

# PhylogeneticEM: An R Package for Change-point Detection on Phylogenetic Trees

C. Ané<sup>1</sup>, Paul Bastide<sup>2</sup>, M. Mariadassou<sup>3</sup>, S. Robin<sup>4</sup>

<sup>1</sup> Department of Statistics and Botany, University of Wisconsin–Madison, USA

<sup>2</sup> Evolutionary and Computational Virology, Rega Institute, KU Leuven, Belgium

<sup>3</sup> MaIAGE, INRA, Jouy-en-Josas, France.

<sup>4</sup> MIA-Paris, INRA - AgroParisTech, Paris, France

8 November 2017



# New World Monkeys

(Aristide et al., 2016)



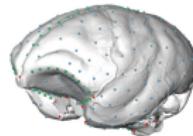
*Callithrix penicillata*

# New World Monkeys

(Aristide et al., 2016)



*Callithrix penicillata*



# New World Monkeys

(Aristide et al., 2016)



*Callithrix penicillata*

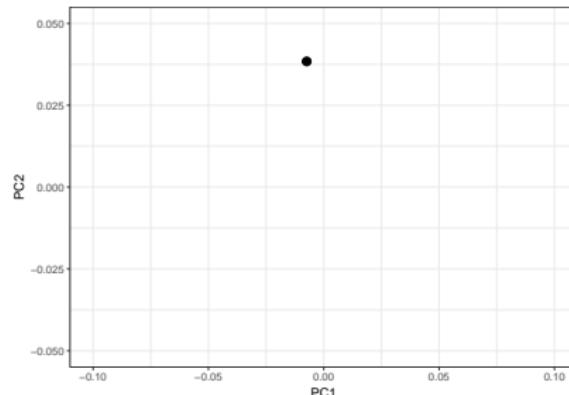
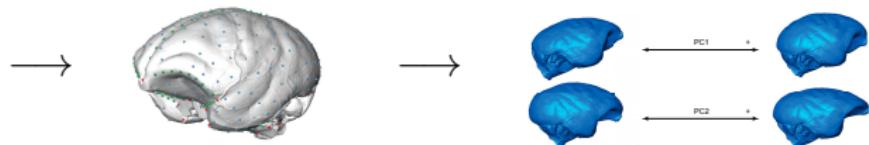


# New World Monkeys

(Aristide et al., 2016)



*Callithrix penicillata*



# New World Monkeys

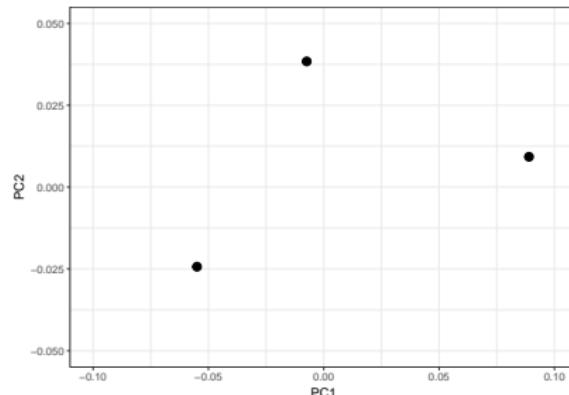
(Aristide et al., 2016)



*Callithrix penicillata*



*Alouatta palliata*



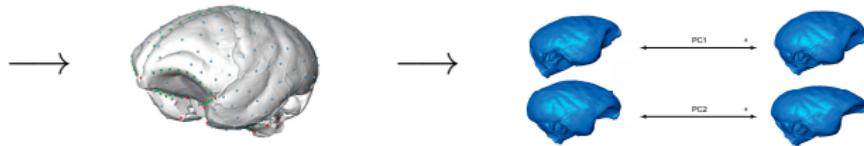
*Saimiri sciureus*

# New World Monkeys

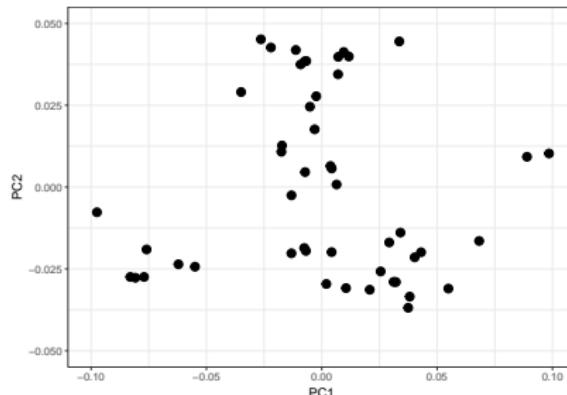
(Aristide et al., 2016)



*Callithrix penicillata*



*Alouatta palliata*



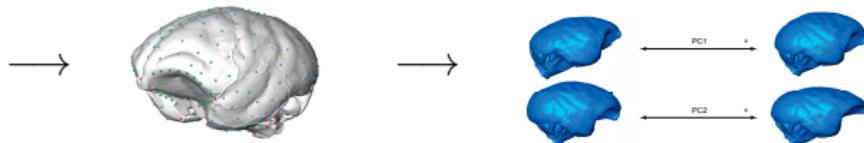
*Saimiri sciureus*

# New World Monkeys

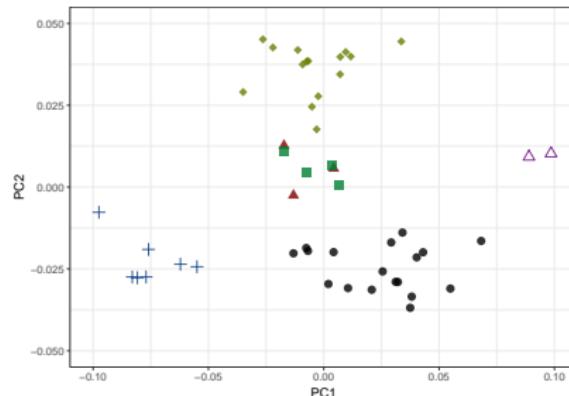
(Aristide et al., 2016)



*Callithrix penicillata*



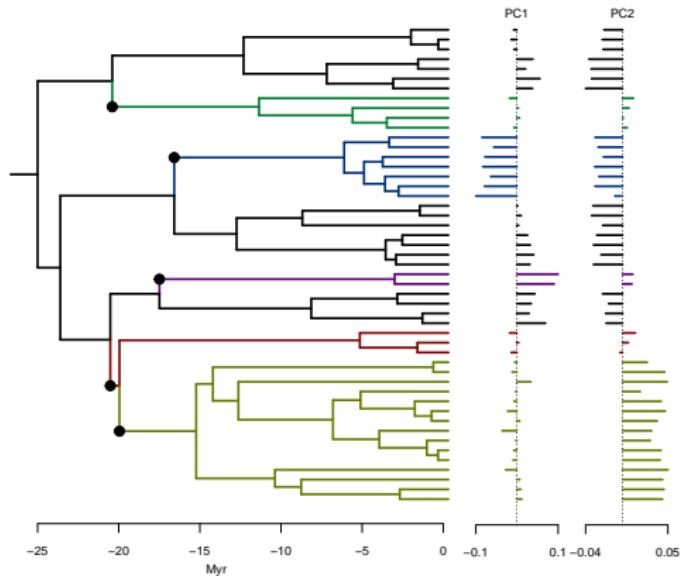
*Alouatta palliata*



*Saimiri sciureus*

# New World Monkeys

(Aristide et al., 2016)



*Alouatta palliata*



*Saimiri sciureus*



*Callithrix penicillata*

# Outline

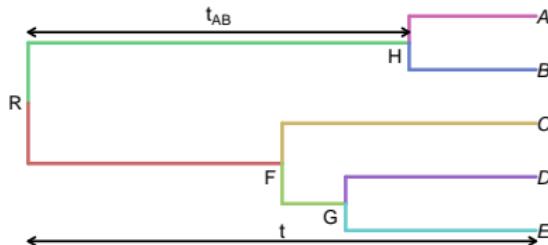
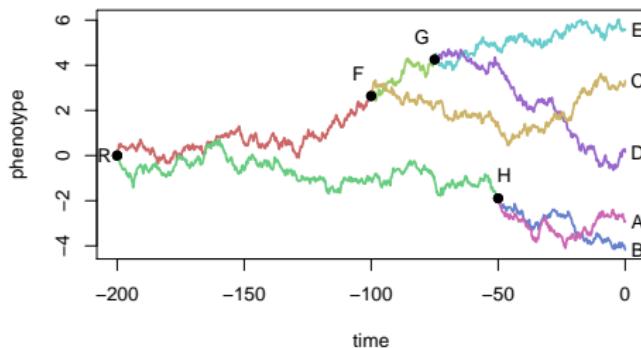
① Stochastic Processes on Trees

② Case Study

③ Advertising

## Stochastic Process on a Tree

(Felsenstein, 1985)

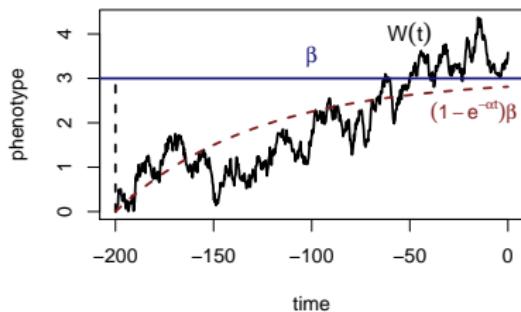
**Known tree.**Only **tips** observed.**Brownian Motion:**

$$\text{Var}[A | R] = \sigma^2 t$$

$$\text{Cov}[A; B | R] = \sigma^2 t_{AB}$$

## OU Modeling

(Hansen, 1997)



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

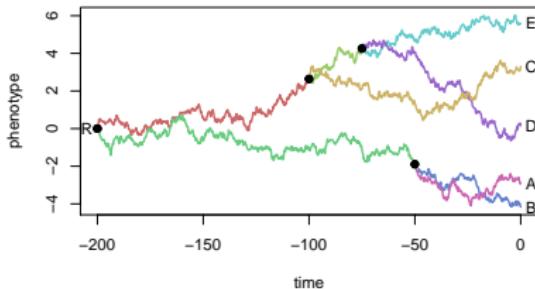
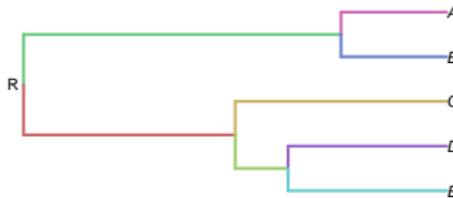
Deterministic part:

- $\beta(t)$ : primary optimum, mechanistically defined.
- $\alpha$ : selection strength.

Stochastic part:

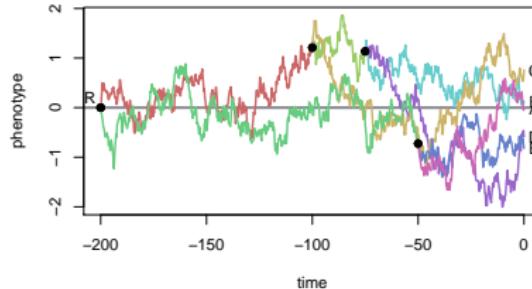
- $W(t)$ : actual optimum (trait value).
- $\sigma dB(t)$  Brownian fluctuations.

# Shifts



**BM Shifts in the mean:**

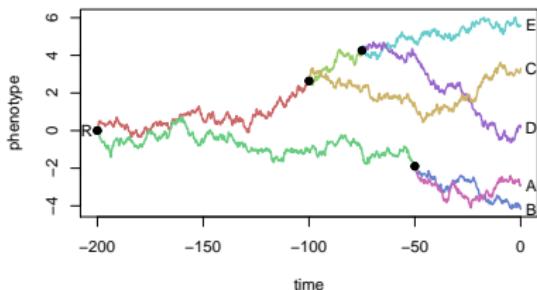
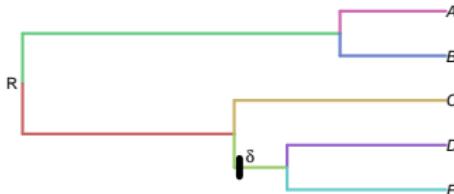
$$m_{\text{child}} = m_{\text{parent}} + \delta$$



**OU Shifts in the optimal value:**

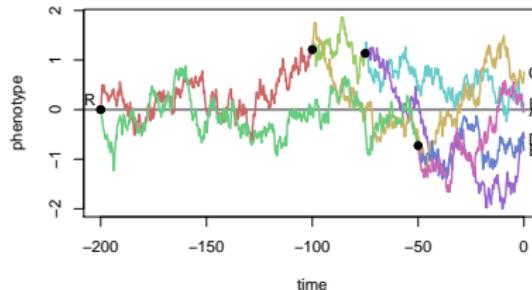
$$\beta_{\text{child}} = \beta_{\text{parent}} + \delta$$

# Shifts



BM Shifts in the **mean**:

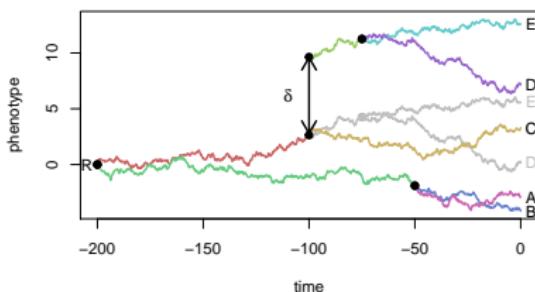
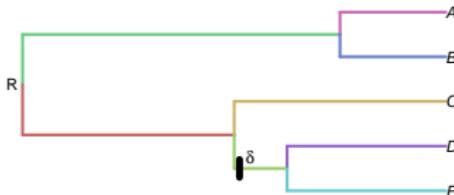
$$m_{\text{child}} = m_{\text{parent}} + \delta$$



OU Shifts in the **optimal value**:

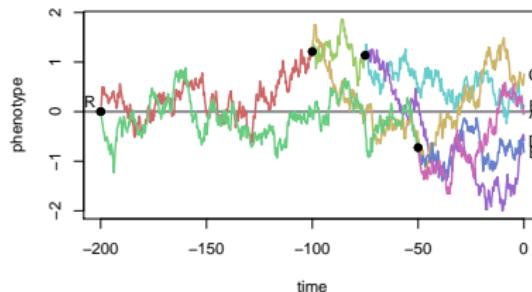
$$\beta_{\text{child}} = \beta_{\text{parent}} + \delta$$

# Shifts



**BM Shifts in the mean:**

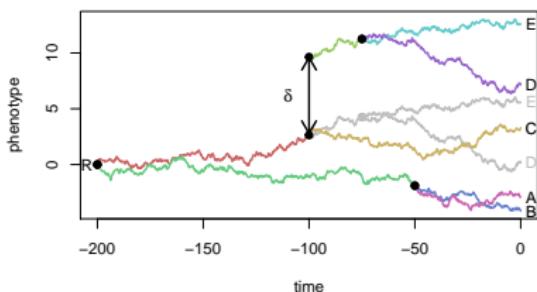
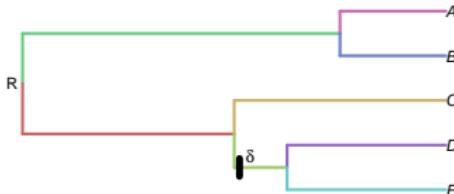
$$m_{\text{child}} = m_{\text{parent}} + \delta$$



**OU Shifts in the optimal value:**

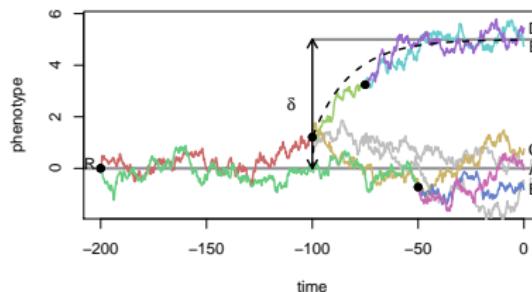
$$\beta_{\text{child}} = \beta_{\text{parent}} + \delta$$

# Shifts



**BM Shifts in the mean:**

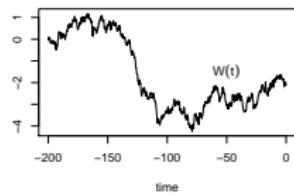
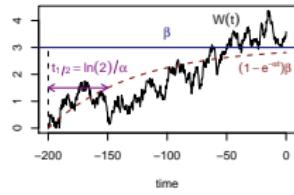
$$m_{\text{child}} = m_{\text{parent}} + \delta$$



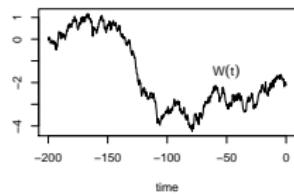
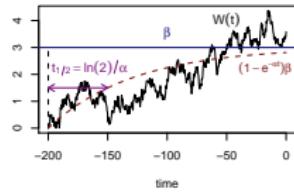
**OU Shifts in the optimal value:**

$$\beta_{\text{child}} = \beta_{\text{parent}} + \delta$$

# Univariate BM vs OU

Equation	$\text{Cov}[Y_i; Y_j]$	Inference
	$dW(t) = \sigma dB(t)$ $t_{ij} \times \sigma^2$ $\therefore$	
	$dW(t) = \sigma dB(t)$ $+ \alpha[\beta - W(t)]dt$ $\frac{1}{2\alpha} e^{-2\alpha h} (e^{2\alpha t_{ij}} - 1) \times \sigma^2$ $\therefore$	

# Univariate BM vs OU

Equation	$\text{Cov}[Y_i; Y_j]$	Inference
	$dW(t) = \sigma dB(t)$ $t_{ij} \times \sigma^2$ $\therefore$	
	$dW(t) = \sigma dB(t)$ $+ \alpha[\beta - W(t)]dt$ $\underbrace{\frac{1}{2\alpha} e^{-2\alpha h} (e^{2\alpha t_{ij}} - 1) \times \sigma^2}_{t'_{ij}(\alpha)}$ $\therefore$	

# Multivariate BM vs OU

→ All the traits shift at the same time

Equation	$\text{Cov} [\mathbf{Y}_i; \mathbf{Y}_j]$	Inference
$d\mathbf{W}(t) = \boldsymbol{\Sigma} d\mathbf{B}_t$	$t_{ij} \times \boldsymbol{\Sigma} \boldsymbol{\Sigma}^T$	:-(
$d\mathbf{W}(t) = -\mathbf{A}(\mathbf{W}(t) - \beta(t))dt + \boldsymbol{\Sigma} d\mathbf{B}_t$	$\int \dots$	:'-(

# Multivariate BM vs scOU

- All the traits shift at the same time
- All the traits shift have the same  $\alpha$ .

Equation	$\text{Cov}[\mathbf{Y}_i; \mathbf{Y}_j]$	Inference
$d\mathbf{W}(t) = \boldsymbol{\Sigma} d\mathbf{B}_t$	$t_{ij} \times \boldsymbol{\Sigma} \boldsymbol{\Sigma}^T$	$\therefore)$
$d\mathbf{W}(t) = -\alpha(\mathbf{W}(t) - \beta(t))dt + \boldsymbol{\Sigma} d\mathbf{B}_t$	$t'_{ij}(\alpha) \times \boldsymbol{\Sigma} \boldsymbol{\Sigma}^T$	$\therefore)$

# Outline

## ① Stochastic Processes on Trees

## ② Case Study

- Simulated Data
- Model Selection
- Monkey Dataset
- Identifiability Problems

## ③ Advertising

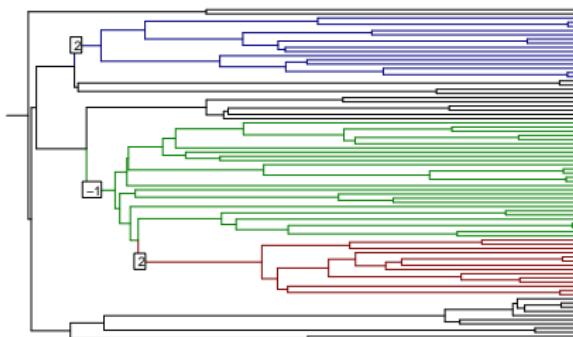
# Simulation: Tree

```
library(PhylogeneticEM)

set.seed(17920902)
ntaxa = 80
tree <- TreeSim::sim.bd.taxa.age(n=ntaxa, numbsim=1, lambda=0.1, mu=0, age=1, mrca=TRUE)[[1]]

params <- params_process("OU",
                           p = 2,                                     ## Process
                           variance = diag(0.5, 2, 2) + 0.5,           ## Dimension
                           selection.strength = 3,                     ## Rate matrix
                           edges = c(29, 25, 127),                    ## Selection Strength
                           values = cbind(c(2, 1), c(-1, 2), c(2, -1))) ## Position of the shifts
                           ## Values of the shifts

plot(params, phylo = tree, traits = 1, value_in_box = TRUE, shifts_bg = "white")
```



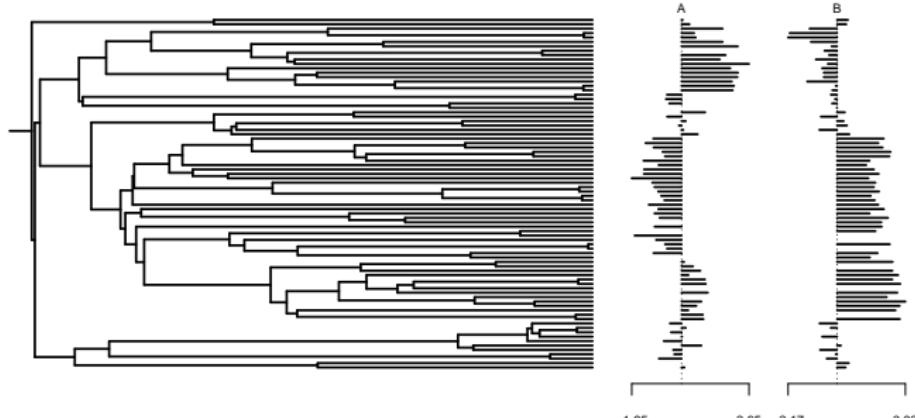
# Simulation: Data

```
sim <- simul_process(params, tree)

data <- extract(sim,           ## A simul_process object
                 what = "states", ## We want the actual values
                 where = "tips") ## Only at the tips of the tree
rownames(data) <- c("A", "B")

nMiss <- floor(ntaxa * 2 * 0.1)
data[sample(c(rep(F, 2*ntaxa - nMiss), rep(T, nMiss)))] <- NA ## 10% of missing data
                                                               ## forget some values

plot(params_BM(p=2), phylo = tree, data = data, edge.width=2)
```



# Inference

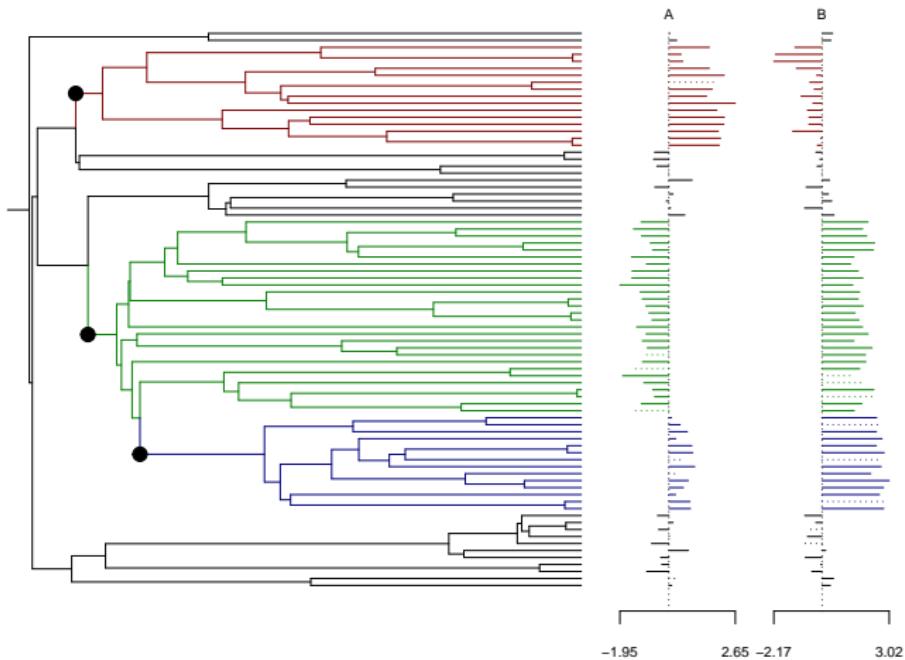
```
## Grid on alpha
alpha_grid <- c(2, 2.5, 3, 3.5)

## Run algorithm
system.time(
res <- PhyloEM(phylo = tree,
               Y_data = data,
               process = "scOU",
               alpha = alpha_grid,
               K_max = 10,
               parallel_alpha = TRUE,
               Ncores = 2)
)

##      user  system elapsed
## 0.536   0.004  57.126
```

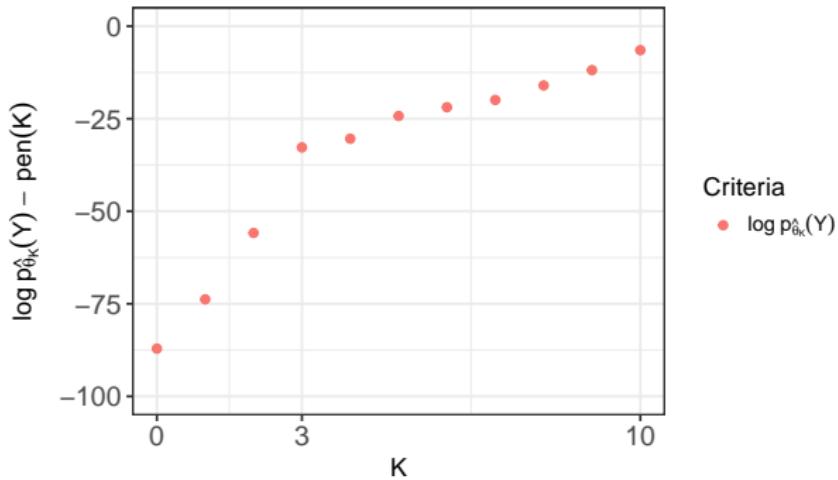
# Analysis

```
plot(res)
```



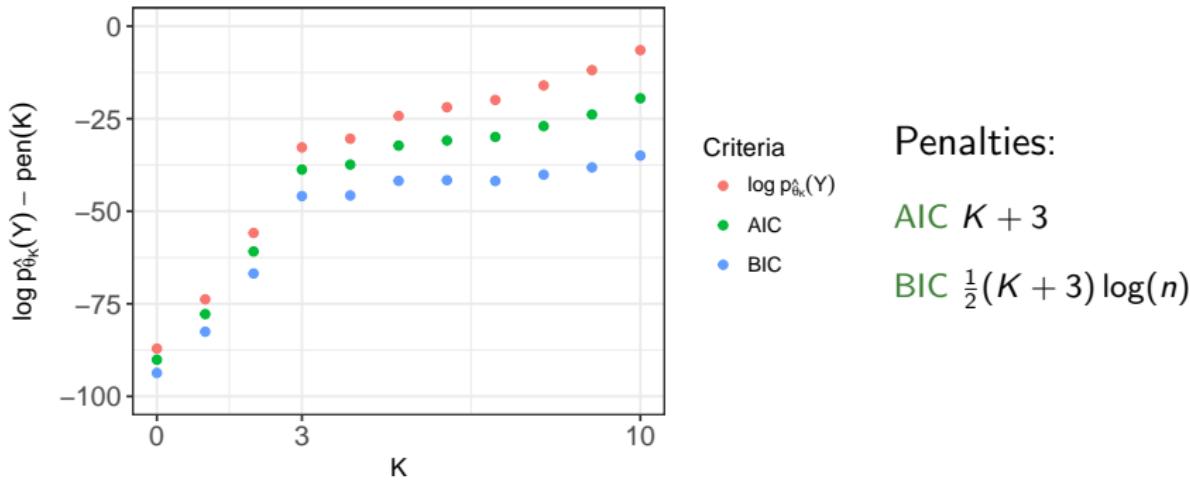
# Model Selection: Penalized Likelihood

Idea  $\hat{K} = \underset{0 \leq K \leq K_{\max}}{\operatorname{argmax}} \left\{ \log p_{\hat{\theta}_K}(Y) - \text{pen}(K) \right\}$  (Univariate)



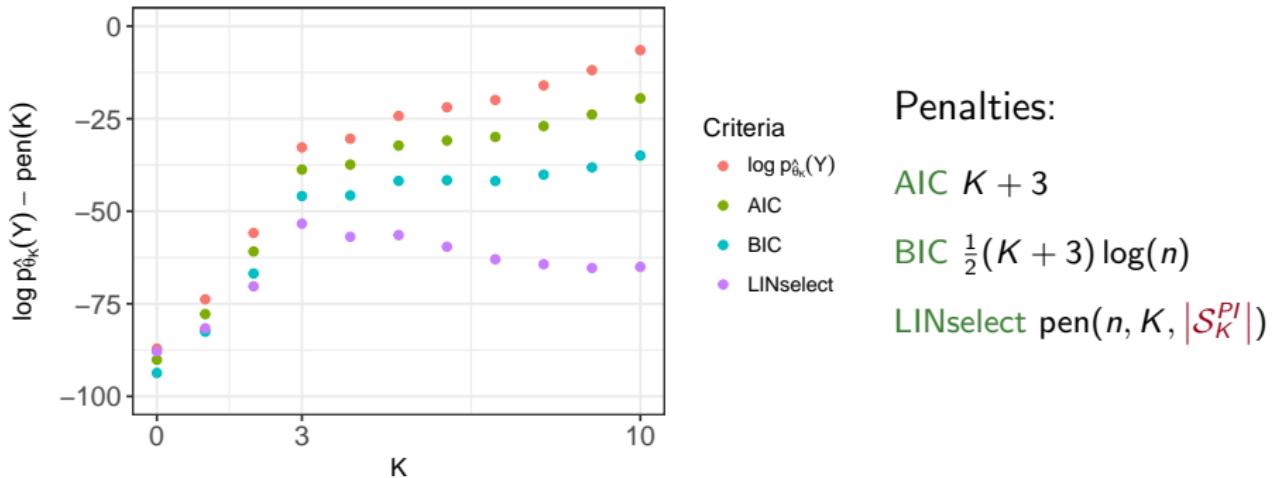
# Model Selection: Penalized Likelihood

Idea  $\hat{K} = \underset{0 \leq K \leq K_{\max}}{\operatorname{argmax}} \left\{ \log p_{\hat{\theta}_K}(Y) - \text{pen}(K) \right\}$  (Univariate)



# Model Selection: Penalized Likelihood

Idea  $\hat{K} = \underset{0 \leq K \leq K_{\max}}{\operatorname{argmax}} \left\{ \log p_{\hat{\theta}_K}(Y) - \text{pen}(K) \right\}$  (Univariate)

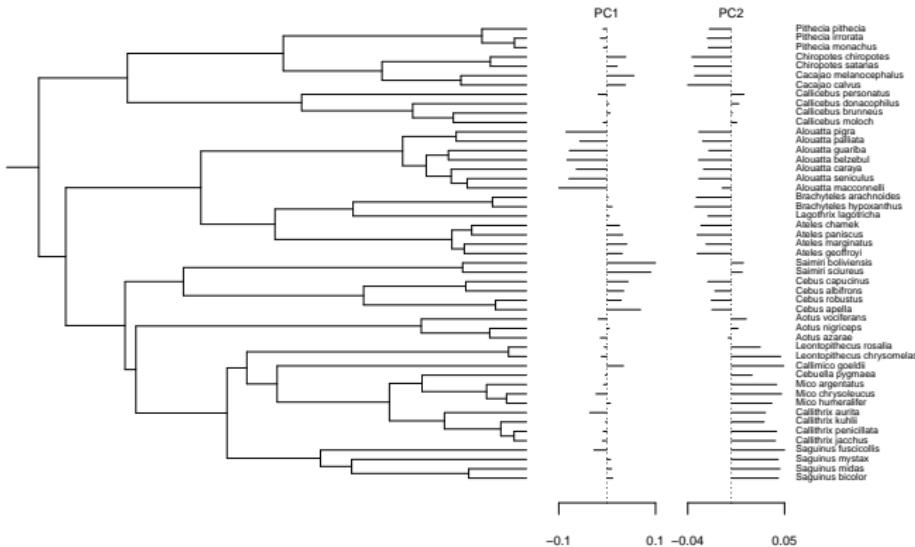


# Monkey Dataset

(Aristide et al., 2016)

```
data(monkeys)
```

```
plot(params_BM(p=2), data = monkeys$dat, phylo = monkeys$phy, show.tip.label = TRUE)
```



# Analysis

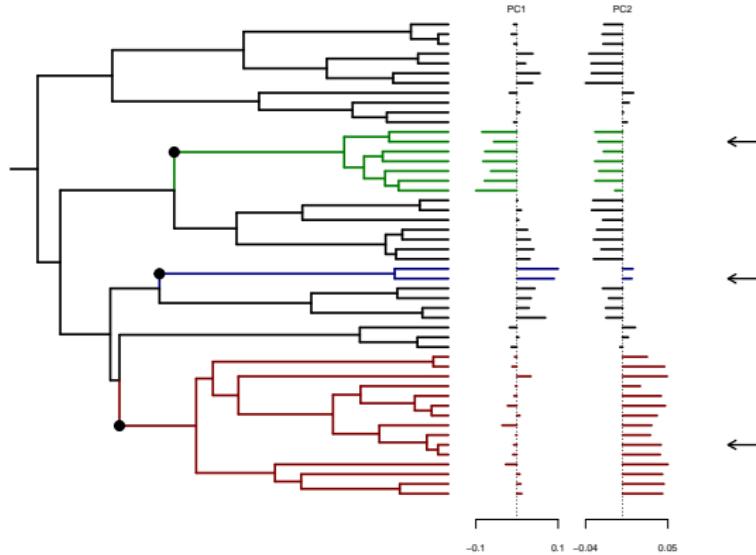
We use function PhyloEM:

```
system.time(
res <- PhyloEM(Y_data = monkeys$dat,           ## data
                 phylo = monkeys$phy,        ## phylogeny
                 process = "scOU",         ## scalar OU
                 K_max = 10,              ## maximal number of shifts
                 nbr_alpha = 4,            ## number of alpha values
                 parallel_alpha = TRUE,   ## parallelize on alpha values
                 Ncores = 2)
##    user  system elapsed
## 0.440  0.004 14.957
```

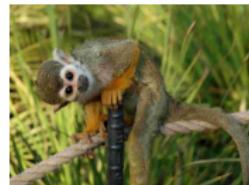
Then plot the solution selected by the default method:

```
plot(res, edge.width = 2)
```

# Result



*Alouatta palliata*



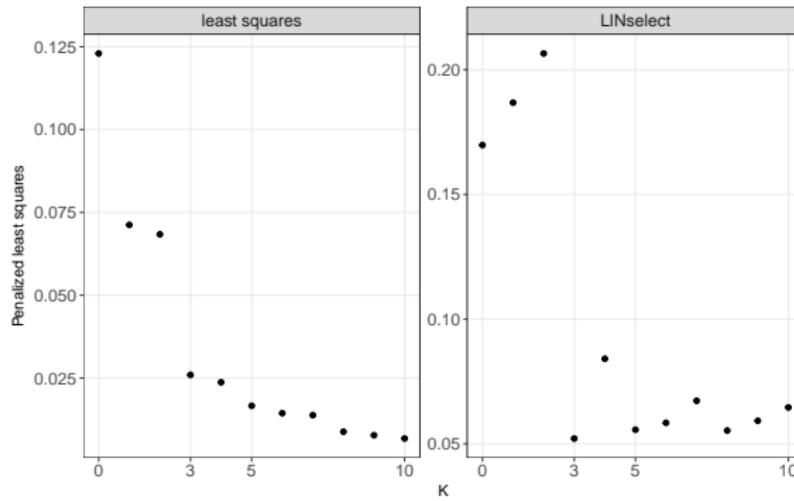
*Saimiri sciureus*



*Callithrix penicillata*

# Model Selection

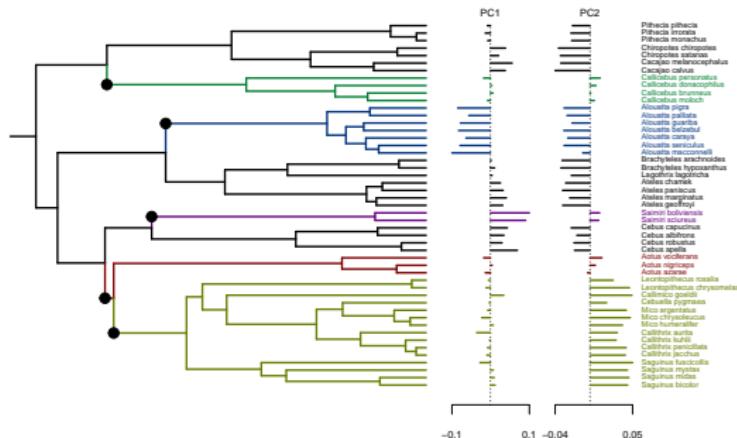
Solution with  $K = 5$  seems to be a good solution too.



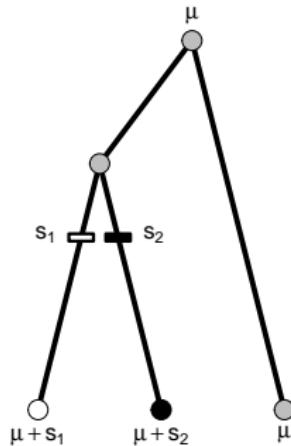
# Solution for $K = 5$

```
plot(res, params = params_process(res, K = 5), edge.width = 2, show.tip.label = TRUE)
```

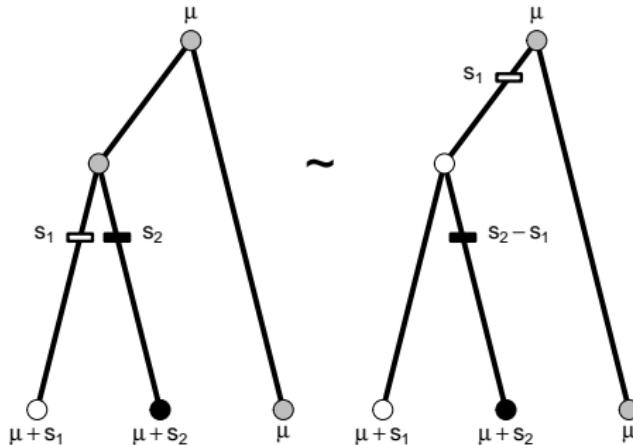
```
## Warning in params_process.PhylоМ(res, K = 5): There are several equivalent solutions for  
this shift position.
```



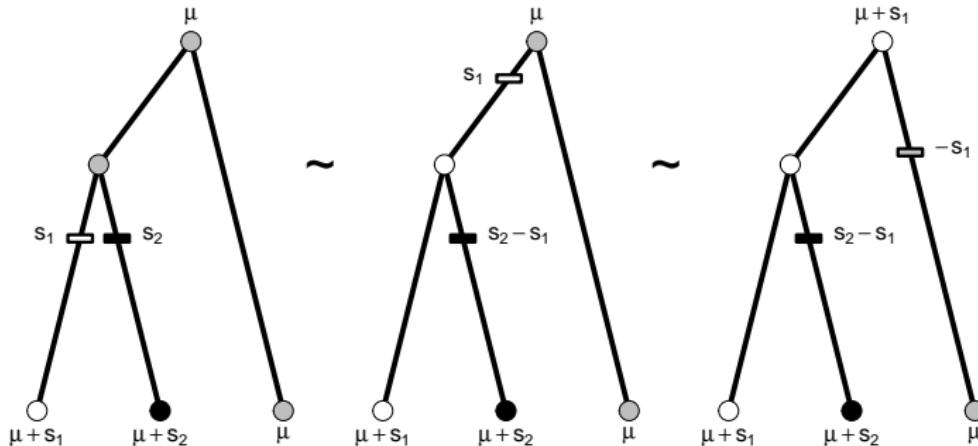
# Equivalencies



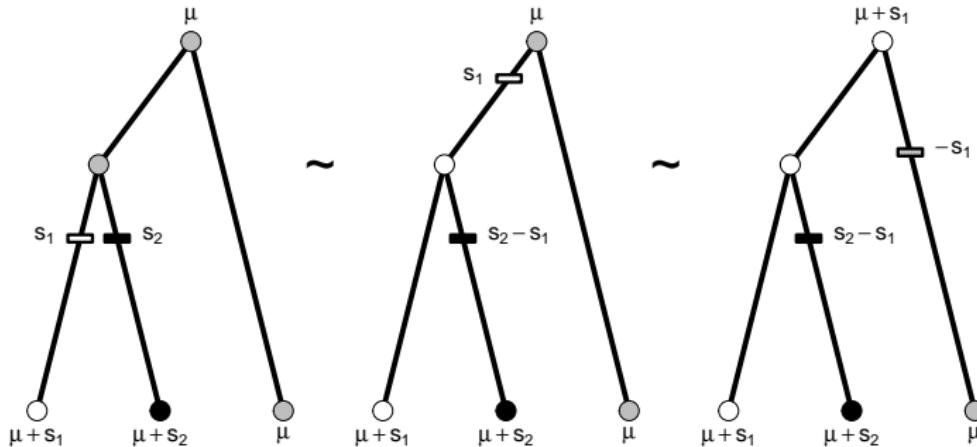
# Equivalencies



# Equivalencies

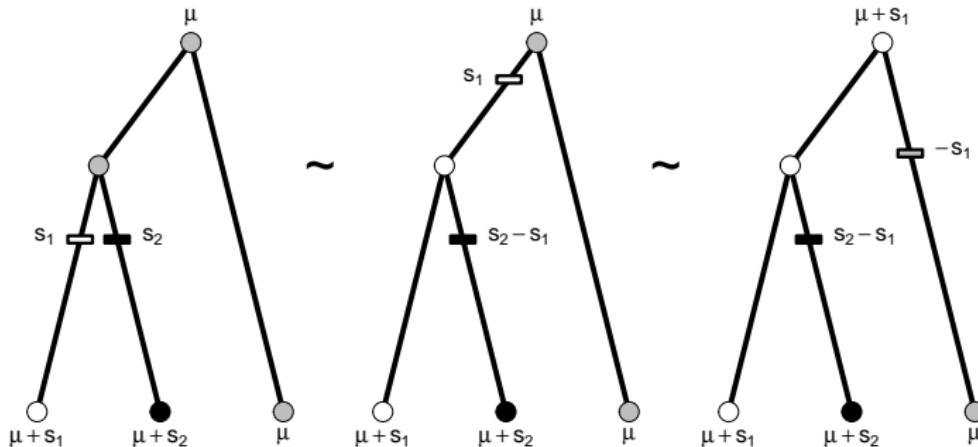


# Equivalencies



Equivalent allocations *cannot* be distinguished from the data.

# Equivalencies



Equivalent allocations *cannot* be distinguished from the data.

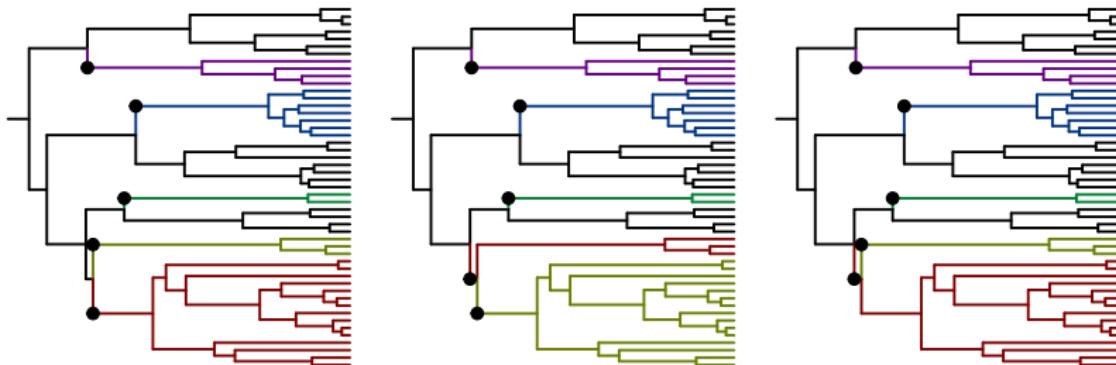
PhylogeneticEM Enumerate all solutions

(Adaptation of Fitch, Sankoff, see Felsenstein, 2004).

# Solution for $K = 5$

```
params_5 <- params_process(res, K = 5)
eq_shifts <- equivalent_shifts(monkeys$phy, params_5)
```

```
plot(eq_shifts, show_shifts_values = FALSE, shifts_cex = 0.5)
```



# Outline

① Stochastic Processes on Trees

② Case Study

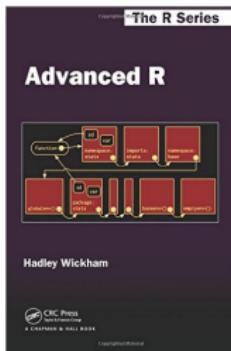
③ Advertising

- Transparency
- Efficiency
- Robustness

# References

I used mainly these two books by Hadley Wickham:

- Advanced R: <http://adv-r.had.co.nz/>
- R Packages: <http://r-pkgs.had.co.nz/>



- Intensive use of devtools and Rstudio.

# Transparency

All the code is available on GitHub:

<https://github.com/pbastide/PhylogeneticEM>

- Package code with version control.
- Works with the CRAN: tag versions.
- Build status with Travis CI.
- Simulations and test cases (reproducibility).
- Automatic documentation with pkgdown.

# Efficiency

- Profiling with `lineprof`.
- Bottleneck: big matrix allocations.
- Solution: use `RcppArmadillo` to code an efficient algorithm (upward-downward).
- Avoid memory leaks with `valgrind`.
- Life saver: the `Armadillo` library.
- Parallel computations with `doParallel` and `foreach`.

# Robustness

- Unitary tests with `testthat`.
- Coverage with `covr` and `codecov`.
- Automated tests with `Travis CI`.

# Conclusion and Perspectives

A general inference framework for trait evolution models.

- |              |   |
|--------------|---|
| Conclusions  | <ul style="list-style-type: none"><li>• A complete maximum likelihood procedure</li><li>• Taking identifiability problems into account</li><li>• With model selection</li></ul>                         |
| R Package    | <ul style="list-style-type: none"><li>• Available on the CRAN and on GitHub</li><li>• Can scale up to big datasets (<math>\sim 1200</math> species)</li></ul>   |
| Perspectives | <ul style="list-style-type: none"><li>• Deal with uncertainty (data, tree, shifts).</li><li>• Phylogenetic networks: See julia package PhyloNetworks.</li><li>• Combine with factor analysis.</li></ul> |

# Bibliography

Bastide P, Mariadassou M, Robin S. 2017. Detection of adaptive shifts on phylogenies by using shifted stochastic processes on a tree. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4), 1067–1093.

Bastide P, Ané C, Robin S, Mariadassou M. 2017. Inference of Adaptive Shifts for Multivariate Correlated Traits. *Systematic Biology, under minor revisions*.

Aristide L, dos Reis SF, Machado AC, Lima I, Lopes RT, Perez SI. 2016. Brain shape convergence in the adaptive radiation of New World monkeys. *Proceedings of the National Academy of Sciences*. 113:2158–2163.

Felsenstein J. 1985. Phylogenies and the Comparative Method. *The American Naturalist*. 125:1–15.

Felsenstein J. 2004. Inferring Phylogenies.

Hansen TF. 1997. Stabilizing Selection and the Comparative Analysis of Adaptation. *Evolution*. 51:1341.

## Photo Credits:

- "Black-tufted marmoset (*Callithrix penicillata*) in Belo Horizonte Zoo, Brazil." Miguelrangeljr - Own work. Licensed under CC BY-SA 3.0

- "Mantled howler in a wildlife sanctuary, Gulf of Dulce, Costa Rica." Steven G. Johnson - Own work. Licensed under CC BY-SA 3.0

- "Squirrel monkey at The Phoenix Zoo. 2.13.06 Phoenix, Arizona." Braboowi - Own work. Licensed under CC BY-SA 3.0

Thank you for listening



[pbastide.github.io](https://pbastide.github.io)