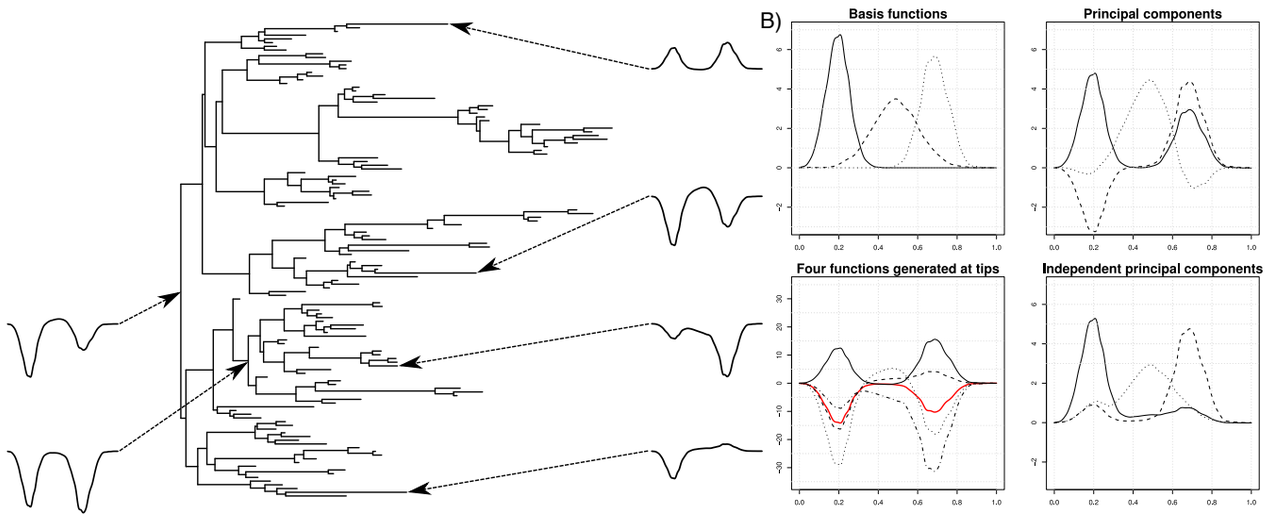
The University of Manchester

Gaussian Process Regression on a Phylogeny

High dimensional and dependent functional data workshop, Bristol 2012

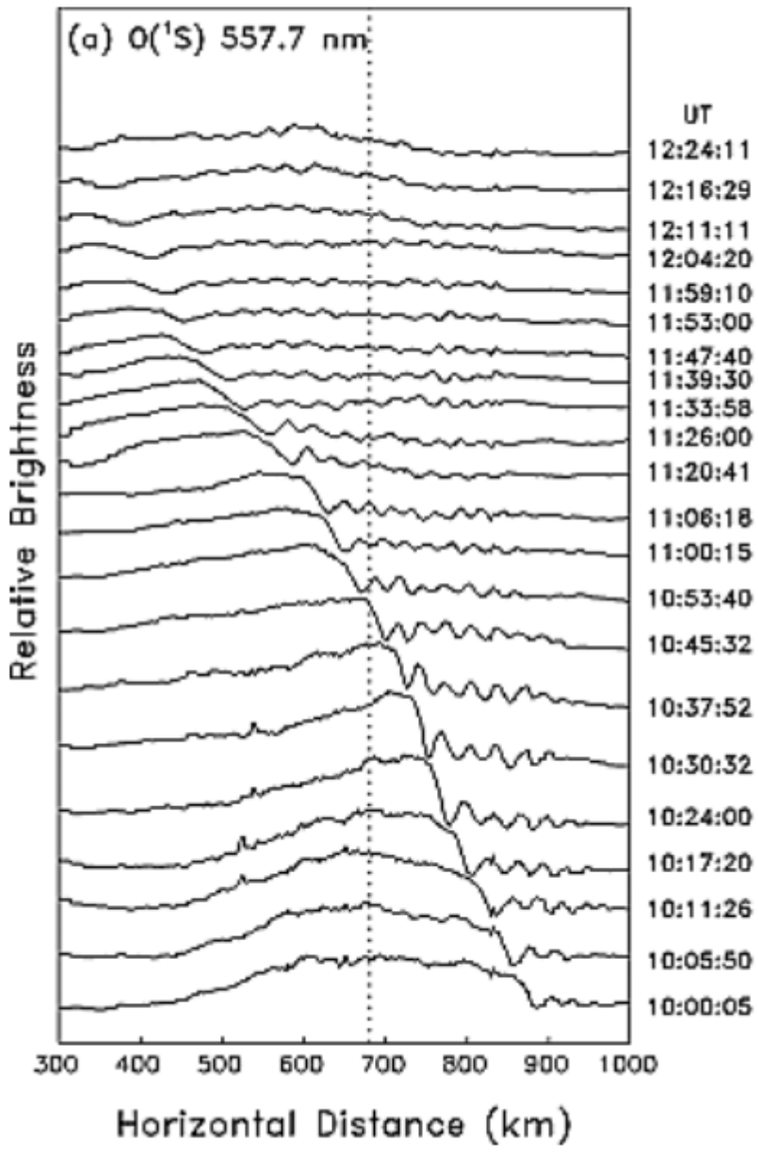
John Moriarty



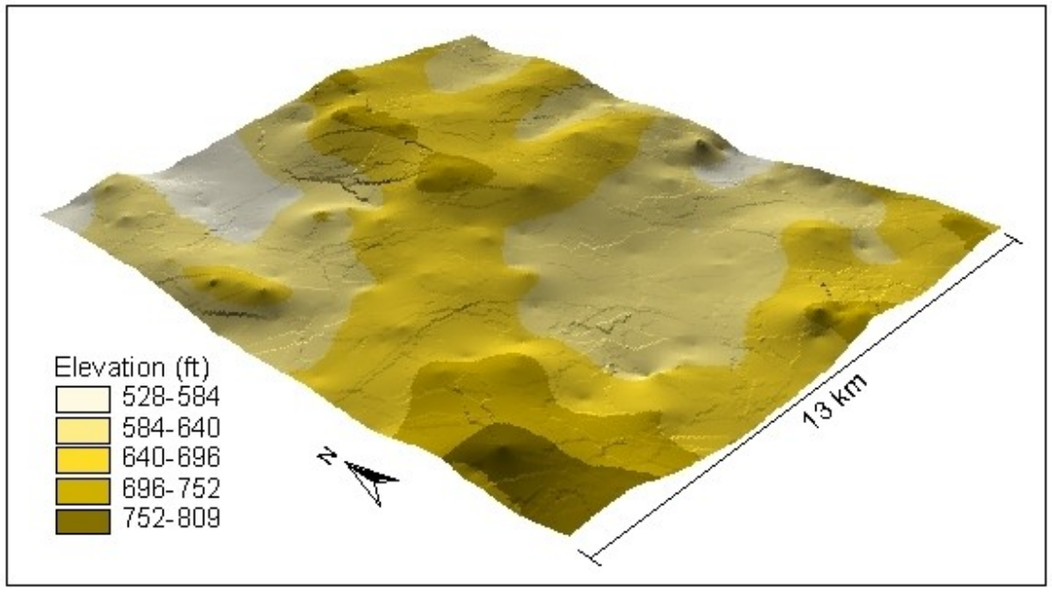
Outline of talk

- The problem: analysis of functional data with phylogenetic dependence
- Ideas for theoretical models
- Constructing a phylogenetic Gaussian process
- Properties of PGP models
- A study on simulated data
- If time allows - future directions

Functional data over time

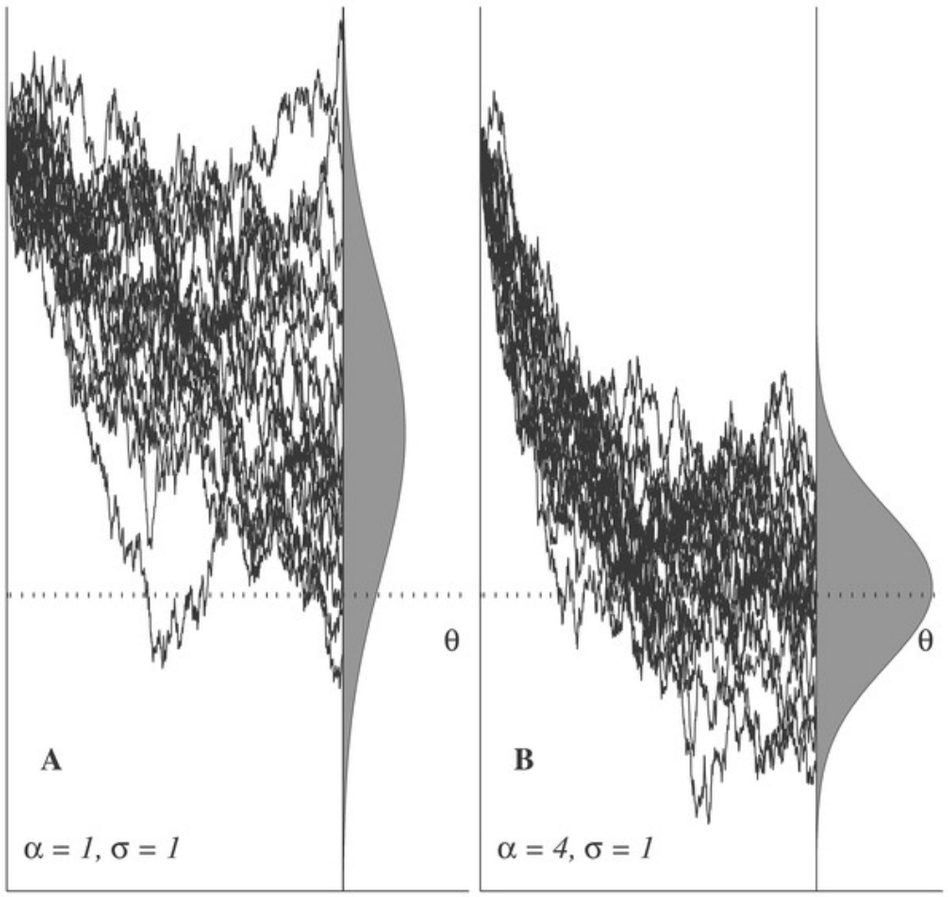


From smith et al., A multidiagnostic investigation of the mesospheric bore phenomenon. JOURNAL OF GEOPHYSICAL RESEARCH, 108 (A2), 1083 (2003)

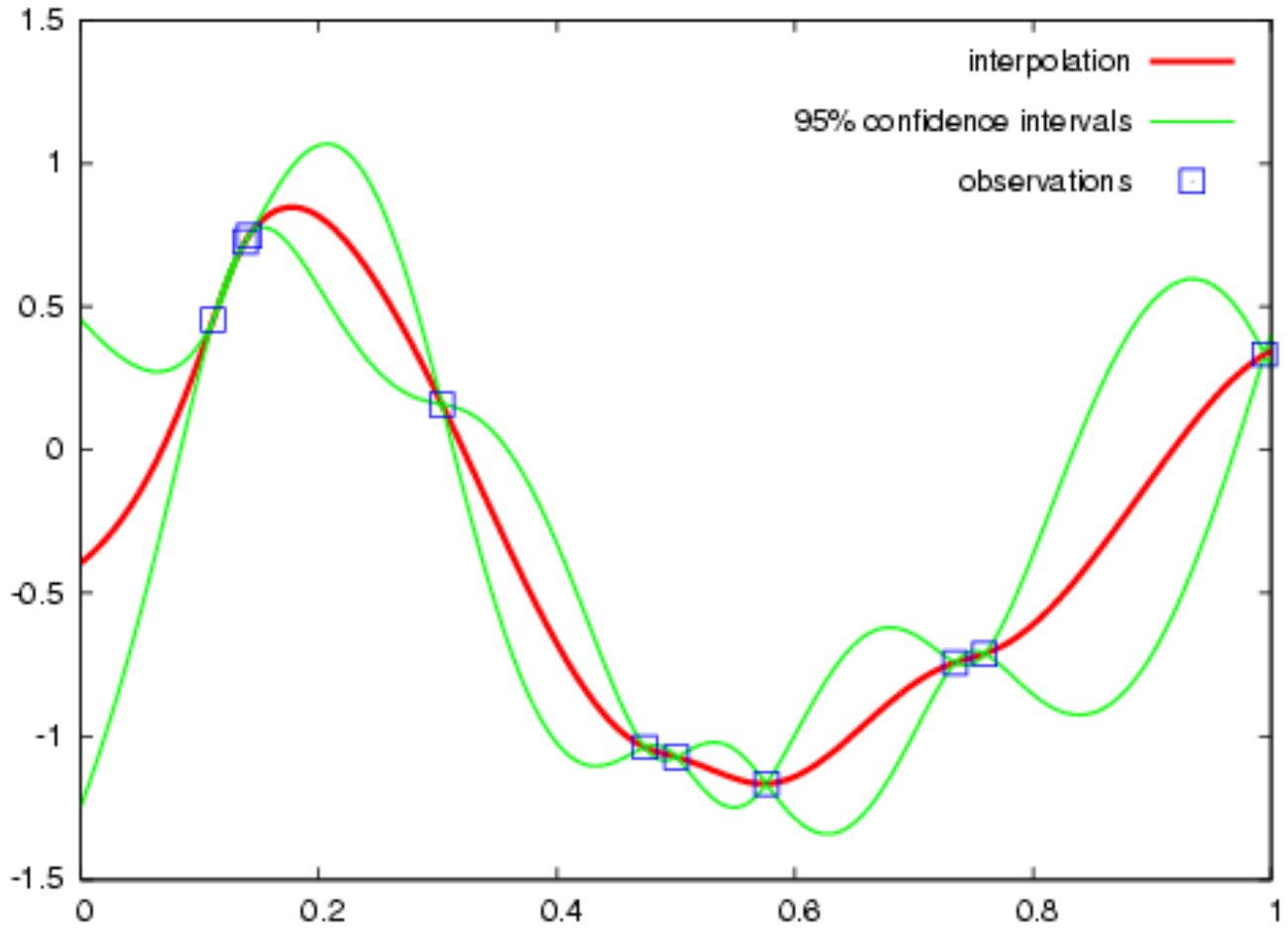


From http://proceedings.esri.com/library/userconf/proc01/professional/papers/pap280/p280.htm

Processes over time and space: quantitative genetics and kriging



Gaussian process over time

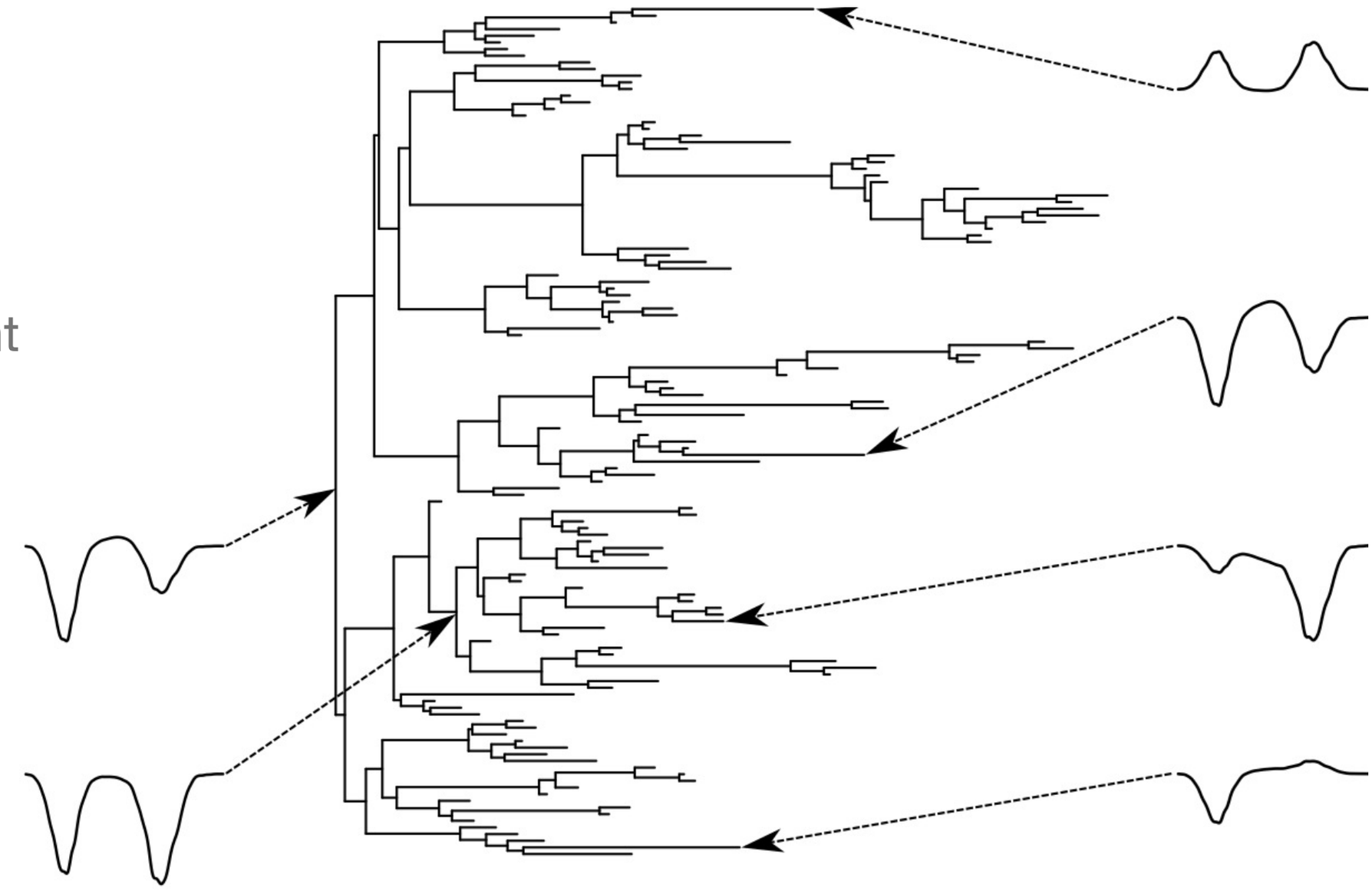


Gaussian process over space

From M. A. Butler and A. A. King, Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. The American Naturalist 164(6), 683-695 (2004)

Processes over time, space *and* phylogeny

Have a random signal at each point in the phylogeny



Statistical model - design specification

- Borrow strength from quantitative genetics: univariate models on phylogenies eg. BM, OU; **extend** to functional data

- Borrow strength from machine learning and kriging: flexible, tractable GP methods for Bayesian functional/ surface regression; **extend** to phylogenies

- Good practical performance when used for inference

Gaussian process regression - in one slide

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Characterised by its covariance function σ, typically depending on a vector θ of parameters. *So we can prove basic results in this framework without functional analysis*

Prior is $f(L) \sim \mathcal{N}(0, \sigma(L, L, \theta))$

GPs are partially observed

Posterior is

$$f(M)|f(L) \sim \mathcal{N}(A, B)$$

where

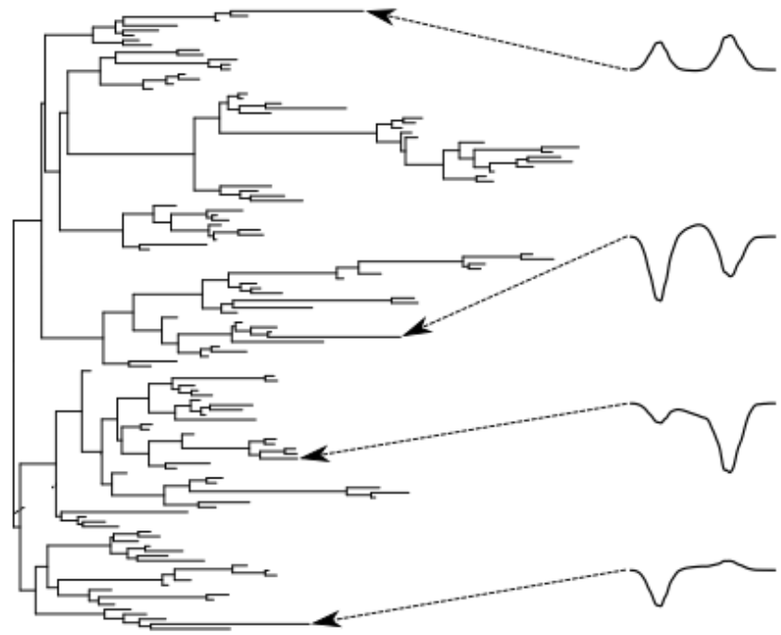
$$\begin{aligned} A &= \sigma(M, L, \theta)\sigma(L, L, \theta)^{-1} f(L), \\ B &= \sigma(M, M, \theta) - \sigma(M, L, \theta)\sigma(L, L, \theta)^{-1}\sigma(M, L, \theta)^T \end{aligned}$$

Phylogenetic Gaussian process (Jones-M., 2011)

To obtain a unique phylogenetic covariance function $\Sigma_{\mathbf{T}}$ we make two time-domain assumptions:

Assumption 1 Conditional on their common ancestors in the phylogenetic tree $\mathbf{T}$, any two signals are statistically independent.

Assumption 2 The statistical relationship between a signal and any of its descendants in $\mathbf{T}$ (the 'marginal process')is independent of the topology of $\mathbf{T}$.



$\mathbf{T}$

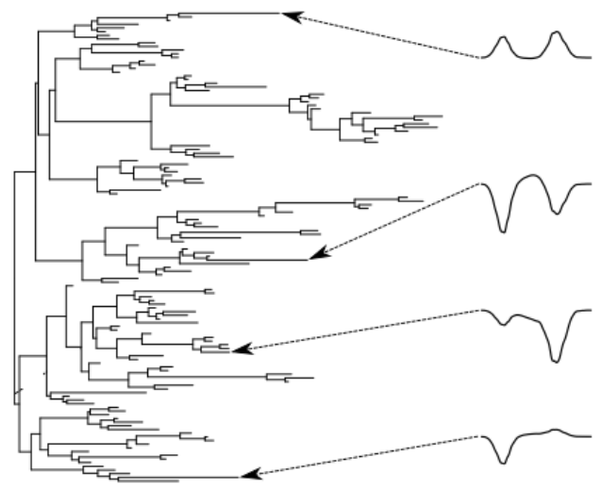
Properties of phylogenetic covariance functions

1. If the marginal covariance function Σ is space-time separable so that

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

then the phylogenetic covariance function $\Sigma_{\mathbf{T}}$ is also space-time separable, i.e.

$$\Sigma_{\mathbf{T}}((x_1, \mathbf{t}_1), (x_2, \mathbf{t}_2)) = K(x_1, x_2)k_{\mathbf{T}}(\mathbf{t}_1, \mathbf{t}_2)$$



$\mathbf{T}$

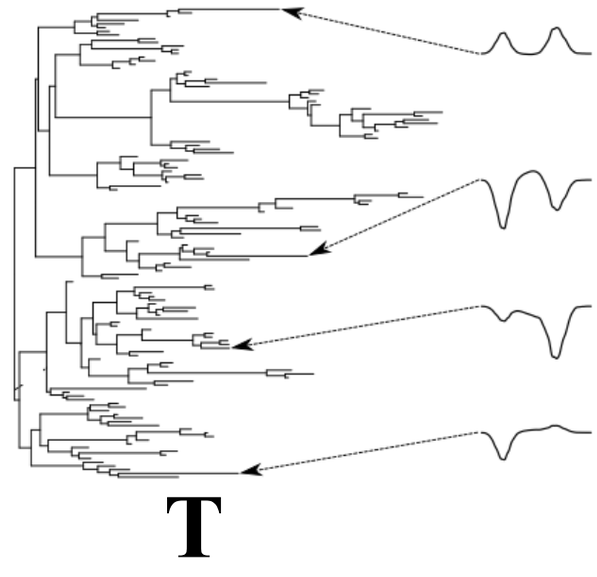
2. If k is Markovian in time, we have the simple expression

$$k_{\mathbf{T}}(\mathbf{t}_1, \mathbf{t}_2) = k(t_1, t_{12})k(t_{12}, t_{12})^{-1}k(t_2, t_{12})$$

where $\mathbf{t}_{12}$ is the most recent common ancestor of $\mathbf{t}_1$ and $\mathbf{t}_2$ and t_{12} is its depth in $\mathbf{T}$.

Properties (cont.)

3. 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 $k_{\mathbf{T}}(\mathbf{t}_1, \mathbf{t}_2)$ is isotropic (meaning a function of the patristic distance between $\mathbf{t}_1$ and $\mathbf{t}_2$ only). In fact, $k_{\mathbf{T}}$ is only isotropic when k is the Ornstein-Uhlenbeck covariance.



4. Let Y be a PGP with separable covariance function $\Sigma_{\mathbf{T}}$. Under weak conditions on K, there exist deterministic functions $\phi_i : S \to \mathbb{R}$ and univariate PGPs X_i, for $i = 1 \ldots n$, such that the Gaussian process given by

Countable representation for PGPs

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

has the same distribution as Y.

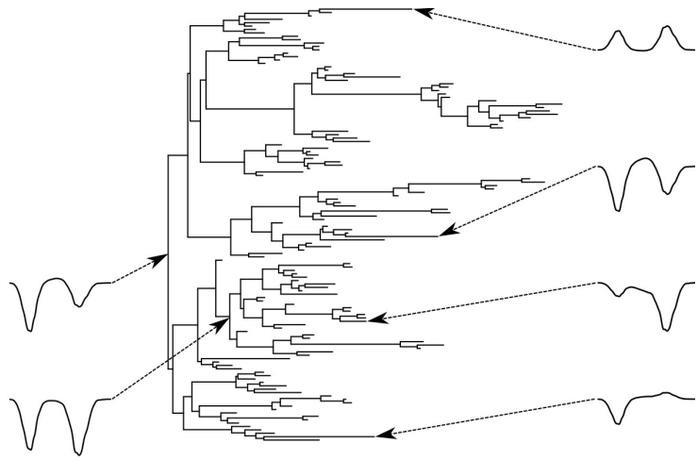
Generating synthetic data (Hadjipantelis et al. 2012)

1. Generated a 128-tip phylogentic tree $\mathbf{T}$ and three fixed spatial basis functions ϕ

2. Three independent *univariate* PGPs used to generate weights for mixing, with covariance

$$k^i_{\mathbf{T}}(\mathbf{t}_1, \mathbf{t}_2) = E[w^i_{\mathbf{t}_1}, w^i_{\mathbf{t}_2}] \quad (1)$$
$$= (\sigma^i_f)^2 exp\left(\frac{-d_T(\mathbf{t}_1, \mathbf{t}_2)}{\lambda^i}\right) + (\sigma^i_n)^2 \delta_{\mathbf{t}_1, \mathbf{t}_2}$$

i	σ^i_f	λ^i	σ^i_n
1	4.5	17.9	0.45
2	0	NA	1
3	3.0	8.95	0.45



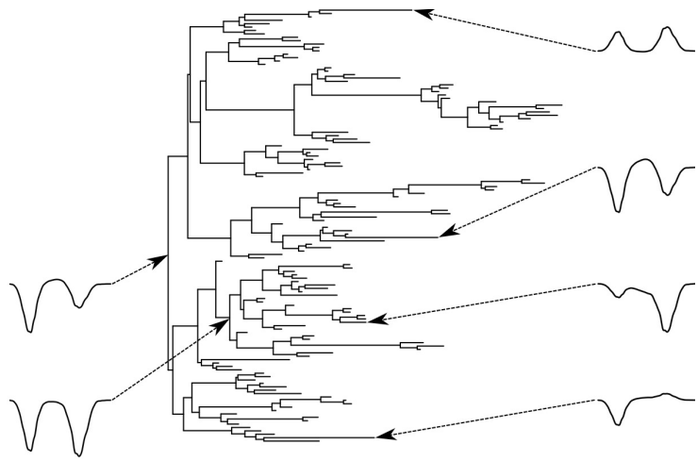
Generate a randomly mixed signal at each tree node

Synthetic data (cont.)

3. Each node in the tree $\mathbf{t} \in \mathbf{T}$ thus had an associated vector $w_\mathbf{t} = (w_\mathbf{t}^1, w_\mathbf{t}^2, w_\mathbf{t}^3)$ of weights for the basis functions. These produced a single function-valued trait $f_\mathbf{t}$ at each node:

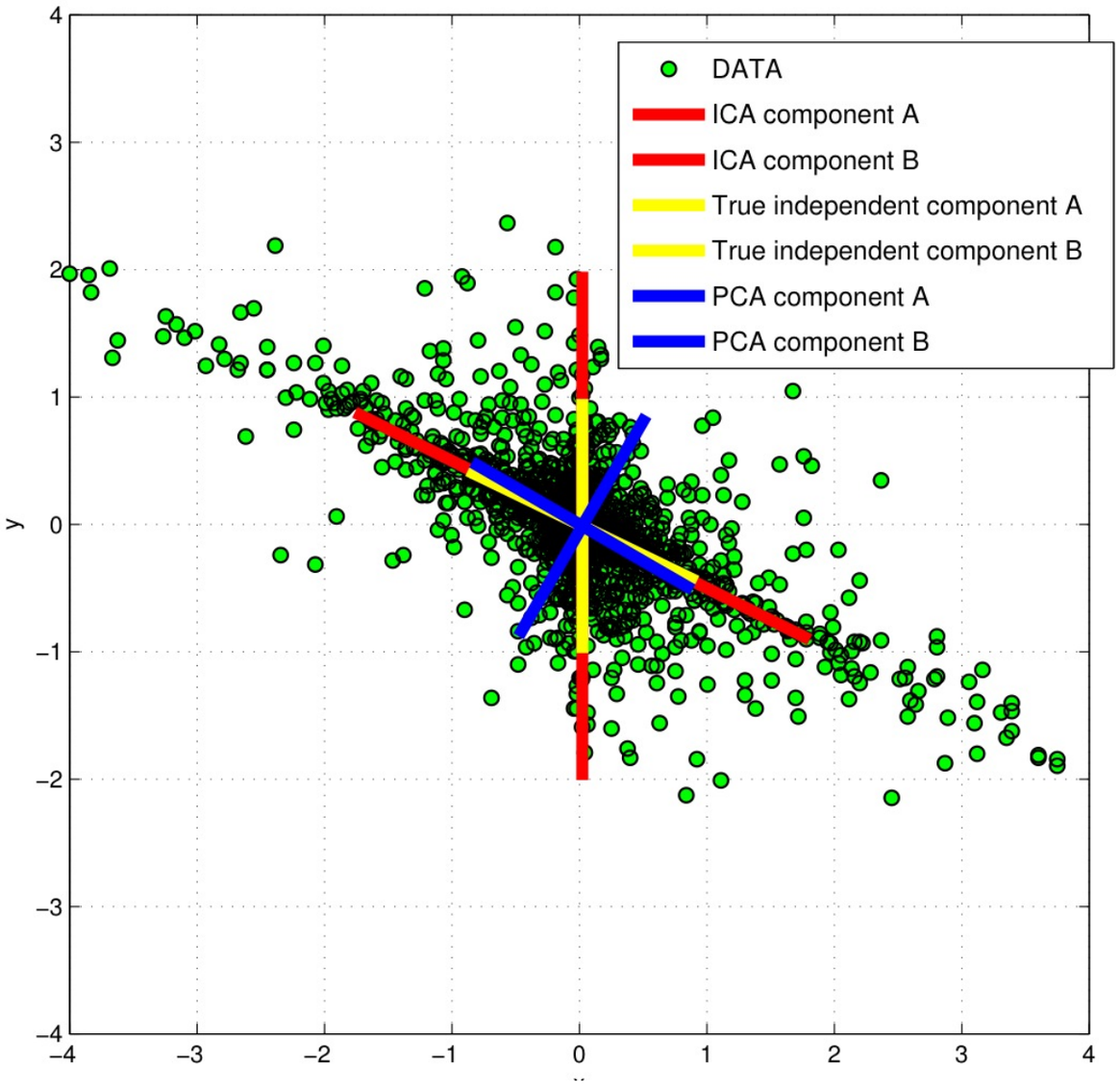
$$f_\mathbf{t} = w_\mathbf{t}^T \phi \tag{2}$$

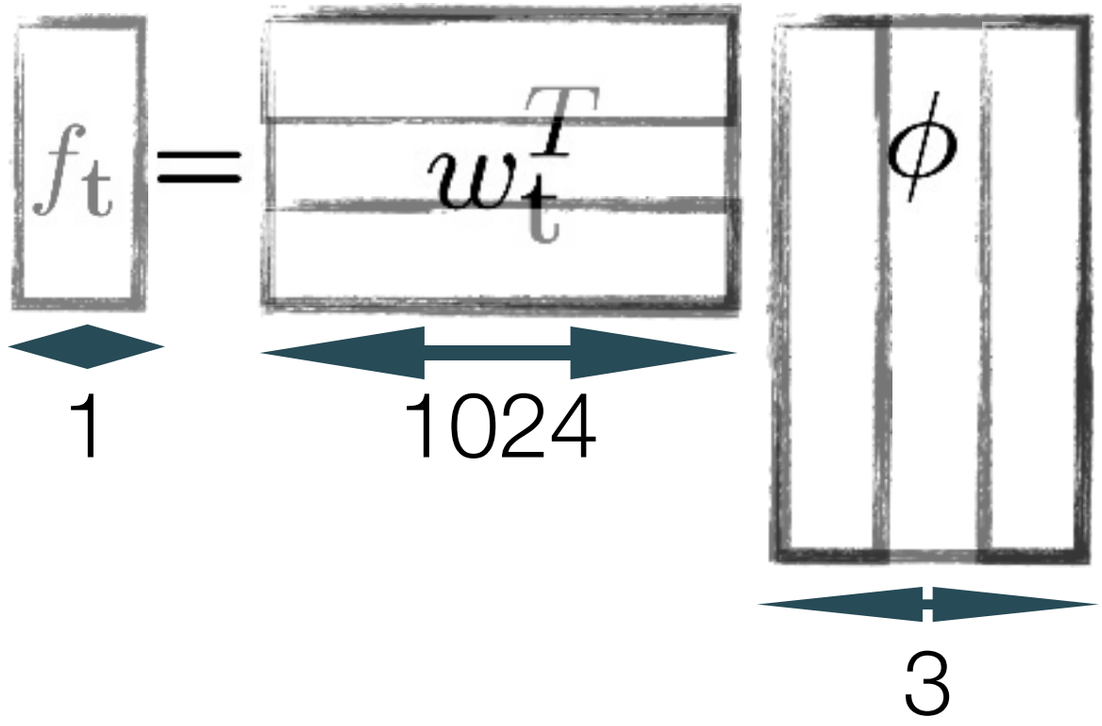
where ϕ is the 3×1024 matrix having each ϕ_i as its rows. The 128 curves at tips of $\mathbf{T}$ were taken as inputs to our regression analysis, and the 127 curves at internal nodes of $\mathbf{T}$ were used for validation of the method.



Generate a randomly mixed signal at each tree node

Reversing the CLT: ICA

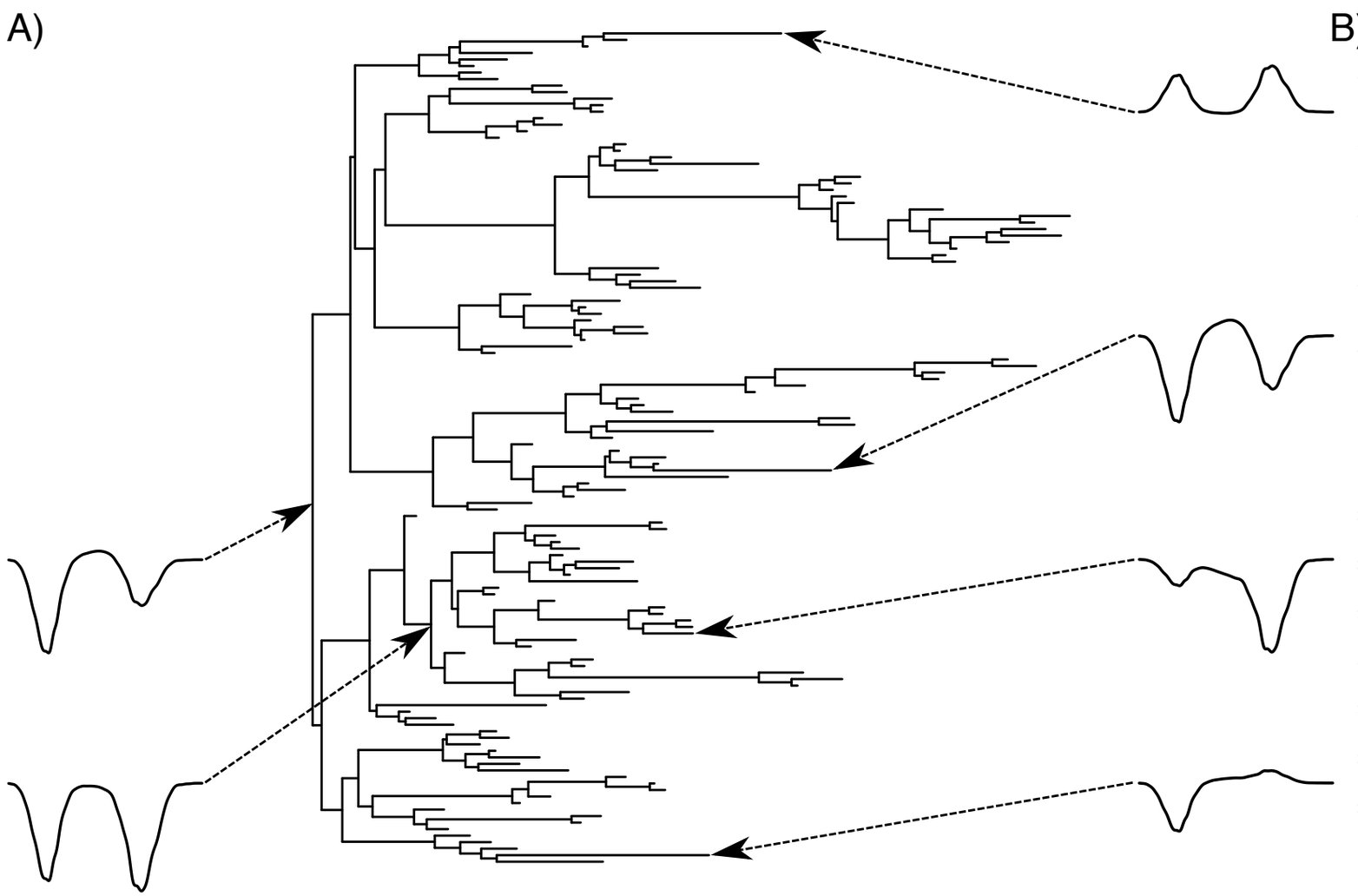




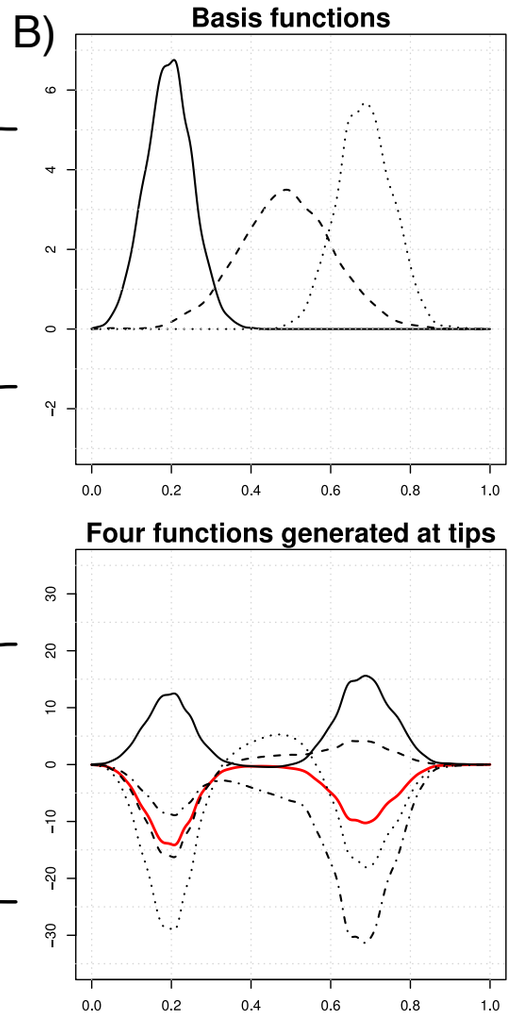
The blue components ('factor scores') are more Gaussian than the red/ yellow components

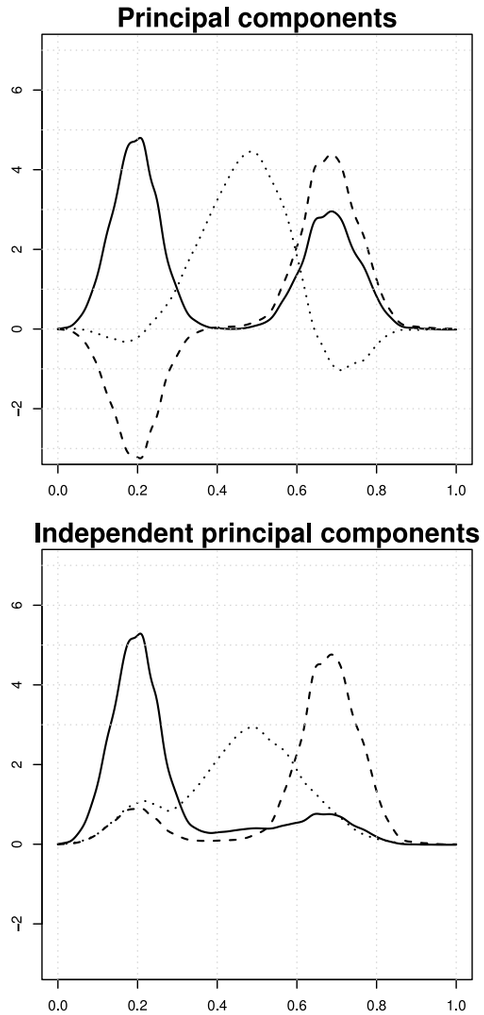
Dimension reduction, source separation: IPCA

A)

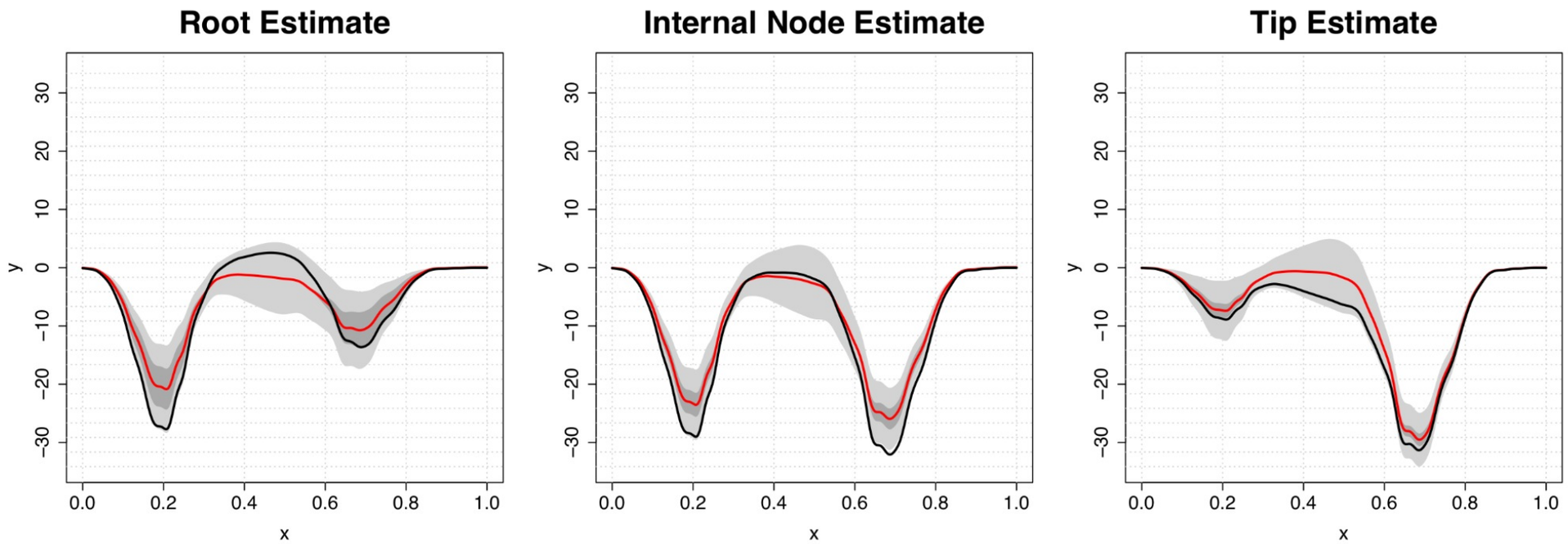


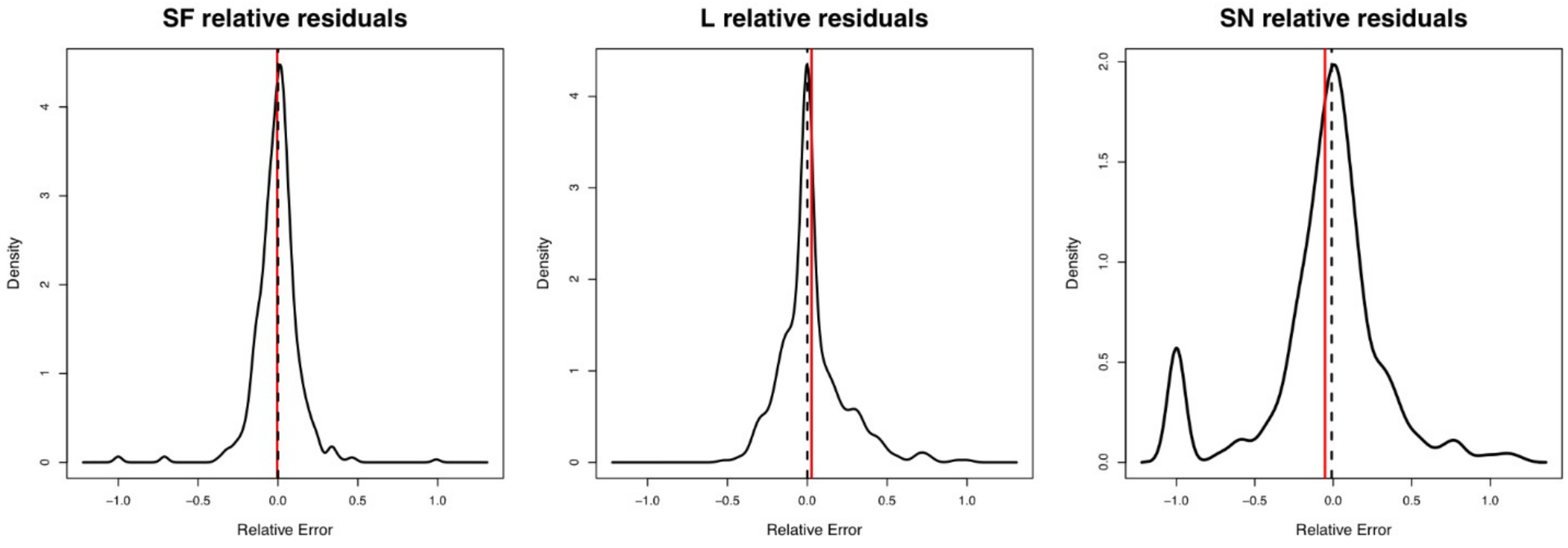
B)





Validation: ancestor, parameter estimation





Conclusion

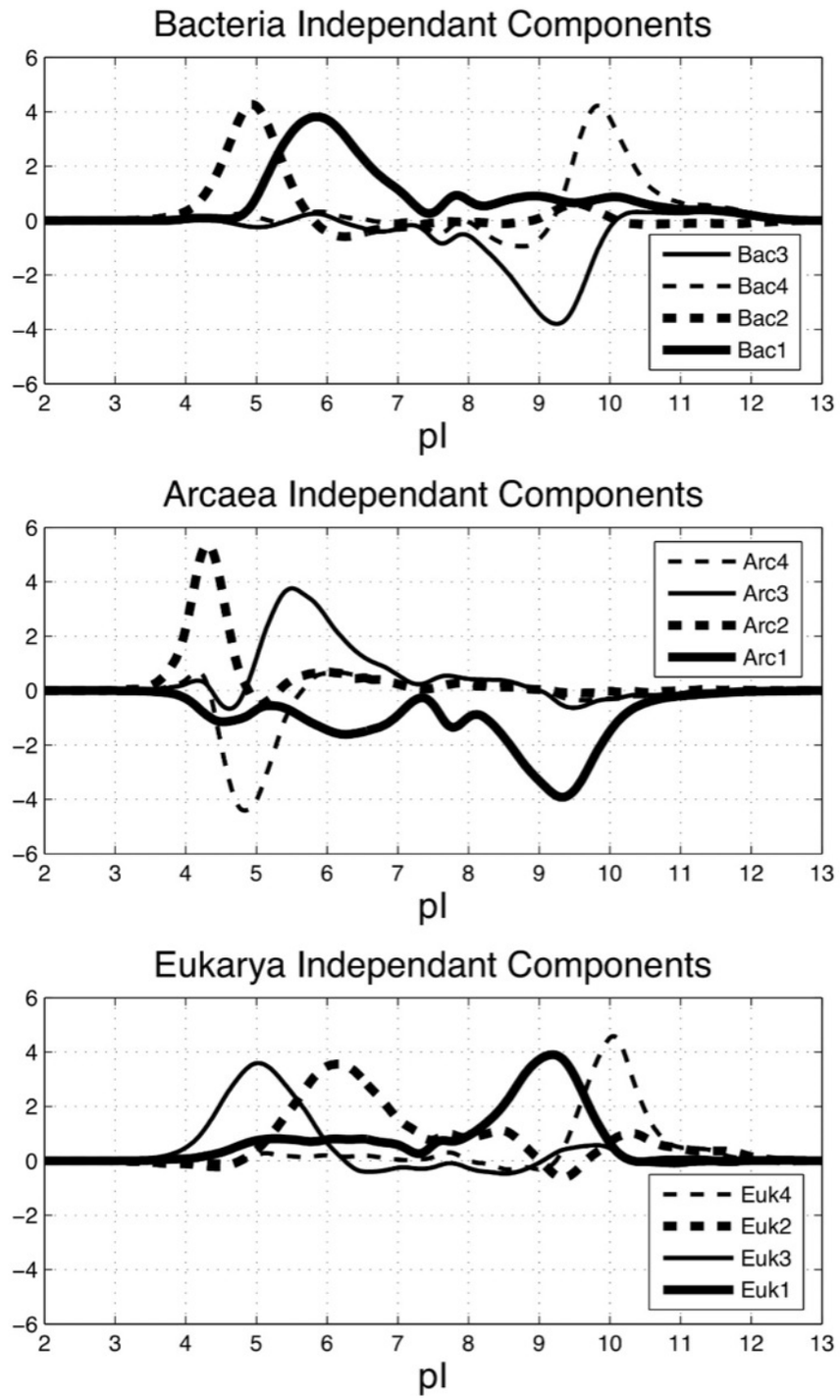
We have developed Phylogenetic Gaussian Process models, usable as priors for Bayesian functional inference in the presence of evolutionary dependence between functional data, and tested the performance of inference methods using IPCA with synthetic data.

Further detail:

Hadjipantelis, P.Z., Jones, N.S., Moriarty, J, Springate, D, Knight, C.G. Ancestral Inference from Functional Data: Statistical Methods and Numerical Examples
arxiv:1208.0628

Nick S. Jones, John Moriarty
Evolutionary Inference for Function-valued Traits: Gaussian Process Regression on Phylogenies
arxiv:1004.4668v3

Future work:
big biological data, 'known' phylogeny



Future work: time warping

Question: How far does GP regression model phase variation in signals?

Toy model: Consider a space-time GP $Z(t, x)$ ($t \in [0, T], x \in [0, 1]$). Apply a 'time warping' function to relabel the domain of each signal Z_t:

$$(t, x) \mapsto (t, F(t, x)). \tag{0.3}$$

If Z is stationary then the compound process $Y(t, x) := Z(t, F(t, x))$ is a GP and so we can still do GP regression! Variation in amplitude and phase are both considered to carry evolutionary signal, and are modelled jointly (NB not individually identifiable).

Future work: time warping (2)

Example result: Choose this covariance function for Z:

$$\sigma_Z((s,x),(t,y)) = \phi(s,t)\phi(x,y) \tag{0.4}$$

where $\phi(a,b) := \exp\left(-\frac{(a-b)^2}{2}\right)$, and model F as a GP with covariance function σ_F. Then

$$\sigma_Y((s,x),(t,y)) \;=\; \phi(s,t)\phi^C(x,y), \tag{0.5}$$

where $\phi^C(x,y) := \frac{1}{\sqrt{1+C}} \exp\left(-\frac{(x-y)^2}{2(1+C)}\right)$, where

$$C \;=\; Var(F(s,x) - F(t,y)) \tag{0.6}$$

$$\;=\; \sigma_F^2(s,x) + \sigma_F^2(t,y) - 2\sigma_F((s,x),(t,y)) \tag{0.7}$$

Conclusion: toy model implies a covariance function with a specific kind of non-separability—the spatial (x,y) dispersion parameter $2(1+C)$ depends on the time separation $|s-t|$.

Mutating Messages!

Biological data objects are often high-dimensional and dependent because of phylogeny