Functional phylogenetic Gaussian process regression:

Statistical Methods and Numerical Examples

P.Z. Hadjipantelis[1], J. Moriarty[2], D. Springate[2] and C.G. Knight[2]

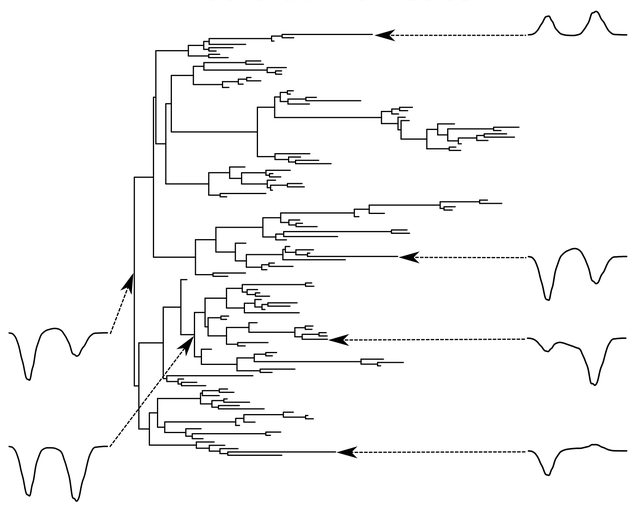
[1]University of Warwick, Coventry, UK [2]University of Manchester, Manchester, UK

THE UNIVERSITY OF WARWICK

Introduction

- Ancestral Inference is an integral part for deducing the evolutionary dynamics behind a phylogeny
  - relates to how fast a trait evolves
  - has strict physiological characteristics
  - most modelling approaches ignore curve/functional nature
- Phylogeny-specific issues
  - information is typically only available for extant organisms
  - a phylogeny describes a complex pattern of non-independence
- Current proposal
  - characterizes the trait we are interested as a curve; the realization of a stochastic gaussian process
  - identifies variations due to non-phylogenetic effects

Random Tree with Function-valued data at the Nodes

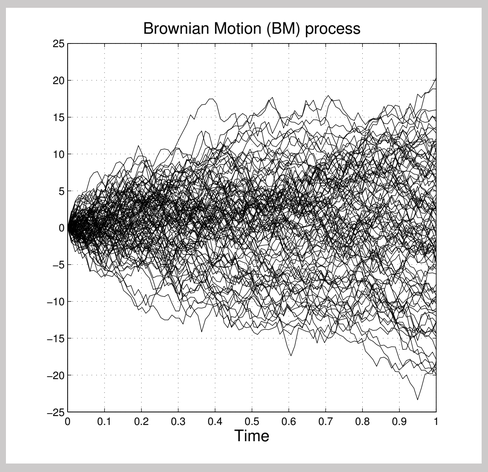


Stochastic Processes in Evolutionary Biology

- Causes of Evolution
  - Natural Selection
  - Drift
  - Founder Effects
- By far, the most popular model evolution is that of Brownian Motion.
- Fundamental Limitation: Unaccountability for Evolution.
- Obvious Solution : Random Walk with a Deterministic Drift

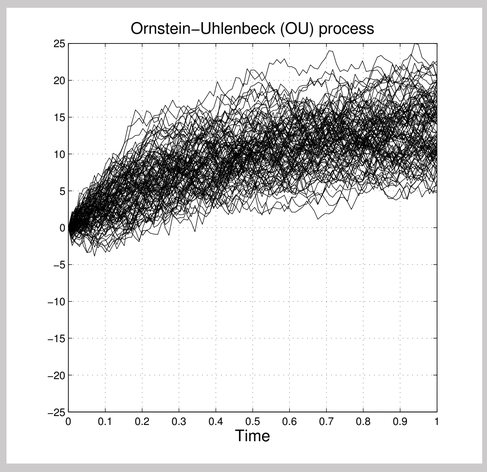
Brownian Motion

- Markovian
- $\mathbf{dX(t) = \sigma dB(t)}$
- σ: intensity of random fluctuations in evolution
- B(s + dt) B(s) $\sim$ N(0, σ dt)
- Cov (B(s), B(t)) = σ^2min(s, t)



Ornstein-Ulhenbeck Process

- Markovian
- $\mathbf{dX(t) = \alpha[\theta - X(t)]dt + \sigma dB(t)}$
- α : strength of selection (if $\alpha = 0 \rightarrow$ BM)
- θ : Optimum trait value
- Cov (X (s), X (t)) = σ^2 exp(-|s − t|α)

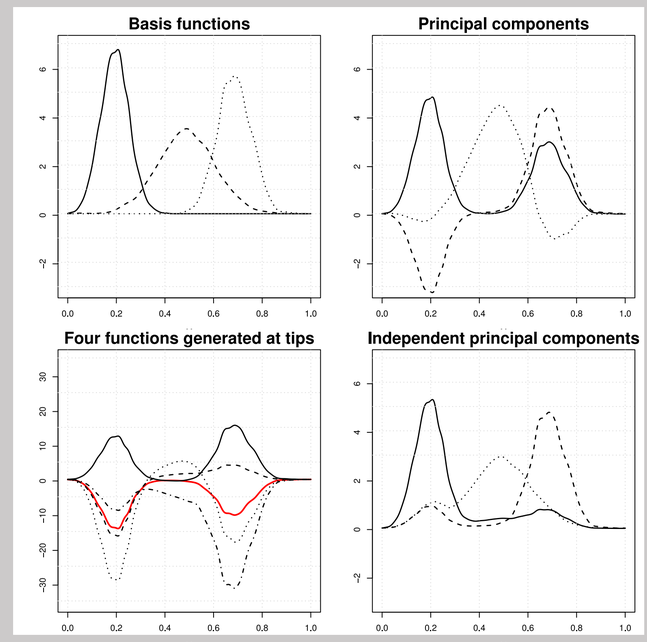


Independent Principal Component Analysis (IPCA)

- Given a Gaussian process $\mathbf{Y(t)}$,$\mathbf{t \epsilon [0,1]}$, sample curves assumed to have
  - $\mathbf{E[Y(t)] = \mu(t)}$
  - $\mathbf{Cov[Y(s), Y(t)] = C(s,t)}$
- Mercer's Theorem for symmetric $\mathbf{C(s,t) = \sum_{n=1}^{\infty} \lambda_n \phi_n(s) \phi_n(t)}$
  - order eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$
  - corresponding eigenfunctions ϕ_n's
- ϕ_n's are by definition orthogonal; *unrealistic assumption* for real data
- Blind Source Seperation
- $\mathbf{\Phi = AS}$ **S :** statistically independent source signals
- $\mathbf{A_{opt}^{-1}} = \text{argmin}_{A^{-1}} |\mathbf{C_{abcd}^{(s)}}|$,
- cumulant tensors $\mathbf{C_{abcd}^{s}}$ of the output data $\mathbf{s_i}$.
  - $\mathbf{C_{ijk}^{(s)} = \langle s_i, s_j, s_k \rangle}$
  - $\mathbf{C_{ijkl}^{(s)} = \langle s_i, s_j, s_k, s_l \rangle - \langle s_i, s_j \rangle \langle s_k, s_l \rangle - \langle s_i, s_k \rangle \langle s_j, s_l \rangle - \langle s_i, s_l \rangle \langle s_j, s_k \rangle}$

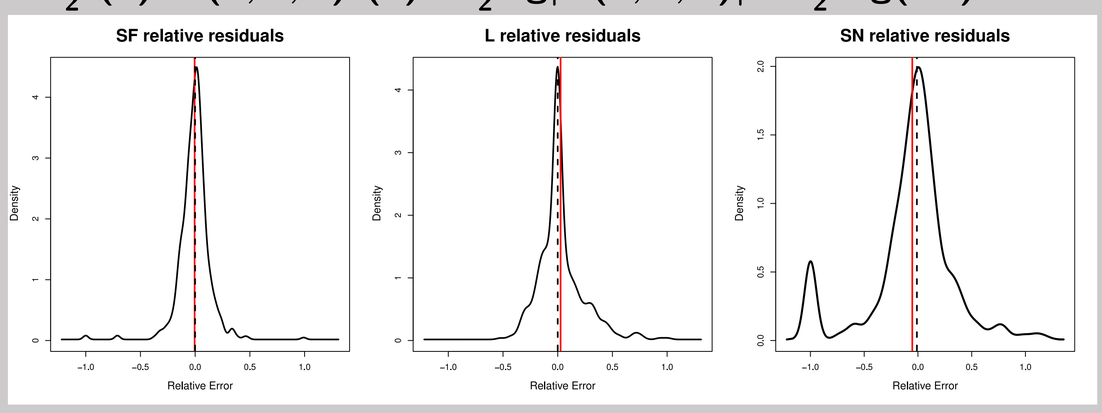
Insights from the IPC Analysis

- IPCA $\mathbf{W_{i,n}}$ scores as:
  - $\mathbf{\hat{W}_{i,n} = \sum_{k=1}^{m} \{Y_i(t_{i,k}) - \hat{m}(t_{i,k})\} \hat{s}_n(t_{i,k}) \Delta_{i,k}}$
- IPCA is a *phylogeny-agnostic*
- Automatic dimensionality determination PCA methodology
- Exclusion of outliers

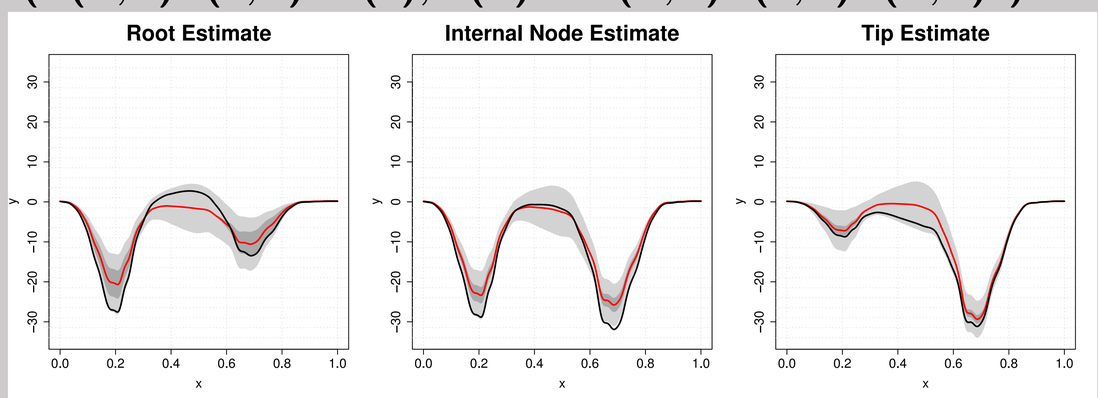


Phylogenetic Gaussian Process Regression

- **Gaussian Process** is defined as a probability distribution over functions $\mathbf{Y(x)}$ such that the set of values of $\mathbf{Y(x)}$ evaluated at an arbitrary set of points $\mathbf{x_1, \ldots, x_N}$ jointly have a Gaussian distribution.
- GPs are completely specified by the second-order statistics
- Assumptions
  - Conditional on their common ancestor any two trait are statistically independent. (Markov)
  - The statistical relationship between a node and any of its descendants is independent of the tree topology.
- Given trait $\mathbf{f(L)}$ on a finite set of co-ordinates $\mathbf{L}$ where $\mathbf{K(L,L,\theta)}$ is the matrix of covariances of pairs ($\mathbf{l_i, l_j}$) with hyperparameters θ then : $\mathbf{f(L) \sim N(0, K(L,L,\theta))}$.
- $\mathbf{K(l_i, l_j) = s_f^2 exp(-|l_i - l_j|/\lambda) + s_n^2 \delta_{l_i, l_j}}$
  - $\mathbf{K(l_i, l_j)}$: Covariance between known points in the phylogeny
  - $\mathbf{s_f^2}$: intensity of random fluctuations in evolution due to balance between the restraining forces / amplitude of function variation
  - $\boldsymbol{\lambda}$: phylogenetic horizon / characteristic length scale
  - $\mathbf{s_n^2}$: interspecies variation, unaccountable from relations conveyed by the phylogeny / noise
- Maximizes the GP LogLikelihood: $\mathbf{logp(f(L)|\theta) = -\frac{1}{2} f(L)^T K(L,L,\theta) f(L) - \frac{1}{2} log|K(L,L,\theta)| - \frac{|L|}{2} log(2\pi)}$



- $\mathbf{f(A)|f(L) \sim N(K(A,L)K(L,L)^{-1} f(L), K(A) - K(A,L)K(L,L)K(A,L)^T)}$



Illustrative Bibliography

- P.Z. Hadjipantelis, N.S. Jones, J. Moriarty, D. Springate and C.G. Knight, 2012.*Ancestral Inference from Functional Data: Statistical Methods and Numerical Examples*
- N.S. Jones and J. Moriarty, 2012. *Evolutionary Inference for Function-valued Traits: Gaussian Process Regression on Phylogenies*

http://go.warwick.ac.uk/hadjipantelis p.z.hadjipantelis@warwick.ac.uk