Shifted stochastic processes evolving on trees: application to models of adaptive evolution on phylogenies

Paul Bastide (1,2), Mahendra Mariadassou (2), Stéphane Robin (1)

(1) UMR 518 MIA, INRA/AgroParisTech, Paris, France.
(2) MaIAGE, INRA, Jouy-en-Josas, France.

Goals and Setting

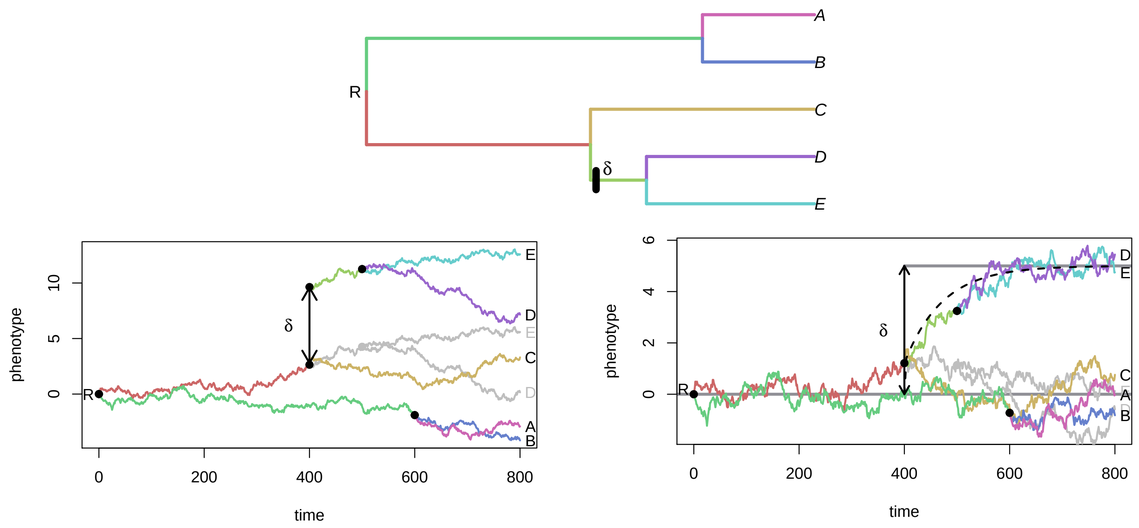
Data

- Measure of one quantitative trait for a set of related extant species.
- A phylogenetic tree, time-calibrated and ultrametric.

Goals

- Explain the observed trait distribution, while accounting for phylogenetic correlations.
- Detect environemental shifts that occured in the past.

Model [2, 4]



$\mathbf{dW(t) = \sigma dB(t)}$

BM Shifts in the **mean**.

$\mathbf{dW(t) = \alpha[\beta(t) - W(t)]dt + \sigma dB(t)}$

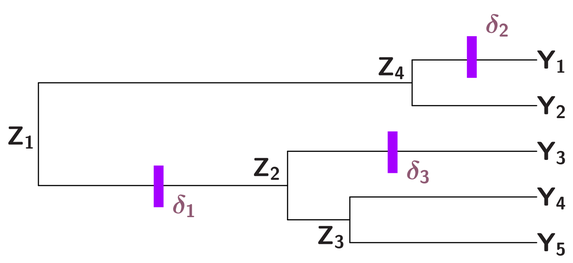
OU Shifts in the **optimal value**.

Incomplete Data Point of View

$$\mathbf{X_j | X_{pa(j)}} \sim \mathcal{N}\left(\mathbf{q_j X_{pa(j)} + r_j} + \mathbf{s_j} \sum_{\mathbf{k}} \mathbb{I}\{\tau_k = b_j\}\delta_k, \sigma_j^2\right)$$

EM Algorithm Maximize $\mathbb{E}_\theta[\log p_\theta(\mathbf{Z, Y}) \mid \mathbf{Y}]$.
E step "Upward-Downward" Algorithm.
M step OU: increase objective function (GM).
Initialization LASSO regression.

Linear Regression Point of View



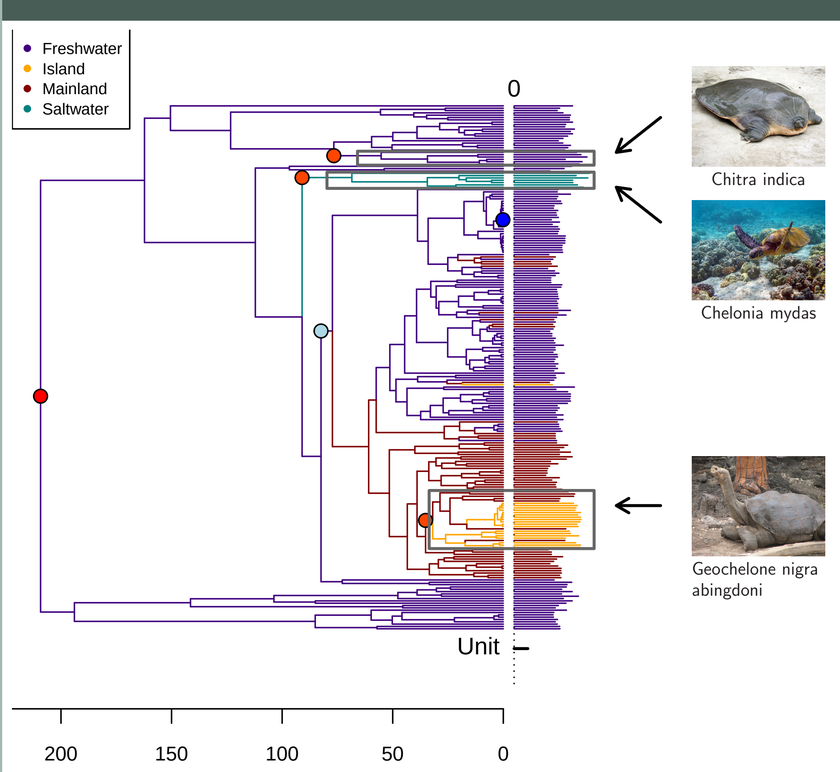
$$\mathbf{\Delta} = \begin{pmatrix} \mu \\ \delta_1 \\ 0 \\ 0 \\ \delta_2 \\ 0 \\ \delta_3 \\ 0 \\ 0 \end{pmatrix} \qquad \mathbf{T\Delta} = \begin{pmatrix} \mu + \delta_2 \\ \mu \\ \mu + \delta_1 + \delta_3 \\ \mu + \delta_1 \\ \mu + \delta_1 \end{pmatrix}$$

$$\mathbf{T} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{array} \begin{array}{c} \begin{array}{ccccccccc} Z_1 & Z_2 & Z_3 & Z_4 & Y_1 & Y_2 & Y_3 & Y_4 & Y_5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

BM : $\mathbf{Y = T\Delta + E}$

OU : $\mathbf{Y = TW(\alpha)\Delta + E}$

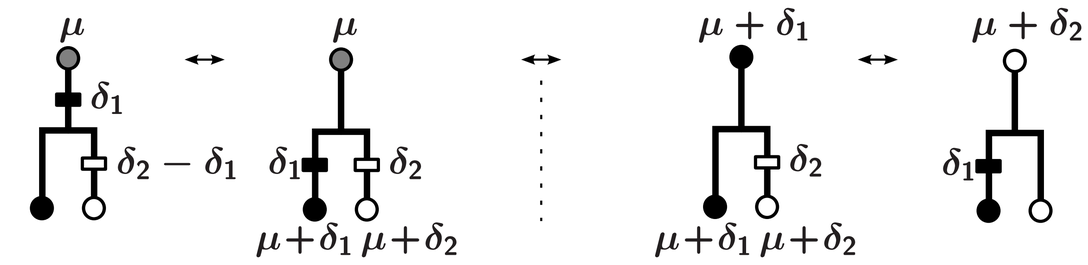
Chelonia Dataset [5]



	Habitat	EM
No. of shifts	16	5
No. of regimes	4	6
lnL	-133.86	-97.59
ln 2 / α (%)	7.44	5.43
γ^2	0.33	0.22
CPU time (min)	65.25	134.49

Colors: habitats. Boxes: selected EM regimes. Positive selected shifts are shown in red, and negative in blue.

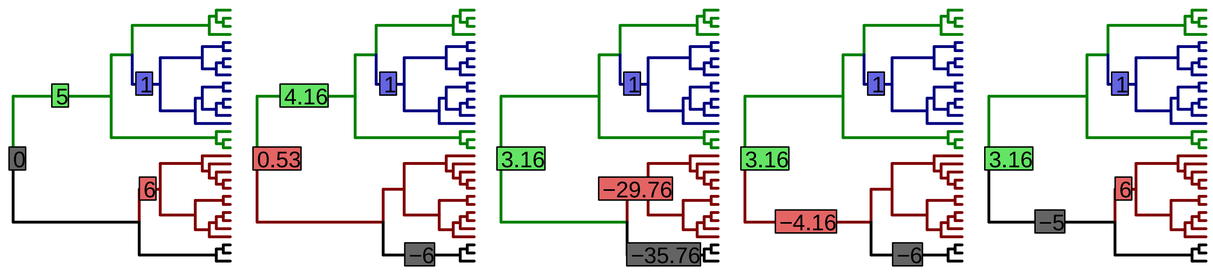
Identifiability Issues



All these shifts allocations give the same tips distribution (under a BM). The two on the right are parsimonious, they are said to be *equivalent*. We discard the two on the left as over-parametrized.

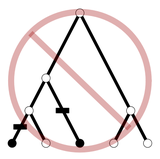
Parsimony and Equivalence Classes

Find one solution : Existing Dynamic Programming algorithms (Fitch, Sankoff) [3].
Enumerate on equivalent class : New recursive algorithm (implemented in R).



These five shifts allocations are equivalent: they are parsimonious and they produce the same tips distribution (under an OU).

Number of Models with K shifts



No Homoplasy : 1 shift = 1 new color.
Proposition : **K** shifts $\iff$ **K + 1** colors.

$$\mathcal{S}_K^{PI} = \{\text{Parsimonious allocations of } K \text{ shifts}\}/\sim$$
$$\mathcal{S}_K^{PI} \simeq \{\text{Coloring of tips in } K+1 \text{ colors}\}$$

Proposition:

- $\left|\mathcal{S}_K^{PI}\right| \leq \binom{m+n-1}{K}$
- $\left|\mathcal{S}_K^{PI}\right|$ depends on the topology of the tree.
- For a binary tree: $\left|\mathcal{S}_K^{PI}\right| = \binom{2n-2-K}{K}$.

Model Selection on K (α known) [1]

Under our setting:

$$\mathbf{Y = R\Delta + \gamma E} \quad \text{with} \quad \mathbf{E \sim \mathcal{N}(0, V)} \quad \text{and} \quad \mathcal{S} = \{S_\eta, \eta \in \mathcal{M}\}, \quad \mathcal{M} = \bigcup_{K \geq 0} \mathcal{S}_K^{PI}$$

Define the following penalty:

$$\mathbf{pen(K) = A\frac{n-K-1}{n-K-2} EDkhi[K+2, n-K-2, e^{-L_K}]}, \quad \mathbf{L_K = \log\left|\mathcal{S}_K^{PI}\right| + 2\log(K+2)}$$

and the estimator: $\hat{\eta} = \underset{\eta \in \mathcal{M}}{\mathbf{argmin}} \|\mathbf{Y} - \hat{\mathbf{s}}_\eta\|_{\mathbf{V}}^2 \left(1 + \frac{\mathbf{pen(K_\eta)}}{\mathbf{n - K_\eta - 1}}\right)$

Under some reasonable technical hypothesis, we get the non-asymptotic bound:

$$\mathbb{E}\left[\frac{\|\mathbf{s} - \hat{\mathbf{s}}_{\hat{\eta}}\|_{\mathbf{V}}^2}{\gamma^2}\right] \leq \mathbf{C(A, \kappa)}\left[\inf_{\eta \in \mathcal{M}}\left\{\frac{\|\mathbf{s} - \mathbf{s}_\eta\|_{\mathbf{V}}^2}{\gamma^2} + \mathbf{D}_\eta(3 + \log(n))\right\} + 1 + \log(n)\right]$$

Conclusion and Perspectives

- We developed a general statistical framework for trait evolution models with unconstrained shifts on ultrametric trees.
- R codes available on GitHub: `https://github.com/pbastide/Phylogenetic-EM`
- Perspectives:
  - Handle multivariate (correlated) traits.
  - Deal with uncertainty (tree, data).
  - Use fossil records (non-ultrametric tree).

References

[1] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with an unknown variance. *Annals of Statistics*, 37(2):630–672, Apr. 2009. doi: 10.1214/07-AOS573.

[2] M. A. Butler and A. A. King. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *The American Naturalist*, 164(6):pp. 683–695, 2004. ISSN 00030147.

[3] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Suderland, USA, 2004.

[4] T. F. Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5): 1341–1351, oct 1997.

[5] A. L. Jaffe, G. J. Slater, and M. E. Alfaro. The evolution of island gigantism and body size variation in tortoises and turtles. *Biology letters*, 2011.