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

Irene Mariñas-Collado[*] Adrian Bowman Vincent Macaulay

June 14, 2019

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; nose shape;

1 State of the art

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.

[*] Irenemc@usal.es

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. During the last few decades several authors have tried to clarify the anthropological aspects of the shape of the human nose (Mladina et al., 2009). Modern humans possess a unique projecting, external nose whose basic structure is reflected in a series of skeletal features (Franciscus and Trinkaus, 1988). The approach proposed here is to use shape information, more specifically facial curves, to study relationships between different ethnic groups and their (our) ancestors. This work is presented as a complement to methods that use genetic information, on which extensive research has been carried out (Page and Holmes, 1998). The statistical modelling of a d-dimensional curve that changes over time usually involves the reduction to low-dimensional summaries or discrete characters. Many authors have considered the motion of curves in the plane through changes in curvature and direction of the normal (Kimmel et al., 1997). The use of shape information, expressed on a continuous and multivariate scale, raises a number of very interesting issues from a methodological perspective. The main challenge is to model the data without sacrificing information, as traditionally happens, for example, in distance-based methods (Singh, 2015), while maintaining computational feasibility. Jones and Moriarty (2013) presented a flexible statistical model, 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.

The statistical analysis of information on shape has been a research topic of considerable interest since the earliest part of the twentieth century, but it has developed considerably in the present century, thanks to advances in computer tools especially. Recent advances in image-capture technologies have made available detailed three-dimensional facial images, which permit a quantitative investigation of different issues of medical and biological interest.

2 Contributions

This work presents an approach for modelling the evolution of d-dimensional curves. Two scales of evolution are considered. First, time is modelled as a linear continuous variable, i.e., one curve that is gradually changing in a particular situation. This can be regarded as a degenerate scenario of the phylogenetic GP model, when, inside the branching structure of a phylogenetic tree, one focuses on the evolution of one single curve along one branch, without taking into account the branching patterns and the ancestors. The model is extended, using the phylogenetic covariance function to allow for branching points in the evolution.

A curve embedded in a d-dimensional space changing over time can be parametrised in terms of its arc-length s (a continuous index that can be rescaled to be from 0 to 1), the set of discrete coordinate labels, $\{c_1, \ldots, c_d\}$, and a time component $\mathtt{t}$. A GP can be then specified as:

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

where the two arguments of the GP refer to the mean and covariance function and where $c, c' \in \{c_1, \ldots, c_d\}$ and $s, s' \in [0, 1]$. The time component differs according to the type of evolution, thus $\mathtt{t} = t \in \mathbb{R}$ when it is considered a linear continuous variable, and $\mathtt{t} = \mathrm{t} \in \mathbf{T}$ is the location in the tree (at a specific time on a specific branch) when considering phylogenetic evolution.

Since separability is assumed, different covariance matrices can be used. The Squared-Exponential function is used for the space-covariance matrix and a $d \times d$ matrix conveys the correlations between coordinates. When modelling time as a linear continuous variable, the process is assumed Markovian and the Ornstein-Uhlenbeck covaraince function used. The phylogenetic covariance function is used for the phylogenetic model. The likelihood functions of the models and predictive distributions are presented, together with a discussion of the model implementation. The predictive distributions provide a powerful tool, specially in the phylogenetic setting, 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. To see the model perform on real data, a small case study is presented, where the nose shape of three different ethnic groups is studied. The diversity of facial features across human populations and the evolutionary reasons for variation in nose shape across human populations have been subject to debate in recent years. Typically, studies are based on DNA analysis, and it could be that the morphology of the nose has evolved differently to adapt to different environmental conditions. Noses adapted to cold weather may function differently from those that evolved in hot and humid climates.

The findings in this work provide a new, powerful, tool for the study of shape combined with genealogical analysis. Even more powerful methods of analysis could be developed through fusions of genetic and shape information. The results presented here are meant just as an illustration of what these models could accomplish and leave an open door for this interdisciplinary field.

References

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

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

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates.

Franciscus, R. G. and Trinkaus, E. (1988). Nasal morphology and the emergence of Homo erectus. *American Journal of Physical Anthropology*, 75(4):517–527.

Gregory, T. R. (2009). Understanding natural selection: essential concepts and common misconceptions. *Evolution: Education and Outreach*, 2(2):156–175.

Jones, N. S. and Moriarty, J. (2013). Evolutionary inference for function-valued traits: Gaussian process regression on phylogenies. *Journal of the Royal Society Interface*, 10(78):20120616.

Kimmel, R., Kiryati, N., and Bruckstein, A. M. (1997). Analyzing and Synthesizing Images by Evolving Curves with the Osher-Sethian Method. *International Journal of Computer Vision*, 24(1):37–55.

Mladina, R., Skitarelić, N., and 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(3):280–283.

Page, R. and Holmes, E. (1998). Molecular evolution: a phylogenetic approach. *Blackwell Science Oxford*.

Singh, G. B. (2015). *Distance Based Methods*, pages 253–260. Springer International Publishing, Cham.