of observed data: in the paper [1], we present a practical methodology for this empirical Bayesian approach. If we assume that the $X_i$ are independent phylogenetic Ornstein–Uhlenbeck processes on $T$ with noise, then their respective covariance functions are:

$$k_T^i(t_1, t_2) = (1 - (\sigma_n^i)^2)\exp\left(-\frac{d_T(t_1, t_2)}{\theta^2}\right) + (\sigma_n^i)^2 \delta_{t_1 t_2}.$$  

(2.15)

The full covariance function of our prior distribution is then:

$$\Sigma_T^{\text{inhom}}((x_1, t_1), (x_2, t_2)) = E[f(x_1, t_1)f(x_2, t_2)]$$

(2.16)

$$= \sum_{i=1}^n k_T^i(t_1, t_2)\phi_i(x_1)\phi_i(x_2).$$

Figure 1a gives a schematic of functional data and the posterior distribution for the root function, when all functions are constructed from the spatial basis $\phi_1, \ldots, \phi_n$.

3. Discussion

In this paper, we have exploited the powerful Bayesian inference architecture provided by Gaussian processes to address evolutionary questions for function-valued data. The closed-form posterior distribution (2.3) offers a straightforward approach to the prediction of unobserved function-valued traits, and the associated explicit likelihood function is an attractive feature for Bayesian inference in the presence of uncertainty over the phylogenetic tree or the evolutionary process. The approach is suitable both for complete (or dense) observations of function-valued traits and for sparsely and even irregularly sampled traits, with missing observations [19].

References


2. Ramsay JO. 1982 When the data are functions. Psychometrika 47, 379–396. (doi:10.1007/BF02291704)


10. Catalano SA, Goloboff PA, Giannini NP. 2010 Phylogenetic morphometrics (I). The use of


