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

C. Ané[1], Paul Bastide[2], M. Mariadassou[3], S. Robin[4]

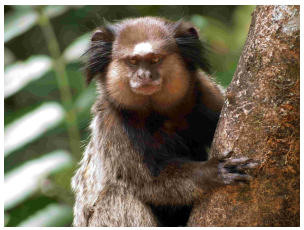
[1] Department of Statistics and Botany, University of Wisconsin–Madison, USA
[2] Evolutionary and Computational Virology, Rega Institute, KU Leuven, Belgium
[3] MaIAGE, INRA, Jouy-en-Josas, France.
[4] MIA-Paris, INRA - AgroParisTech, Paris, France

8 November 2017



New World Monkeys

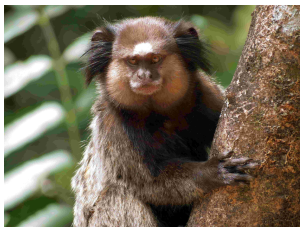
(Aristide et al., 2016)



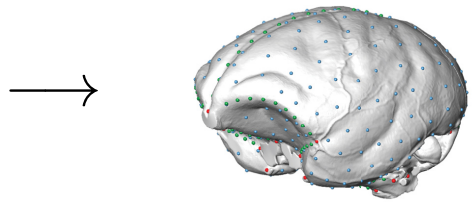
Callithrix penicillata

New World Monkeys

(Aristide et al., 2016)

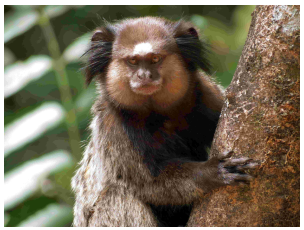


Callithrix penicillata

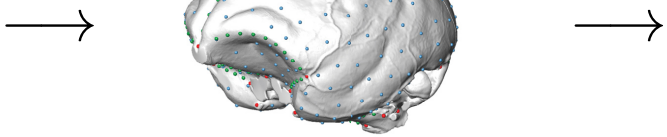


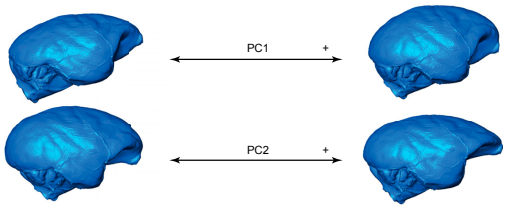
New World Monkeys

(Aristide et al., 2016)



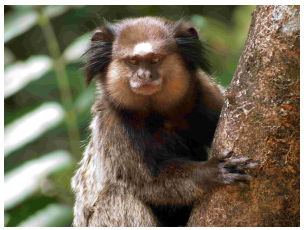
Callithrix penicillata



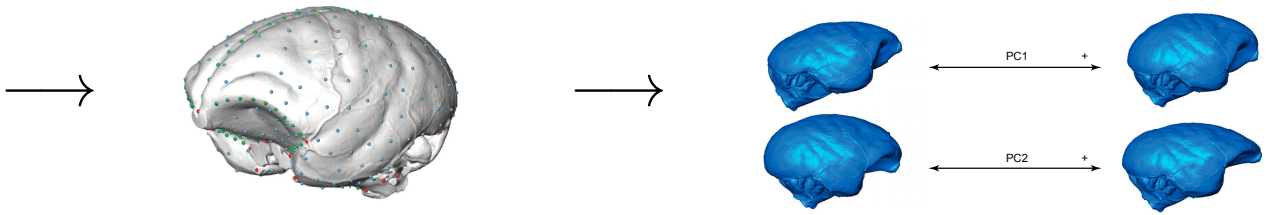


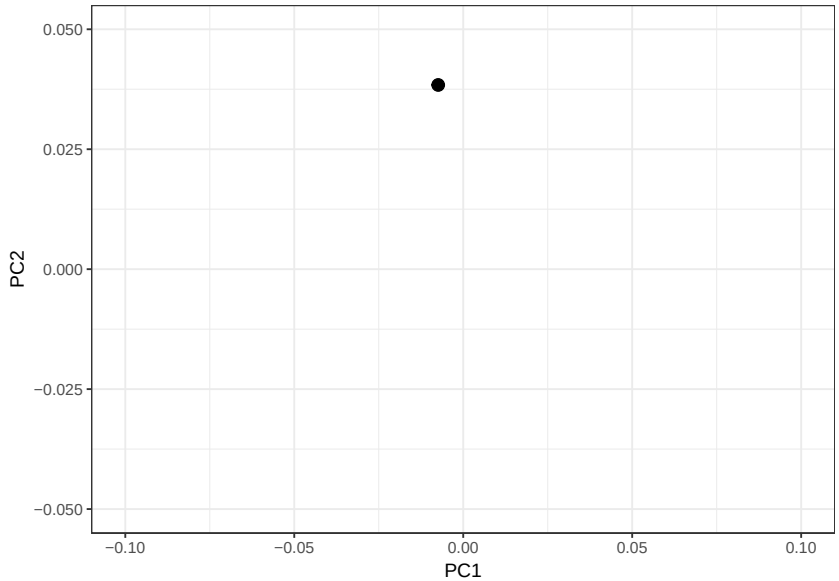
New World Monkeys

(Aristide et al., 2016)



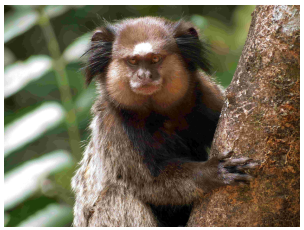
Callithrix penicillata



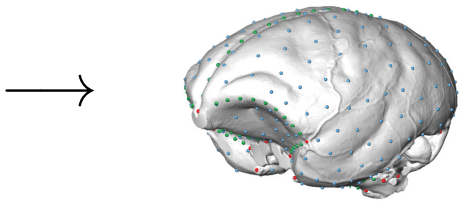


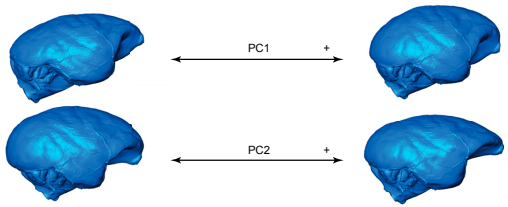
New World Monkeys

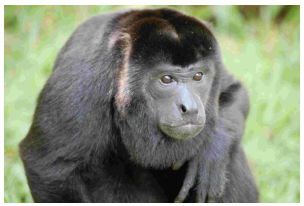
(Aristide et al., 2016)



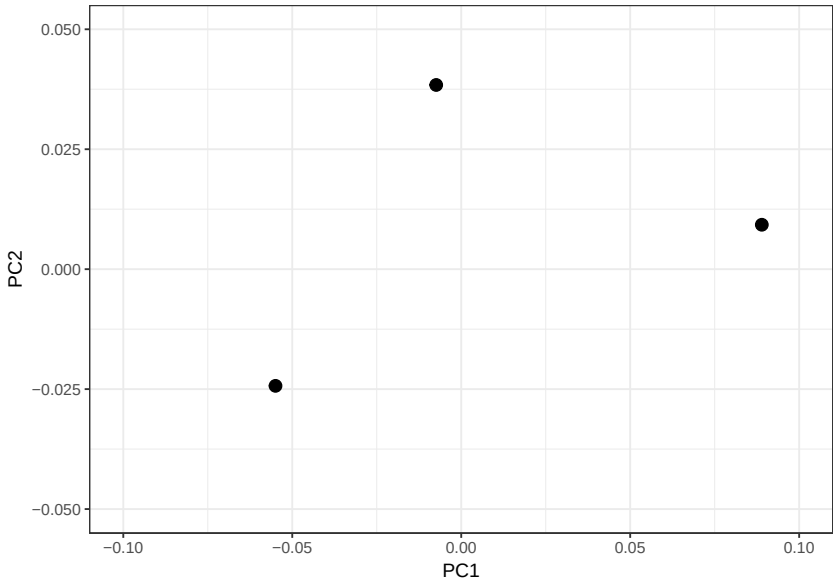
Callithrix penicillata

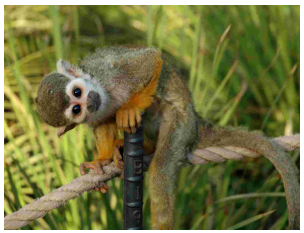






Alouatta palliata

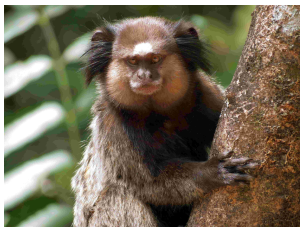




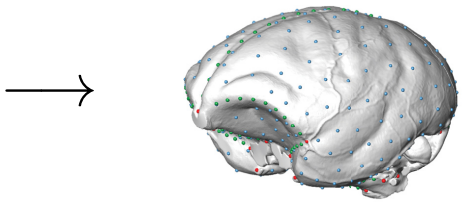
Saimiri sciureus

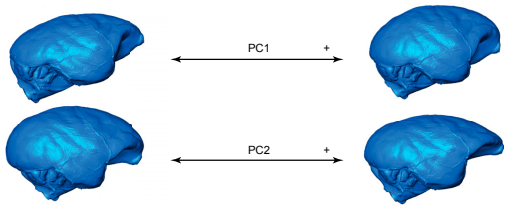
New World Monkeys

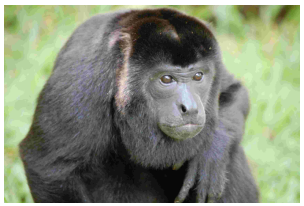
(Aristide et al., 2016)



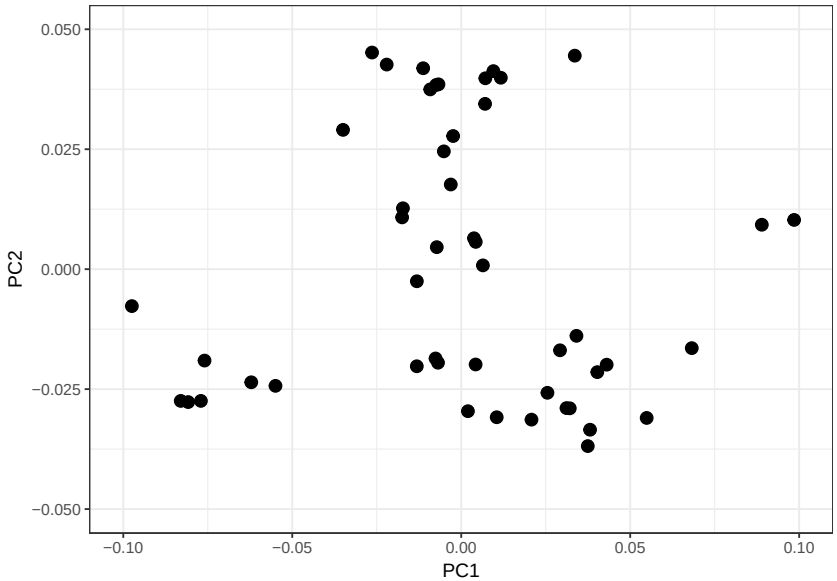
Callithrix penicillata

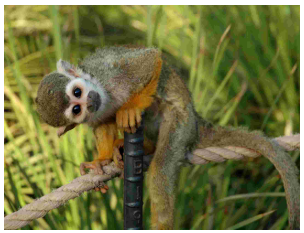






Alouatta palliata

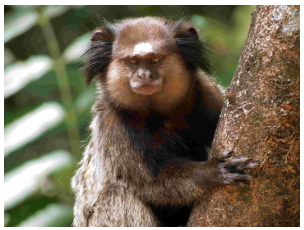




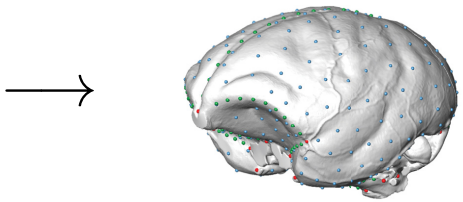
Saimiri sciureus

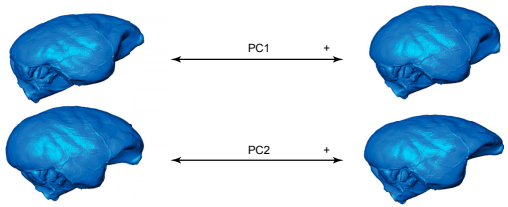
New World Monkeys

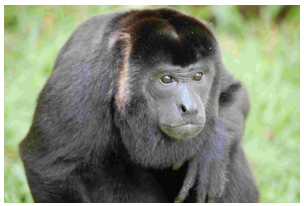
(Aristide et al., 2016)



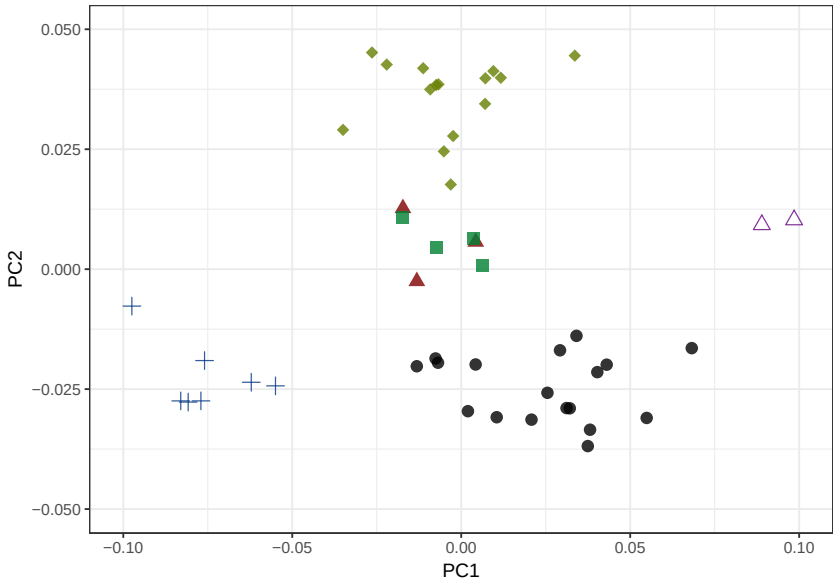
Callithrix penicillata

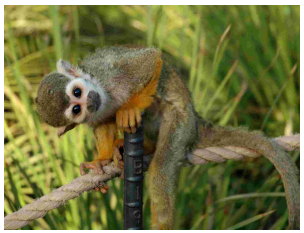






Alouatta palliata

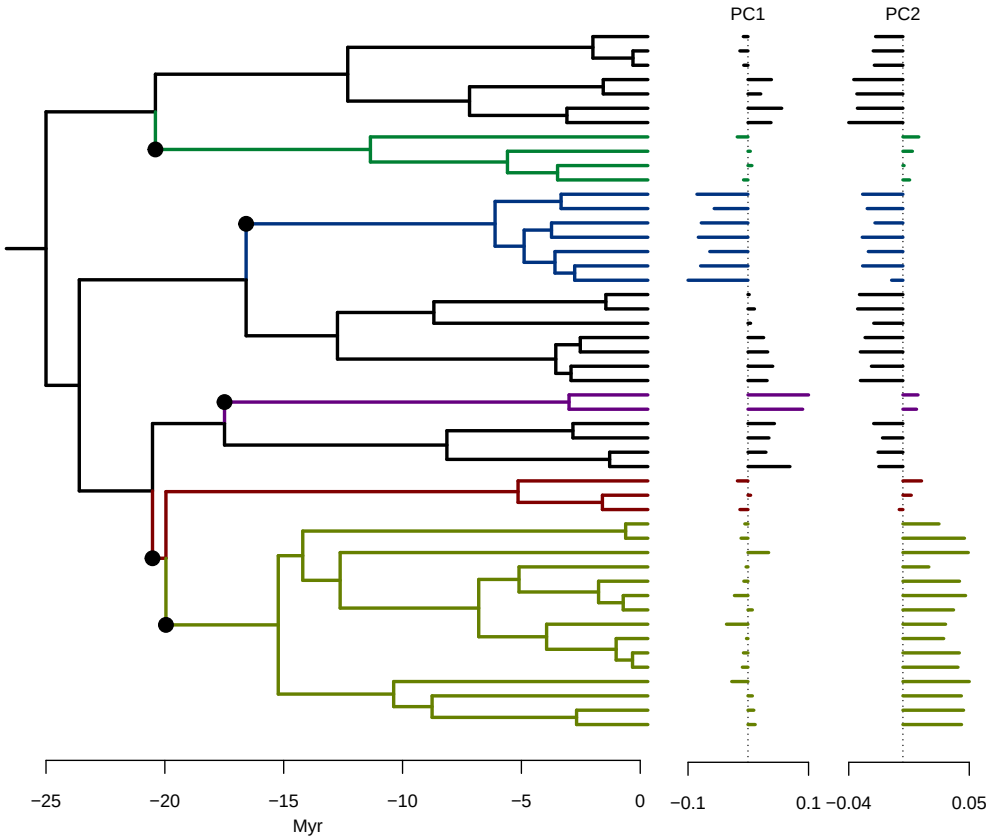


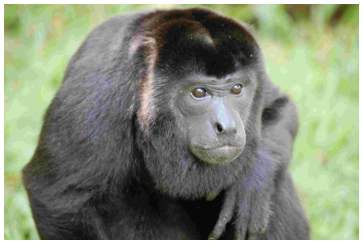


Saimiri sciureus

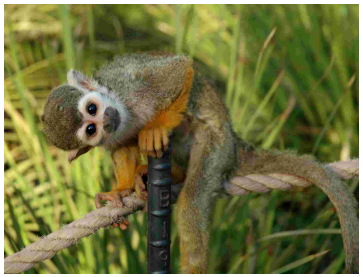
New World Monkeys

(Aristide et al., 2016)

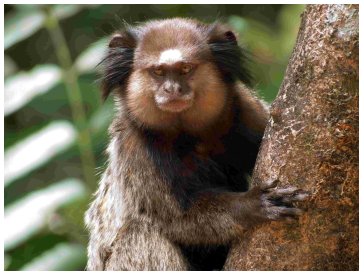




Alouatta palliata



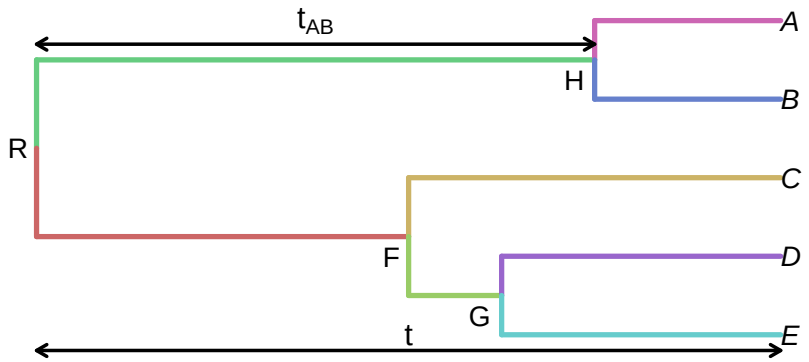
Saimiri sciureus

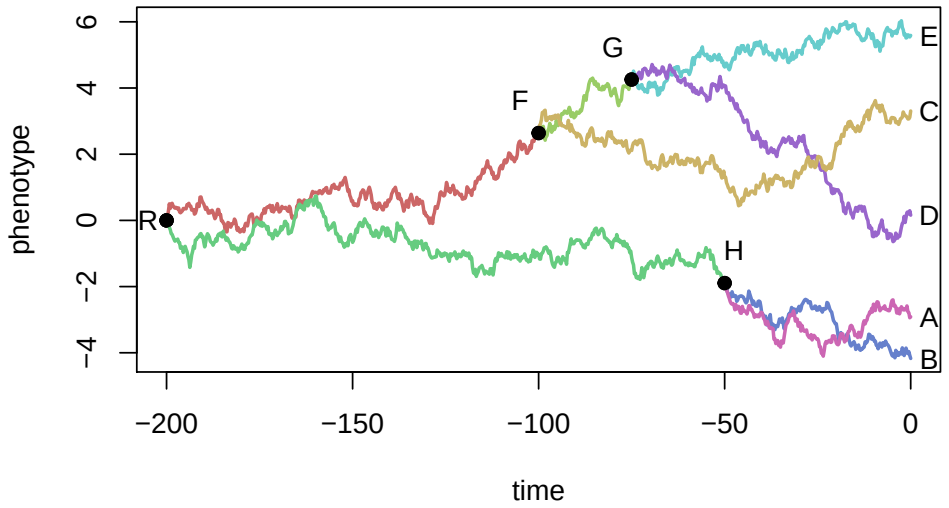


Callithrix penicillata

Outline

Stochastic Process on a Tree (Felsenstein, 1985)





Known tree.

Only **tips** observed.

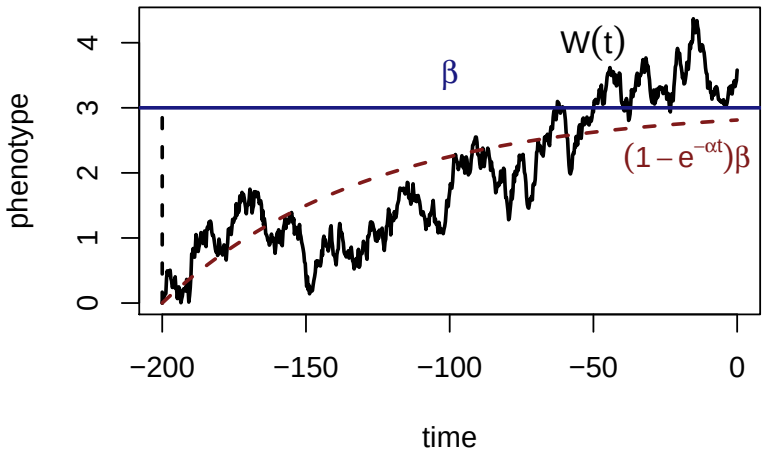
Brownian Motion:

$$\mathbb{V}\mathrm{ar}\left[A \mid R\right] = \sigma^2 t$$

$$\mathbb{C}\mathrm{ov}\left[A; B \mid R\right] = \sigma^2 t_{AB}$$

OU Modeling

(Hansen, 1997)



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

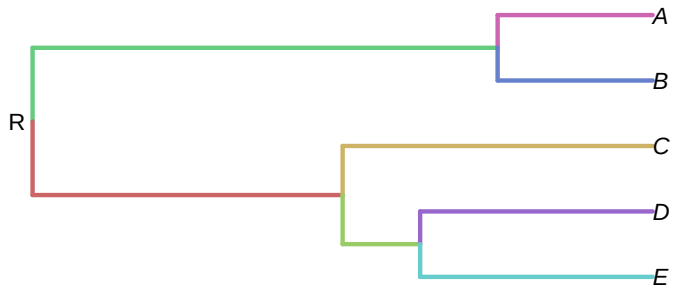
Deterministic part:

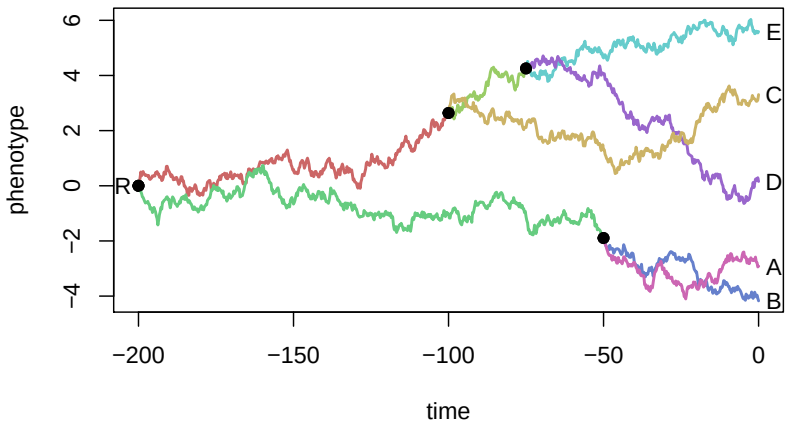
- $\beta(t)$: primary optimum, mechanistically defined.
- α: selection strength.

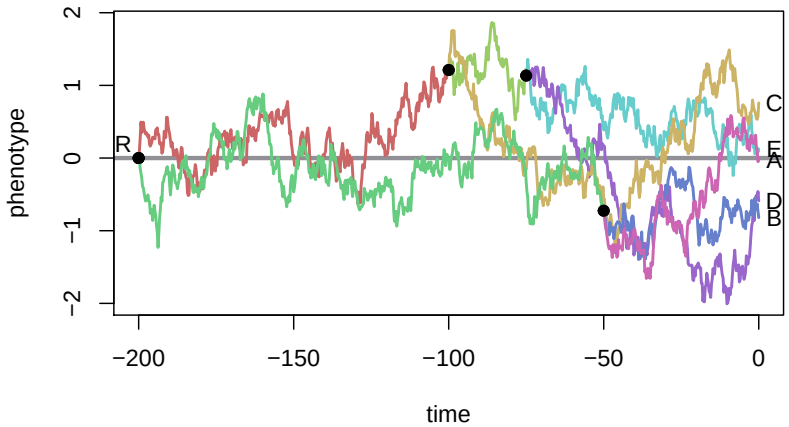
Stochastic part:

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

Shifts







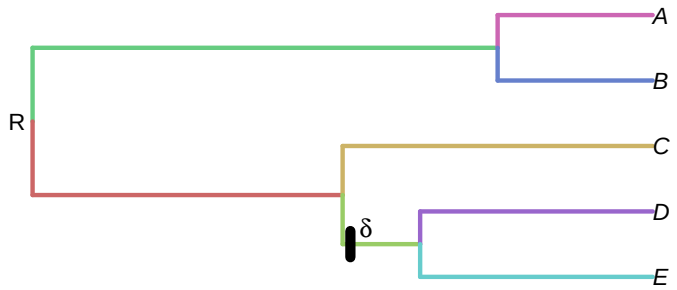
BM Shifts in the **mean**:

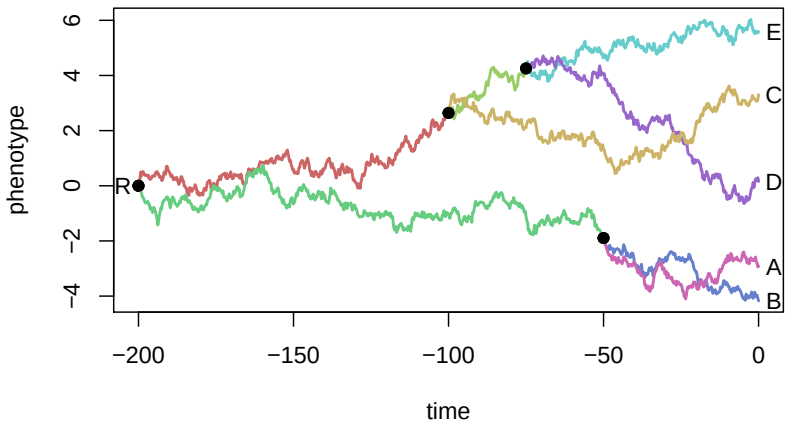
$$m_{\text{child}} = m_{\text{parent}} + \delta$$

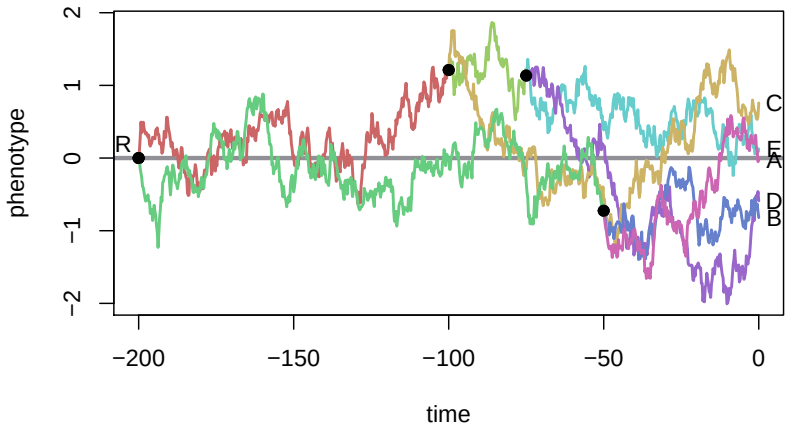
OU Shifts in the **optimal value**:

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

Shifts







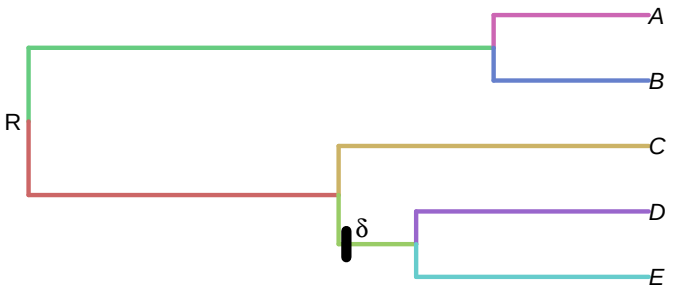
BM Shifts in the **mean**:

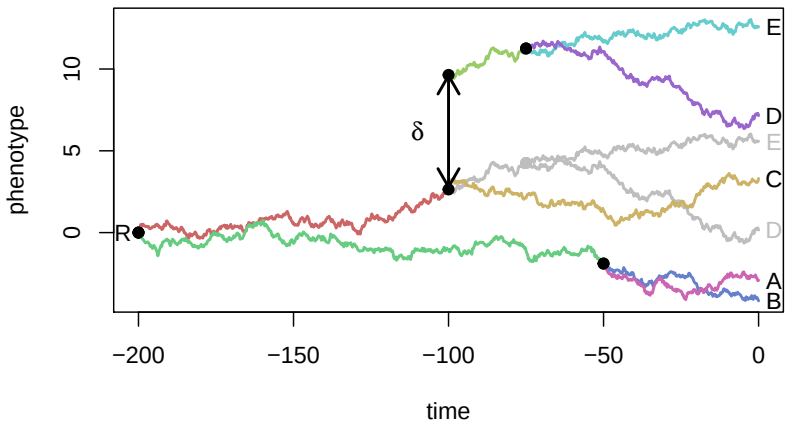
$$m_{\text{child}} = m_{\text{parent}} + \delta$$

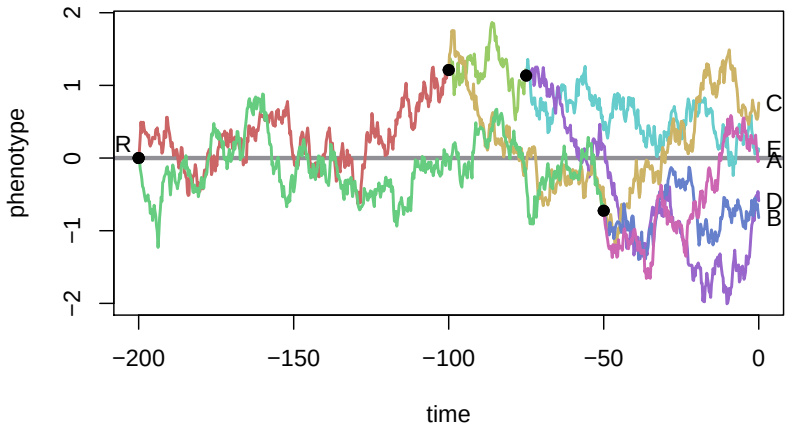
OU Shifts in the **optimal value**:

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

Shifts







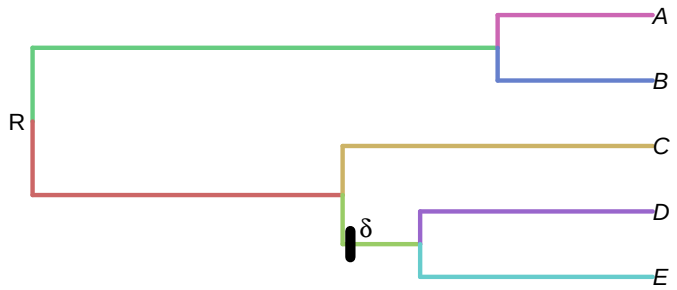
BM Shifts in the **mean**:

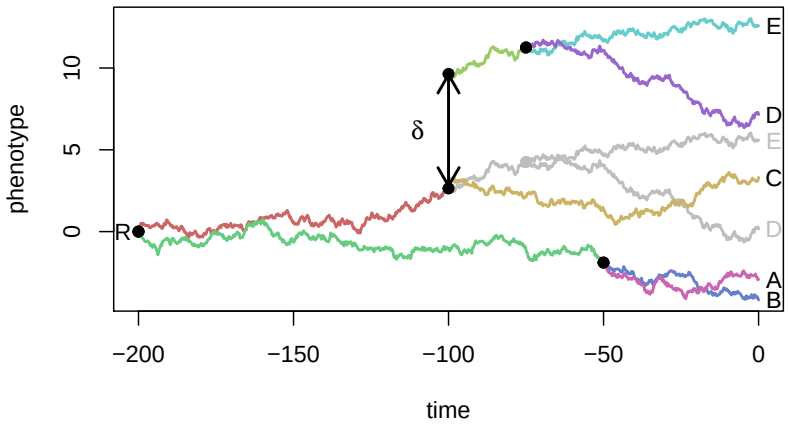
$$m_{\text{child}} = m_{\text{parent}} + \delta$$

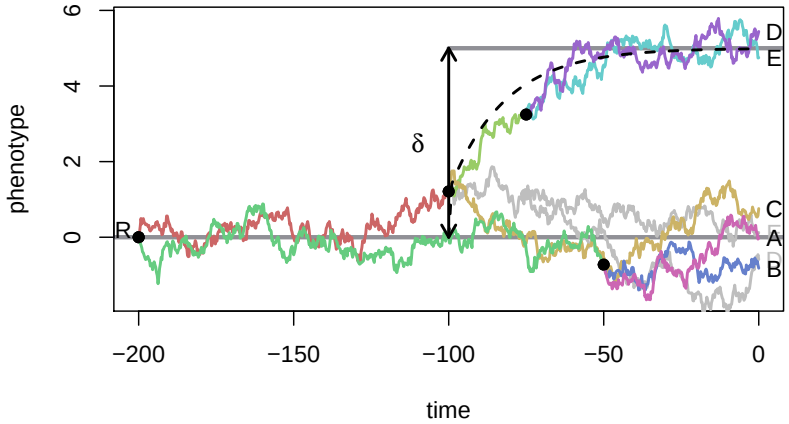
OU Shifts in the **optimal value**:

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

Shifts







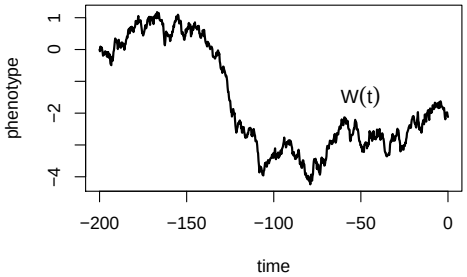
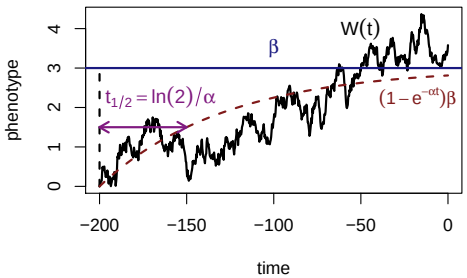
BM Shifts in the **mean**:

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

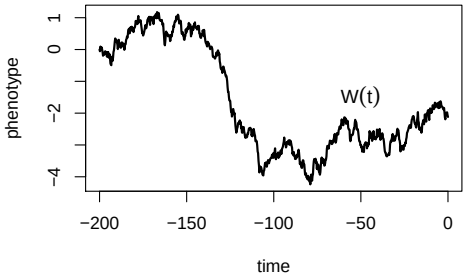
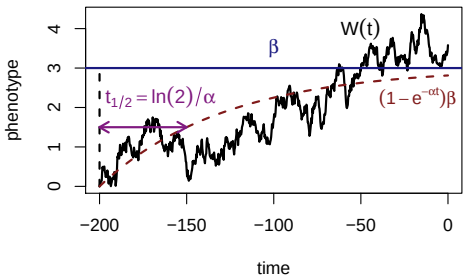
OU Shifts in the **optimal value**:

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

Univariate BM vs OU

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

Univariate BM vs OU

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

Multivariate BM vs OU

$\mapsto$ All the traits shift at the same time

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

Multivariate BM vs **scOU**

$\mapsto$ All the traits shift at the same time

$\mapsto$ All the traits shift have the same α.

Equation	$\mathbb{C}\text{ov}\left[\mathbf{Y}_i; \mathbf{Y}_j\right]$	Inference
$d\mathbf{W}(t) = \mathbf{\Sigma} d\mathbf{B}_t$	$t_{ij} \times \mathbf{\Sigma}\mathbf{\Sigma}^T$	:-)
$d\mathbf{W}(t) = -\alpha(\mathbf{W}(t) - \beta(t))dt + \mathbf{\Sigma} d\mathbf{B}_t$	$t'_{ij}(\alpha) \times \mathbf{\Sigma}\mathbf{\Sigma}^T$	:-)

Outline

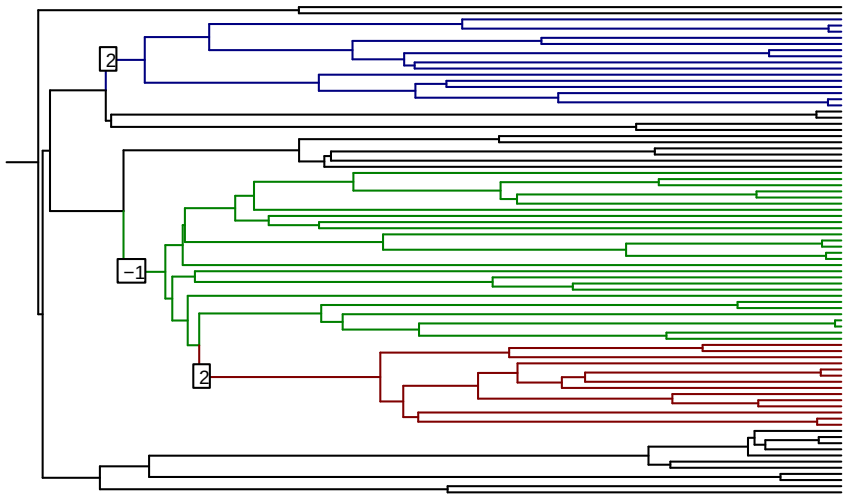
Simulation: Tree

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

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



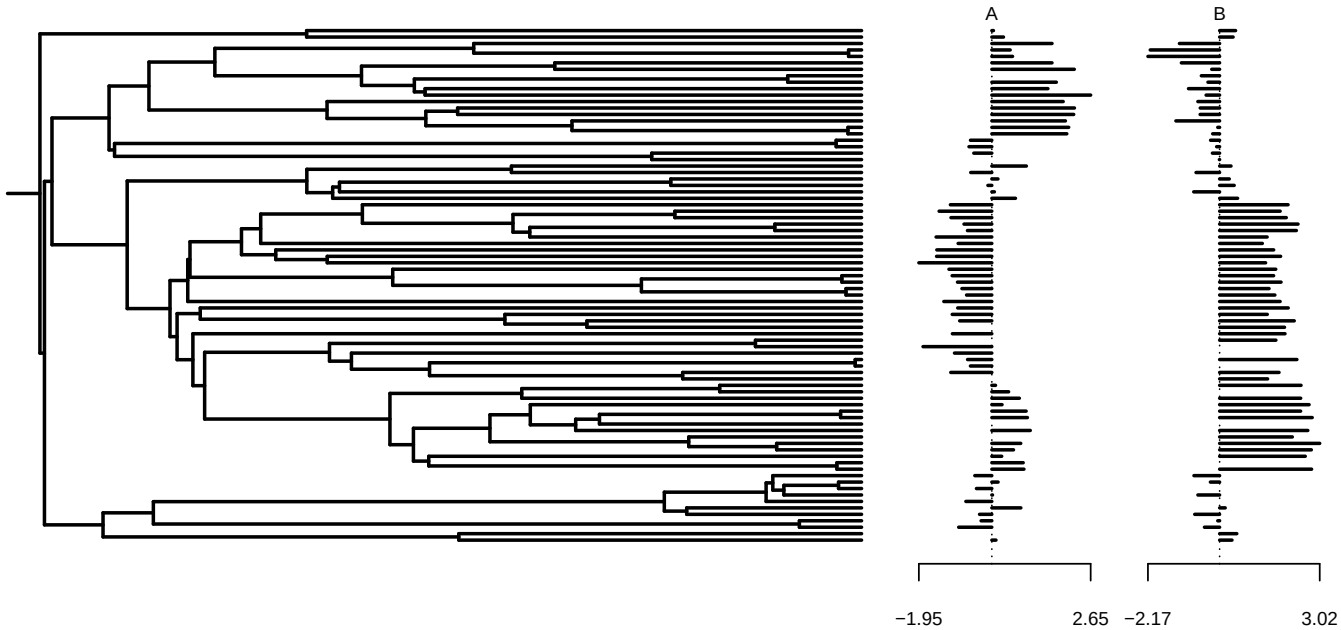
Simulation: Data

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

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

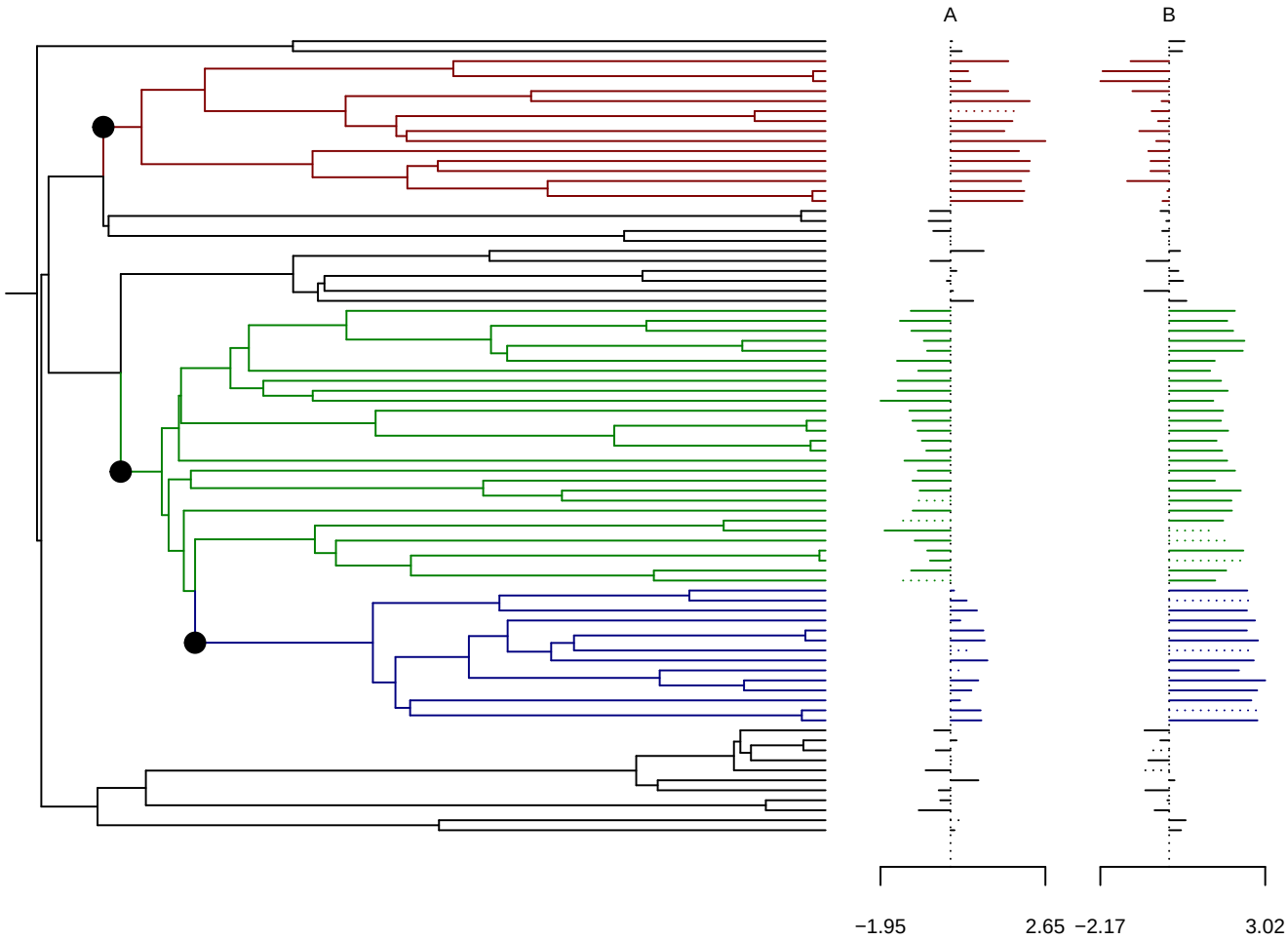


Inference

```
## 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",                ## scalar OU model
               alpha = alpha_grid,              ## On a grid of alpha
               K_max = 10,                      ## Maximal number of shifts
               parallel_alpha = TRUE,           ## This can be set to TRUE for
               Ncores = 2)                      ## parallel computations
)


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

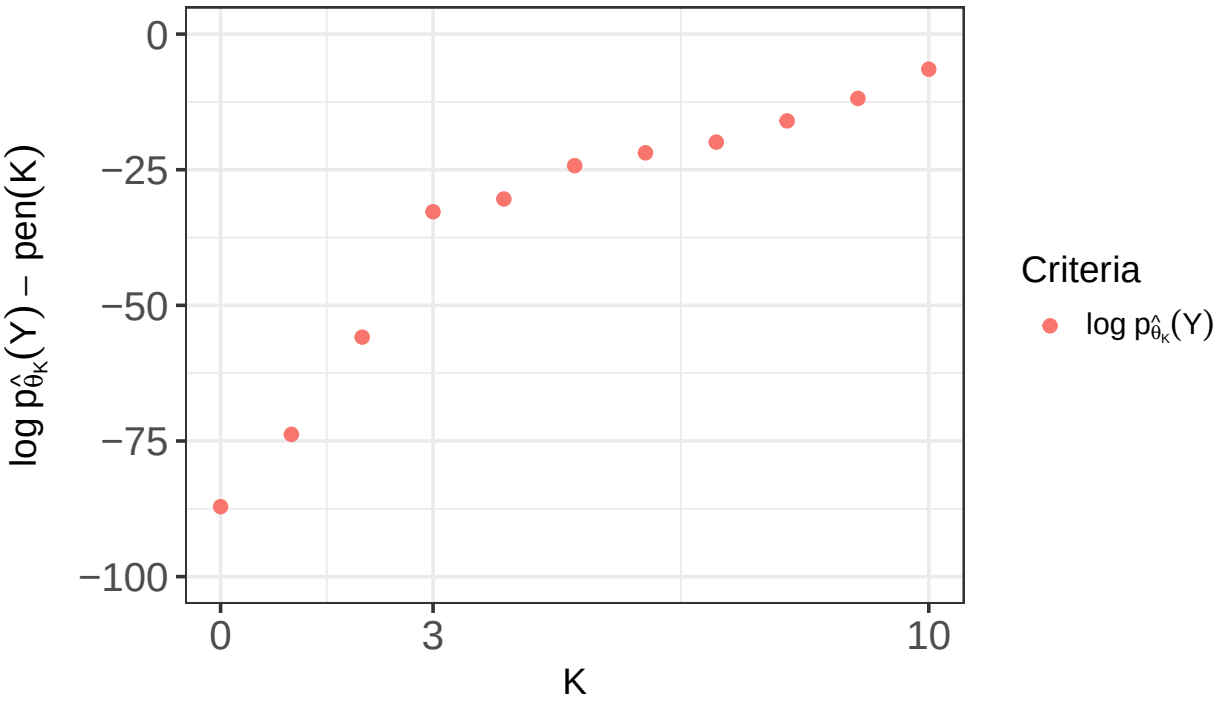
Analysis

```
plot(res)
```



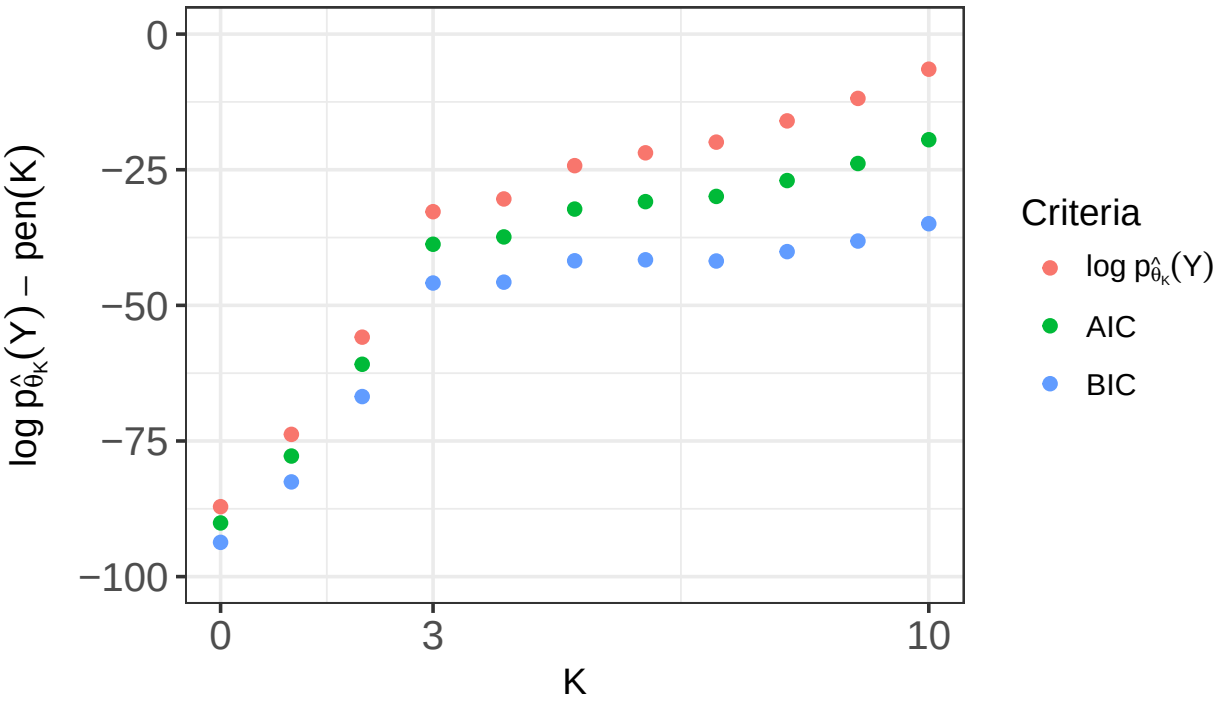
Model Selection: Penalized Likelihood

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)



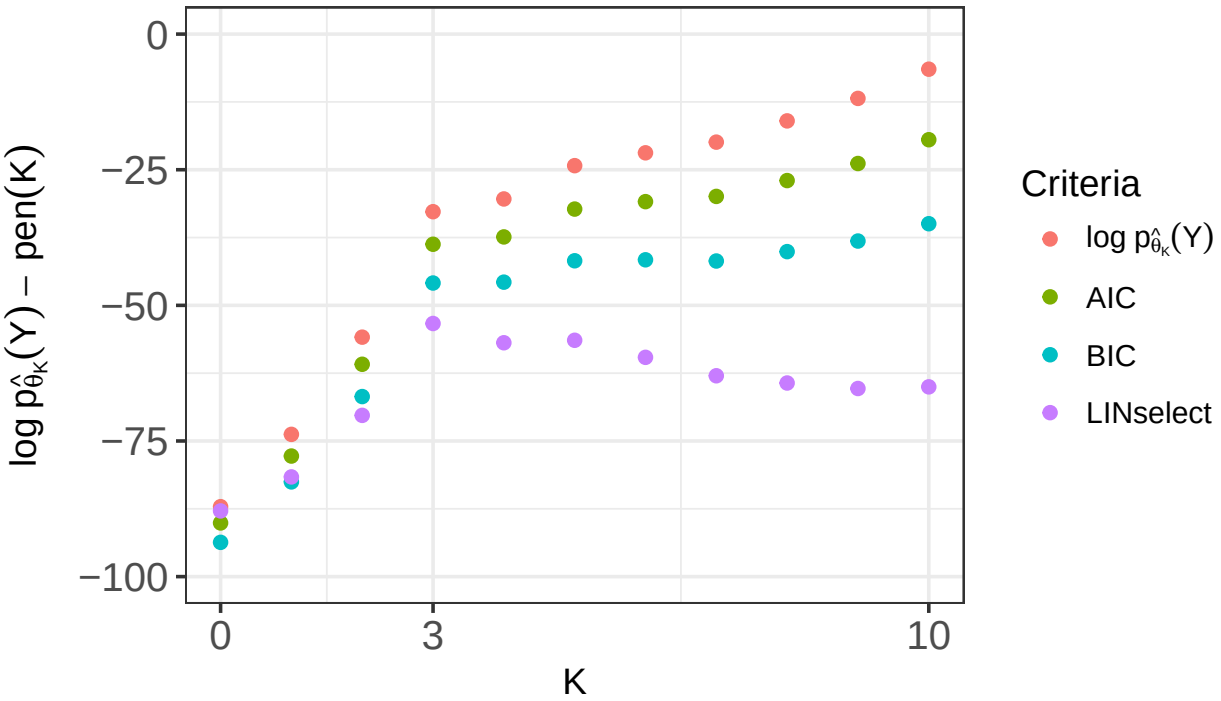
Penalties:

AIC $K + 3$

BIC $\frac{1}{2}(K + 3)\log(n)$

Model Selection: Penalized Likelihood

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



Penalties:

AIC $K + 3$

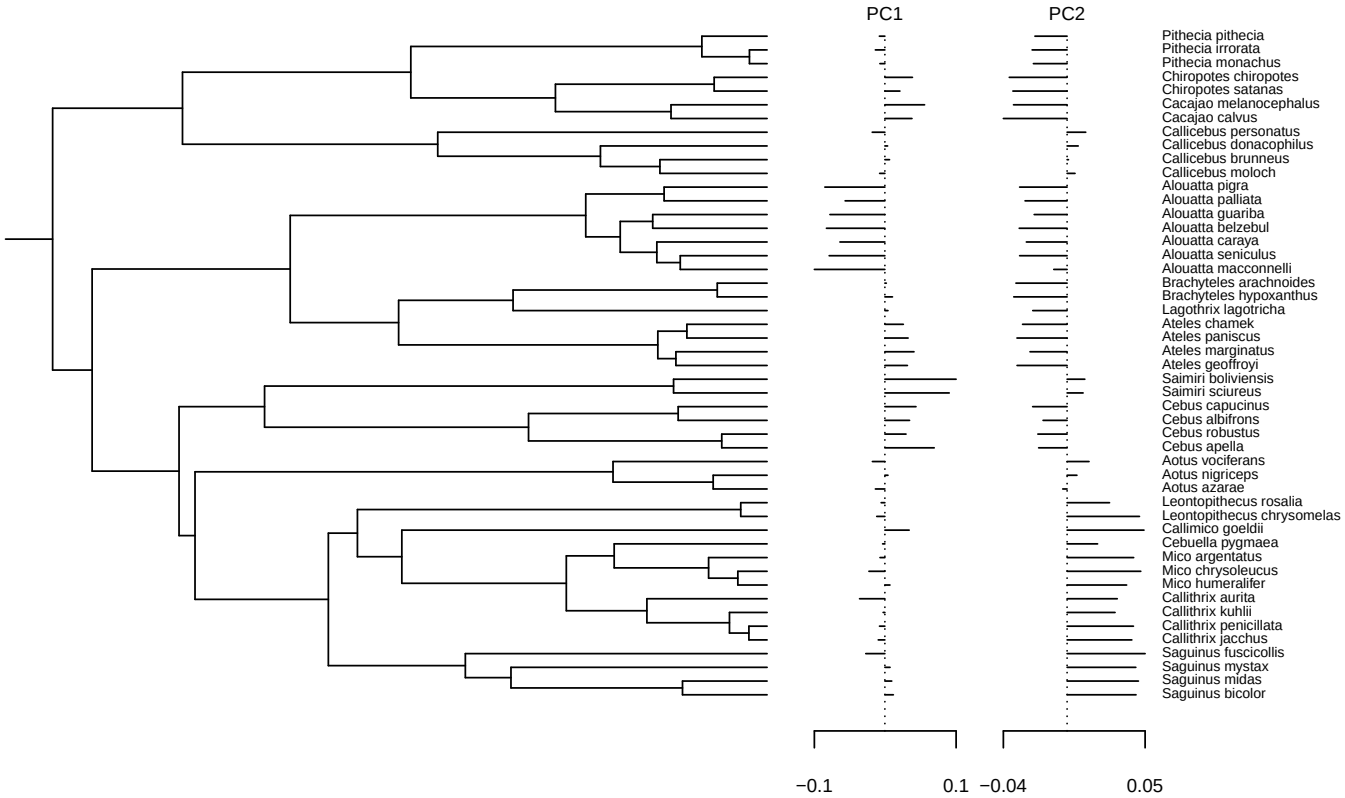
BIC $\frac{1}{2}(K+3)\log(n)$

LINselect $\mathrm{pen}(n, K, |\mathcal{S}_K^{PI}|)$

Monkey Dataset

(Aristide et al., 2016)

```
data(monkeys)

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



Analysis

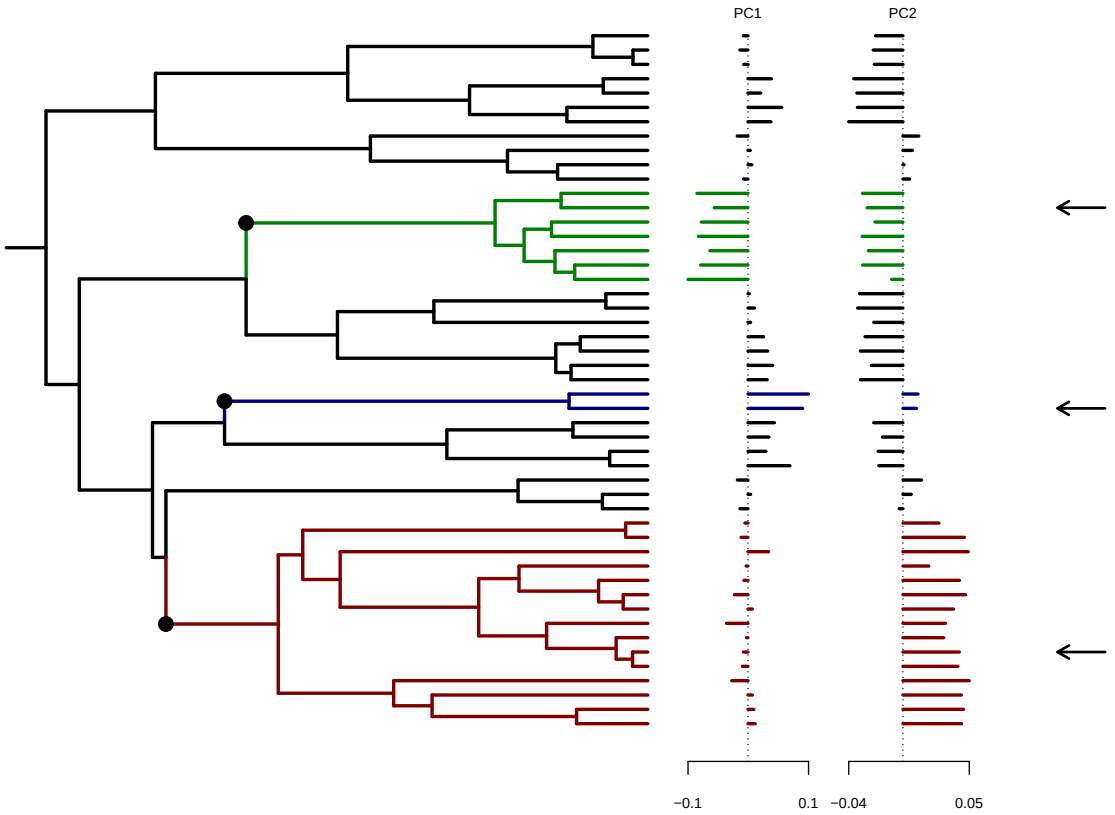
We use function `PhyloEM`:

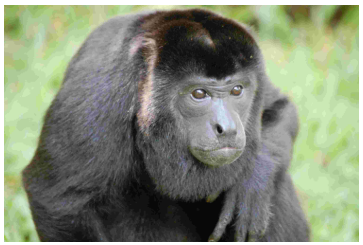
```
system.time(
res <- PhyloEM(Y_data = monkeys$dat,          ## data
               phylo = monkeys$phy,           ## phylogeny
               process = "scOU",              ## scalar OU
               K_max = 10,                    ## maximal number of shifts
               nbr_alpha = 4,                 ## 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:

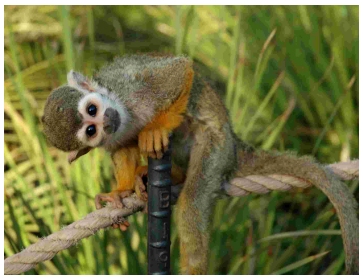
```
plot(res, edge.width = 2)
```

Result

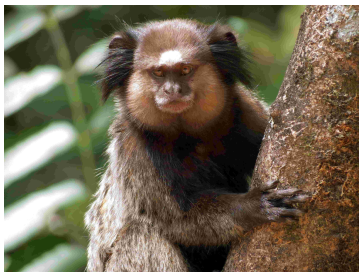




Alouatta palliata



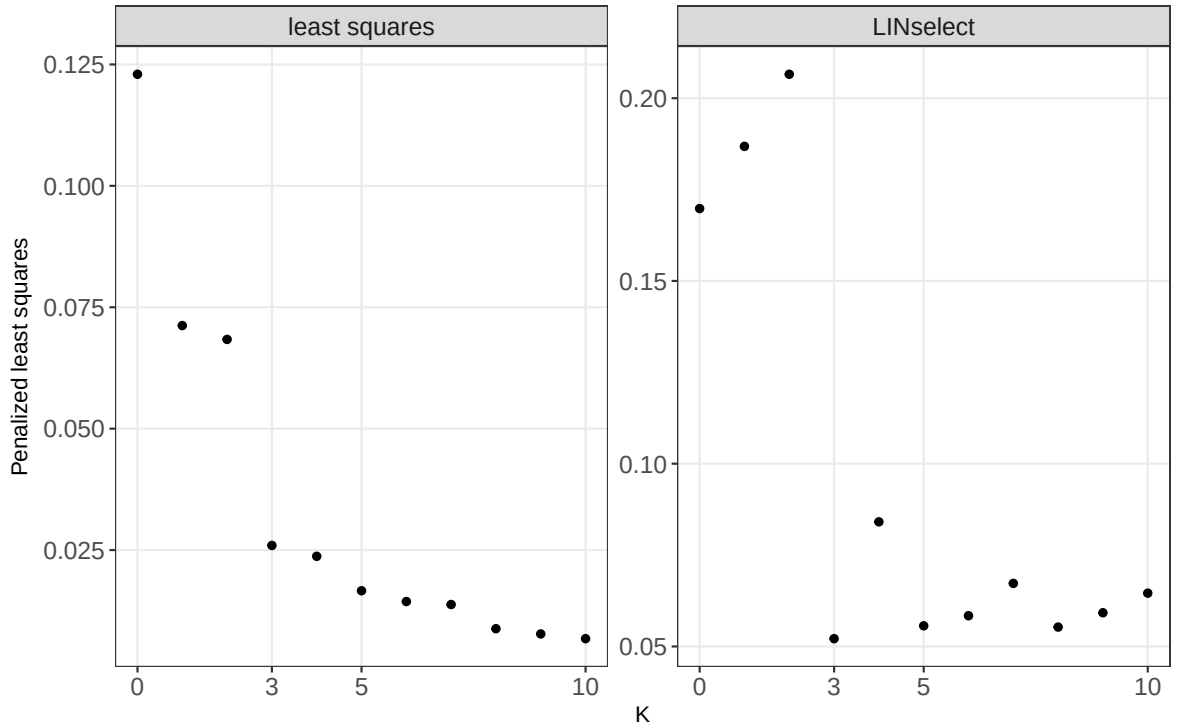
Saimiri sciureus



Callithrix penicillata

Model Selection

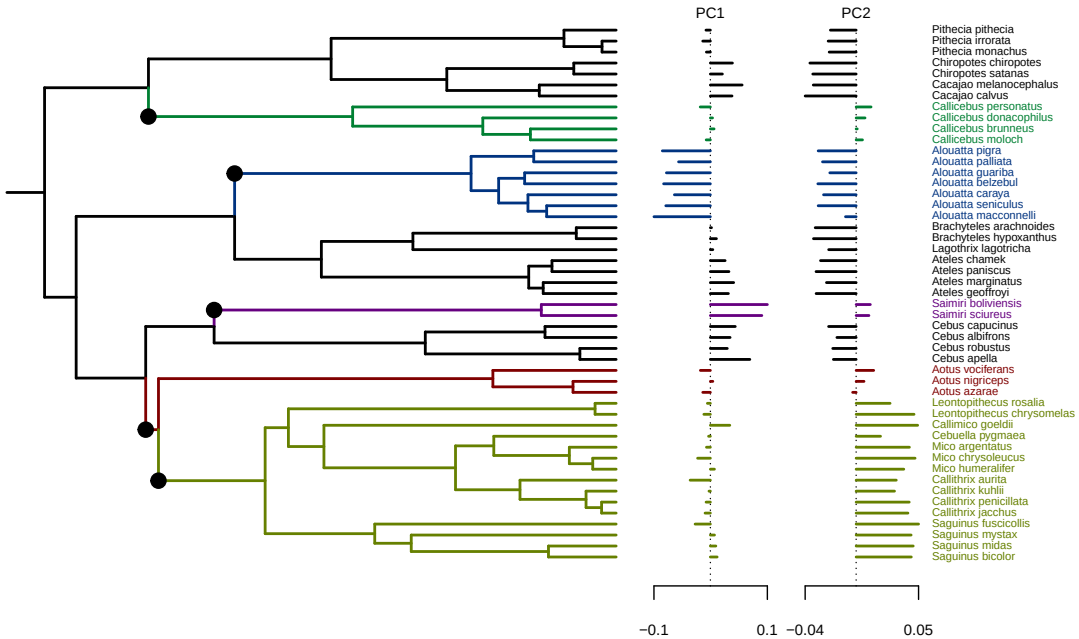
Solution with $K = 5$ seems to be a good solution too.



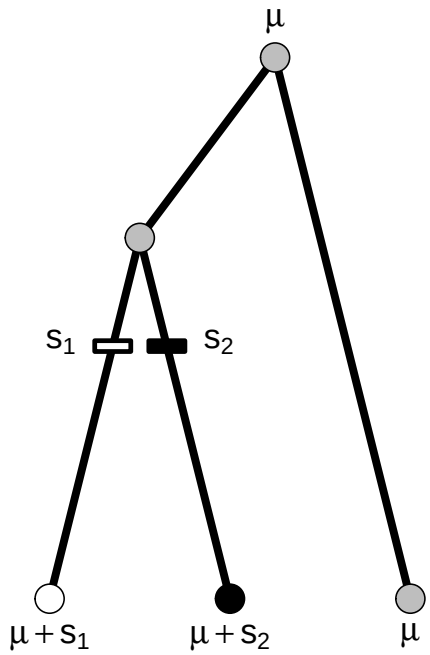
Solution for $K = 5$

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

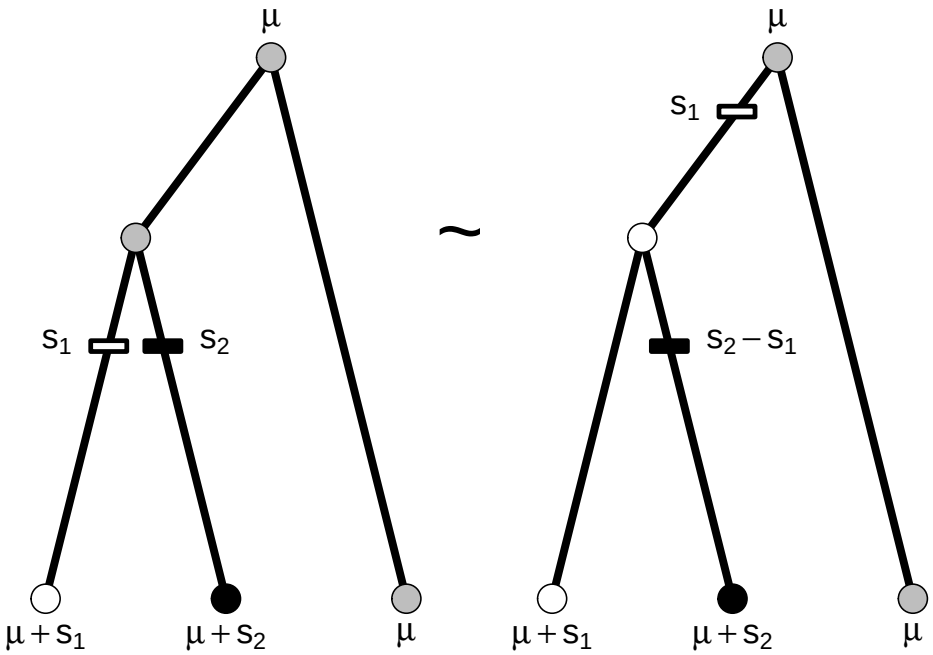
```
## Warning in params_process.PhyloEM(res, K = 5): There are several equivalent solutions for
this shift position.
```



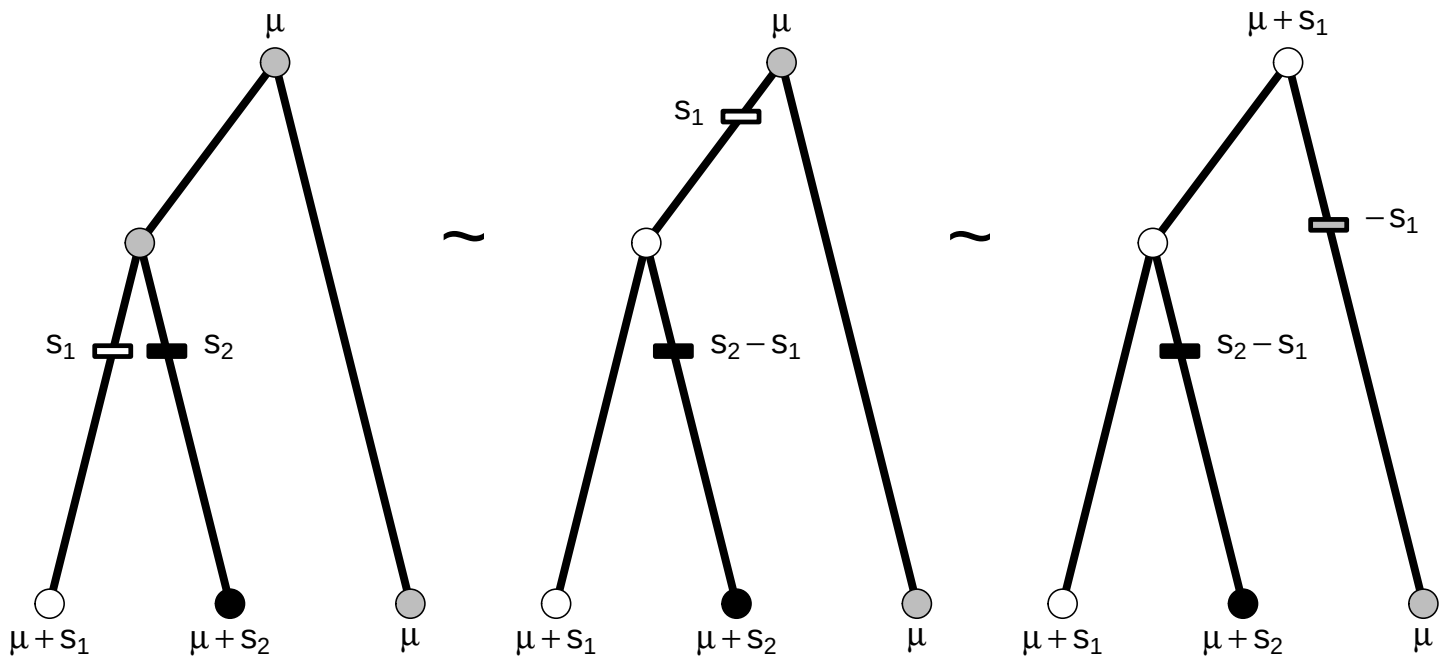
Equivalencies



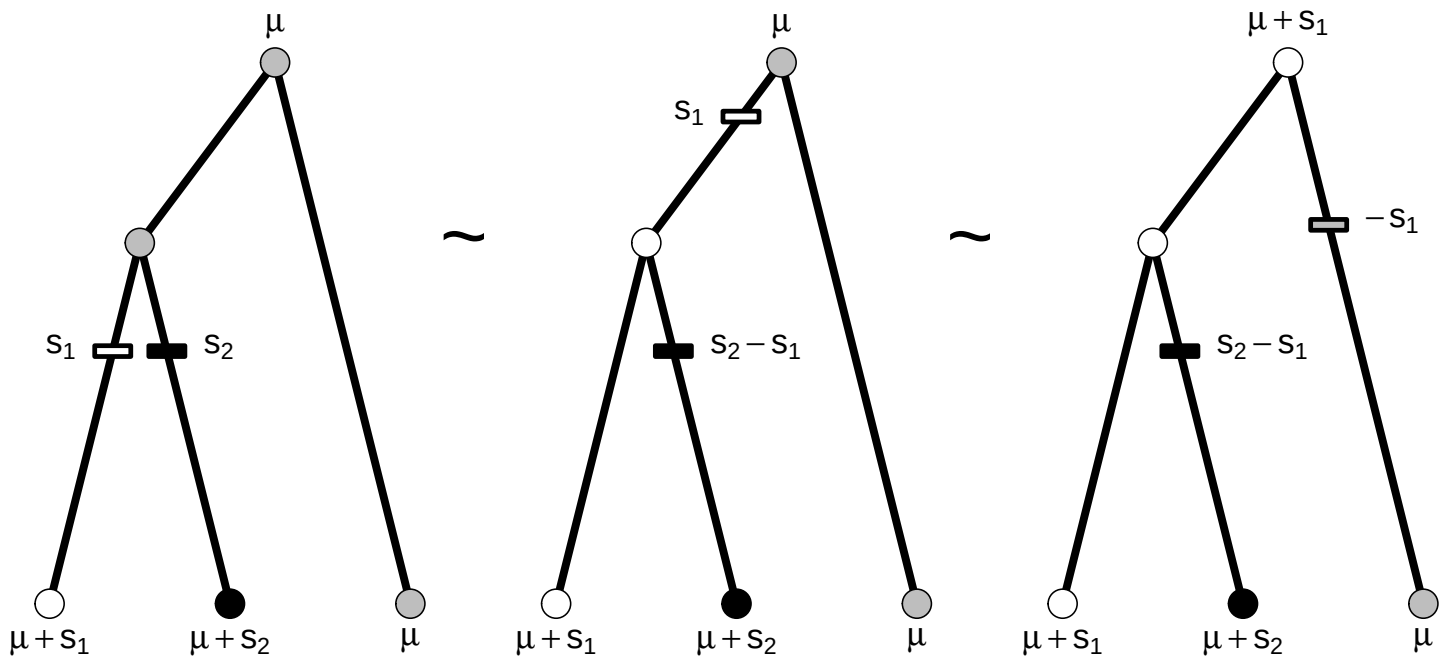
Equivalencies



Equivalencies

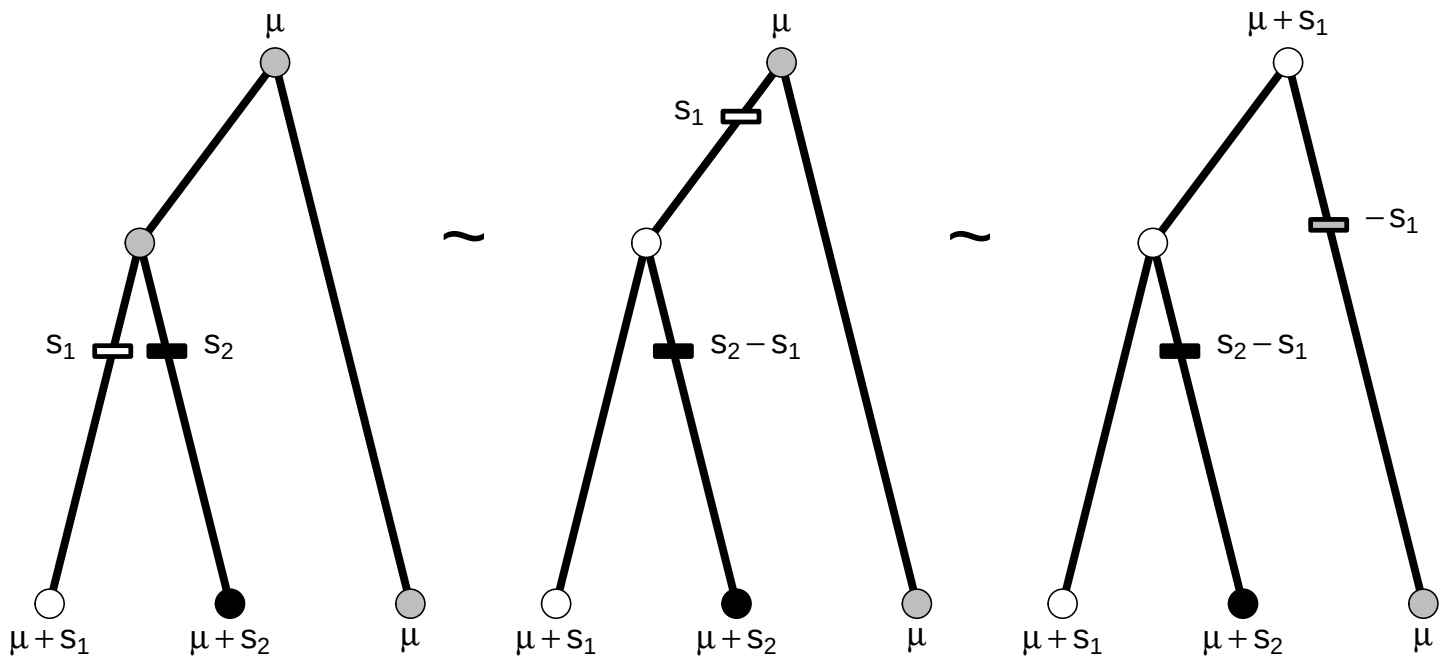


Equivalencies



Equivalent allocations *cannot* be distinguished from the data.

Equivalencies



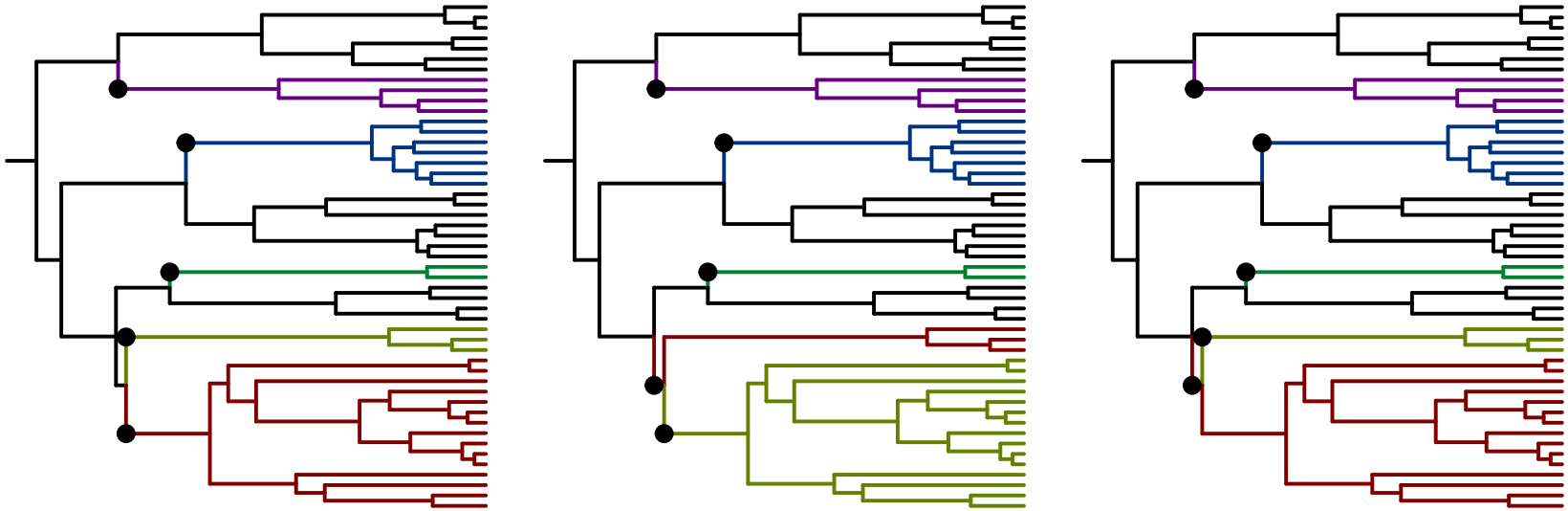
Equivalent allocations *cannot* be distinguished from the data.

`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)
```

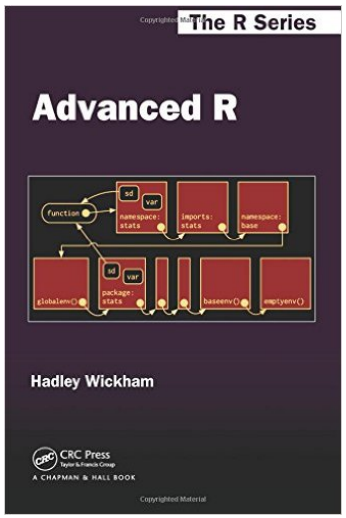


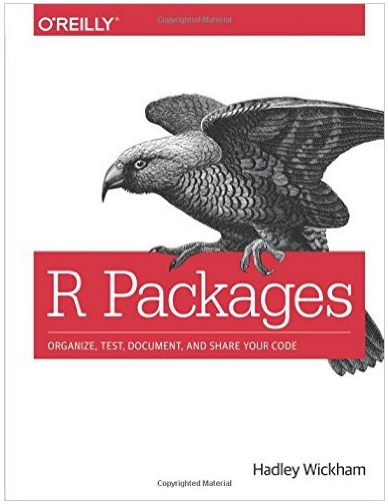
Outline

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`.

Implementation

Transparency

GitHub	`https://github.com/pbastide/PhylogeneticEM`
CRAN	Version `1.1.0`
Doc	Automatically build with `pkgdown`.

Efficiency

Profiling	Find bottleneck with `lineprof`
Optimizing	Speed up with `C++` and `RcppArmadillo`
	Avoid memory leaks with `valgrind`.
	Use the `Armadillo` library.

Robustness

Tests	Unitary tests with `testthat`.
CI	Continuous Integration with `Travis CI`.
Coverage	With `covr` and `codecov`.

Conclusion and Perspectives

A general inference framework for trait evolution models.

Conclusions

- A complete maximum likelihood procedure
- Taking identifiability problems into account
- With model selection

R Package

- Available on the CRAN and on GitHub
- Can scale up to big datasets ($\sim$ 1200 species)

Perspectives

- Deal with uncertainty (data, tree, shifts).
- Phylogenetic networks: See `julia` package `PhyloNetworks`.
- Combine with factor analysis.

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