

# **factorMerger: hierarchical clustering and model visualization**

Agnieszka Sitko

University of Warsaw, MI<sup>2</sup> Group

useR!2017 | 06-07-2017

## Problem 1

Given a factor  $C$  (with  $k^*$  levels) and a numeric response  $y^*$  analyze the differences among group means of  $y$ .

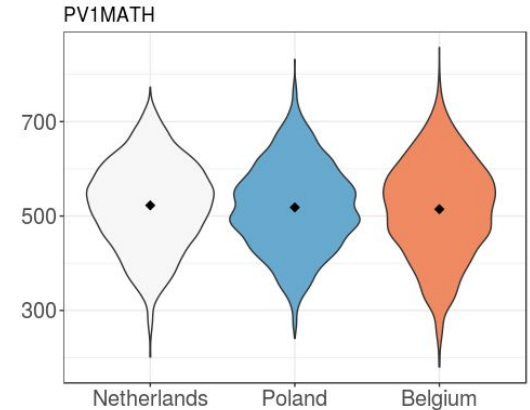
	PV1MATH	CNT
1	427.9561	Belgium
2	411.5984	Poland
3	471.1092	Poland
4	526.8032	Netherlands
5	721.4597	Poland
6	540.9020	Poland

## Problem 1

Given a factor  $C$  (with  $k^*$  levels) and a numeric response  $y^*$  analyze the differences among group means of  $y$ .

- \*  $k$  is greater than 2,
- \*  $y$  is normally distributed.

	PV1MATH	CNT
1	427.9561	Belgium
2	411.5984	Poland
3	471.1092	Poland
4	526.8032	Netherlands
5	721.4597	Poland
6	540.9020	Poland



## Problem 1

Given a factor  $C$  (with  $k^*$  levels) and a numeric response  $y^*$  analyze the differences among group means of  $y$ .

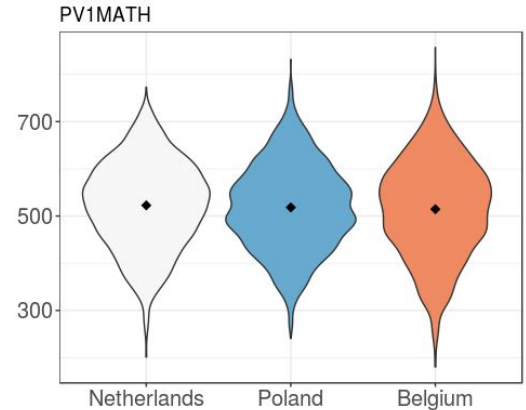
## Solution

That's easy!

Let's run ANOVA and then post-hoc tests.

- \*  $k$  is greater than 2,
- \*  $y$  is normally distributed.

	PV1MATH	CNT
1	427.9561	Belgium
2	411.5984	Poland
3	471.1092	Poland
4	526.8032	Netherlands
5	721.4597	Poland
6	540.9020	Poland



## **Solution**

Let's run ANOVA and then post-hoc tests.

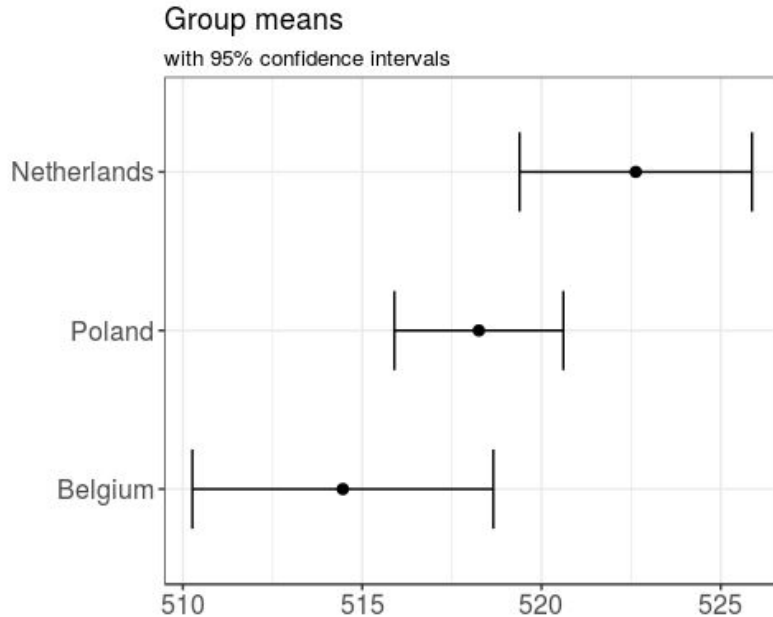
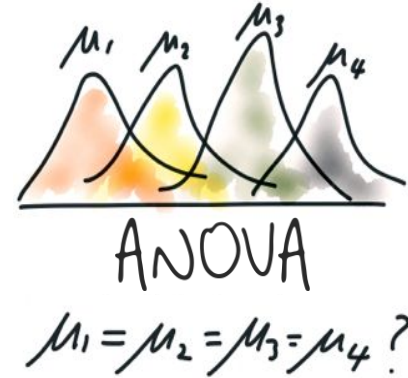
## **Solution**

Let's run ANOVA and then post-hoc tests.

Let's try this out.

# Solution

Let's run ANOVA and then post-hoc tests.

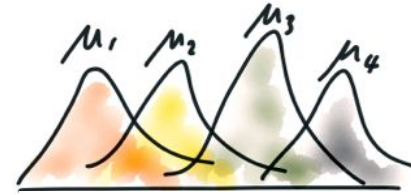


```
pisaAOV <- aov(PV1MATH ~ CNT, pisaNPB)
summary(pisaAOV)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
> CNT	2	84272	42136	4.836	0.00796 **
> Residuals	11113	96829278	8713		
> ---					
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1					

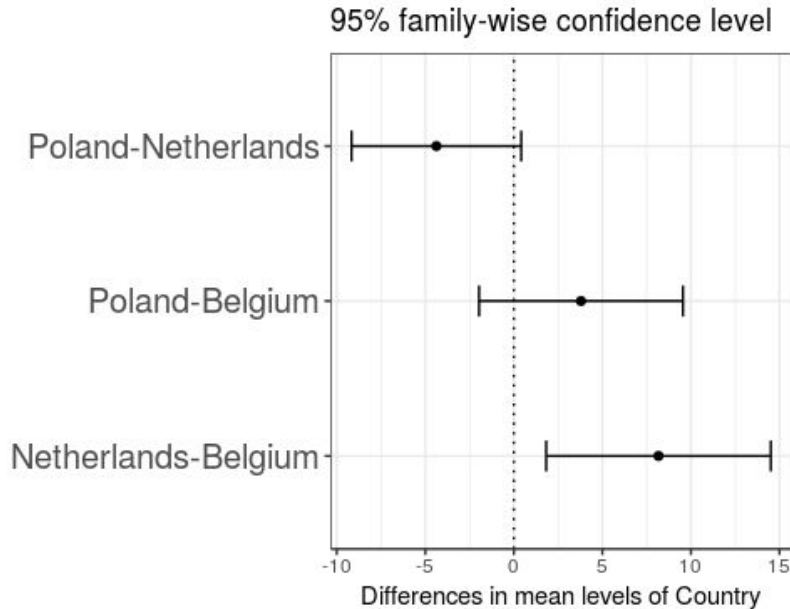
# Solution

Let's run ANOVA and then **post-hoc tests**.



post hoc

$\mu_1 = \mu_2?$   $\mu_1 = \mu_4?$   $\mu_1 = \mu_3?$   
 $\mu_2 = \mu_3?$   $\mu_3 = \mu_4?$   $\mu_2 = \mu_4?$



TukeyHSD(pisaA0V)

```
#> $CNT
```

```
#>
```

	diff	lwr	upr	p adj
#> Netherlands-Belgium	8.168042	1.827307	14.5087766	0.0071606
#> Poland-Belgium	3.793268	-1.962187	9.5487229	0.2700258
#> Poland-Netherlands	-4.374774	-9.166789	0.4172409	0.0819931

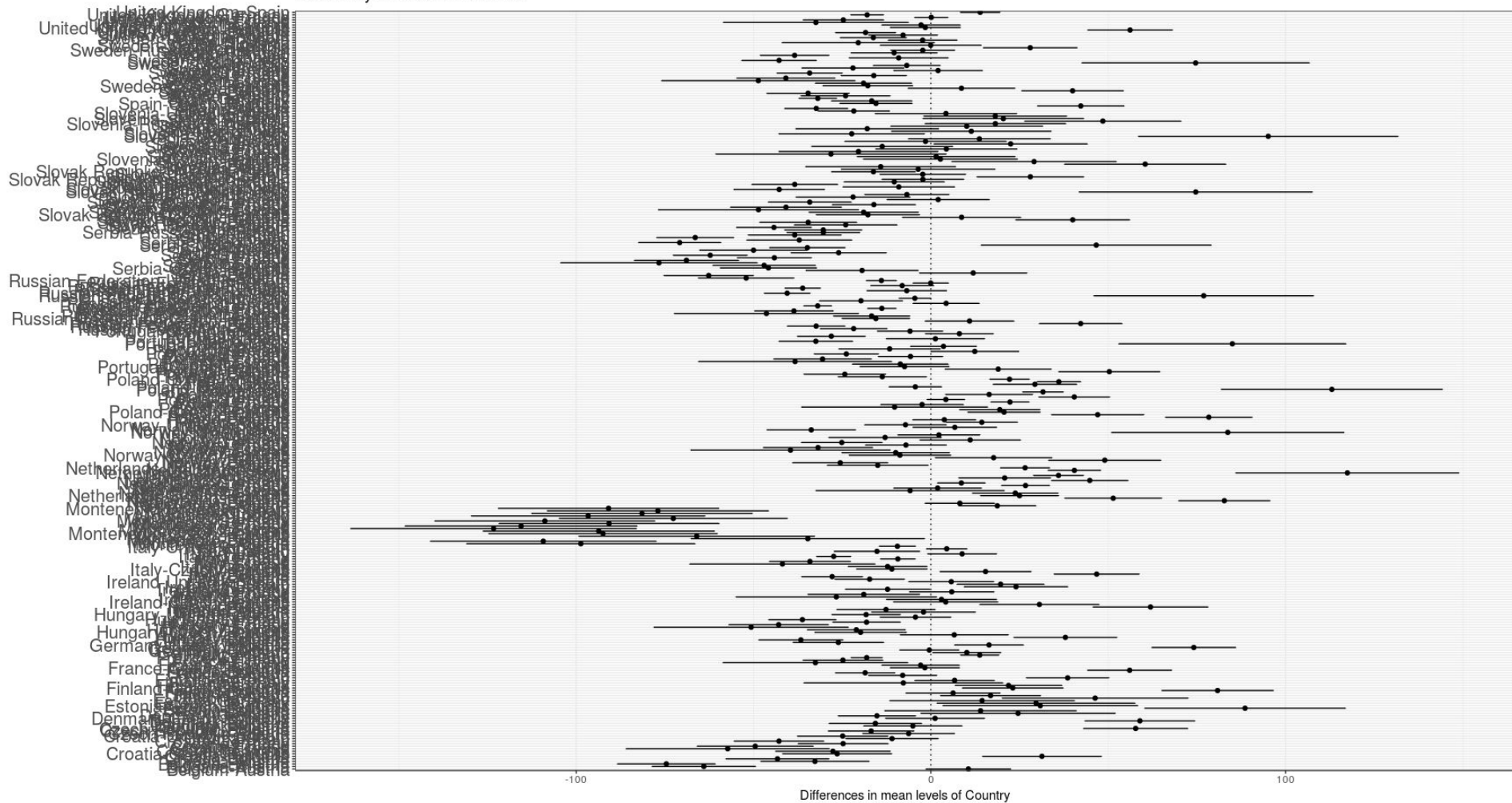


## Problem 2

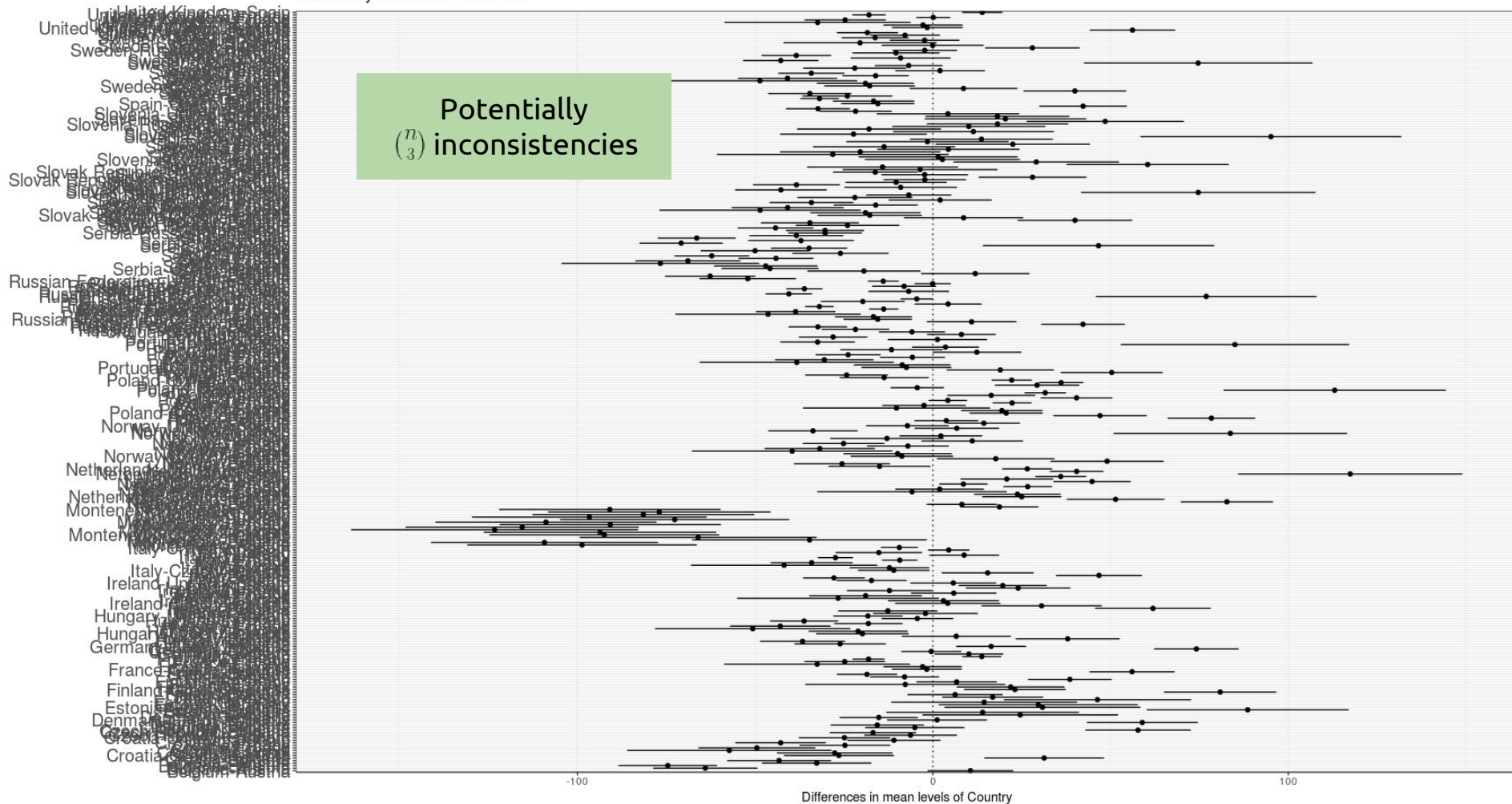
Given a factor  $C$  (with  $k^*$  levels) and a numeric response  $y^*$  analyze the differences among group means of  $y$ .

Group levels of  $C$  into non-overlapping clusters.

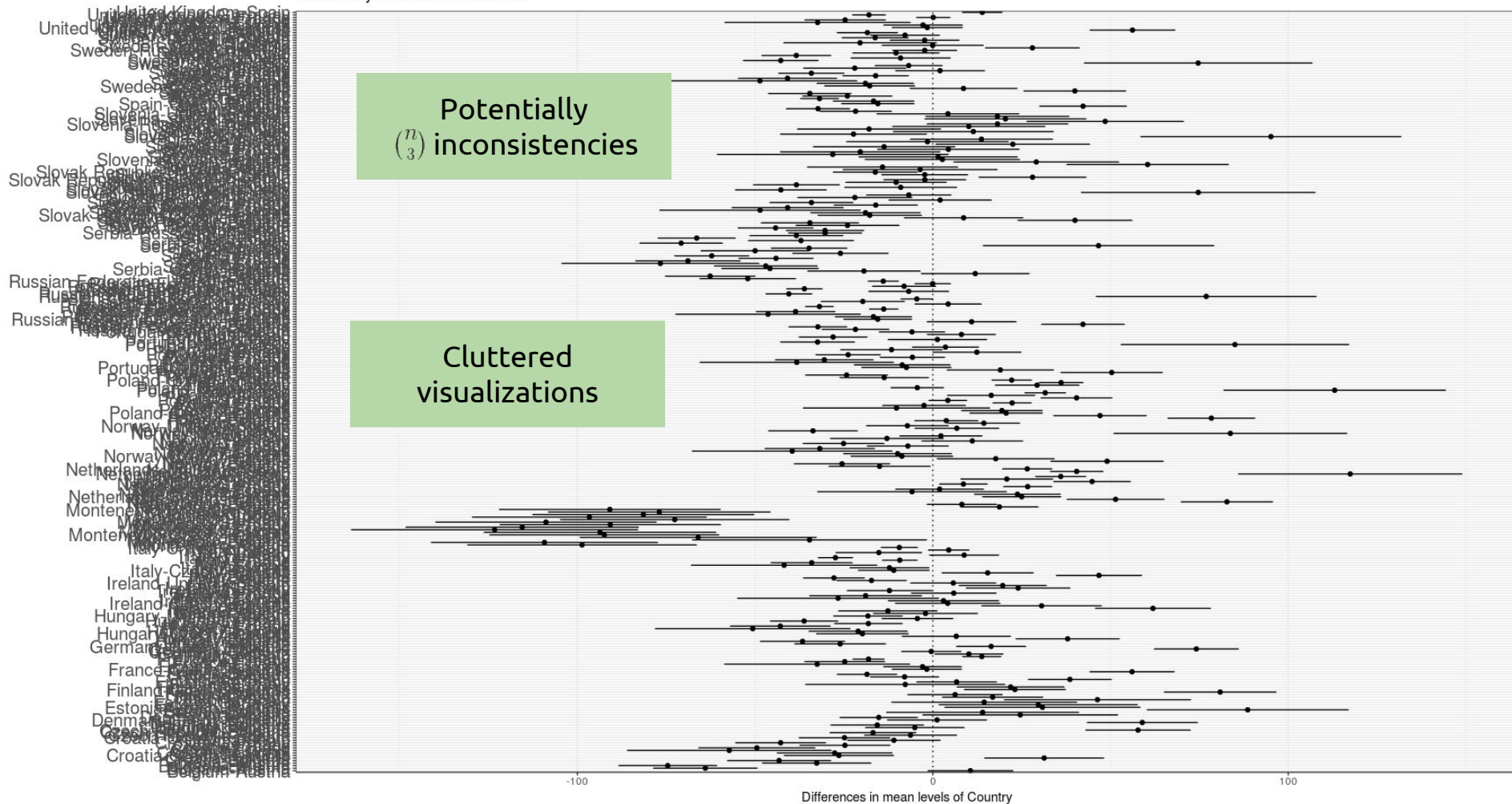
95% family-wise confidence level



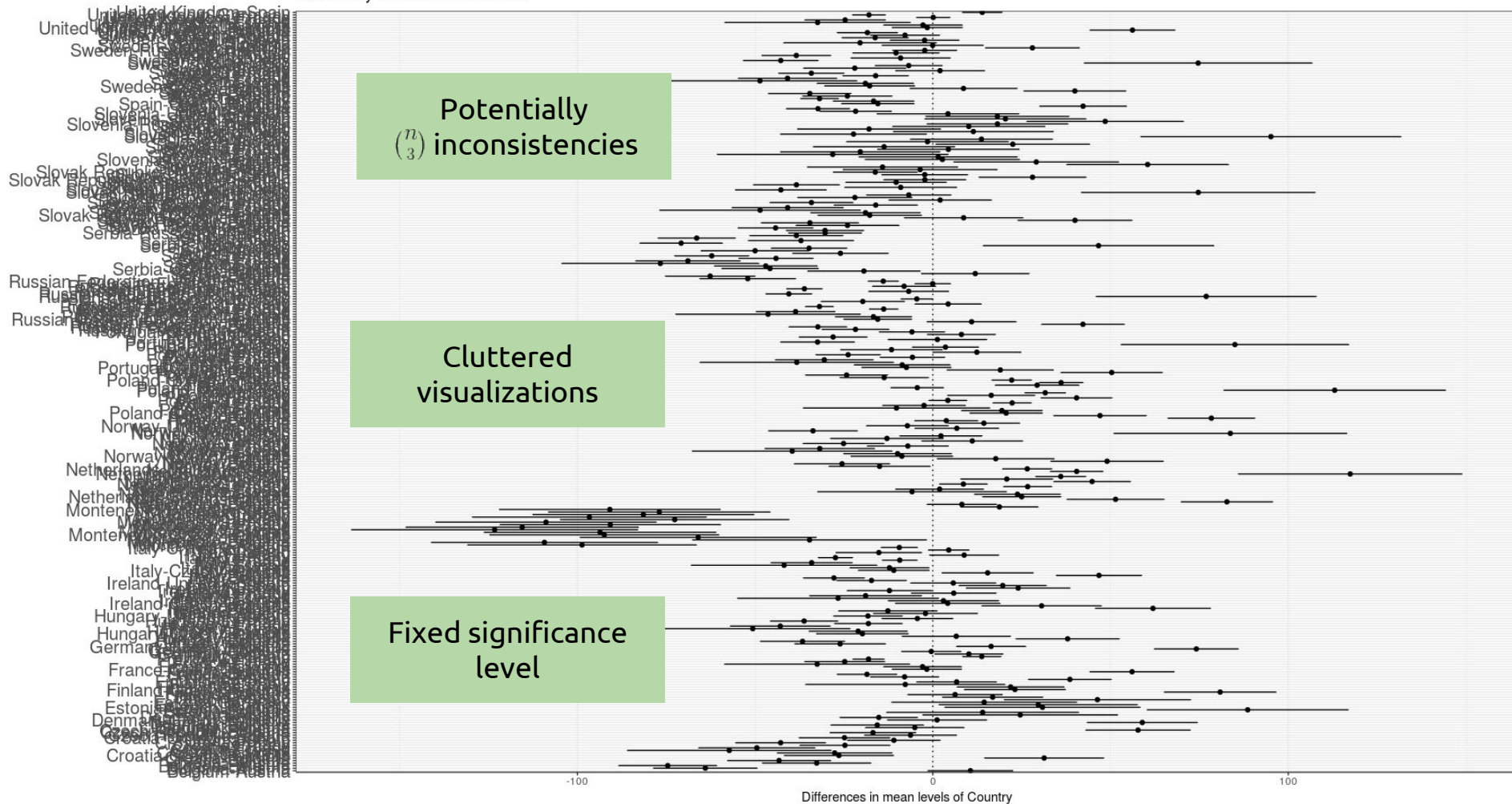
95% family-wise confidence level



95% family-wise confidence level



95% family-wise confidence level



**Time for  
factorMerger**

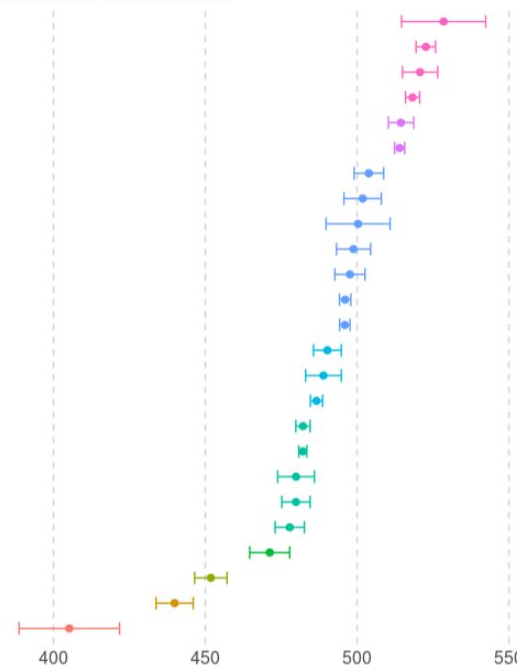
### Results in mathematics by country



1101991

1099551  
1099376

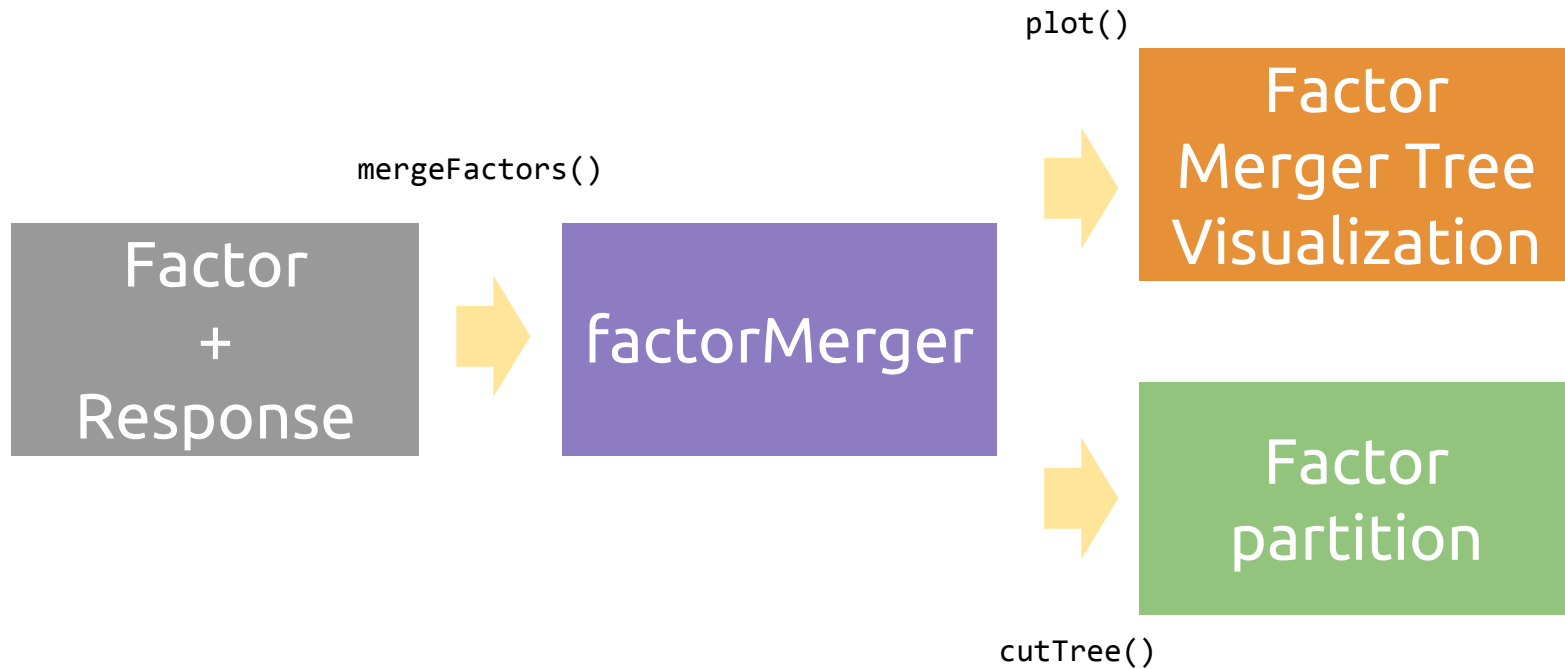
## with 95% confidence intervals



ANOVA table

	Df	F	p-value
factor	25	105939.6	< 2.2e-16
Res	92411		

# Working with factorMerger





# Merge

1. Likelihood Ratio Tests
2. Delete or Merge Regressors

```
factorMerger::mergeFactors(response = myResponse,  
                             factor = myFactor,  
                             method = "LRT")
```

```
factorMerger::mergeFactors(response = myResponse,  
                             factor = myFactor,  
                             method = "hclust",  
                             successive = TRUE)
```

# Merge

1. Likelihood Ratio Tests
2. Delete or Merge Regressors

---

## Algorithm 1 Merging with LRT

---

```
function MERGEFACTORS(response, factor, successive)
2:   pairsSet := generatePairs(response, factor, successive)
    $M_0$  := full model
4:   while levels(factor) > 1 do
        $toBeMerged := \operatorname{argmax}_{pair \in pairsSet} l(updateModel(M_0, pair))$ 
6:        $M_0 := updateModel(M_0, toBeMerged)$ 
       factor := mergeLevels(factor, pair)
8:       pairsSet := pairsSet \ pair
   end while
10: end function
```

---

# Merge

1. Likelihood Ratio Tests
2. Delete or Merge Regressors

---

**Algorithm 2** Merging with agglomerative clustering

---

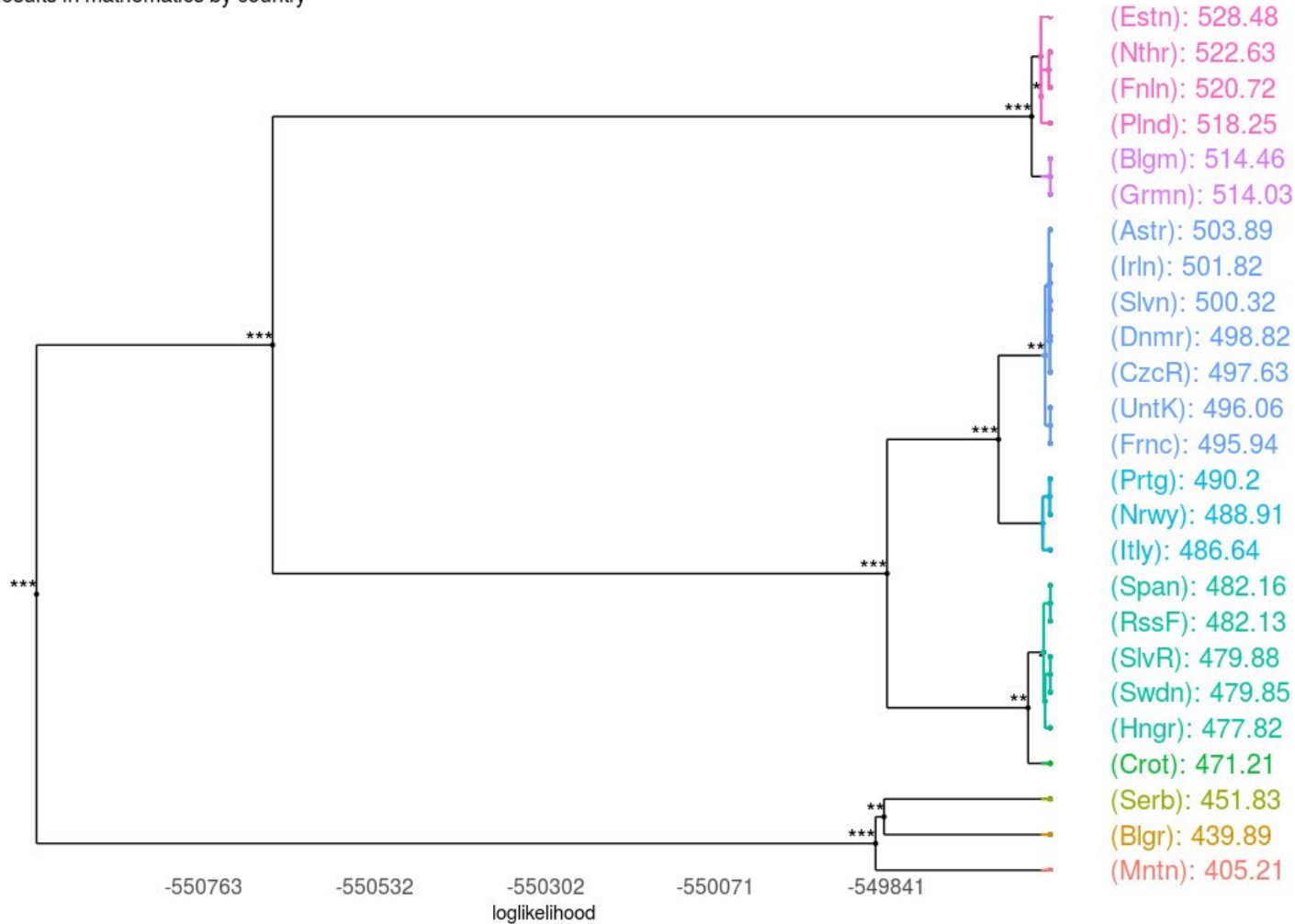
```
function MERGEFACTORS(response, factor, successive)
2:   pairsSet := generatePairs(response, factor, successive)
   dist := set of distances
4:   for all pair  $\in$  pairsSet do
        $h := \{\mu_{pair_1} = \mu_{pair_2}\}$   $\triangleright$  hypothesis under which pair is merged
6:       dist[pair] =  $LRT(M_h|M_0)$ 
   end for
8:   if successive then
       hClust(dist, method = "single")
10:  else
       hClust(dist, method = "complete")
12:  end if
end function
```

---

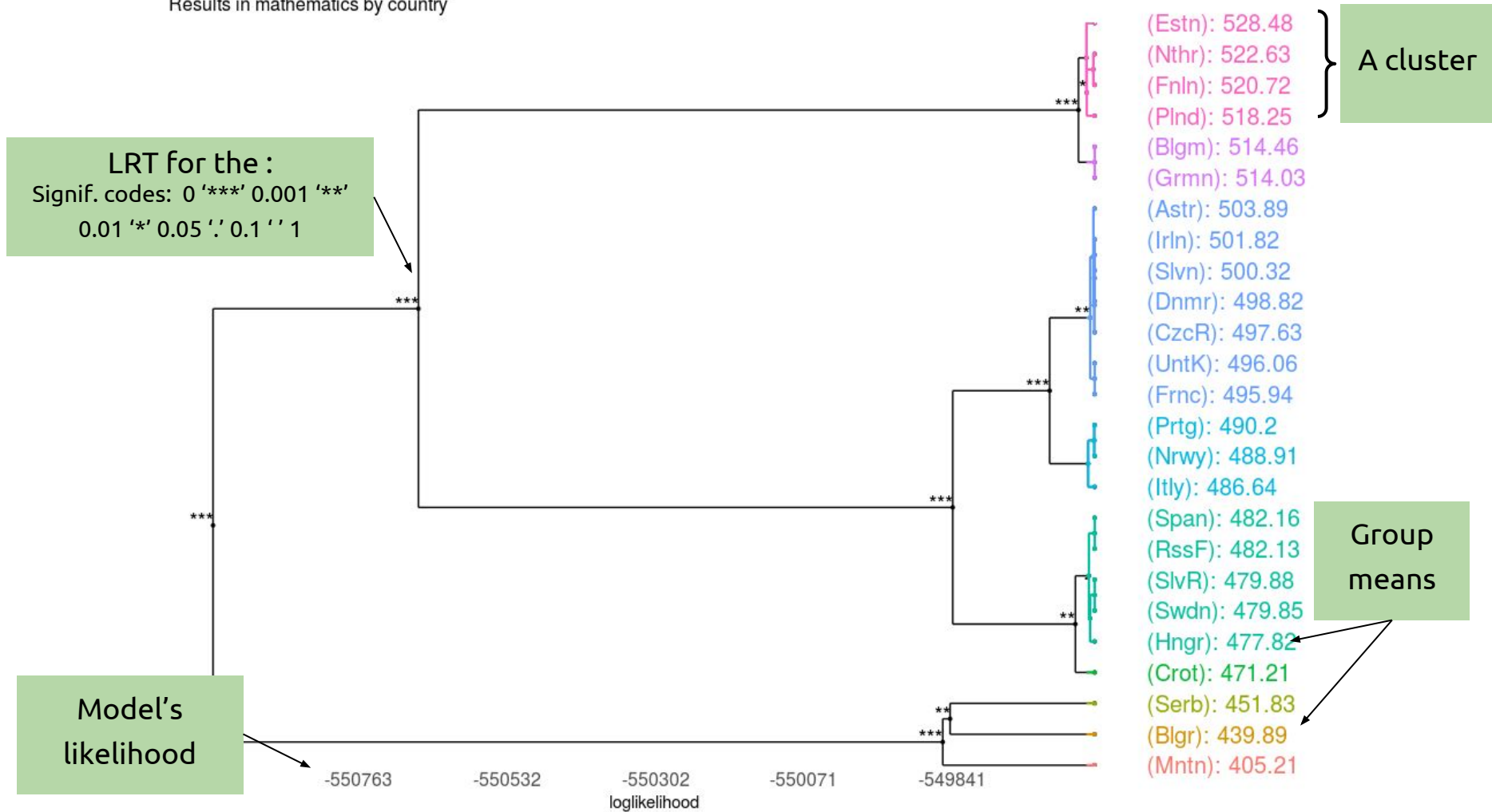
More about the DMR algorithm: <https://arxiv.org/abs/1505.04008>

# PISA 2012

Results in mathematics by country

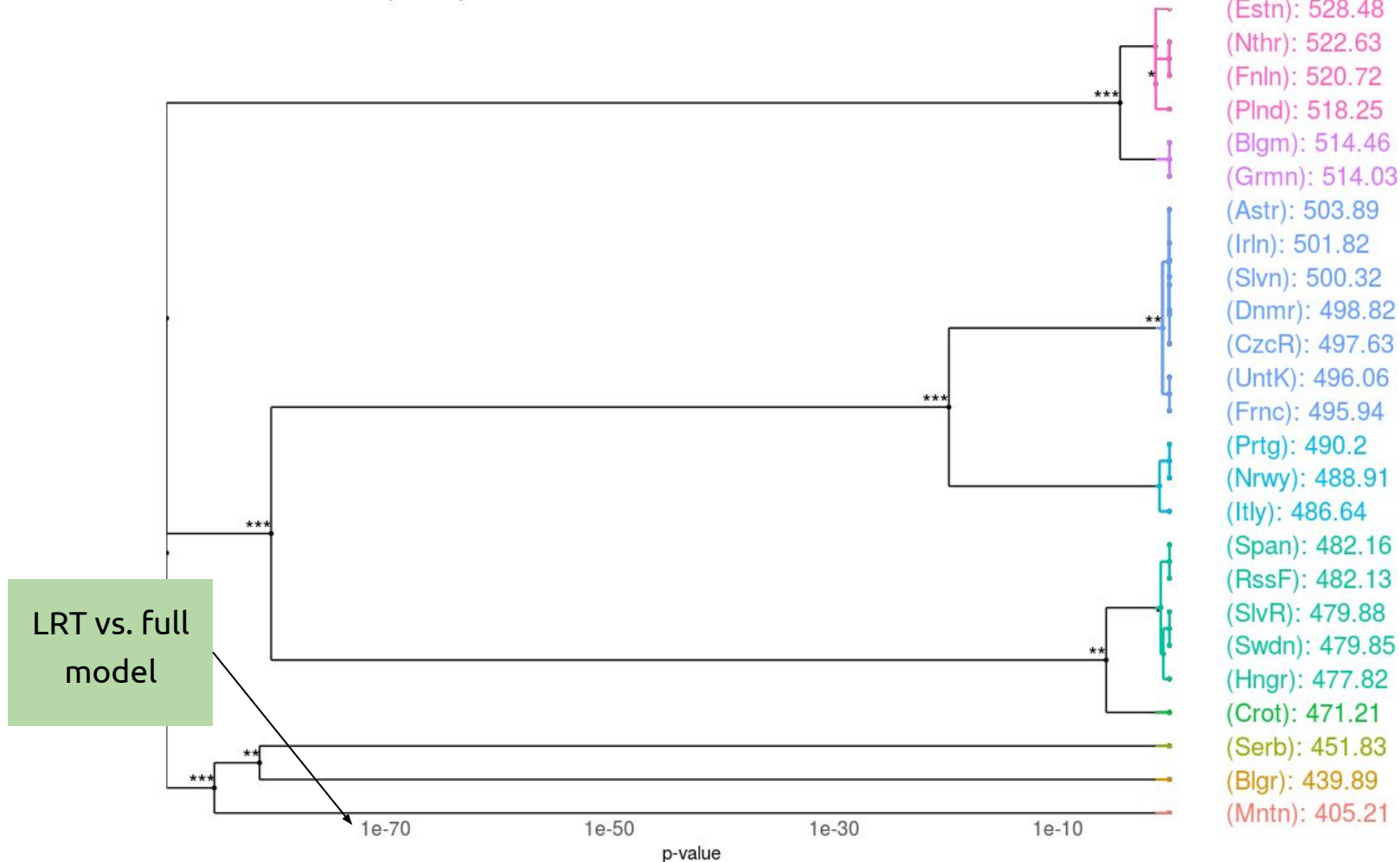


PISA 2012  
Results in mathematics by country

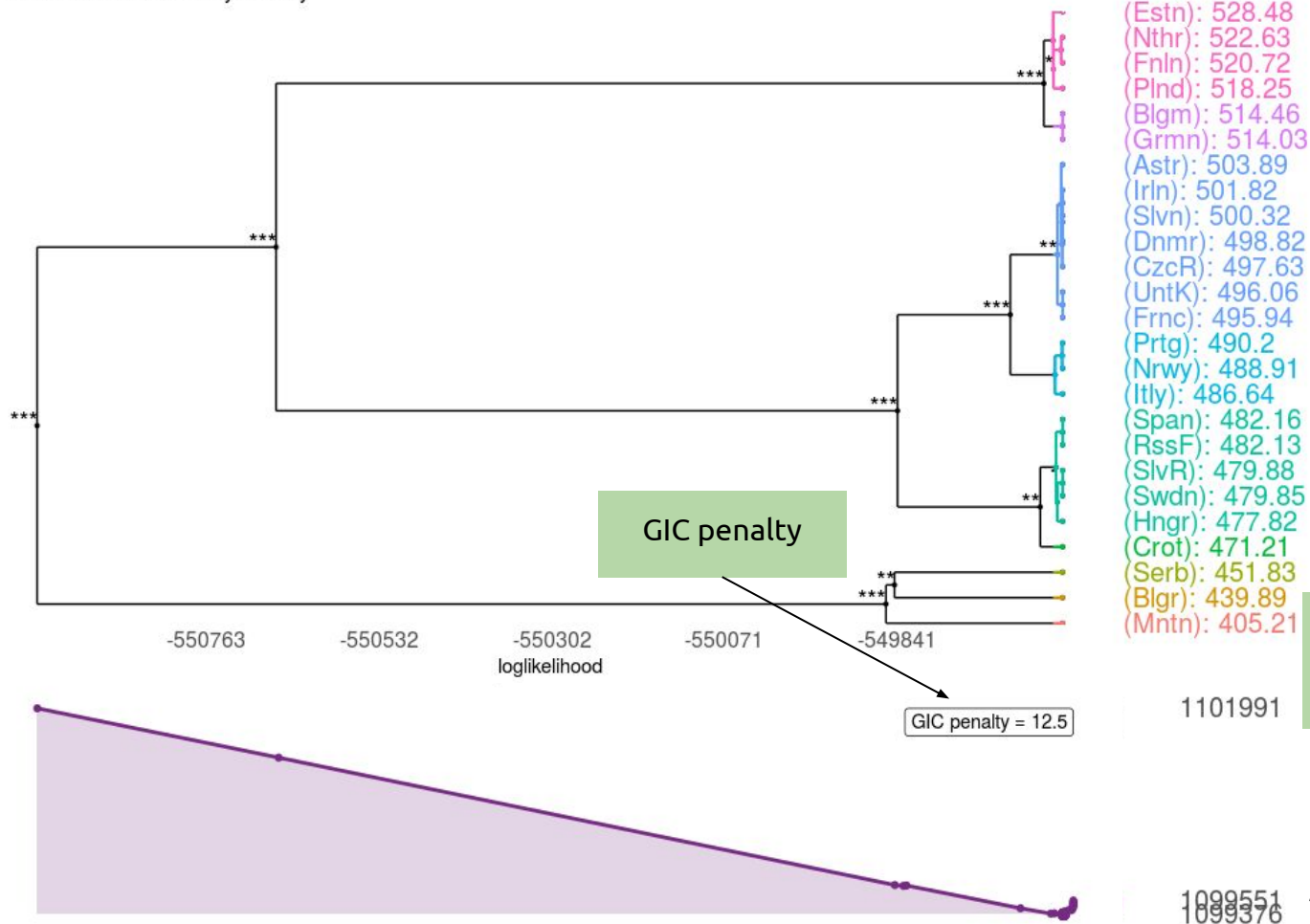


# PISA 2012

Results in mathematics by country

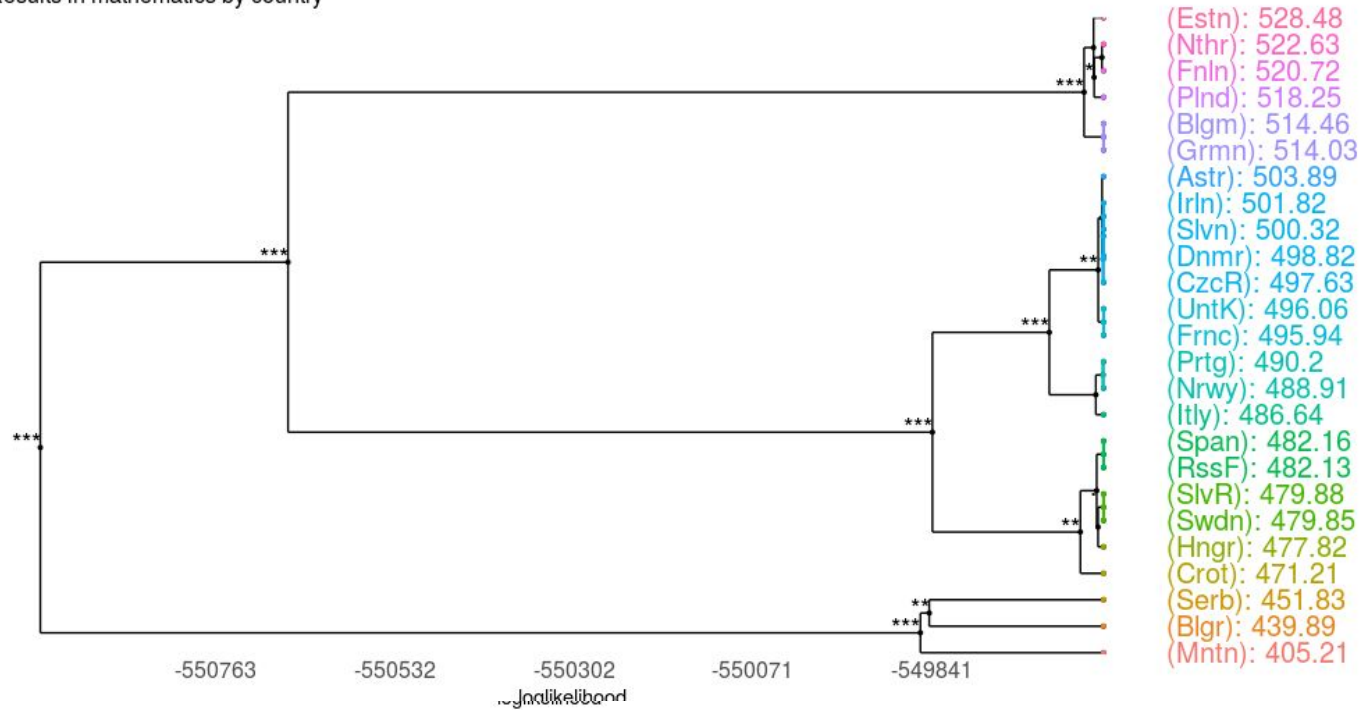


### Results in mathematics by country



# PISA 2012

Results in mathematics by country

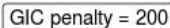


1101979

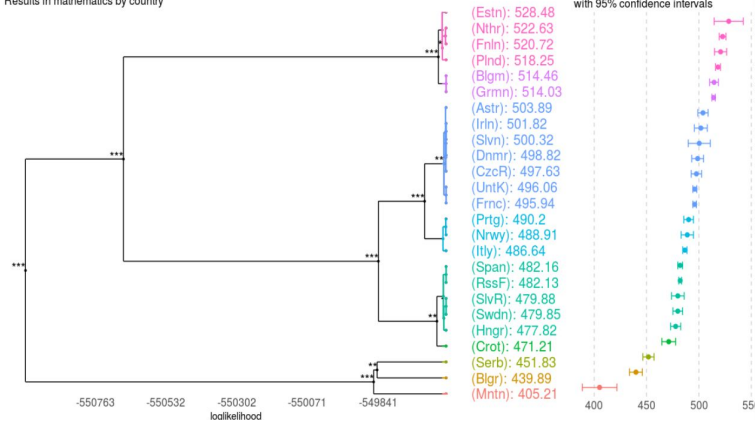
1099266



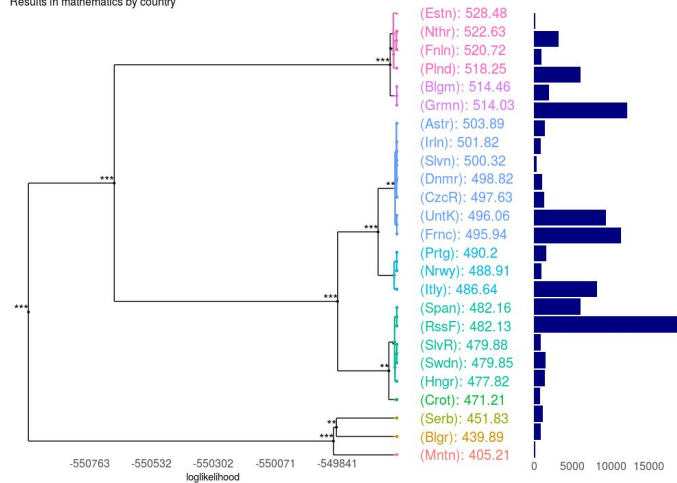
### Results in mathematics by country



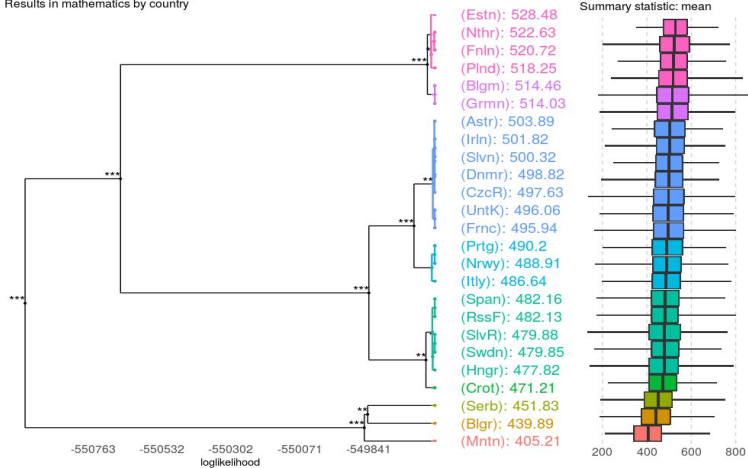
PISA 2012  
Results in mathematics by country



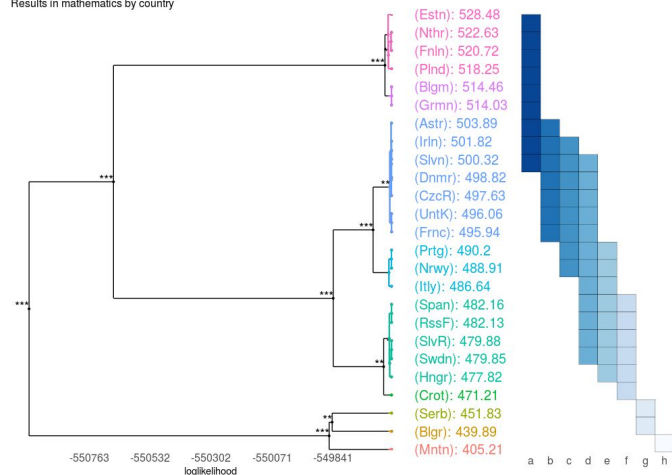
PISA 2012  
Results in mathematics by country



PISA 2012  
Results in mathematics by country



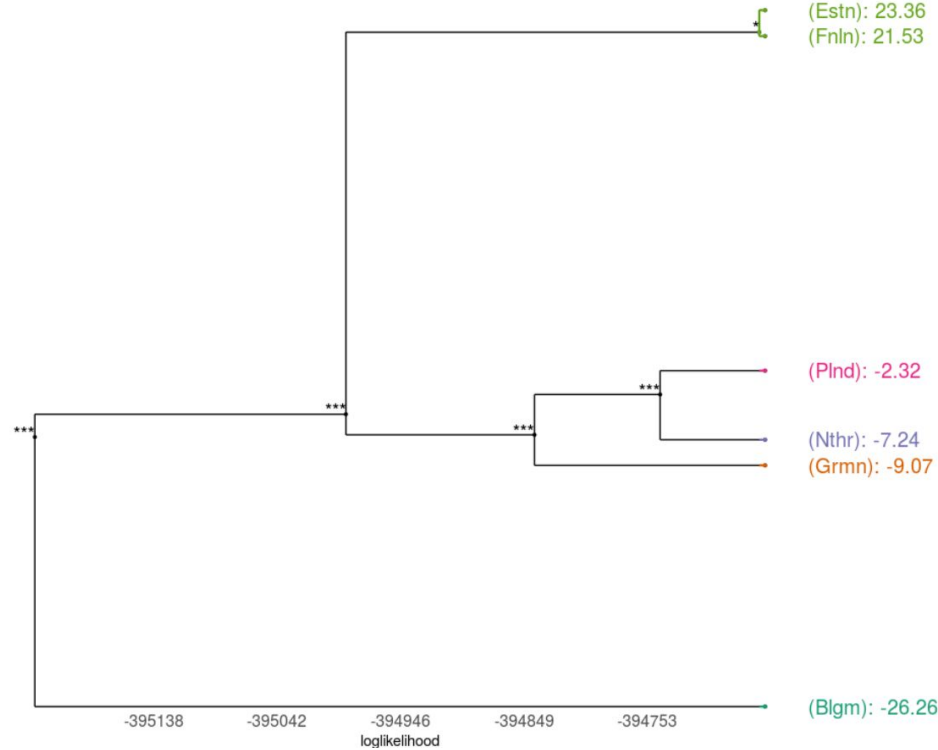
PISA 2012  
Results in mathematics by country



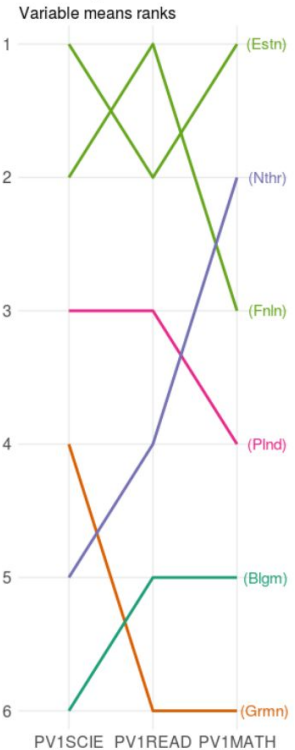
# Other parametric models

1. multi dimensional Gaussian model,
2. binomial model,
3. survival model.

PISA 2012 - students' performance



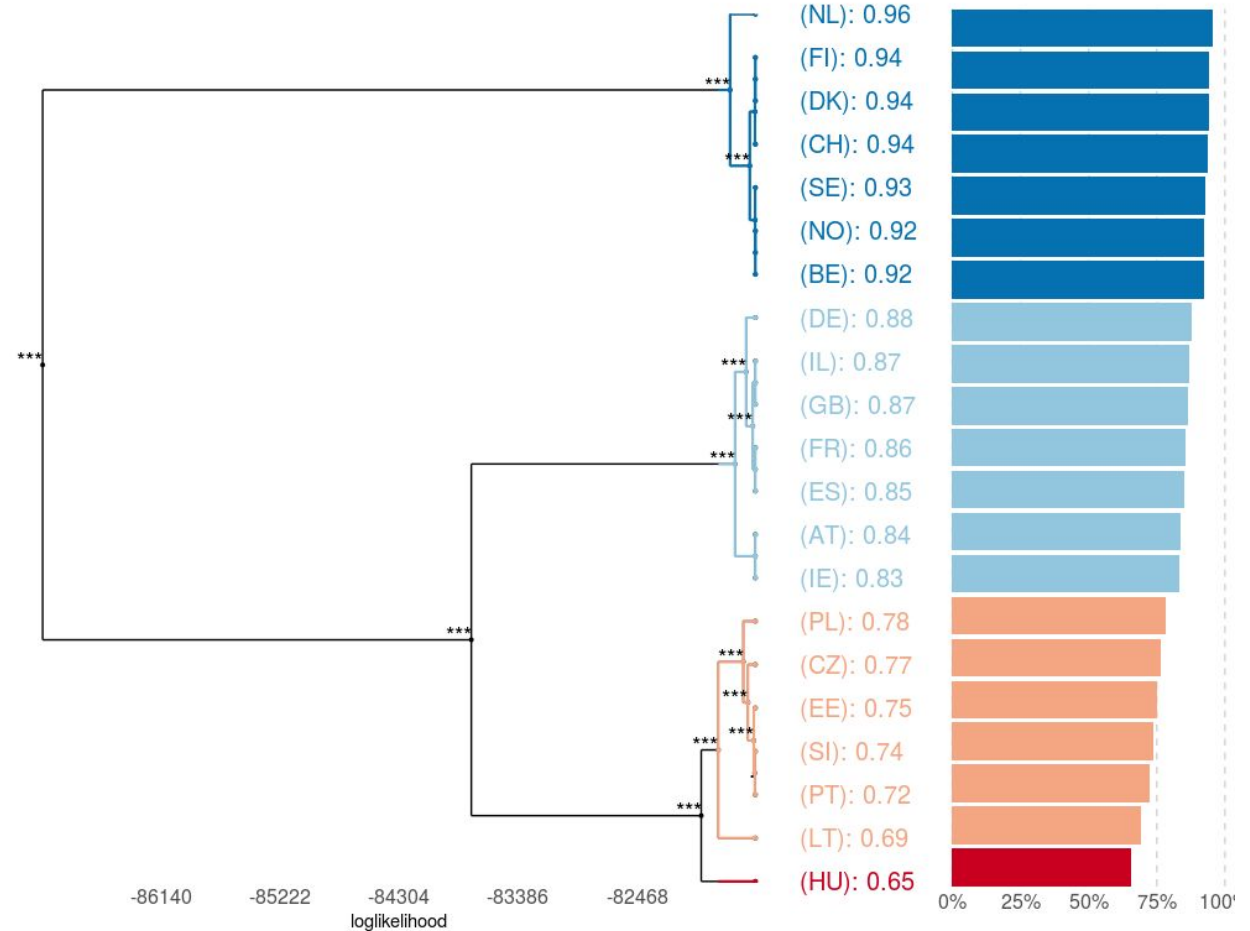
Profile plot



# Other parametric models

1. multi dimensional Gaussian model,
2. binomial model,
3. survival model.

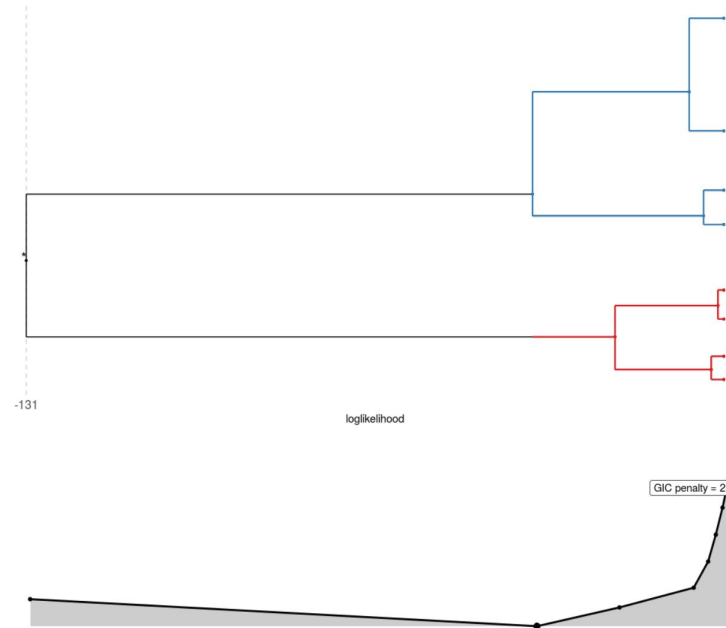
European Social Survey 2014 - HOW HAPPY ARE YOU?



# Other parametric models

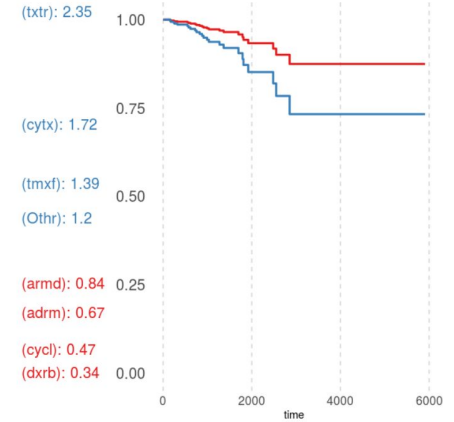
1. multi dimensional Gaussian model,
2. binomial model,
3. survival model.

Breast Cancer - treatment



Survival plot

Adjusted survival curves for coxph model



ANOVA table

	loglik	Chisq	Df	p-value
NULL	-131.2			
factor	-128.4	5.4	7	0.606233

# Install and use the package

```
install.packages("factorMerger")
```

} CRAN

```
if (!require(devtools)) install.packages("devtools")  
devtools::install_github("geneticsMiNIng/factorMerger")
```

} Github


```
library(factorMerger)  
fm <- mergeFactors(response = myResponse,  
                    factor = myFactor,  
                    family = "survival",  
                    successive = TRUE,  
                    method = "LRT")
```

Find more: <https://github.com/geneticsMiNIng/factorMerger>




# The aim of the factorMerger package

1. Create an algorithm which outputs an unequivocal data partition.
2. Improve visualizations.
3. Include other parametric models:
  - a. multi dimensional Gaussian model,
  - b. binomial model,
  - c. survival model.

# geneticsMiNIng

 This organization

[Pull requests](#) [Issues](#) [Marketplace](#) [Gist](#)



geneticsMiNIng: Research group from Warsaw University of Technology and University of Warsaw

 **Repositories**

 People **10**

 Teams **2**

 Projects **0**

Type: **All** ▾

Language: **All** ▾

 **New**

## MLGenSig

Machine Learning for Genetic Signatures

 HTML  2 Updated 2 hours ago



## factorMerger

Set of tools to support results from post hoc testing

 HTML  1 Updated 7 hours ago



## BlackBoxOpener

Set of tools to understand what is happening inside 'BlackBox' classifiers like Random Forest / Gradient Boosting

 HTML Updated 14 hours ago



### Top languages

 HTML  R

### People

10 >





**Any questions?**

Agnieszka Sitko

[ag.agnieszka.sitko@gmail.com](mailto:ag.agnieszka.sitko@gmail.com)

06-07-2017