

# Inference of Adaptive Shifts for Multivariate Correlated Traits



Paul Bastide (1,2), Cécile Ané (3), Stéphane Robin (1), Mahendra Mariadassou (2)

AgroParisTech

(1) UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France

(2) MaIAGE, INRA, Université Paris-Saclay, 78352 Jouy-en-Josas, France

(3) Department of Statistics and Botany, University of Wisconsin-Madison, WI, 53706, USA

MaIAGE

## Goals and Setting

### Goals

- Detect environmental shifts that occurred in the past.

- Account for phylogenetic and process correlations.

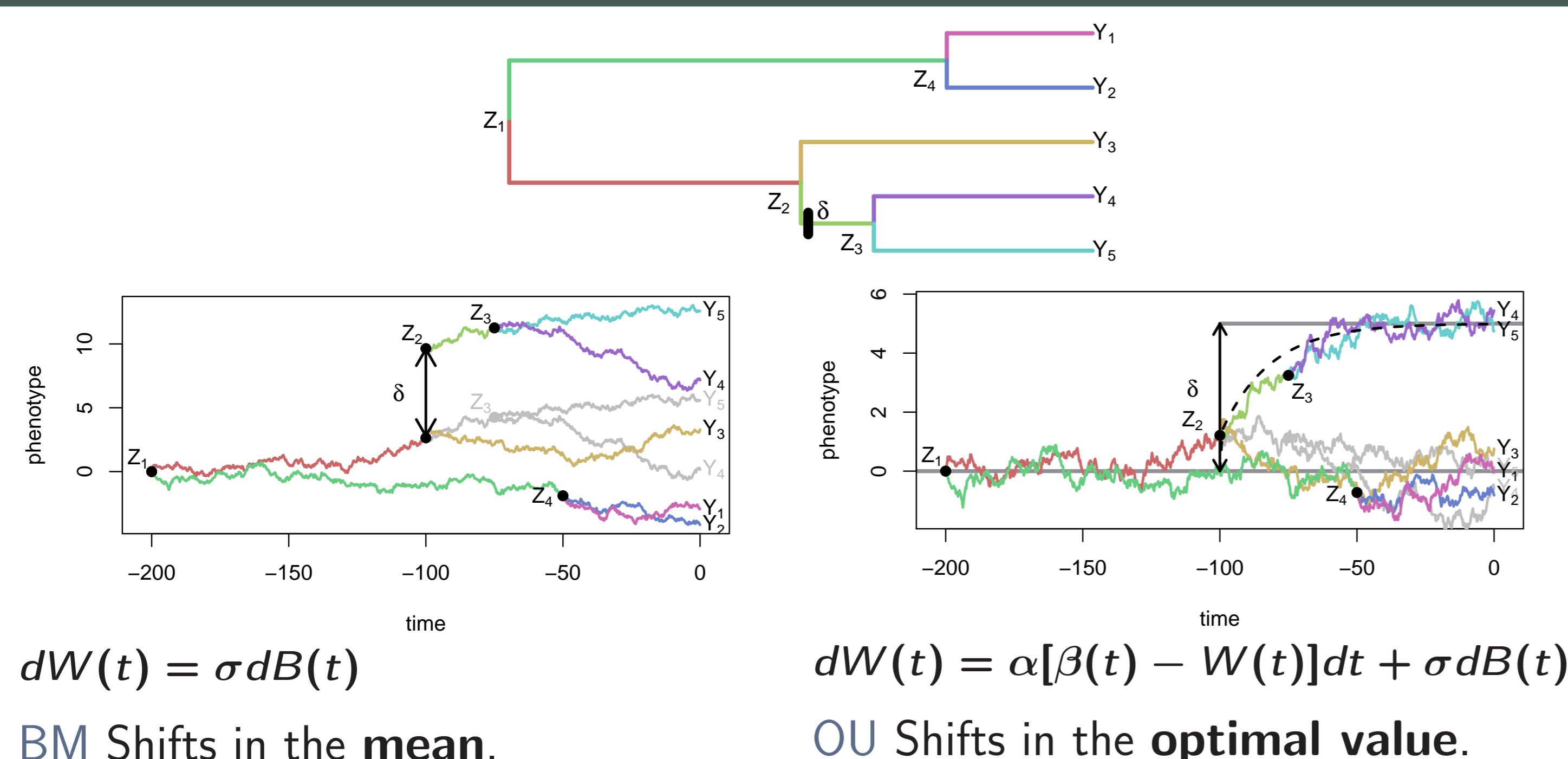
### Data

- Measure of several quantitative traits for a set of related extant species.

- A phylogenetic tree, time-calibrated and ultrametric.

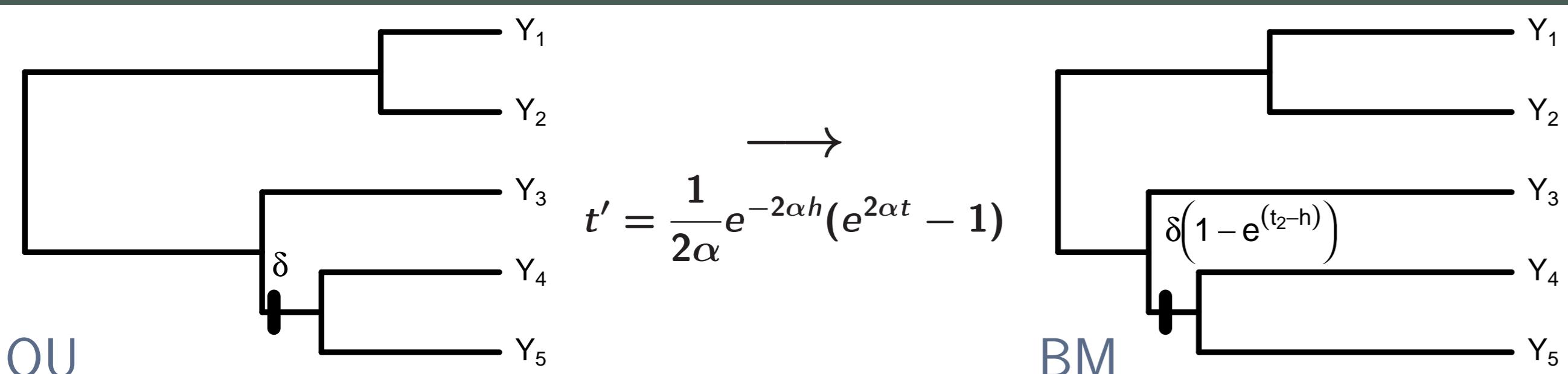
## Model

[3,7]



## OU $\iff$ BM

[5,6]



## Multivariate: Assumptions

[5,8]

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

- All traits shift at the same time.
- Previous works: independent traits.  $\mathbf{A}$  and  $\Sigma$  diagonal.  $\rightarrow \ell 1$  ou
- $\hookrightarrow$  Pre-processing pPCA ?
- This work: correlated traits. scOU: scalar  $\mathbf{A}$ , full  $\Sigma$ .  $\hookrightarrow$  Re-scaling of the tree.  $\rightarrow$  PhylogeneticEM

## pPCA is Biased (BM with shifts)

[5,9]

BM with shifts

$$\mathbf{Y} \sim \mathcal{MN}(\mathbf{a}, \mathbf{C}, \mathbf{R})$$

$$\mathbf{a} \neq \mathbf{1}_n \mu^T$$

$$\mathbf{C} = [c_{ij}]_{1 \leq i,j \leq n}$$

$$\mathbf{R} = \Sigma^T \Sigma$$

Expectation Matrix ( $n \times p$ )

Tree Correlations ( $n \times n$ )

Traits Correlations ( $p \times p$ )

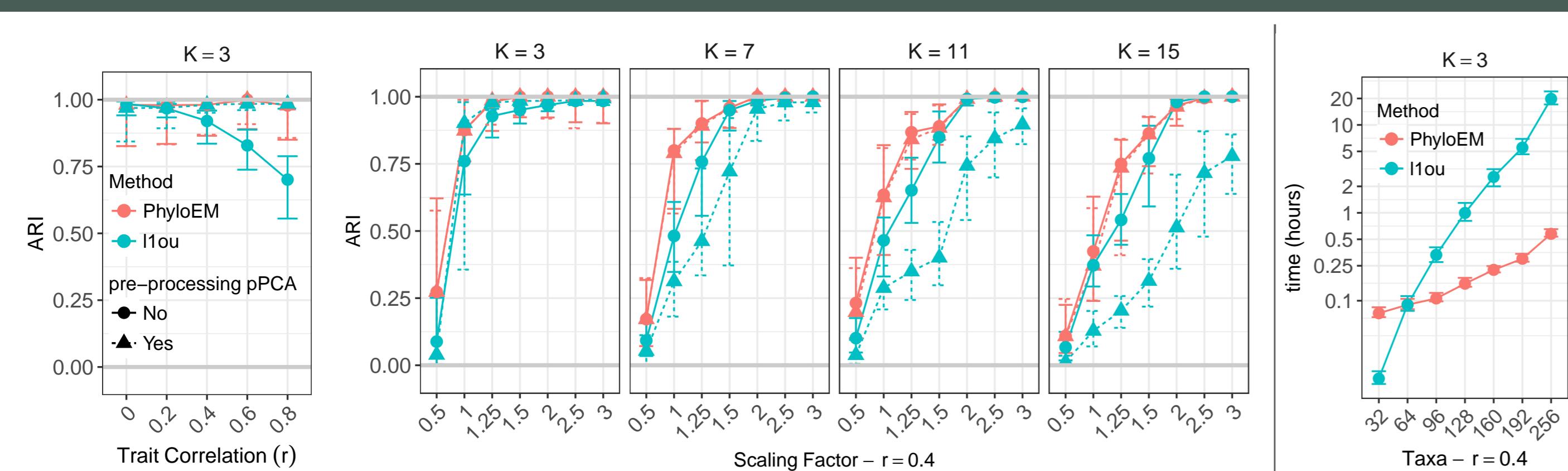
Bias:

$$\mathbb{E}[\hat{\mathbf{R}}] = \mathbf{R} + \mathbf{B} \quad \text{where} \quad \mathbf{B} = (\mathbf{a} - \mathbf{1}_n \bar{\mathbf{a}}^T)^T \mathbf{C}^{-1} (\mathbf{a} - \mathbf{1}_n \bar{\mathbf{a}}^T) / (n-1)$$

$=$  Between group variance

## Simulations on 160 taxa

[5]



## References

- [1] Aristide et al. (2016). PNAS, 113(8):2158-2163. [2] Baraud et al. (2009). Annals of Statistics, 37(2):630-672. [3] Bartoszek et al. (2012). J. Th. Biology, 314, 204-204. [4] Bastide et al. (2016). J. R. Stat. Soc. B. [5] Bastide et al. (2017). Submitted. [6] Blomberg et al. (2003). Evolution, 57(4):717-754. [7] Hansen (1997). Evolution, 51(5):1341-1351 [8] Khabbazian et al. (2016). Methods Ecol Evol, 7(7), 811-824. [9] Revell (2009). Evolution, 63(12), 3258-3268.

## Incomplete Data Point of View

[4,5]

$$\mathbf{X}^j | \mathbf{X}^{\text{pa}(j)} \sim \mathcal{N}(\mathbf{x}^{\text{pa}(j)} + \Delta^j, \ell_j \mathbf{R})$$

EM Algorithm Maximize  $\mathbb{E}_{\theta}[\log p_{\theta}(\mathbf{Z}, \mathbf{Y}) | \mathbf{Y}]$ .

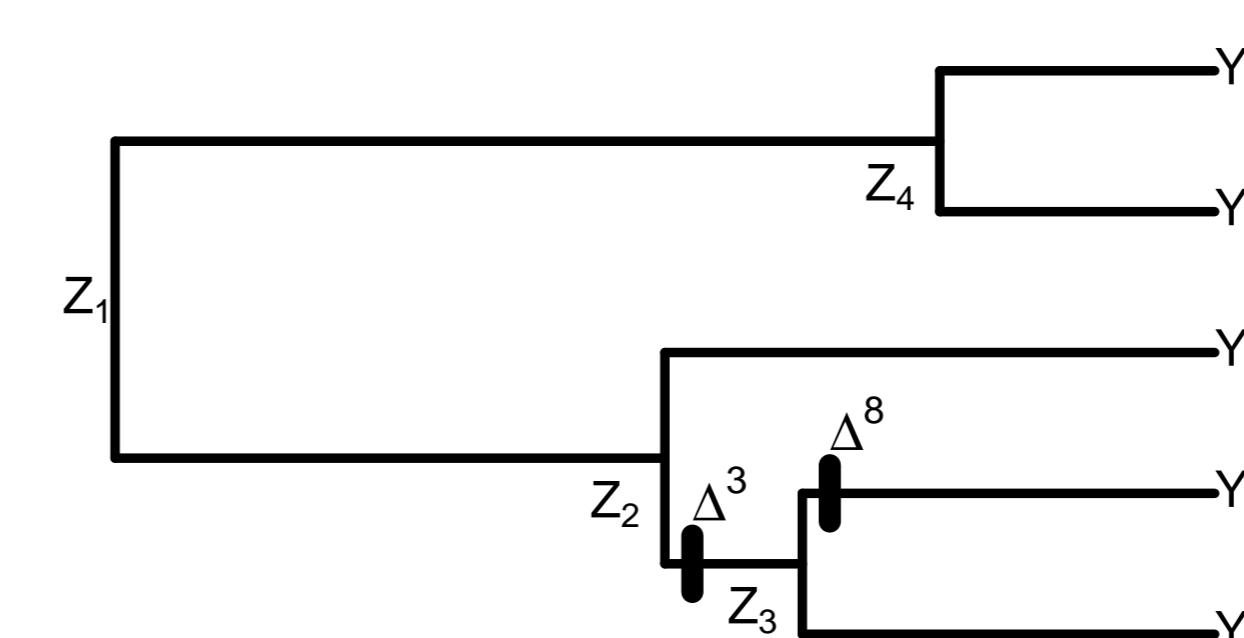
E step "Upward-Downward" Algorithm (with missing data).  $\rightarrow$  C++

M step Explicit formulas.

Initialization LASSO regression.

## Linear Regression Point of View

[4,5]



$$\Delta = \begin{pmatrix} 1 & \cdot & \cdot \\ 2 & \cdot & \cdot \\ 3 & \Delta^3 & \cdot \\ 4 & \cdot & \cdot \\ 1 & \cdot & \cdot \\ 2 & \cdot & \cdot \\ 3 & \cdot & \cdot \\ 4 & \Delta^3 & \Delta^8 \\ 5 & \cdot & \cdot \end{pmatrix} \quad T\Delta = \begin{pmatrix} 1 & \cdot & \cdot \\ 2 & \cdot & \cdot \\ 3 & \cdot & \cdot \\ 4 & \cdot & \cdot \\ 5 & \Delta^3 & \Delta^8 \end{pmatrix}$$

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

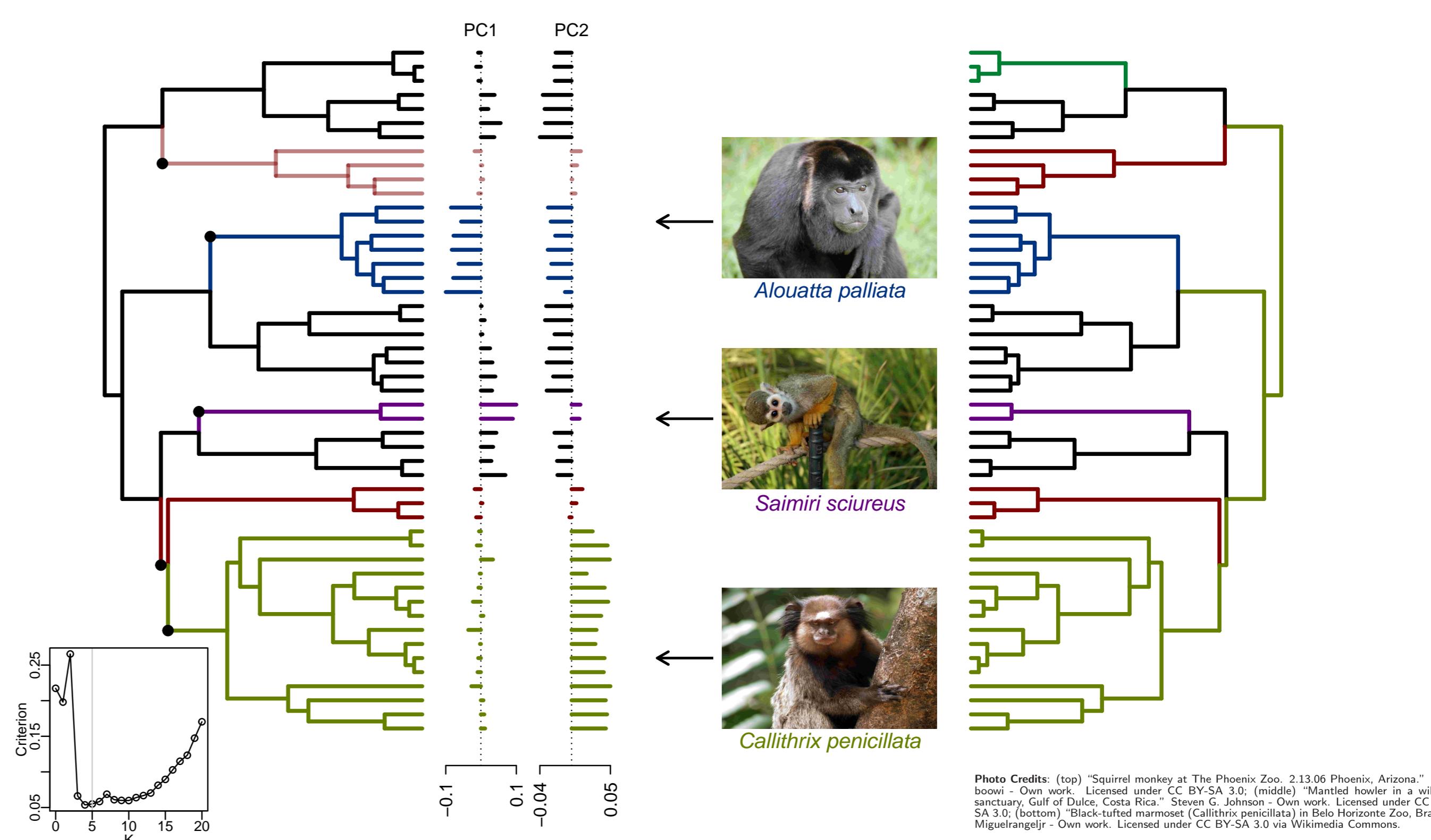
## Model Selection on $K$ ( $\alpha$ known)

[2,4,5]

$$\hat{K} = \underset{K \geq 0}{\operatorname{argmin}} \| \mathbf{Y} - \hat{\mathbf{Y}}_K \|^2_{\mathbf{C}(\alpha)^{-1}} \left( 1 + \frac{\text{pen}(K)}{n - K - 1} \right)$$

## New World Monkeys: Brain Shape

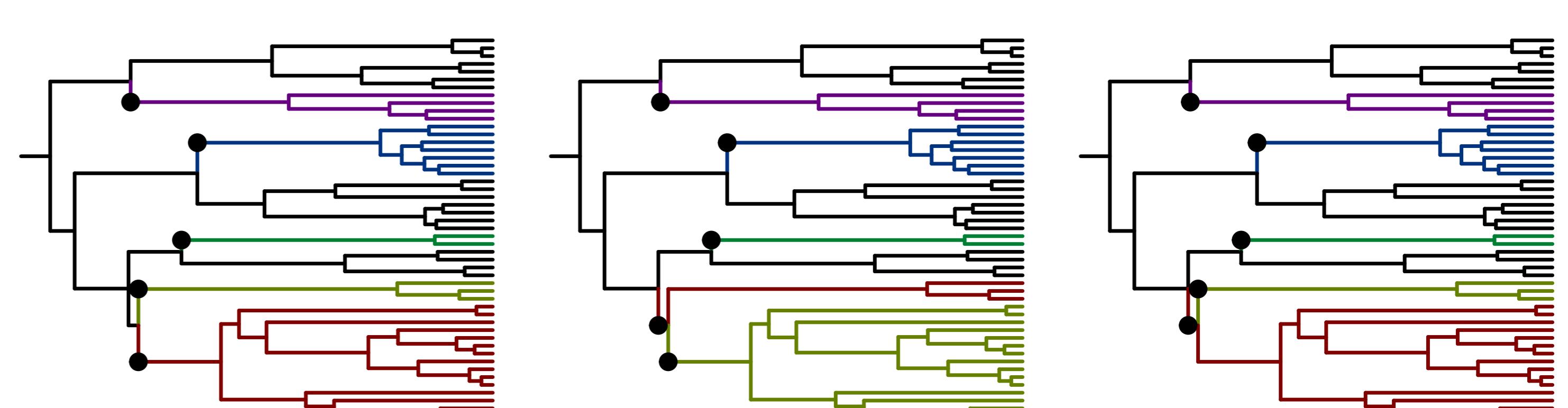
[1,5]



Right: PhylogeneticEM. Left: Habitats.

## Equivalent Solutions

[4,5]



## Conclusion and Perspectives

A general statistical framework for trait evolution models with unconstrained shifts on ultrametric trees.

- R Package PhylogeneticEM (on the CRAN).

### Perspectives:

- Use fossil records (non-ultrametric tree).
- Study convergence of regimes.
- Deal with uncertainty (tree, data).

