Marinas-Collado, I., Bowman, A. and Macaulay, V. (2019) A Phylogenetic Gaussian process model for the evolution of curves embedded in d-dimensions. *Computational Statistics and Data Analysis*, 137, pp. 285-298. (doi:10.1016/j.csda.2019.03.002)

This is the author's final accepted version.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

http://eprints.gla.ac.uk/181828/

Deposited on: 15 March 2019

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

A Phylogenetic Gaussian process model for the evolution of curves embedded in d-dimensions

Irene Mariñas-Collado[a,1,*], Adrian Bowman[b], Vincent Macaulay[b]

[a] *Dpto. de Estadística, Facultad de Ciencias, Universidad de Salamanca. Plaza de los Caidos s/n 37008, Salamanca (Spain)*
[b] *School of Mathematics and Statistics, University of Glasgow. University Avenue, Glasgow (UK) G12 8QQ*

Abstract

Statistical methods which enable shape information on organisms to be used to construct a phylogenetic tree and to learn how shape evolves are developed. In particular, this allows the evolution of facial curves to be used in studying relationships between and within different ethnic groups and their ancestors. The main challenge is to exploit the details of surface shape, while maintaining computational feasibility. A Gaussian process approach is adopted.

Keywords: Gaussian processes, phylogenetic evolution, shape analysis, facial curves, morphometrics, nose shape, functional phylogenetics, ancestral reconstruction

1. Introduction

The statistical modelling of evolution is a topic of sizeable interest with an extended range of applications. Phylogenies, or evolutionary trees, are the basic structures necessary to think clearly about differences between species and to analyse those differences statistically. The use of both metaphors and models comparing evolution to branching trees goes back at least to Darwin's *On the Origin of Species* (Darwin, 1859) but the major advances from a statistical, computational and algorithmic point of view have mostly been made in the past 50 years. Felsenstein (2004) summarised well the major advances that had been achieved in the course of the previous four decades.

This paper builds on the idea of developing statistical methods through which shape information on organisms can be used to infer a phylogenetic tree and to learn how shape evolves. In recent years, genetic information has most commonly been used for this purpose, typically DNA sequences observed in present-day organisms, but the underlying principle in Darwin's idea of 'descent with modification' was based on physical features (the size and shape of different body parts, the presence or absence of different physical characteristics and so on). External physical traits such as facial shape or skin pigmentation are likely to have been influenced by natural selection (Gregory, 2009). How selection may have affected facial shape, which is also quite variable between populations, has received less attention than its effect on other traits. The approach proposed here is to use shape information, more specifically facial curves, to study relationships between different ethnic groups and their (our) ancestors. The use of shape information, expressed as a continuous and multivariate data type, raises a number of very interesting issues from a methodological perspective.

The raw facial data are in the form of three-dimensional point clouds, usually of size around $100,000$, which characterize each facial surface. A traditional starting point for the analysis of shape is to identify a

[*] Corresponding author
Email addresses: Irenemc@usal.es (Irene Mariñas-Collado), Adrian.Bowman@glasgow.ac.uk (Adrian Bowman), Vincent.Macaulay@glasgow.ac.uk (Vincent Macaulay)

[1] I. Mariñas-Collado is grateful to the School of Mathematics and Statistics at the University of Glasgow for her research studentship. Thanks are offered to all the people who kindly donated their facial images to the project.

number of landmarks, which are reproducible and anatomically meaningful points on the surface. Methods for the statistical analysis of landmarks are well developed, with an excellent description given by Dryden and Mardia (1998). However, the expectation of this approach is that the number of landmarks will be small, which does not allow the full surface to be represented adequately. An alternative approach is to identify anatomical curves, rather than landmarks. The motivation for this is strengthened by the observation that the key features of the face can be viewed as a set of ridges and valleys. For example, the mid-line of the lip is a valley, while the nose profile is a ridge. The essential information for characterising the curvature across the surface is provided by the principal curvatures and their associated directions (Tanaka et al., 1998). Methods of curve estimation are described by Vittert et al. (2017) and Mariñas–Collado (2017). Figure 1 shows a human face with a set of anatomical curves superimposed. The nasal curves in red are used in an application in Section 5.

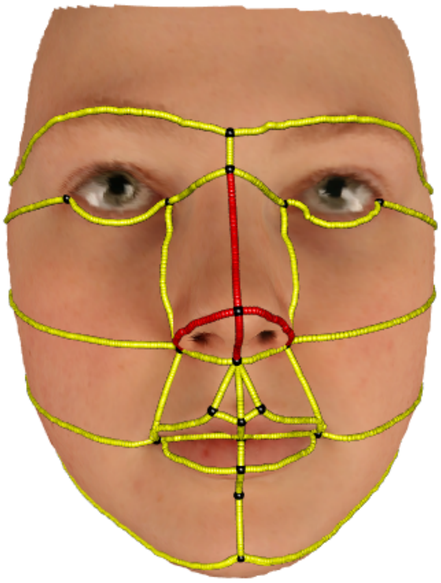


Figure 1: A human face with anatomical landmarks (black) and anatomical curves (yellow). The curves in red are used in the application in Section 5.

Facial curves change in many different contexts. They vary between and within population subgroups and they have changed over the evolution of modern humans. During the last few decades, several authors have tried to clarify the anthropological aspects of the shape of the human nose (Franciscus and Trinkaus, 1988; Mladina et al., 2009). It seems that the erectile posture of *H. sapiens* caused remarkable morphological changes to skull shape. Erectile posture enabled man to see around more effectively, to recognise potential dangers, enemies, sources of food, etc. more easily. The olfactory function of the nose, so important for quadrupeds, started therefore slowly to diminish over time. Its respiratory function became the leading driver in man's nose physiology. This paper uses the evolution of nose curves across three ethnic groups as a case study (Section 5).

The general machinery proposed is that of Gaussian processes (GPs). A GP is a collection of random variables any finite number of which have a multivariate Gaussian (normal) distribution. The random variables will represent the coordinates of the three-dimensional points that define a facial curve, as a function of their location along the curve's arc-length and the time point in the phylogenetic tree. A GP is defined by its mean function (widely assumed to be zero (Rasmussen and Williams, 2006)) and its covariance function. One of the main advantages in the use of GPs to model the spatio-temporal characteristics of facial features is that GPs allow one to specify various types of covariance functions that can capture complex data structures. This is important to be able to model the interactions among the different coordinates, as well as the space and time locations of the points. Facial curves are functions rather than single numbers or vectors and they can be correlated due to phylogenetic relationships in the evolution of facial morphology within and between families or ethnic groups. Jones and Moriarty (2013) presented a flexible statistical model for such data by combining assumptions from phylogenetics with GPs. Their approach generalizes the Brownian motion and Ornstein-Uhlenbeck models of continuous-time evolution from quantitative genetics (Felsenstein, 1985). This paper extends their model to data in the form of points on curves embedded in d-dimensions, where the covariance of the different coordinates needs to be modelled, in addition to the spatial and phylogenetic covariances.

Inside the branching structure of a phylogenetic tree, one can focus on the evolution of one single curve along one branch, without taking into account the branching patterns and the ancestors. This is equivalent to modelling the curve evolving over time as a linear continuous variable, which can be regarded as a degenerate scenario of the phylogenetic GP model and, therefore, is the first model introduced (Section 2). The model is later extended, using the phylogenetic covariance function to allow for branching points in the evolution in Section 3. The challenges encountered when implementing the model are discussed in Section 4. Finally, in Section 5, we present a case study to compare nose shape between and within three broad ethnic groups: Sub-Saharan Africans, White British and Chinese.

2. Gaussian process model for evolving curves

A curve embedded in a d-dimensional space can be parametrised in terms of its arc-length s (a continuous index that can be rescaled to be from 0 to 1) and the set of discrete coordinate labels, $\{c_1, \ldots, c_d\}$. The curve might, for example, represent a ridge or a valley on a surface embedded in those d dimensions. If the curve is, moreover, changing over time, a time dependence t can be added to the model. A GP can be then specified as:

$$w(t, c, s) \sim GP\big(m(t, c, s), k(t, t', c, c', s, s')\big), \tag{1}$$

where the two arguments of the GP refer to the mean and covariance function and where $t, t' \in \mathbb{R}$; $c, c' \in \{c_1, \ldots, c_d\}$ and $s, s' \in [0, 1]$.

Let $\mathbf{s} = (s_1, ..., s_n)^{\mathrm{T}}$ represent a choice of n values of s. Each coordinate can be represented as a function of the arc-length and the time, i.e., $w(t, c_i, s) \equiv c_i(t, s)$. Then, a curve at time t can then be notated as:

$$\mathbf{W}(t) = \left[\mathbf{c}_1(t), \ldots, \mathbf{c}_d(t)\right]^{\mathrm{T}}, \tag{2}$$

where $\mathbf{c}_i(t) = \big(c_i(t, s_1), \ldots, c_i(t, s_n)\big)^{\mathrm{T}}$. The vector $\mathsf{W} = \big(\mathbf{W}(t_1), \ldots, \mathbf{W}(t_T)\big)^{\mathrm{T}}$ for a choice of T values of t, $\mathbf{t} = (t_1, ..., t_T)^{\mathrm{T}}$, can then be written as:

$$\mathsf{W} \sim N_{Tdn}\left(\mathbf{m}, \mathbf{K}\right), \tag{3}$$

where $\mathbf{m}$ is the mean, assumed to be zero, and $\mathbf{K}$ the covariance matrix. Separability is assumed, so that $k(t, t', c, c', s, s') = k_t(t, t')k_c(c, c')k_s(s, s')$ (Rasmussen and Williams, 2006), i.e., $\mathbf{K} = \mathbf{K}_t \otimes \mathbf{K}_c \otimes \mathbf{K}_s$.

- If the process is assumed Markovian, the Ornstein-Uhlenbeck (OU) covariance function can be used, i.e., $k_t(t, t') = \exp(-|t - t'|/\mu)$, with hyperparameter μ, the time-scale. Crudely, μ explains how wiggly the function is in time. A large μ corresponds to a process where change is slow. $\mathbf{K}_t$ represents the covariance of curves at different time points, with the $(i, j)^{th}$ element equal to $k_t(t_i, t_j)$.

- For the $d \times d$ matrix $\mathbf{K}_c$, $k(c, c') = \kappa_{cc'}$ is used, with $\kappa_{cc'}$ representing the correlation between coordinates c and c'. Up to $d(d-1)/2$ hyperparameters can be specified to capture the relationships between coordinates. For example, for curves embedded in three dimensions, with coordinates x, y and z, one could specify three hyperparameters:

$$\mathbf{K}_c = \begin{pmatrix} 1 & \kappa_{xy} & \kappa_{xz} \\ \kappa_{xy} & 1 & \kappa_{yz} \\ \kappa_{xz} & \kappa_{yz} & 1 \end{pmatrix}.$$

If two coordinates behave in a very similar manner, one can assume they have similar correlations with the remaining coordinates. For example, if y and z behave in a similar fashion, then one can specify a single hyperparameter for the correlation between x and y or z:

$$\mathbf{K}_c = \begin{pmatrix} 1 & \kappa_{xyz} & \kappa_{xyz} \\ \kappa_{xyz} & 1 & \kappa_{yz} \\ \kappa_{xyz} & \kappa_{yz} & 1 \end{pmatrix}.$$

In both cases, the hyperparameters are restricted to values which produce positive-definite covariance matrices.

- The space-covariance function used is $k_s(s, s') = \sigma_f^2 \exp\left(-\frac{1}{2}(s - s')^2/\lambda^2\right)$, the Squared-Exponential (SE), with hyperparameters σ_f^2, the signal variance, and λ, the length-scale. λ explains how wiggly the function is in space. $\mathbf{K}_s$ is the covariance matrix for the n arc-length inputs, with $(i, j)^{th}$ element equal to $k_s(s_i, s_j)$.

2.1. Likelihood

From (3), the log-likelihood of the hyperparameters, $\boldsymbol{\theta} = (\sigma_f, \lambda, \mu, \kappa_1, \ldots, \kappa_{d(d-1)/2})$, by the Markov property in time, can be decomposed as:

$$\log p\big(\mathbf{W} \mid \boldsymbol{\theta}\big) = \log p\big(\mathbf{W}(1) \mid \boldsymbol{\theta}\big) + \sum_{i=2}^{T} \log p\big(\mathbf{W}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}\big). \tag{4}$$

The time difference between adjacent time points is assumed to be constant along the sequence; specifically, it is assumed to be 1. Defining $\kappa_t = \exp(-1/\mu)$,

$$\begin{array}{rcl} \mathbf{W}(1) & \sim & N_{dn}(\mathbf{0}, \mathbf{K}_c \otimes \mathbf{K}_s), \\ \mathbf{W}(t) \mid \mathbf{W}(t-1) & \sim & N_{dn}\big(\kappa_t \mathbf{W}(t-1), (1-\kappa_t^2)\mathbf{K}_c \otimes \mathbf{K}_s\big), \quad (t \geq 2). \end{array} \tag{5}$$

Therefore, the marginal log-likelihood for the first curve is:

$$\log p\big(\mathbf{W}(1) \mid \boldsymbol{\theta}\big) = -\frac{dn}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_c \otimes \mathbf{K}_s| - \frac{1}{2}\mathbf{W}(1)^{\mathrm{T}} \left[\mathbf{K}_c \otimes \mathbf{K}_s\right]^{-1} \mathbf{W}(1), \tag{6}$$

and the conditional log-likelihood for the tth curve given the $(t-1)$th is:

$$\begin{aligned} \log p\big(\mathbf{W}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}\big) = & -\frac{dn}{2}\log(2\pi) - \frac{dn}{2}\log(1-\kappa_t^2) - \frac{1}{2}\log|\mathbf{K}_c \otimes \mathbf{K}_s| \\ & - \frac{1}{2(1-\kappa_t^2)}\big(\mathbf{W}(t) - \kappa_t \mathbf{W}(t-1)\big)^{\mathrm{T}} \left[\mathbf{K}_c \otimes \mathbf{K}_s\right]^{-1} \big(\mathbf{W}(t) - \kappa_t \mathbf{W}(t-1)\big). \end{aligned} \tag{7}$$

Moreover, at each time point, the total log-likelihood can be calculated as the sum of the log-likelihood of each coordinate, using conditional distributions.

Because of the complexity of the likelihood function, maximum likelihood estimates are more effectively located by an initial grid search, followed by local refinement. To ease the optimisation process, the number of hyperparameters can be reduced by finding the signal variance, σ_f^2, that maximises the log-likelihood function analytically:

$$\hat{\sigma}_f^2 = \frac{\mathbf{W}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{W}}{Tdn}. \tag{8}$$

The profile log-likelihood, given the remaining hyperparameters $\boldsymbol{\theta'} = (\lambda, \mu, \kappa_1, \ldots, \kappa_{d(d-1)/2})$, is:

$$\log p(\mathbf{W} \mid \boldsymbol{\theta'}) = -\frac{Tdn}{2}\log(2\pi) - \frac{Tdn}{2}\log\left(\frac{\mathbf{W}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{W}}{Tdn}\right) - \frac{1}{2}\log|\mathbf{K}| - \frac{Tdn}{2}, \tag{9}$$

which can be decomposed as in (4).

The standard errors for $\hat{\boldsymbol{\theta}}'$ can be computed from the second derivative information in the usual way. The (conditional) standard error of the estimated variance can be calculated from the square root of the negative of the reciprocal of the second derivative of the full log-likelihood at $\hat{\boldsymbol{\theta}}$:

$$\mathrm{SE}(\hat{\sigma}_f) = \sqrt{-\left[\frac{Tdn}{\sigma_f^2} - \frac{3\mathbf{W}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{W}}{\sigma_f^4}\right]^{-1}}. \tag{10}$$

2.2. *Predictive distributions*

Marginal predictions at time $q \in \mathbb{R}$ can be made at a set of test points (arc-lengths) $\mathbf{s}^* = (s_1^*, \ldots, s_{n^*}^*)$ for each coordinate, using the observed training points $\mathbf{s} = (s_1, ..., s_n)^{\mathrm{T}}$. The joint distribution for a curve at time q, $\mathbf{W}^*(q)$, and the whole sequence, $\mathbf{W}$, can be written as:

$$\begin{bmatrix} \mathbf{W}^*(q) \\ \mathbf{W} \end{bmatrix} \sim N_{dn^*+Tdn}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_c \otimes \mathbf{K}_{s^*} & \mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s} \\ (\mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s})^{\mathrm{T}} & \mathbf{K}_t \otimes \mathbf{K}_c \otimes \mathbf{K}_s \end{bmatrix}\right), \tag{11}$$

where

$$\mathbf{L} = \left[\exp\left(-\frac{|q-1|}{\mu}\right), \quad \exp\left(-\frac{|q-2|}{\mu}\right), \quad \ldots, \quad \exp\left(-\frac{|q-T|}{\mu}\right)\right]. \tag{12}$$

Then

$$\begin{aligned} \mathbf{W}^*(q) \mid \mathbf{W} \sim N_{dn^*} \big(& [\mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s}][\mathbf{K}_t \otimes \mathbf{K}_c \otimes \mathbf{K}_s]^{-1}\mathbf{W}, \\ & \mathbf{K}_c \otimes \mathbf{K}_{s^*} - [\mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s}][\mathbf{K}_t \otimes \mathbf{K}_c \otimes \mathbf{K}_s]^{-1}[\mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s}]^{\mathrm{T}} \big). \end{aligned}$$

The matrix $\mathbf{L}$ depends on the value of q. $\mathbf{K}_{ss^*}$ denotes the $n \times n^*$ matrix of spatial covariances evaluated at all pairs of training and test points, with $(i,j)^{th}$ element equal to $k_s(s_i, s_j^*)$, $\mathbf{K}_{s^*s}$ is its transpose and $\mathbf{K}_{s^*}$ represents the covariance matrix for the test points.

3. Gaussian process model for curves evolving along a phylogenetic tree

A tree can be defined as a mathematical structure which is used to model the evolutionary history of a group of objects, such as organisms, DNA sequences, or, in our case, curves. The actual pattern of historical relationships is the phylogeny or evolutionary tree for which estimation is the aim (Page and Holmes, 1998). A tree consists of nodes connected by edges (Figure 2). In most cases, the terminal nodes (or leaves) represent the objects for which data are available, usually at the present time. Internal nodes represent hypothetical ancestors and the ancestor of all the objects that comprise the tree is the root. The nodes and edges of a tree may have various kinds of information associated with them and one of the issues of interest is to reconstruct the (missing) data at each hypothetical ancestor. Most methods also try to estimate the amount of change that takes place between each pair of nodes, which is represented as an edge length.

The particular branching pattern of a tree is called its topology. This does not represent the distance or time between nodes. A widely-used shorthand notation for the topology of a tree is the Newick format (Huson et al., 2010), where each internal node is represented by a pair of parentheses that enclose all descendants of that node. In this notation, the tree from Figure 2 would be written as $(A, (B, C))$.

Where there are models of evolution for the data, standard statistical methods can be used to make estimates of the phylogeny, such as maximum likelihood estimation (Felsenstein, 2004). Depending on the different types of data that can be observed at the leaves of the tree, an appropriate likelihood function can be associated with the tree. This likelihood function can be maximised with respect to the tree topology, the branch lengths, and other model parameters. Because there exist finitely many tree topologies, it is possible, in principle, to optimize branch lengths and model parameters for every possible tree topology and choose the tree that has the highest likelihood value as the maximum likelihood tree. However this

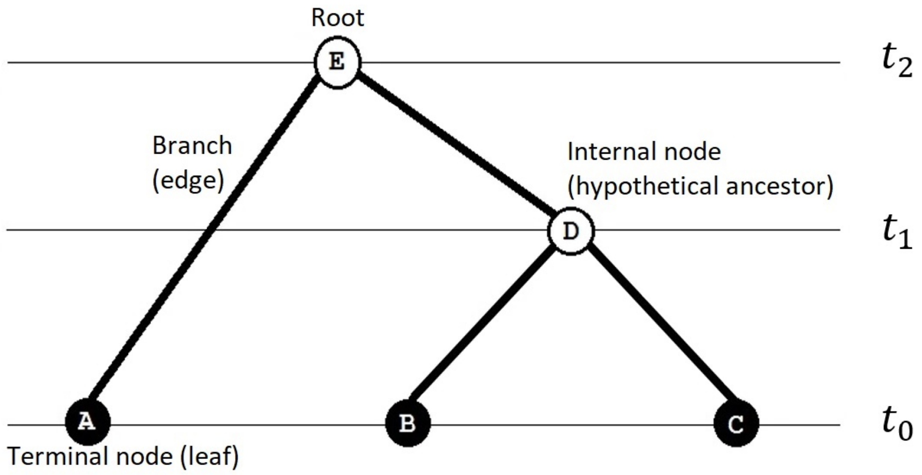


Figure 2: A simple tree and associated terms and times.

approach is only viable for trees with a small number of leaves and a correspondingly small number of possible topologies. For trees with a larger number of leaves the number of possible topologies increases exponentially, so exhaustively searching over this tree space is computationally infeasible. Various heuristics can be used to find the topology that has the highest likelihood, such as local modifications of a previously visited tree topology. For example, common methods traverse the tree topology space greedily by comparing the likelihood values between modified trees and by choosing the topology that increases the likelihood the most; the procedure will end if there are no trees that increase the likelihood (Dhar and Minin, 2016).

3.1. Phylogenetic covariance function

Consider a set of observations such that each corresponds to a point $(s, \mathfrak{t})$, where s is a continuous-spatial index, the arc-length, and $\mathfrak{t}$ is the location in the tree (at a specific time on a specific branch). To construct a covariance function for this type of data, two assumptions natural to the context of evolution are made. *Assumption 1*: Conditional on their common ancestors in the phylogenetic tree $\mathbf{T}$, any two function-valued traits are statistically independent. *Assumption 2*: The statistical relationship between a function-valued trait and any of its descendants in $\mathbf{T}$ is independent of the topology of the tree. In most cases, the covariance function is assumed to be separable in s and $\mathfrak{t}$, so that we need to specify just a spatial and a phylogenetic covariance function, as done by Jones and Moriarty (2013). The phylogenetic covariance function can be defined as:

$$k_{\mathbf{T}}(\mathfrak{t}_i, \mathfrak{t}_j) = \exp\left(-\frac{d_{\mathbf{T}}(\mathfrak{t}_i, \mathfrak{t}_j)}{\mu}\right), \tag{13}$$

where $d_{\mathbf{T}}(\mathfrak{t}_i, \mathfrak{t}_j)$ denotes the 'patristic' distance between $\mathfrak{t}_i$ and $\mathfrak{t}_j$, i.e., the distance along the path from $\mathfrak{t}_i$ to $\mathfrak{t}_j$. Consider the tree in Figure 2, and let t_0 be the time at present, i.e. $t_0 = 0$. The patristic distance between (terminal) nodes B and C is twice the age of their most common recent ancestor, i.e., the time back to node D: $d_{\mathbf{T}}(B, C) = 2t_1$. The time between A and D is the sum of the times from each of them to their most common recent ancestor, E: $d_{\mathbf{T}}(A, D) = t_2 + (t_2 - t_1) = 2t_2 - t_1$. The matrix of patristic distances for the tree is:

$$\begin{array}{c|c} & \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} & \begin{bmatrix} 0 & 2t_2 & 2t_2 & 2t_2 - t_1 & t_2 \\ 2t_2 & 0 & 2t_1 & t_1 & t_2 \\ 2t_2 & 2t_1 & 0 & t_1 & t_2 \\ 2t_2 - t_1 & t_1 & t_1 & 0 & t_2 - t_1 \\ t_2 & t_2 & t_2 & t_2 - t_1 & 0 \end{bmatrix} \end{array}. \tag{14}$$

The hyperparameter μ specifies the characteristic time scale for the evolutionary dynamics in the tree. As previously, the overall process variance is included in the spatial covariance function. Note that other non-exponential covariance functions built in terms of patristic distances can be employed (Anderes et al., 2017).

3.2. The model

Each curve can be represented as a GP which describes the joint distribution over arc-length s, indexing space, the discrete label c, indexing coordinates, and $\mathfrak{t}$, indexing the position in the tree. The GP can then be defined as:

$$\omega(\mathfrak{t}, c, s) \sim GP\big(m(\mathfrak{t}, c, s), k(\mathfrak{t}, \mathfrak{t}', c, c', s, s')\big). \tag{15}$$

For a series of spatial points $\mathbf{s} = (s_1 \cdots s_n)^{\mathrm{T}}$ and using the notation from previous section, points on a curve at node $\mathfrak{t}$ can be notated as $\mathbf{W}(\mathfrak{t}) = [\mathbf{c}_1(\mathfrak{t}), \ldots, \mathbf{c}_d(\mathfrak{t})]^{\mathrm{T}}$.

If l represents the number of terminal nodes in the tree, then the total number of nodes, m, is $2l - 1$ (in a fully resolved rooted bifurcating tree). If data are available at all nodes, the joint distribution for the set of points of all the curves in the tree is:

$$\mathbf{W_T} = \big[\mathbf{W}(\mathfrak{t}_1), \quad \ldots, \quad \mathbf{W}(\mathfrak{t}_m)\big]^{\mathrm{T}} \sim N_{mdn}\left(\mathbf{m}, \mathbf{K}\right), \tag{16}$$

where $\mathbf{m}$ represents the mean, again assumed to be zero and $\mathbf{K}$ is the covariance matrix. Once again, separability is assumed, so that $\mathbf{K} = \mathbf{K_T} \otimes \mathbf{K}_c \otimes \mathbf{K}_s$. Whilst $\mathbf{K}_c$ and $\mathbf{K}_s$ remain as before, $\mathbf{K_T}$ is the covariance matrix of points on curves at different nodes. The phylogenetic covariance function is used and, hence, this matrix has $(i, j)^{th}$ element equal to $k_{\mathbf{T}}(\mathfrak{t}_i, \mathfrak{t}_j)$, where $\mathfrak{t}_i$ is the position of the i^{th} node in the tree. If only data at the l terminal nodes are available, the distribution has its dimensions reduced, such that the set of curves at the leaves $\mathbf{W_L} \sim N_{ldn}\left(\mathbf{m}, \mathbf{K}\right)$. The covariance matrix $\mathbf{K}$ will have the same structure. It is important to note that the phylogenetic covariance matrix, although calculated only for the leaves, will reflect the relationship among all nodes, since it takes into account internal nodes to calculate the patristic distances between the leaves. The log-likelihood function and the profile log-likelihood follow from (16) as in Section 2.1.

3.3. Identifiability

The model proposed above assumes that the times of the nodes are known and the evolution rate μ is to be estimated. In practice, both the times and μ are unknown. The likelihood function only depends on these unknowns via the product of μ with the node times. So, deeper trees with larger values of μ result in the same likelihood value as shallower trees with smaller μ. The model is therefore non-identifiable and the likelihood surface contains ridges of equal (maximal) likelihood. One solution is to fix μ so that the node times are effectively being measured in some arbitrary units, not years or generations, as would be ideal (Yang, 2006). An independent calibration of μ would permit a conversion of node times into such units. Note that assuming that μ is constant across the tree implies that the rate of evolution does not change across the phylogeny. This is analogous to the 'molecular clock hypothesis' of molecular evolution (Thorpe, 1982).

3.4. Predictive distributions

In practice, the most common scenario is to have data available only at the leaves of the tree. In which case, it is natural to try to infer curves at internal nodes (ancestors). Predictions can be made both spatially and temporally. For the set of points on curves at the l leaves $\mathbf{W_L} \sim N_{ldn}\left(\mathbf{m}, \mathbf{K}\right)$, with $\mathbf{K} = \mathbf{K_T} \otimes \mathbf{K}_c \otimes \mathbf{K}_s$, marginal predictions at node $\mathfrak{q}$ can be made at a set of test points $\mathbf{s}^* = (s_1^*, \ldots, s_{n^*}^*)$ using the posterior predictive distribution $\mathbf{W}^*(\mathfrak{q}) \mid \mathbf{W_L}$ (conditioned on the maximum likelihood values of the hyperparameters):

$$\begin{aligned} \mathbf{W}^*(\mathfrak{q}) \mid \mathbf{W_L} \sim N_{dn^*}\Big(& \big(\big[\mathbf{L}\mathbf{K}_{\mathbf{T}}^{-1}\big] \otimes \mathbf{I_3} \otimes \big[\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\big]\big)\, \mathbf{W_L}, \\ & \mathbf{K}_c \otimes \mathbf{K}_{s^*} - \big(\, \big[\mathbf{L}\mathbf{K}_{\mathbf{T}}^{-1}\mathbf{L}^{\mathrm{T}}\big] \otimes [\mathbf{I_3}\mathbf{K}_c] \otimes \big[\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\big]\,\big)\Big), \end{aligned} \tag{17}$$

where $\mathbf{L}$ is the covariance matrix between the terminal nodes and node q:

$$\mathbf{L} = \left[\exp\left(-\frac{d_{\mathbf{T}}(\mathsf{t}_q, \mathsf{t}_1)}{\mu}\right), \quad \exp\left(-\frac{d_{\mathbf{T}}(\mathsf{t}_q, \mathsf{t}_2)}{\mu}\right), \quad \ldots, \quad \exp\left(-\frac{d_{\mathbf{T}}(\mathsf{t}_q, \mathsf{t}_l)}{\mu}\right)\right]. \tag{18}$$

4. Implementation

When fitting the model to real data, several problems may arise. The question of whether the hyperparameters should be optimised or drawn from their posterior distribution is a manifestation of the longstanding debate whether Bayesian or frequentist methods are more desirable. Applying Bayesian methods is not always straightforward. The difficulty with priors lies in the challenging process of assigning prior probabilities when prior beliefs may be very hard to express in terms of the models (Gibbs, 1998). Choosing the hyperparameters via maximum likelihood estimation (MLE) has become a widely used approach in the literature, partly because of the intuitive motivation of maximizing the probability of occurrence and partly because of its strong asymptotic properties (consistency and efficiency) (Robert, 2001). MLE was recommended, analysed and widely popularized by Ronald Fisher between 1912 and 1922 (Aldrich, 1997). Since then (Mardia and Marshall, 1984), it has been widely used for the estimation of hyperparameters in Gaussian Processes. In this research hyperparameters are chosen by MLE.

Points in facial curves tend to be highly correlated (due to their smoothness). This causes the covariance matrix to be potentially ill-conditioned. One possible approach is to use spectral decomposition, also known as eigen-decomposition. By expressing a positive semi-definite matrix in terms of its eigenvalues and eigenvectors, the inverse and determinant can be approximated by disregarding those eigenvalues that are very small. This produces the Moore-Penrose pseudoinverse (Ben-Israel and Greville, 2003). However, in some cases, maximisation of the likelihood can still be rather unstable even with the use of the spectral decomposition. Another option is to include a noise term in the model, as this causes the ratio between the largest and the smallest eigenvalue to decrease and hence the correlation matrix is no longer ill-conditioned. This is a reasonable strategy since the data acquisition process is not noise-free in practice.

As previously mentioned, to find the MLE of the parameters, an initial grid search was carried out to locate the local region where the maximum of the likelihood is located. The size of the grid grows exponentially with the dimension of the distribution so it is important to restrict the number of possible values of $\boldsymbol{\theta}$ at which the log-likelihood is computed, while assuring the coverage of the relevant parts of the distribution (Pietilainen, 2010). To reduce the number of hyperparameters, the optimal value of the signal variance that maximises the log-likelihood function (see (8) and (9)) was calculated. As the signal variance is taken out of the Squared Exponential covariance function, where measurement noise was added as an additional variance σ_n^2 on the diagonal, it is now accounted for by specifying a noise-to-signal ratio σ_n^2/σ_f^2 as an additive term on the diagonal of the new $\mathbf{K}_s$.

5. A case study: the evolution of nose shape within and between ethnic groups

To apply the model to real data, a small case study was conducted. Using the ©*Di3D* 3-dimensional surface-imaging device (`http://www.di4d.com/`), facial images were collected from volunteer subjects recruited in the local community. Ethical permissions were obtained from the Ethics Committee of the College of Science and Engineering, University of Glasgow. Three broad ethnic groups were selected for the study: African, European and Asian. The subjects consisted of 12 Sub-Saharan African, 20 British and 12 Chinese males. For each subject, two curves chosen to identify the nose were extracted. These curves are: the *midline nasal profile* (ridge points from the nasal root along the dorsum of the nose and the columella, defined by 28 equally-spaced points) and the *nasal bridge* (which outlines the width of the nose from one alar facial groove to the other, defined by 13 equally-spaced points). To be able to properly study the similarities and dissimilarities of the curves, Generalised Procrustes Analysis (GPA) was used to register the set of shapes into a common coordinate system. This is a standard precursor to exploration of the shape variability in a dataset. To compare the shape of two or more objects, the objects must first be optimally ‘superimposed’.

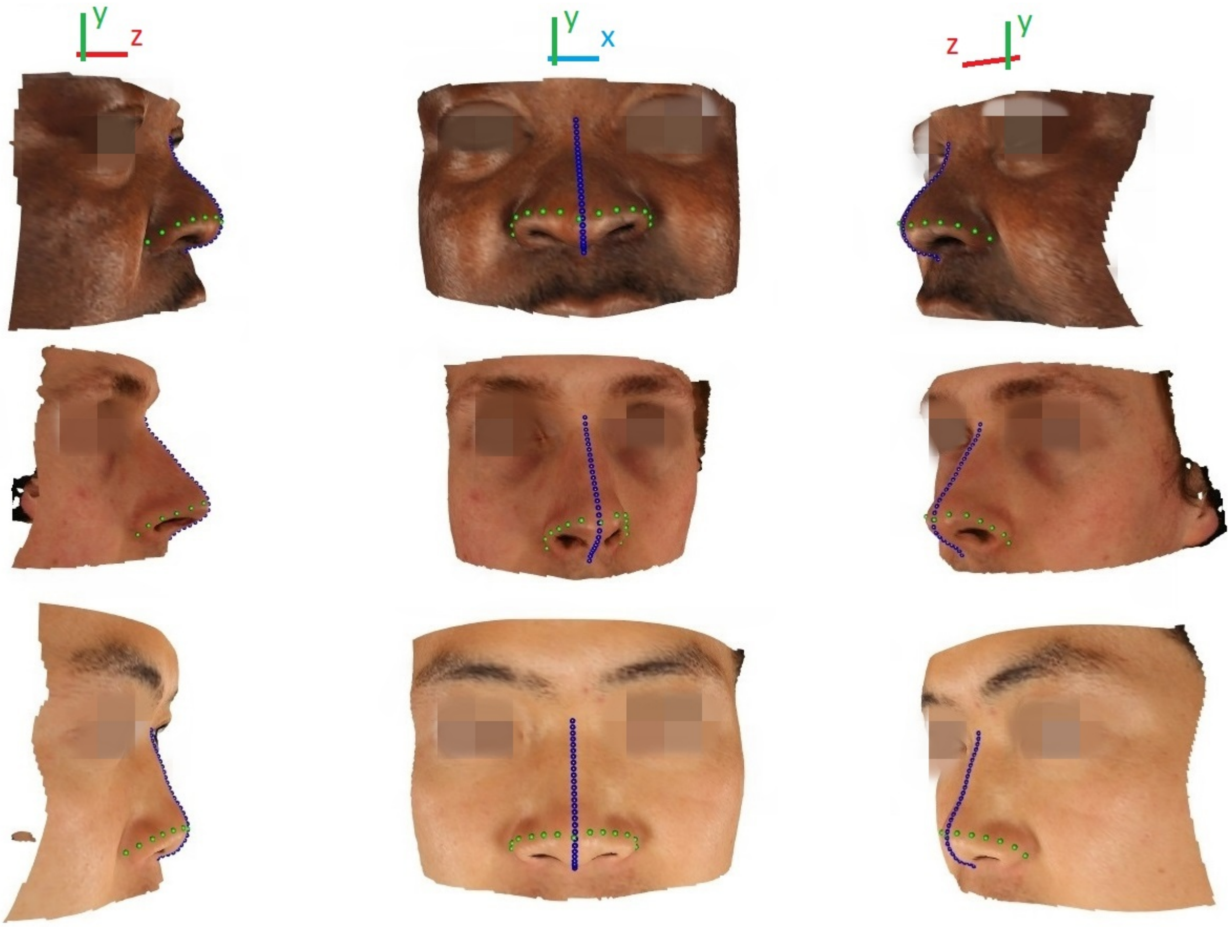


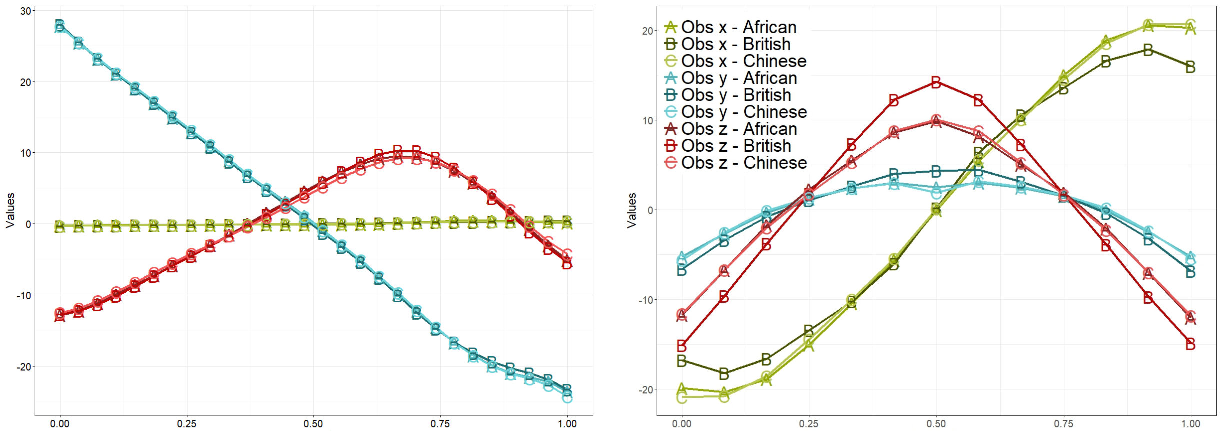
Figure 3: Nose curves: mid-line nasal profile (blue) and nasal bridge (green). From top to bottom: African, British and Chinese subjects.

GPA involves the superimposition of all configurations 'on top of each other' in optimal positions by translating, rotating and rescaling each figure so as to minimize the sum of squared Euclidean distances between corresponding points. The concept of GPA was originally proposed by Kristof and Wingersky (1971). A full description of the method can be found in Dryden and Mardia (1998). An example of the nose curves from one participant from each ethnic group is shown in Figure 3. The curves are superimposed on facial surfaces rotated at different angles for a better appreciation of the three-dimensional curves.

5.1. Evolution of mean nose shape

A natural starting point is a model for the evolution of the mean shape of the nose through a phylogenetic tree. For this, the corresponding three-dimensional points of each curve were averaged for each group, resulting in sets of mean points from the three ethnic groups for the two nasal features. These means can be plotted for each coordinate as a function of the arc-length rescaled from 0 to 1. Moreover, the points on each coordinate curve have the mean over that curve subtracted to match the assumption of zero mean in the GP model. The profile means are shown in Figure 4(a), with each group represented by its first letter, conveniently, A-African, B-British and C-Chinese. By the nature of the nose profile and the frontal orientation of the images, the x coordinate varies very little. For this reason, it was decided to model the data of the nose profiles as curves embedded in two dimensions, omitting coordinate x, while the data for the bridge curves is modelled as 3D curves. The means of the nasal bridges for each group are shown in Figure 4(b).

In both cases, for the two- and three-dimensional models, data are only available at the terminal nodes. Given there are three leaves, the three possible (rooted) topologies are $(A,(B,C))$, $(B,(A,C))$ and $(C,(A,B))$. Calling the root of the tree E and the remaining internal node D (Figure 2), and assigning the data at the leaves time zero, the times of the nodes A-E are $(0,0,0,t_D,t_E)$, where time increases going into the past. The hyperparameter μ has been set to one, to make the model identifiable. The model is reparametrised in terms of differences in node times: t_1, representing the time between node D and the leaves and t_2, representing the time from the root to node D.



(a) Mean mid-line nasal profile

(b) Mean nasal bridge

Figure 4: Points on the mean nose curves.

Let the sets of hyperparameters be $\boldsymbol{\theta}_{2\text{D}} = (\sigma_{f2\text{D}}, \lambda_{2\text{D}}, \kappa_{yz2\text{D}}, t_{12\text{D}}, t_{22\text{D}})$ and $\boldsymbol{\theta}_{3\text{D}} = (\sigma_{f3\text{D}}, \lambda_{3\text{D}}, \kappa_{xy3\text{D}}, \kappa_{yz3\text{D}}, \kappa_{xz3\text{D}}, t_{13\text{D}}, t_{23\text{D}})$ for the nose profile curves and the nasal bridges, respectively. Data from the two nasal curves can be modelled separately or jointly, by adding up the log-likelihood values of the two- and three-dimensional datasets and optimising the hyperparameters simultaneously. If both models are optimised simultaneously, there are three possible scenarios:

1. Assume the nose profile curves and the nasal bridges have evolved at the same rate, and have diverged at the same time points in history, i.e., $\boldsymbol{\theta} = (\sigma_{f2\text{D}}, \lambda_{2\text{D}}, \kappa_{yz2\text{D}}, \sigma_{f3\text{D}}, \kappa_{xy3\text{D}}, \kappa_{yz3\text{D}}, \kappa_{xz3\text{D}}, t_1, t_2)$.
2. Use the same time differences for both sets but allow for a scaling parameter multiplying the rate of change: μ is fixed to 1 for the 3D data, and a hyperparameter $\mu_{2\text{D}}$ is introduced, representing the relative rate of change of profile to bridge: $\boldsymbol{\theta} = (\sigma_{f2\text{D}}, \lambda_{2\text{D}}, \mu_{2\text{D}}, \kappa_{yz2\text{D}}, \sigma_{f3\text{D}}, \kappa_{xy3\text{D}}, \kappa_{yz3\text{D}}, \kappa_{xz3\text{D}}, t_1, t_2)$.
3. Allow for each set of curves to have different times t_1 and t_2, i.e., model each set of curves independently: $\boldsymbol{\theta} = (\boldsymbol{\theta}_{2\text{D}}, \boldsymbol{\theta}_{3\text{D}})$, which is equivalent to modelling them separately.

Hyperparameters were optimised by maximum likelihood for the three possible topologies and the topology with the largest log-likelihood value chosen. For every possible scenario and for modelling both sets of curves independently, the topology producing the highest values was $(B,(A,C))$. That is, the African and Chinese mean curves have a common ancestor more recently than each with the British.

Optimal hyperparameter values are shown in Table 1. Note the small differences in the values of the log-likelihood, particularly between the model for the same times and the model that introduces the scaling parameter $\mu_{2\text{D}}$. That scaling parameter is estimated to be close to one and its approximate 95% CI contains one. This implies that there is indeed no significant difference between the rate of change of the nose profiles and the nasal bridges. An error ratio η was added to the diagonal of the covariance matrices to accommodate errors in the observed values, defined as $\eta = \sigma_n^2/\sigma_f^2$. This was fixed to $\eta = 0.01$. Given the resulting optimal values for $\hat{\sigma}_{f\mathtt{2D}}$ and $\hat{\sigma}_{f\mathtt{3D}}$, the final additive normal error have standard deviations (σ_n) of approximately 0.74 mm and 0.86 mm, respectively.

Model selection can be performed to select the best scenario. The Bayesian information criterion (BIC) was calculated for each model: 169.71 for scenario 1, 175.35 for scenario 2 and 175.01 for scenario 3, favouring the model with fewest parameters (scenario 1).

1. Same times

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_{f\mathtt{2D}}$	$\hat{\lambda}_{\mathtt{2D}}$	$\hat{\kappa}_{\mathtt{22D}}$	$\hat{\sigma}_{f\mathtt{3D}}$	$\hat{\lambda}_{\mathtt{3D}}$	$\hat{\kappa}_{\mathtt{13D}}$	$\hat{\kappa}_{\mathtt{23D}}$	$\hat{\kappa}_{\mathtt{33D}}$	$\hat{t}_1$	$\hat{t}_2$
MLE	7.59	0.15	0.14	6.86	0.11	0.01	0.75	−0.01	0.0073	0.0107
SE	0.41	0.00	0.12	0.45	0.00	0.16	0.07	0.16	0.0017	0.0032
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	−56.59313									

2. Same times + Scaling parameter for μ

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_{f\mathtt{2D}}$	$\hat{\lambda}_{\mathtt{2D}}$	$\hat{\mu}_{\mathtt{2D}}$	$\hat{\kappa}_{\mathtt{22D}}$	$\hat{\sigma}_{f\mathtt{3D}}$	$\hat{\lambda}_{\mathtt{3D}}$	$\hat{\kappa}_{\mathtt{13D}}$	$\hat{\kappa}_{\mathtt{23D}}$	$\hat{\kappa}_{\mathtt{33D}}$	$\hat{t}_1$	$\hat{t}_2$
MLE	7.58	0.15	0.95	0.14	6.96	0.11	0.01	0.76	−0.02	0.0069	0.0104
SE	0.48	0.00	0.38	0.12	0.45	0.00	0.16	0.06	0.16	0.0024	0.0036
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	−56.58461										

3. Different times

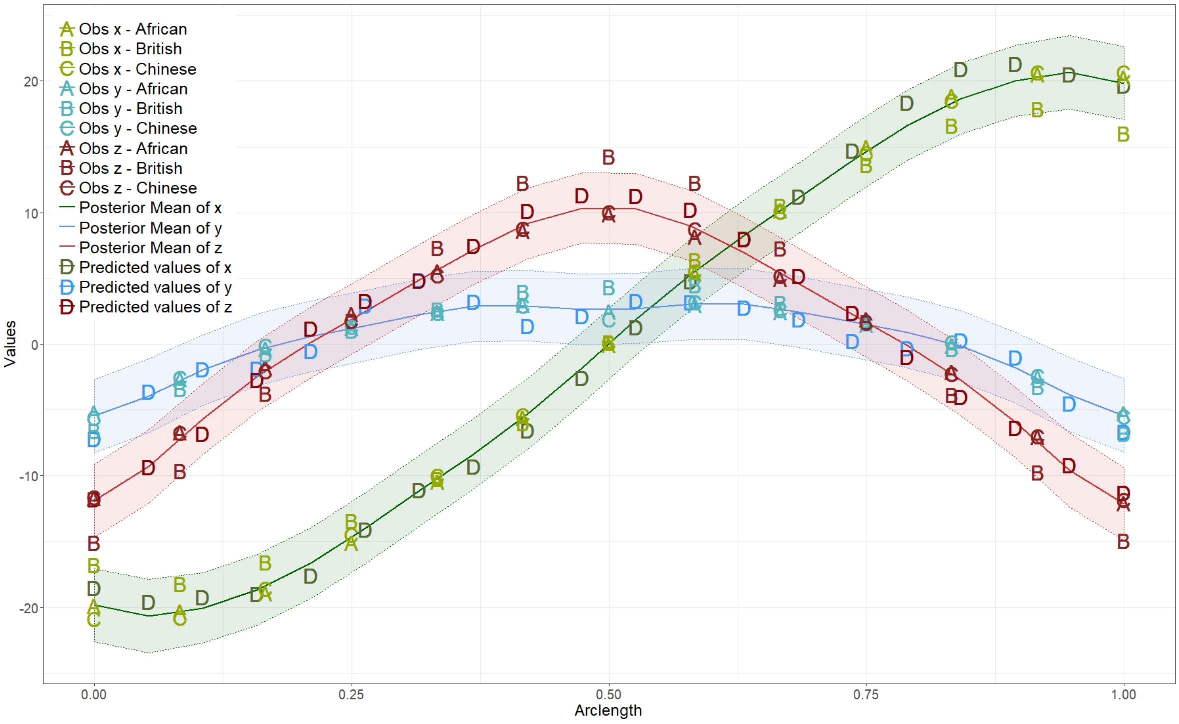
$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_{f\mathtt{2D}}$	$\hat{\lambda}_{\mathtt{2D}}$	$\hat{\kappa}_{\mathtt{22D}}$	$\hat{t}_{\mathtt{12D}}$	$\hat{t}_{\mathtt{22D}}$	$\hat{\sigma}_{f\mathtt{3D}}$	$\hat{\lambda}_{\mathtt{3D}}$	$\hat{\kappa}_{\mathtt{13D}}$	$\hat{\kappa}_{\mathtt{23D}}$	$\hat{\kappa}_{\mathtt{33D}}$	$\hat{t}_{\mathtt{13D}}$	$\hat{t}_{\mathtt{23D}}$
MLE	7.37	0.15	0.07	0.0091	0.0062	6.87	0.11	0.03	0.74	−0.00	0.0036	0.0149
SE	0.46	0.00	0.12	0.0026	0.0035	0.45	0.00	0.16	0.07	0.16	0.0012	0.0052
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	−53.59132											

Table 1: Optimal hyperparameters for the three scenarios.

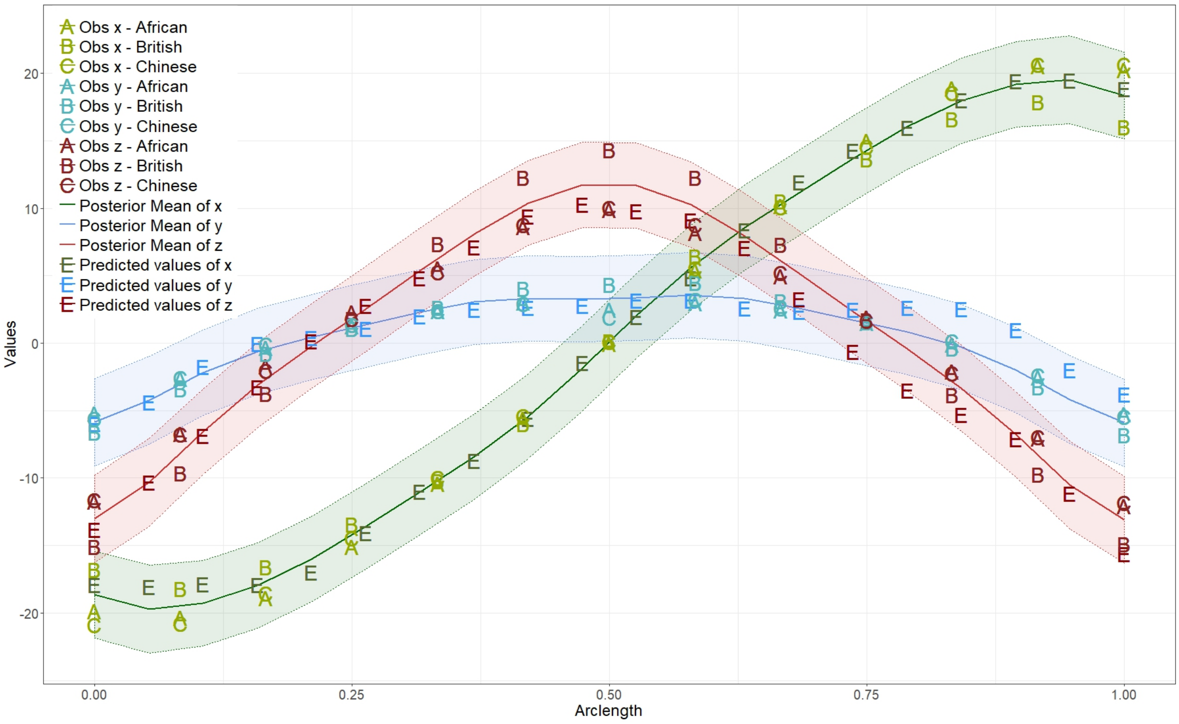
The topology $(B,(A,C))$, which had the maximal log-likelihood value, was used to reconstruct unobserved ancestral mean curves at the internal node D and at the root, E. The maximum likelihood estimates from scenario 1 were used to make predictions at 30 equally-spaced spatial points, $\mathbf{s}^*_{\mathtt{2D}}$, for the profile curves and 20 spatial points, $\mathbf{s}^*_{\mathtt{3D}}$, also equally spaced, for the nasal bridges. As there is little variation among the nose profile curves, only the reconstructed values for the nasal bridges are shown here.

The predicted curves at node D, the common ancestor of A (African) and C (Chinese), are shown in Figure 5(a). The posterior predictive mean is much closer to A and C than to B (British). The observed values of B differ from A, C and the predictions at D considerably, with the middle spatial-points not even contained in the confidence bands of D. The reconstructions for the root of the tree, node E, are shown in Figure 5(b). Note how much wider the confidence bands are in this case, reflecting the fact that extrapolations are being made even further back in time from the observed values. Note also how the posterior mean of D is closer to the observed values of A and C, than the posterior mean of E is to all the observed values. Nonetheless, the data are measured in millimetres, and therefore, even the widest bands represent uncertainty of no more than 4 mm. The predicted values shown for D and E are one random draw using the predictive distributions (17).

Using Procrustes analysis, the posterior means can be translated back to the original 3D space of the nose curves, allowing a more comprehensive exploration of the ancestors' nose shape. These shapes are illustrated in Figure 6. Different angles of view can be seen when viewed in digital form (using Adobe reader `https://get.adobe.com/uk/reader/`). The means of the three ethnic groups are displayed in blue, from the African in the lightest blue, from the British in an intermediate shade and from the Chinese in the darkest shade of blue.

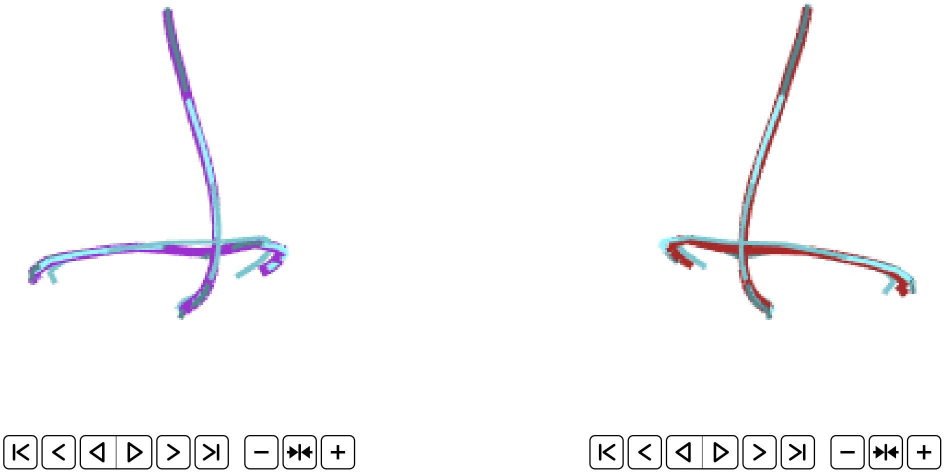


(a) Node D



(b) Node E

Figure 5: Observations along with posterior means, one realization of predicted values and 2-standard deviation credible bands for nasal bridge curves at nodes D and E.



(a) Node D

(b) Node E

Figure 6: Posterior mean for reconstruction at ancestral nodes D and E, displayed with the means of the original data. (www.stats.gla.ac.uk/~adrian/phylogenetic-noseD.mp4 and www.stats.gla.ac.uk/~adrian/phylogenetic-noseE.mp4)

5.2. Inter- and intra-group variation in nose shape

In the study described above, the mean nose curves for each ethnic group have been modelled. This approach does not address the variability within groups. Although the sample size in each group is small for a comprehensive analysis, a model for all the data is described here. All the curves in each ethnic group are assumed to have simultaneously diverged from one common ancestor, different between groups. These assumptions were made to avoid a very challenging search over tree space, which would be a natural further development. Let the common ancestral node of Africans $A_1, \ldots, A_{12}$ be A, of British $B_1, \ldots, B_{20}$ be B and of Chinese $C_1, \ldots, C_{12}$ be C. As before, to select the best topology, optimal hyperparameters were found by maximum likelihood for each of the three possible topologies. From these results, the topology with the highest log-likelihood value is chosen. Two trees were studied, one for the evolution of nasal bridge curves and one for the nose profile curves. In both cases the topology with the highest log-likelihood was, as with the mean curves, $(B,(A,C))$. For the nasal bridges, the optimal hyperparameters were $\hat{\boldsymbol{\theta}}_{\texttt{3D}} = (\hat{\sigma}_{f\texttt{3D}}, \hat{\lambda}_{\texttt{3D}}, \hat{\kappa}_{xy\texttt{3D}}, \hat{\kappa}_{yz\texttt{3D}}, \hat{\kappa}_{xz\texttt{3D}}, \hat{t}_A, \hat{t}_B, \hat{t}_C, \hat{t}_D, \hat{t}_E) = (7.17, 0.08, 0.11, 0.02, -0.02, 0.0049, 0.0055, 0.0028, 0.0055, 0.0114)$. The corresponding tree is displayed in Figure 7.

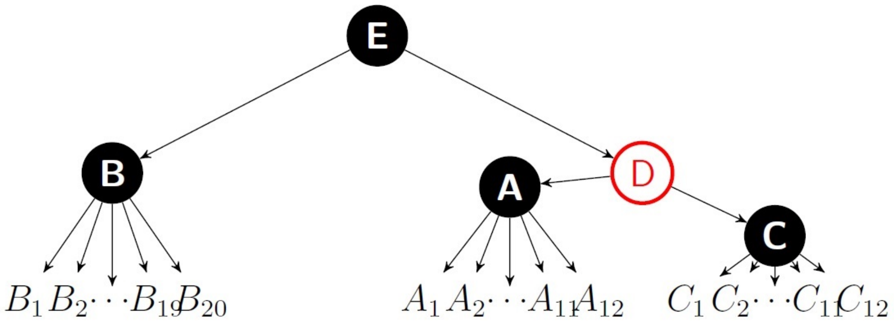
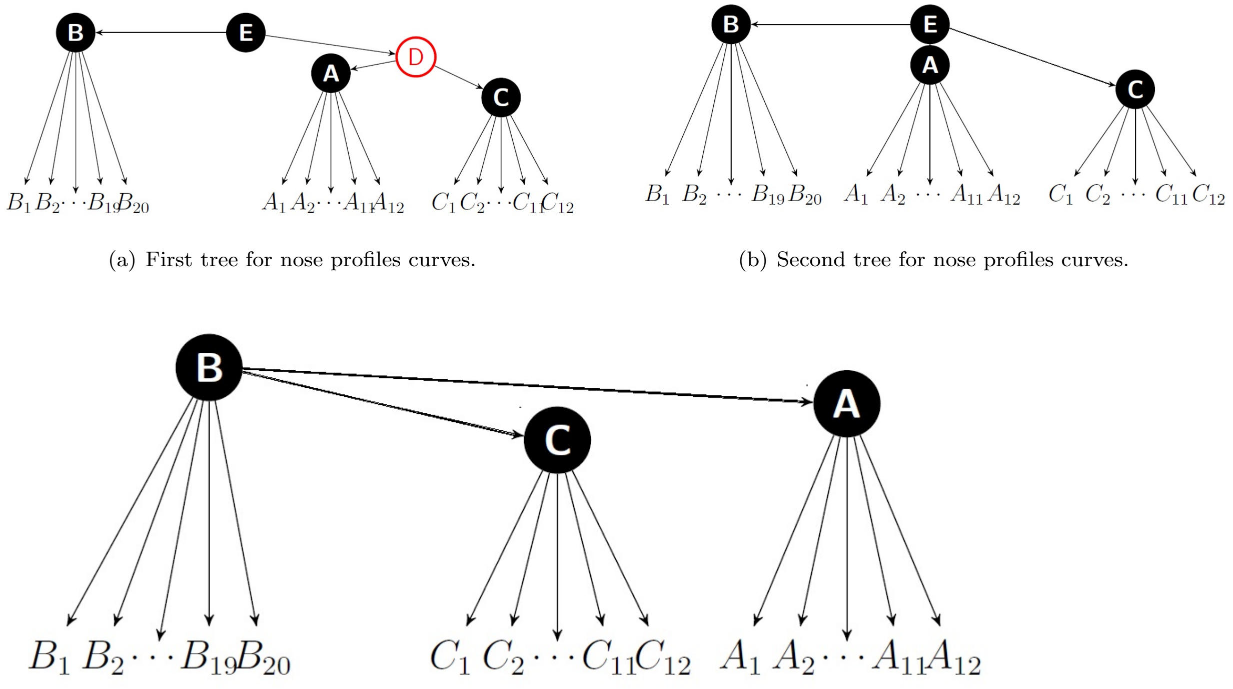

E
B
A
D
C
$B_1 B_2 \cdots B_{19} B_{20}$
$A_1 A_2 \cdots A_{11} A_{12}$
$C_1 C_2 \cdots C_{11} C_{12}$

Figure 7: Tree for nasal bridges, with branch lengths corresponding to the fitted model.

When fitting the multifurcating tree for the nose profiles, problems arose due to the very small variability observed between the curves. The differences between node times are even smaller than those found for the nasal bridges. The root of the tree had the same estimated time as that of the common ancestor for the British curves, i.e., $\hat{t}_B = \hat{t}_E$, (see Figure 8(a)). The estimated time difference between D and E lay on the boundary set by the requirement that the root cannot be younger than any other node. In addition, if an estimate lies on a boundary, the SEs calculated from the Hessian matrix are not reliable. Proceeding heuristically, with the aim of avoiding a model with parameter estimates on the boundary, and since the branch from D to E was estimated as short as it could be, a second multifurcating tree with no internal node D was explored. In this scenario, nodes B and E were still estimated at the same time (see Figure 8(b)). Since both trees had estimated $\hat{t}_B = \hat{t}_E$, it was decided to set node B to be the root of the tree. This final multifurcating tree is shown in Figure 8(c), with the branch lengths, as usual, proportional to the optimal time differences. The final optimal hyperparameters were $\hat{\boldsymbol{\theta}}_{\mathtt{2D}} = (\hat{\sigma}_{f\mathtt{2D}}, \hat{\lambda}_{\mathtt{2D}}, \hat{\kappa}_{yz\mathtt{2D}}, \hat{t}_A, \hat{t}_B, \hat{t}_C) = (7.15, -0.0532, -0.0888, 0.0021, 0.0024, 0.0018)$, which are not on the boundary.

The estimates for the hyperparameters not related to time remained stable across the three different trees. Moreover, the age of B was stable and the age of A always larger than the age of C. This has the interpretation of greater variability amongst the British curves than within the other ethnic groups, which agrees with the topologies previously estimated that had the British evolving more separately. The Chinese group has the smallest variability in nose profile. The final maximised log-likelihood value was around 3 units smaller than in the original tree, but there are two hyperparameters fewer in the later tree. In a likelihood ratio test, the simpler model with fewer hyperparameters cannot be rejected. Furthermore, there are no estimates lying on the boundary.


B
E
D
A
C
$B_1\, B_2 \cdots B_{19} B_{20}$
$A_1\, A_2 \cdots A_{11} A_{12}$
$C_1\, C_2 \cdots C_{11} C_{12}$
B
E
A
C
$B_1\, B_2 \cdots B_{19} B_{20}$
$A_1\, A_2 \cdots A_{11} A_{12}$
$C_1\, C_2 \cdots C_{11} C_{12}$
B
C
A
$B_1\, B_2 \cdots B_{19} B_{20}$
$C_1\, C_2 \cdots C_{11} C_{12}$
$A_1\, A_2 \cdots A_{11} A_{12}$

(a) First tree for nose profiles curves.

(b) Second tree for nose profiles curves.

(c) Final tree for nose profiles curves.

Figure 8: Model selection for topology of tree for nose profiles.

5.3. *Further models*

The particular configuration of facial curves studied here has a common point at the tip of the nose. This raises the question of whether a new model which allows correlation between the curves would be beneficial. This requires a likelihood which is constructed jointly, rather than as a product of independent components. As defined in Section 3, each curve evolving along a phylogenetic tree is defined as

$$\mathsf{W}_{\mathbf{T}} = \begin{bmatrix}\mathsf{W}(\mathsf{t}_1) & \dots & \mathsf{W}(\mathsf{t}_m)\end{bmatrix}^{\mathrm{T}} \sim N_{mdn}\left(\mathbf{m}, \mathbf{K}_{\mathbf{T}} \otimes \mathbf{K}_c \otimes \mathbf{K}_s\right). \tag{19}$$

Let $\mathsf{B}(\mathsf{t})$ denote the nasal bridge curve (3D) at position t in the tree and $\mathsf{P}(\mathsf{t})$, the mid-line nasal profile (2D), and let $\mathbf{c}$ be the common point of the two nose curves, located at arc-length s_p in the mid-line nasal profile and s_b in the nasal bridge. Then:

$$\begin{bmatrix}\mathsf{B}(\mathsf{t})\\ \mathsf{P}(\mathsf{t})\end{bmatrix} \sim N_{3n_b+2n_p}\left(\mathbf{0}, \mathbf{K}_{cs}\right), \tag{20}$$

where $\mathbf{K}_{cs}$ is defined as:

$$\mathbf{K}_{cs} = \begin{bmatrix}\mathbf{K}_{c_b} \otimes \mathbf{K}_{s_b} & \mathbf{K}_{c_{bp}} \otimes \mathbf{K}_{s_{bp}}\\ \left[\mathbf{K}_{c_{bp}} \otimes \mathbf{K}_{s_{bp}}\right]^{\mathrm{T}} & \mathbf{K}_{c_p} \otimes \mathbf{K}_{s_p}\end{bmatrix}, \tag{21}$$

with:

- $\mathbf{K}_{c_b} = \begin{pmatrix}1 & \kappa_1 & \kappa_2\\ \kappa_1 & 1 & \kappa_3\\ \kappa_2 & \kappa_3 & 1\end{pmatrix}$.
- $\mathbf{K}_{s_b}$ has $(i,j)^{th}$ element equal to $\exp\left(-\frac{1}{2}(s_i - s_j)^2/\lambda_b^2\right)$.
- $\mathbf{K}_{c_p} = \begin{pmatrix}1 & \kappa_4\\ \kappa_4 & 1\end{pmatrix}$.
- $\mathbf{K}_{s_p}$ has $(i,j)^{th}$ element equal to $\exp\left(-\frac{1}{2}(s_i - s_j)^2/\lambda_p^2\right)$.
- $\mathbf{K}_{c_{bp}} = \begin{pmatrix}\kappa_c & \kappa_c\\ \kappa_c & \kappa_c\\ \kappa_c & \kappa_c\end{pmatrix}$.
- $\mathbf{K}_{s_{bp}}$ has $(i,j)^{th}$ element equal to $\exp\left(-\frac{1}{2}\left[\frac{(s_i - s_b)^2}{\lambda_b^2} + \frac{(s_j - s_p)^2}{\lambda_p^2}\right]\right)$.

The key point is that the elements of the between-curves covariance matrix $\mathbf{K}_{s_{bp}}$ are based on the sum of the distances from point i on one curve to the common point $\mathbf{c}$, and from $\mathbf{c}$ to the point j on the other curve. To contain the rise in the number of hyperparameters, the correlation between coordinates from different curves in $\mathbf{K}_{c_{bp}}$ are set to a common value, κ_c.

The joint sequence of curves along the phylogenetic tree can then be written as:

$$\begin{bmatrix}\mathsf{B}(\mathsf{t}_1) & \mathsf{P}(\mathsf{t}_1) & \dots & \mathsf{B}(\mathsf{t}_m) & \mathsf{P}(\mathsf{t}_m)\end{bmatrix}^{\mathrm{T}} \sim N_{m(3n_b+2n_p)}\left(\mathbf{0}, \mathbf{K}_{\mathbf{T}} \otimes \mathbf{K}_{cs}\right). \tag{22}$$

This model was fitted to the mean curves presented in Section 5.1, with the common point set at $s_b = 0.5$ and $s_p = 0.67$. Previously, as there were two separated likelihood components, the signal variance was calculated independently for each curve. An assumption of one overall σ_f is reasonable, as the confidence intervals for variances of the two curves overlap. The hyperparameters in the new joint model were optimised for the three possible topologies and, as previously, the topology with the highest value was $(B,(A,C))$. The optimal hyperparameters are shown in Table 2.

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_f$	$\hat{\lambda}_s$	$\hat{\lambda}_p$	$\hat{\kappa}_1$	$\hat{\kappa}_2$	$\hat{\kappa}_3$	$\hat{\kappa}_4$	$\hat{\kappa}_c$	$\hat{t}_1$	$\hat{t}_2$
MLE	6.64	0.09	0.01	0.03	-0.001	0.38	0.02	0.06	0.0021	0.0061
SE	0.278	0.004	0.003	0.108	0.092	0.112	0.082	0.155	0.0002	0.0013
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	-334.2679									

Table 2: Optimal hyperparameters for the model with one common point.

The hyperparameter estimate $\hat{\kappa}_c = 0.06$ has SE $= 0.155$ and so is not significantly different from 0. This does not provide convincing evidence of correlation among the coordinates of the different curves. In addition, in comparison with the model of Scenario 1 in Section 5.1, which has the same number of hyperparameters, the log-likelihood for the joint model is very much smaller. So, while the new model demonstrates how correlation between curves can be incorporated, there is no evidence that this is necessary or helpful in the current example.

6. Discussion

The use of GP models combined with a phylogenetic covariance function provides a new, powerful, tool for the study of shape combined with genealogical analysis. In this work, the GP (prior) mean has been assumed to be zero. This assumption is not a drastic limitation and could be relatively easily relaxed. In any case, the covariance function controls how rapidly (and where) deviations from the mean function occur and so a more flexible covariance function can, to some extent, substitute for a more structured mean function; where to put the effort is somewhat a matter of taste. Further, separability is assumed for the covariance function. Whilst there has been much development in the use of cross-covariance functions for multivariate data, the theoretical characterization of the allowable classes of multivariate covariances is still ambiguous and it is not clear, given a certain number of marginal covariances, what is the valid class of possible cross-covariances that still results in a non-negative definite structure (Genton and Kleiber, 2015). Further lines of investigation might include a study of non-separable covariance functions, such as convolved GPs (Shi and Choi, 2011).

In simulations (not shown), the model performed well, capturing adequately the covariance structure in space and time, for curves embedded in two or three dimensions. In particular, the model performs well even when data are available only at the leaves, which will be the case in most applications. This is shown with the study of the evolution of nose shape. The predictive distributions provide a powerful tool to estimate ancestral shapes. Spatial marginal predictions to interpolate the data at any measured node can also be made but most interest lies in being able to reconstruct data which one could never directly obtain. In the second study, where all the curves at the leaves are studied, prediction could also be carried out for the common ancestor of each ethnic group. Given the small amount of data available, the predictions in this case are not expected to differ largely from the mean of the curves in each.

When studying the structure of the tree for the mean nose curves for the three ethnic groups A-African, B-British and C-Chinese, the first thing observed is that the optimal topology links A and C with one common ancestor more recent than the common ancestor of the three groups. From the genetic study of ethnic groups, there is general agreement that the human lineage evolved in Africa and then spread to southern Eurasia as *H. erectus*. After the evolution of modern humans in Africa, a second expansion occurred out of Africa between 60000-80000 years ago that resulted in a global replacement (Macaulay et al., 2005). Therefore, it might have been expected to find a topology that links B and C under one more recent ancestor, having A evolving on their own. These results are based on analyses of DNA variation assuming no strong role for natural selection. However, it could well be that the morphology of the nose has evolved to adapt to different environmental conditions. Noses adapted to cold weather may function differently from those that evolved in hot and humid climates. People of African descent typically have shorter noses, with wider nostrils, whilst people of northern European descent typically have longer, thinner, noses. Individuals from cold, dry climates have higher and narrower nasal cavities than those from hot, humid climates (Noback et al., 2011). The clustering of African and Asian noses then would reflect convergent evolution of nasal shape rather than an ancestral population relationship.

This topology was found to be optimal for both sets of nose curves for the mean of each group, as well as when using the within-group variation of the nasal bridges. The equivalent topology for the nose profiles (Figure 8(c)) looks at first sight distinct, with the British ancestor sitting at the root. However the multifurcation of all British with the ancestor of Chinese and sub-Saharan Africans is most naturally interpreted as reflecting uncertainly in the tree topology due to the limited information in the variation of nasal profiles and indeed that tree can be resolved into the corresponding tree of nasal bridges. The proposal to model both nasal curves together (Section 5.3), recognizing that they must become more correlated as they near their crossing point, even if it did not prove a significantly better fit, at least led to the same tree topology.

The results presented here provide an illustration of what these models can accomplish. A larger study with more subjects, that takes into account more ethnic and sub-ethnic groups, would permit one to test the idea that nose shape is correlated with climate condition. The models could clearly also be applied to other facial curves. If data could be collected from various members of a family, it would be interesting to model the facial morphology within its members. Even more powerful methods of analysis could be developed through fusions of genetic and shape information. The results presented here open the door to this interdisciplinary field.

References

Aldrich, J., 1997. R. A. Fisher and the making of maximum likelihood 1912-1922. Statist. Sci. 12, 162–176. URL: http://dx.doi.org/10.1214/ss/1030037906, doi:10.1214/ss/1030037906.

Anderes, E., Møller, J., Rasmussen, J.G., 2017. Isotropic covariance functions on graphs and their edges. ArXiv e-prints arXiv:1710.01295.

Ben-Israel, A., Greville, T.N., 2003. Generalized inverses: theory and applications. volume 15. Springer Science & Business Media.

Darwin, C., 1859. On the Origin of Species by Means of Natural Selection. J. Murray.

Dhar, A., Minin, V., 2016. Maximum likelihood phylogenetic inference, in: Kliman, R.M. (Ed.), Encyclopedia of Evolutionary Biology. Academic Press, pp. 499–506.

Dryden, I., Mardia, K., 1998. Statistical Shape Analysis. John Wiley.

Felsenstein, J., 1985. Phylogenies and the comparative method. The American Naturalist 125, 1–15.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates.

Franciscus, R.G., Trinkaus, E., 1988. Nasal morphology and the emergence of Homo erectus. American Journal of Physical Anthropology 75, 517–527.

Genton, M.G., Kleiber, W., 2015. Cross-covariance functions for multivariate geostatistics. Statist. Sci. 30, 147–163. URL: https://doi.org/10.1214/14-STS487, doi:10.1214/14-STS487.

Gibbs, M., 1998. Bayesian Gaussian Processes for Regression and Classification. Ph.D. thesis. University of Cambridge, England.

Gregory, T.R., 2009. Understanding natural selection: essential concepts and common misconceptions. Evolution: Education and Outreach 2, 156–175. URL: https://doi.org/10.1007/s12052-009-0128-1.

Huson, D.H., Rupp, R., Scornavacca, C., 2010. Phylogenetic Networks: Concepts, Algorithms and Applications. Cambridge University Press.

Jones, N.S., Moriarty, J., 2013. Evolutionary inference for function-valued traits: Gaussian process regression on phylogenies. Journal of the Royal Society Interface 10, 20120616.

Kristof, W., Wingersky, B., 1971. Generalization of the orthogonal procrustes rotation procedure for more than two matrices. Proceedings of the Annual Convention of the American Psychological Association 6, 89–90.

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., Blackburn, J., Semino, O., Scozzari, R., Cruciani, F., et al., 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308, 1034–1036.

Mardia, K.V., Marshall, R.J., 1984. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika 71, 135–146. URL: http://www.jstor.org/stable/2336405.

Mariñas–Collado, I., 2017. Statistical Models for the Evolution of Facial Curves. Ph.D. thesis. University of Glasgow.

Mladina, R., Skitarelić, N., Vuković, K., 2009. Why do humans have such a prominent nose? The final result of phylogenesis: a significant reduction of the splanchocranium on account of the neurocranium. Medical Hypotheses 73, 280–283. URL: http://www.sciencedirect.com/science/article/pii/S0306987709002588.

Noback, M.L., Harvati, K., Spoor, F., 2011. Climate-related variation of the human nasal cavity. American Journal of Physical Anthropology 145, 599–614.

Page, R., Holmes, E., 1998. Molecular evolution: a phylogenetic approach. Blackwell .

Pietilainen, V., 2010. Approximations for integration over the hyperparameters in gaussian processes .

Rasmussen, C., Williams, C., 2006. Gaussian Processes for Machine Learning. MIT Press.

Robert, C., 2001. The Bayesian Choice: from Decision-theoretic Foundations to Computational Implementation. Springer Texts in Statistics, Springer, New York. URL: http://opac.inria.fr/record=b1097199.

Shi, J.Q., Choi, T., 2011. Gaussian process regression analysis for functional data. Chapman and Hall/CRC.

Tanaka, H.T., Ikeda, M., Chiaki, H., 1998. Curvature-based face surface recognition using spherical correlation. principal directions for curved object recognition, in: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 372–377. doi:10.1109/AFGR.1998.670977.

Thorpe, J.P., 1982. The molecular clock hypothesis: biochemical evolution, genetic differentiation and systematics. Annual Review of Ecology and Systematics 13, 139–168.

Vittert, L., Bowman, A., Katina, S., 2017. Statistical models for manifold data with applications to the human face. arXiv:1701.07328 .

Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press.