VariSel : An a R package to perform variable selection in linear models

Marie

3 février 2019

Introduction

Statistical modelling

- **Dataset description:**
  - $\boldsymbol{X}$: $n \times p$ design matrix
  - $\boldsymbol{Y}$: $n \times q$ response matrix
- **Question:** Which variables influence the responses?
- **Approach:**
  - Variable selection in
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E},$$
where
    - $\boldsymbol{B}$: $p \times q$ **sparse** coefficients matrix
    - $\boldsymbol{E}$: $n \times q$ error matrix with
$$\forall i \in \{1, \dots, n\},\ (E_{i,1}, \dots, E_{i,q}) \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma}_q)$$
  - We take the dependence into account by estimating $\boldsymbol{\Sigma}_q$.

Differents penalties : for different point of view

- **Lasso** : *select variables without taking into account potential links.*
$$\widehat{b}_L = \text{Argmin}_b \left\{ ||y - \mathcal{X}b||_2^2 + \lambda ||b||_1 \right\},$$
- **Group-Lasso** : *select a group of variables.*
$$\widehat{b}_G = \text{Argmin}_{b_1,\dots,b_L} \left\{ ||y - \sum_{1 \leq \ell \leq L} \mathcal{X}_{(\ell)} b_{(\ell)}||_2^2 + \lambda \sum_{1 \leq \ell \leq L} \sqrt{p_\ell} ||b_\ell||_2 \right\},$$
- **Fused-Lasso** : *influence a group of variables to have the same coefficient.*
$$\widehat{b}_F = \text{Argmin}_b ||y - \mathcal{X}b||_2^2 + \left\{ \lambda_1 \sum_{(i,j) \in \mathcal{G}} |b_i - b_j| + \lambda_2 ||b||_1 \right\},$$

This differents penalties are here in univariate. In order to use them and in order to take into account the dependance that may exist among variables we propose the following transformation :

$\boldsymbol{Y}\widehat{\boldsymbol{\Sigma}}_q^{-1/2} = \boldsymbol{X}\boldsymbol{B}\widehat{\boldsymbol{\Sigma}}_q^{-1/2} + +\boldsymbol{E}\widehat{\boldsymbol{\Sigma}}_q^{-1/2}$

$$\begin{aligned}\mathcal{Y} &= vec(\boldsymbol{Y}\widehat{\boldsymbol{\Sigma}}_q^{-1/2}) = vec(\boldsymbol{X}\boldsymbol{B}\widehat{\boldsymbol{\Sigma}}_q^{-1/2}) + vec(\boldsymbol{E}\widehat{\boldsymbol{\Sigma}}_q^{-1/2}) \\ &= ((\widehat{\boldsymbol{\Sigma}}_q^{-1/2})' \otimes \boldsymbol{X})vec(\boldsymbol{B}) + vec(\boldsymbol{E}\widehat{\boldsymbol{\Sigma}}_q^{-1/2}) \\ &= \mathcal{X}\mathcal{B} + \mathcal{E}.\end{aligned}$$

This transformation need the estimation of the square root of the inverse of the covariance matrix $\boldsymbol{\Sigma}_q$.

Package Installation

```
devtools::install_github("Marie-PerrotDockes/VariSel")
```

First exemples : Iris dataset

```
library(car)
iris %>%
  head() %>%
  kable()
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Associtaion between sepal characteristics and petal characteristics

The aim of this exemple is to select association between the Sepal (Length and width) and the Petal (Length and Width).

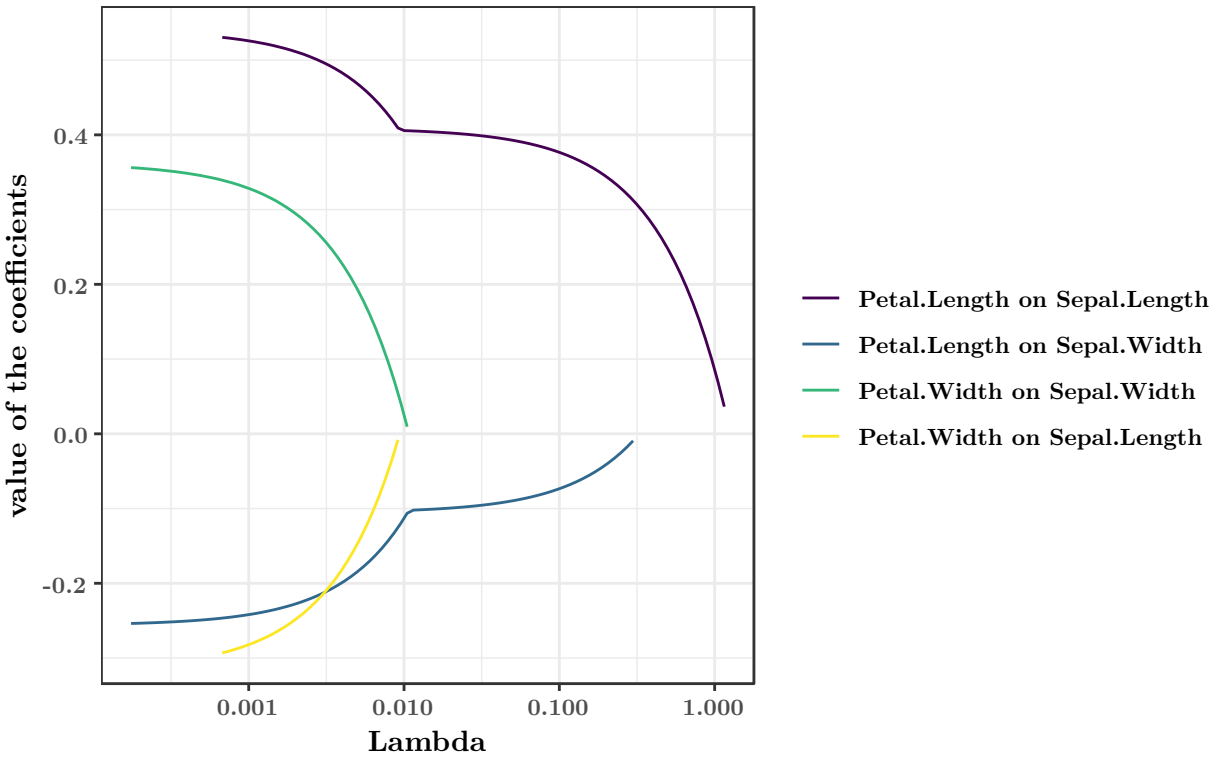
Construction of the matrices :

```
Y <- iris %>% select( starts_with("Sepal"))
Petal_char <- iris %>% select( starts_with("Petal"))
```

Coefficient estimation using a lasso criterion:

```
mod <-  train_VariSel( Y = Y,
                       X = Petal_char,
                       type ="lasso_univ")
```

```
plot(mod)
```



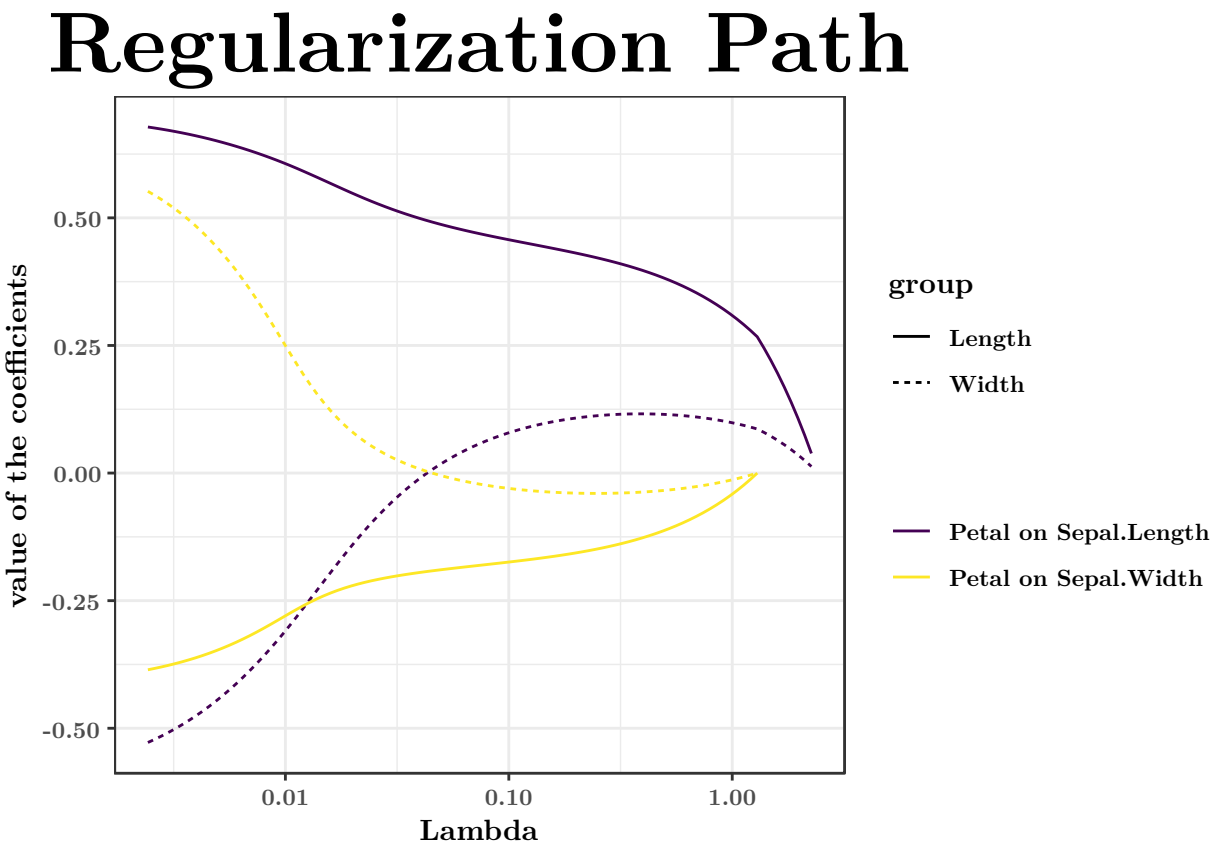
It start by selecting an association between the petal length and the sepal length and then an association between the petal lenght and the sepal widthand so on..

Let us now see what happened when we use group lasso to force the association to be between one sepal characteristics with all petal characteristics.

```
mod_group <-  train_VariSel( Y = Y,
                        X = Petal_char,
                        type ="group_multi_regr",
                        sepx = "\\.")
```

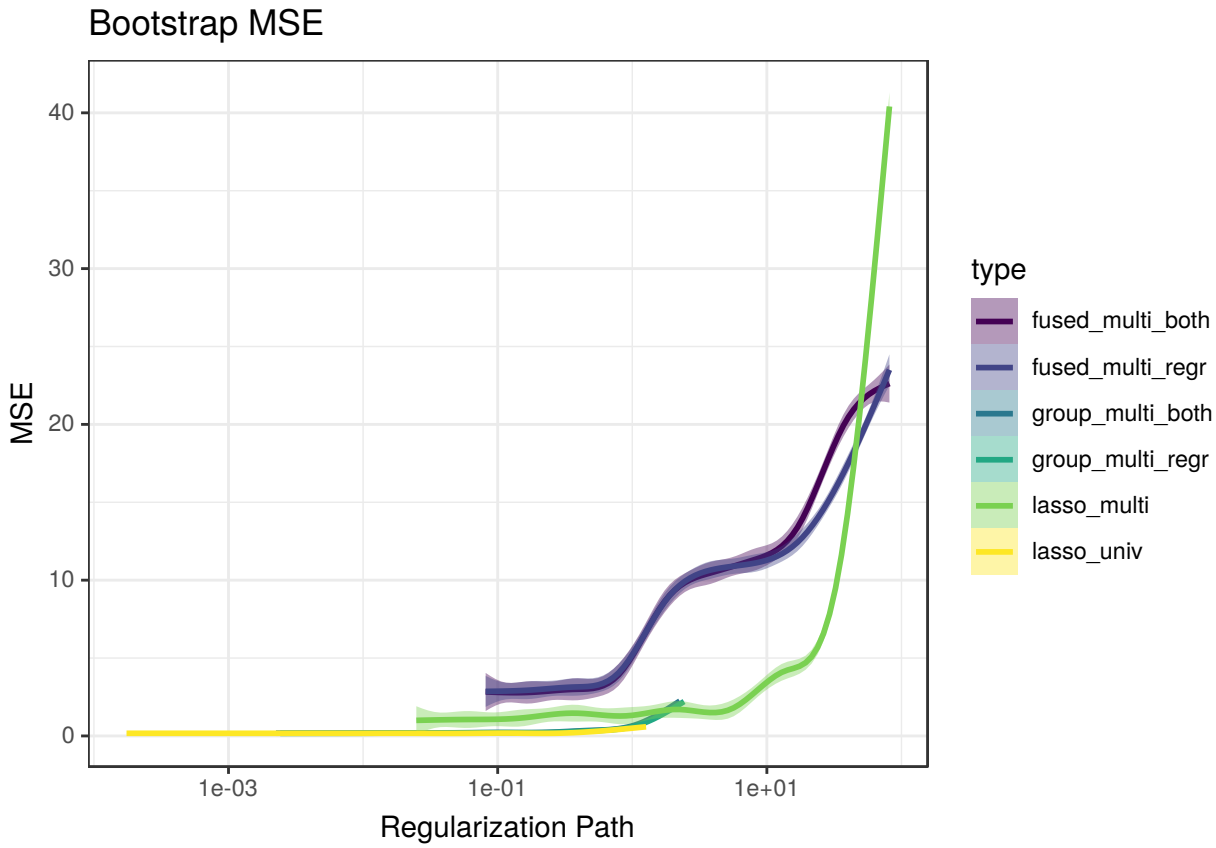
The argument ~sepx = "\."~ mean that the name of X will be seperate into a group name and a characteristics name. Two variables having the same group name will be selected together. Here both Petal.length and Petal.Width start with Petal hence they will be selected together.

```
plot(mod_group)
```



Let's now compare differents types grouping or not the petal characteristics and influence it or not to have the same coefficients.

```
 plan(multiprocess)
ct <- compar_type( Y = Y, X = Petal_char,
  types = c("group_multi_regr" , "group_multi_both" ,
  "fused_multi_regr", "fused_multi_both",
  "lasso_multi", "lasso_univ" ), times = 10)
plot_ct(ct)
```

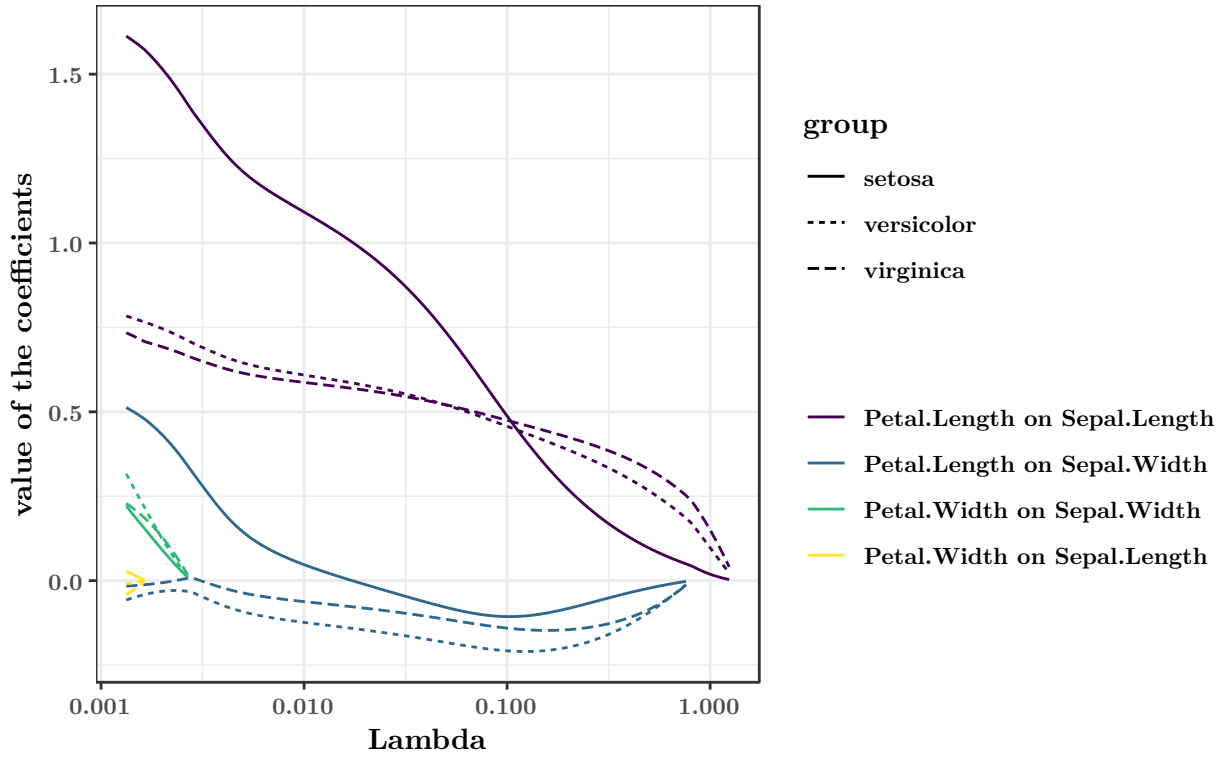


We can also propose different associations between petal characteristics and sepal charcateristics depending on the species.

```
mod_group_species <-  train_VariSel( Y = Y,
                       regressors =  Petal_char,
                       group = iris$Species,
                       type ="group_multi_regr")
```

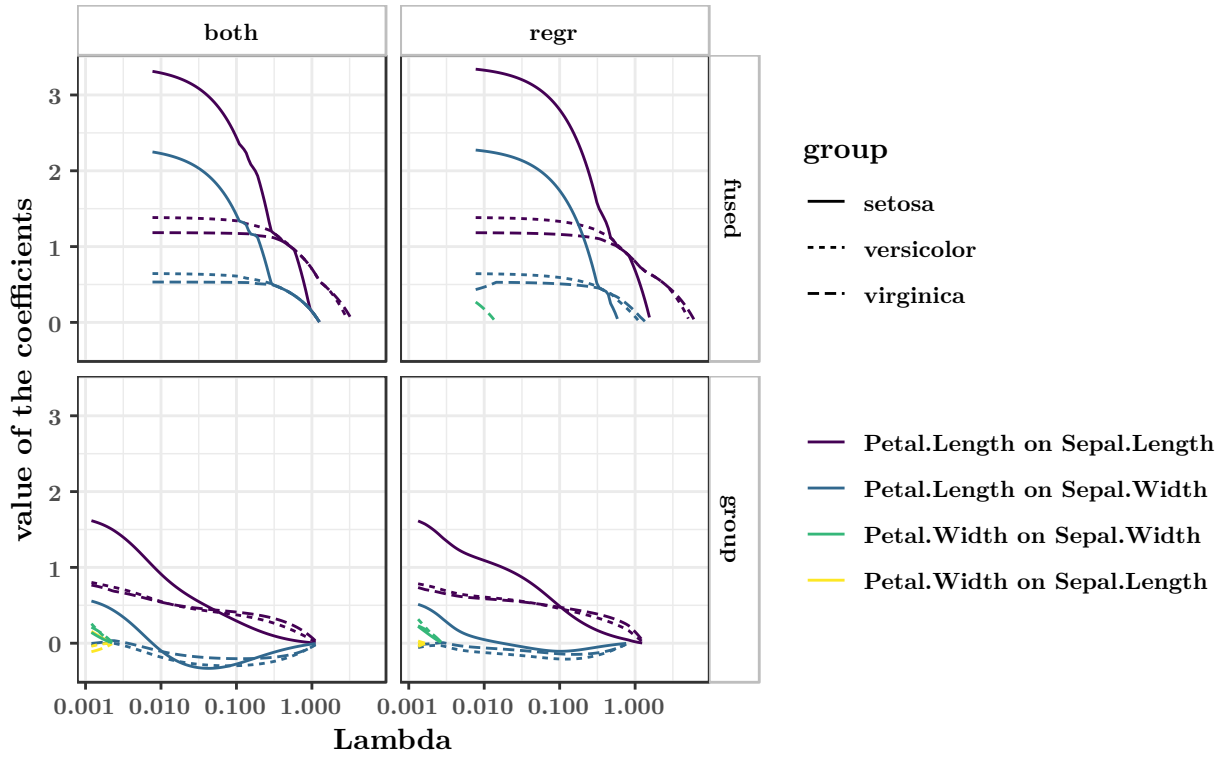
```
plot(mod_group_species)
```

Regularization Path



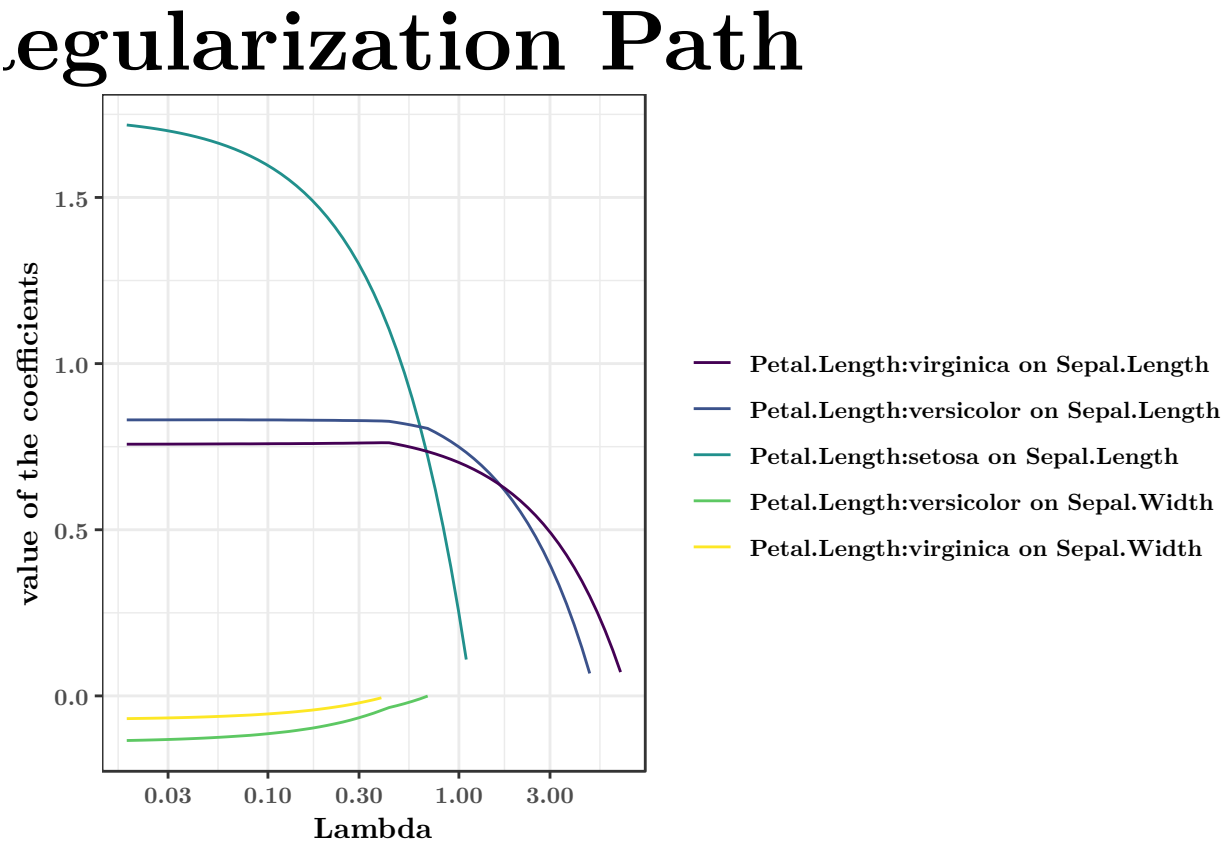

```
compar_path(mods = list(mod_group_species,
                        mod_group_resp_species,
                        mod_fused_multi_species,
                        mod_fused_multi_species_resp))
```

Regularization Path




```
mod_lasso <-   train_VariSel( Y = Y,
                        regressors = Petal_char,
                        group = iris$Species,
                        type ="lasso_multi")
```

```
plot(mod_lasso)
```



A more serious exemple

we will now use the dataset available and described in "A global dataset of CO2 emissions and ancillary data related to emissions for 343 cities"

The data are available in https://www.nature.com/articles/sdata2018280?WT.feed_name=subjects_environmental-social-sciences

```
CO2 <-read.csv(file="Data/D_FINAL.csv", dec =",")
dim(CO2)
```

```
## [1] 343 180
```

```
cc <-complete.cases(CO2[,c("Scope.1.GHG.emissions..tCO2.or.tCO2.eq.",
                           "Total.emissions..CDP...tCO2.eq.",
                           "Population..CDP.",
                           "Average.annual.temperature..CDP...degrees.Celsius.",
                           "Average.altitude..m.")])
CO2_full <- CO2[cc, ]
CO2_quant <- CO2_full %>% select_if(is.numeric)
CO2_quant_full <- CO2_quant[,complete.cases(t(CO2_quant))]
Region <- as.character(CO2_full$Region)
Region[grepl('Asia', Region)] <- "Asia"
kable(table(Region))
```

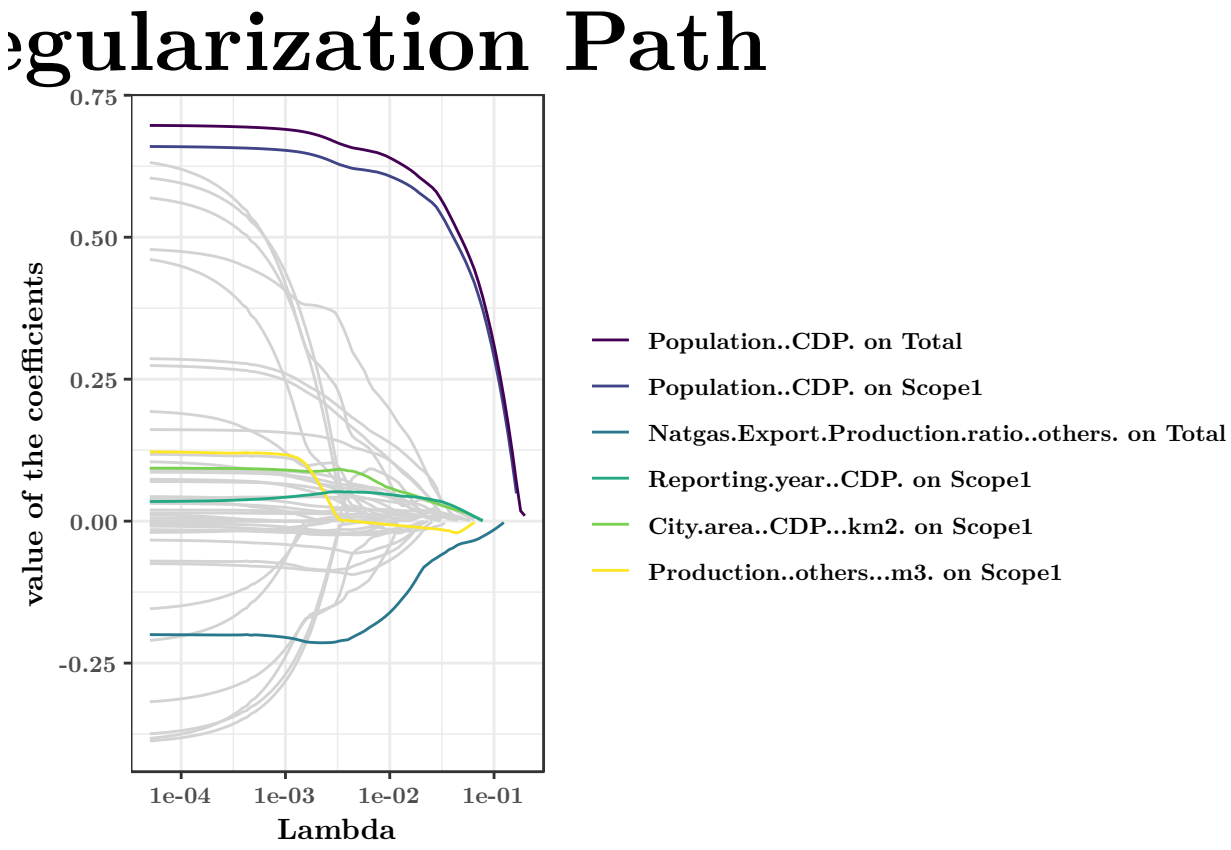
Region	Freq
Africa	4

Region	Freq
Asia	17
Europe	29
Latin America & Caribbean	17
North America	64
Oceania	5

```
Y <- CO2_quant_full[,c("Scope.1.GHG.emissions..tCO2.or.tCO2.eq.",
                       "Total.emissions..CDP...tCO2.eq.")] %>%
  scale()
colnames(Y) <- c("Scope1", "Total")
CO2_quant_full_exp <-CO2_quant_full %>%
  select(-Scope.1.GHG.emissions..tCO2.or.tCO2.eq.,
    -Total.emissions..CDP...tCO2.eq.) %>% scale()
```

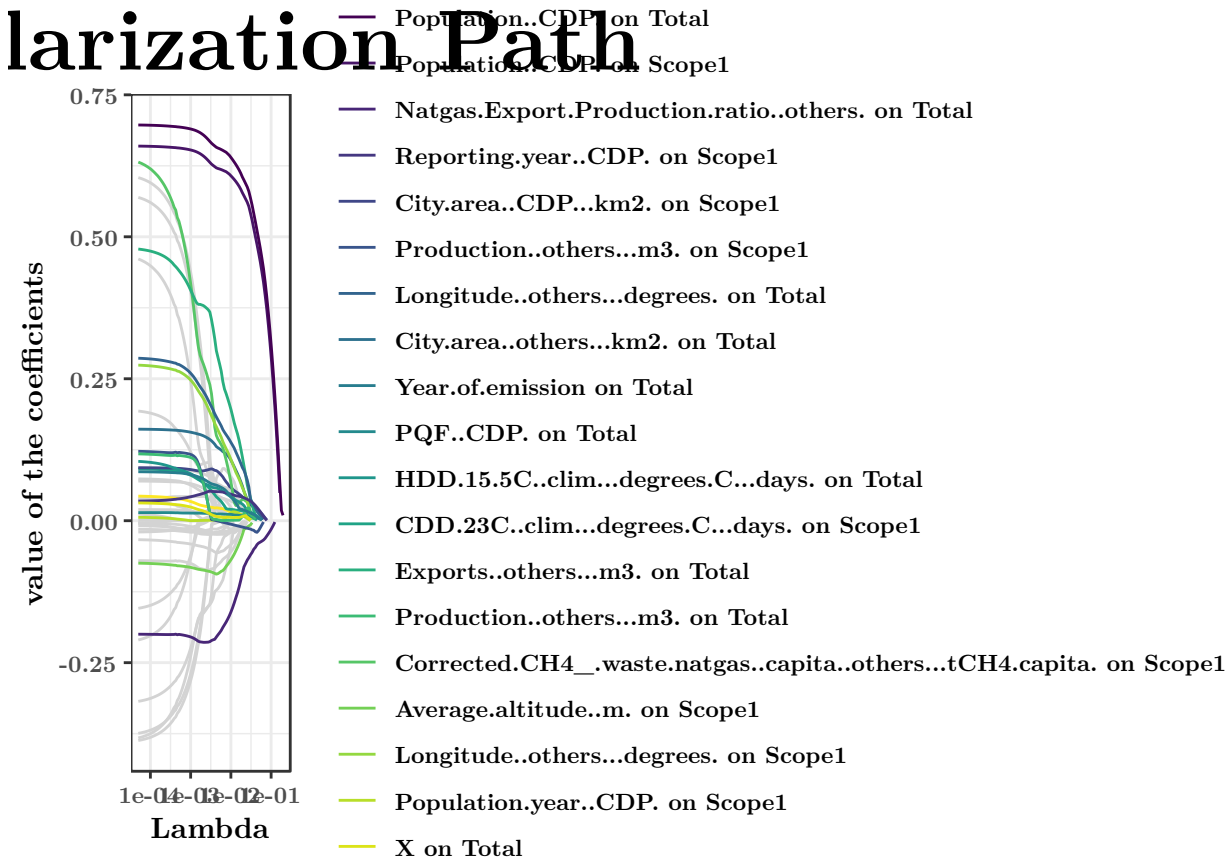
First model : association between in different region

```
mod_lasso <- train_VariSel(Y = Y,
                           X = CO2_quant_full_exp,
                           type = "lasso_multi")
plot(mod_lasso)
```



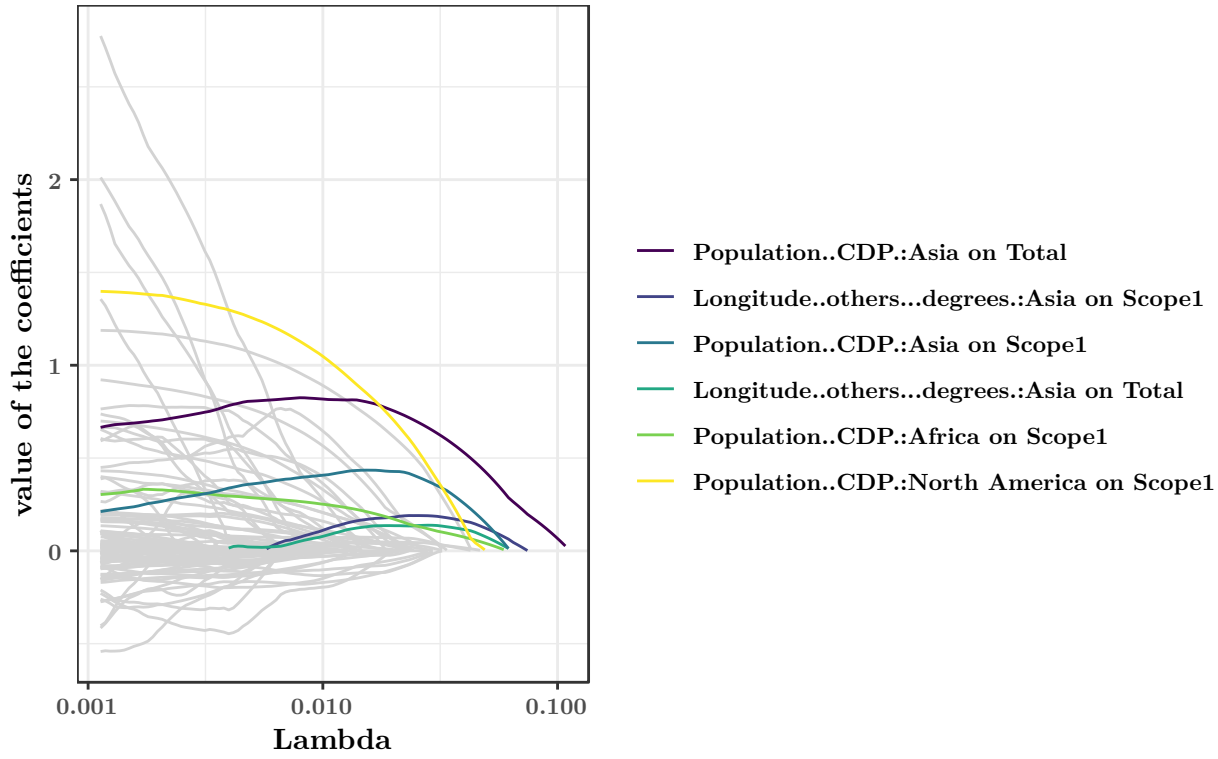
You can choose to change the number of color variable (the one we can see as selected)

```
plot(mod_lasso, n=20)
```

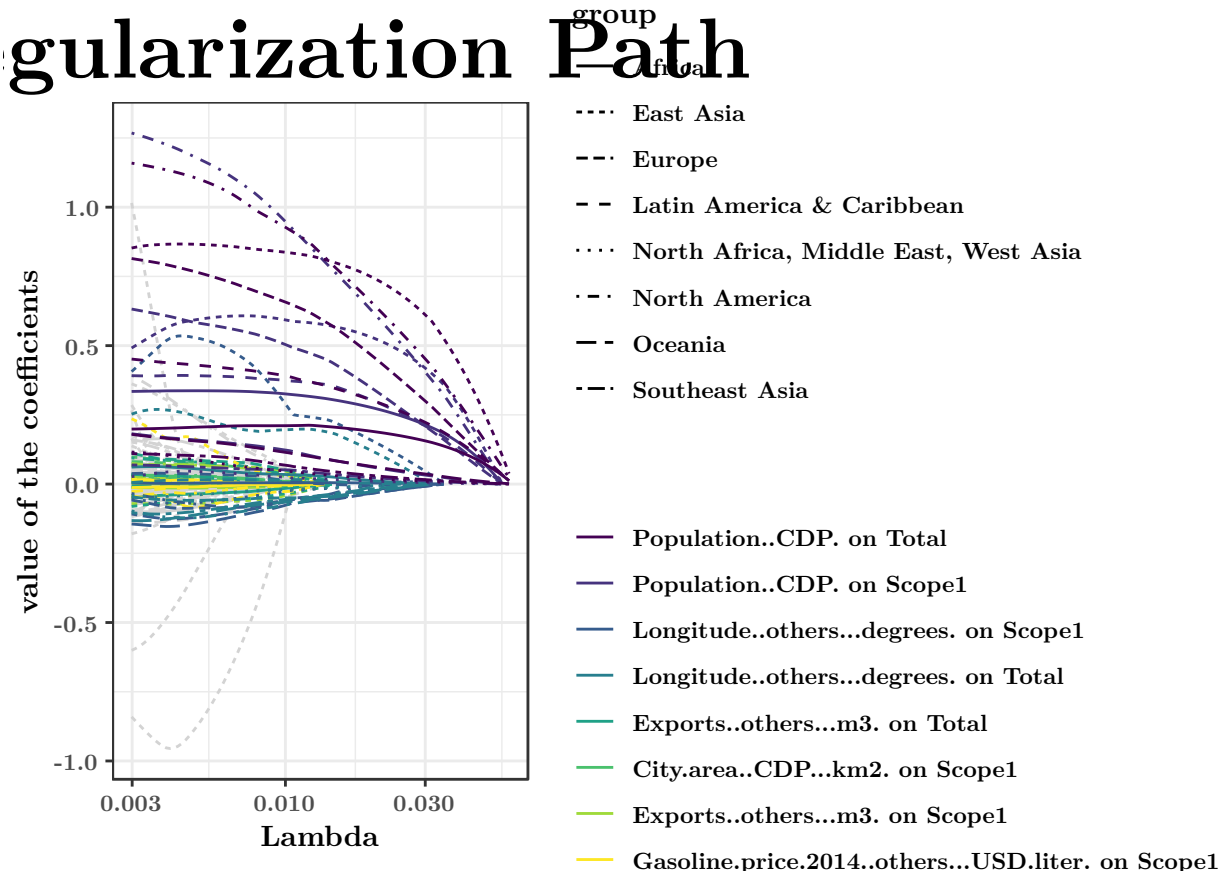



```
mod_lasso_reg <- train_VariSel(Y = Y,
                               regressors = CO2_quant_full_exp,
                               group = Region,
                               type = "lasso_multi")
plot(mod_lasso_reg)
```

egularization Path




```
mod_group_multi_regr<- train_VariSel(Y = Y,
                           regressors = CO2_quant_full_exp,
                           group = as.character(CO2_full$Region),
                           type = "group_multi_regr")
plot(mod_group_multi_regr, n =8)
```



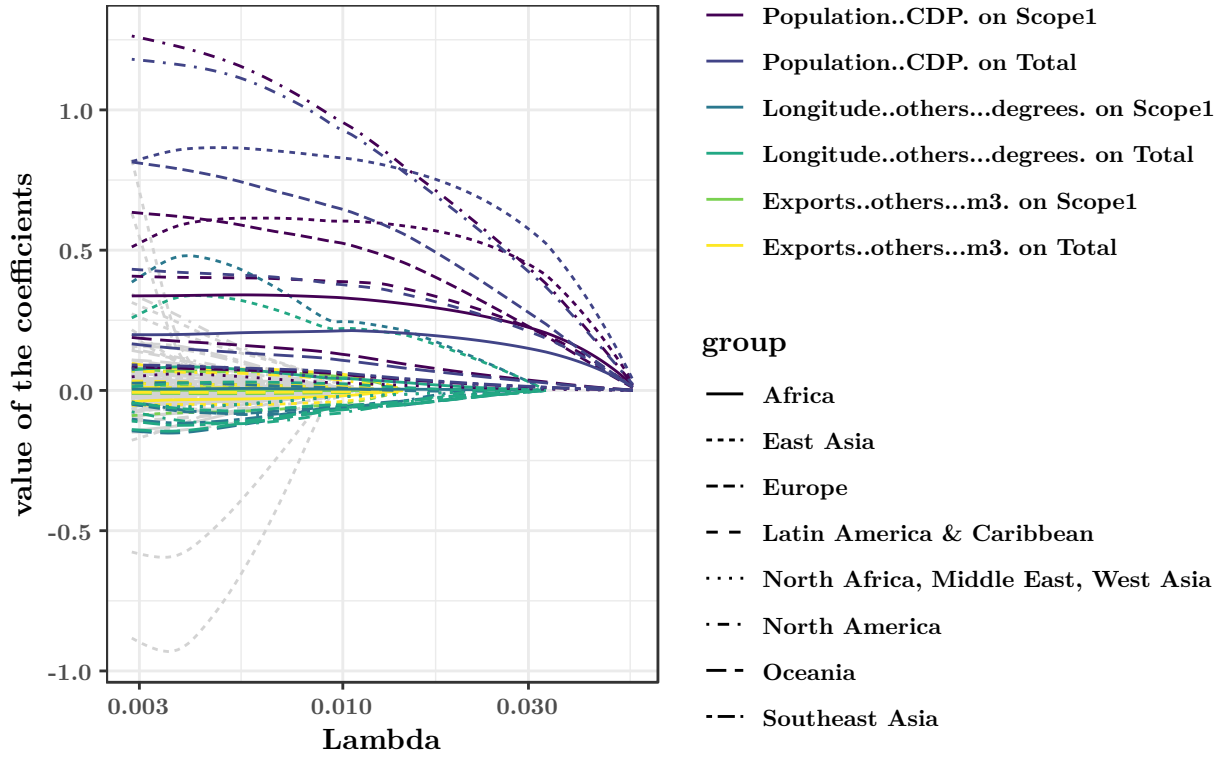
we also easily have acces to coefficient :

```
kable(head(coef(mod_group_multi_regr)))
```

Lambda	num_lambda	Trait	Reg	group	value
0.0579564	2	Total	Population..CDP.	Africa	0.0119366
0.0579564	2	Total	Population..CDP.	East Asia	0.0380006
0.0579564	2	Total	Population..CDP.	Europe	0.0122879
0.0579564	2	Total	Population..CDP.	Latin America & Caribbean	0.0122957
0.0579564	2	Total	Population..CDP.	North Africa, Middle East, West Asia	0.0003086
0.0579564	2	Total	Population..CDP.	North America	0.0217811

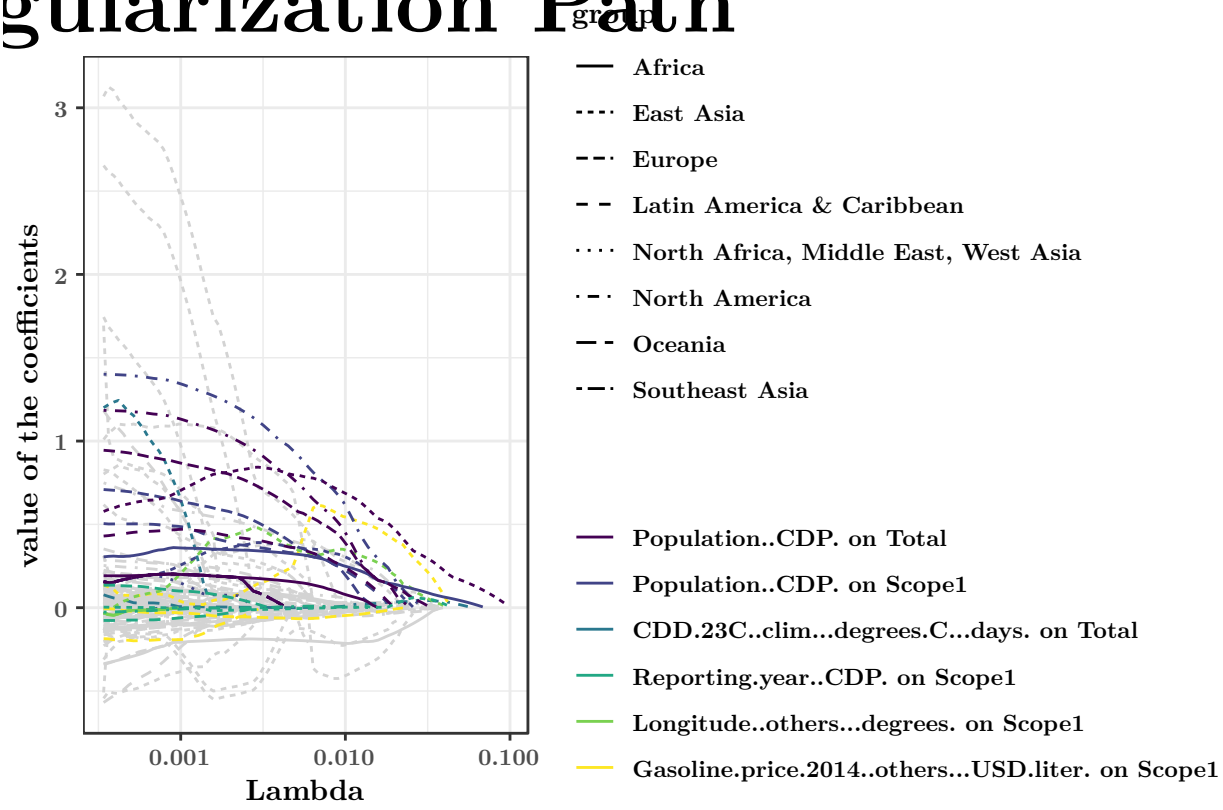
```
mod_lasso_reg <- train_VariSel(Y = Y,
                               regressors = CO2_quant_full_exp,
                               group = as.character(CO2_full$Region),
                               type = "group_multi_both")
plot(mod_lasso_reg, n =6)
```

Regularization Path

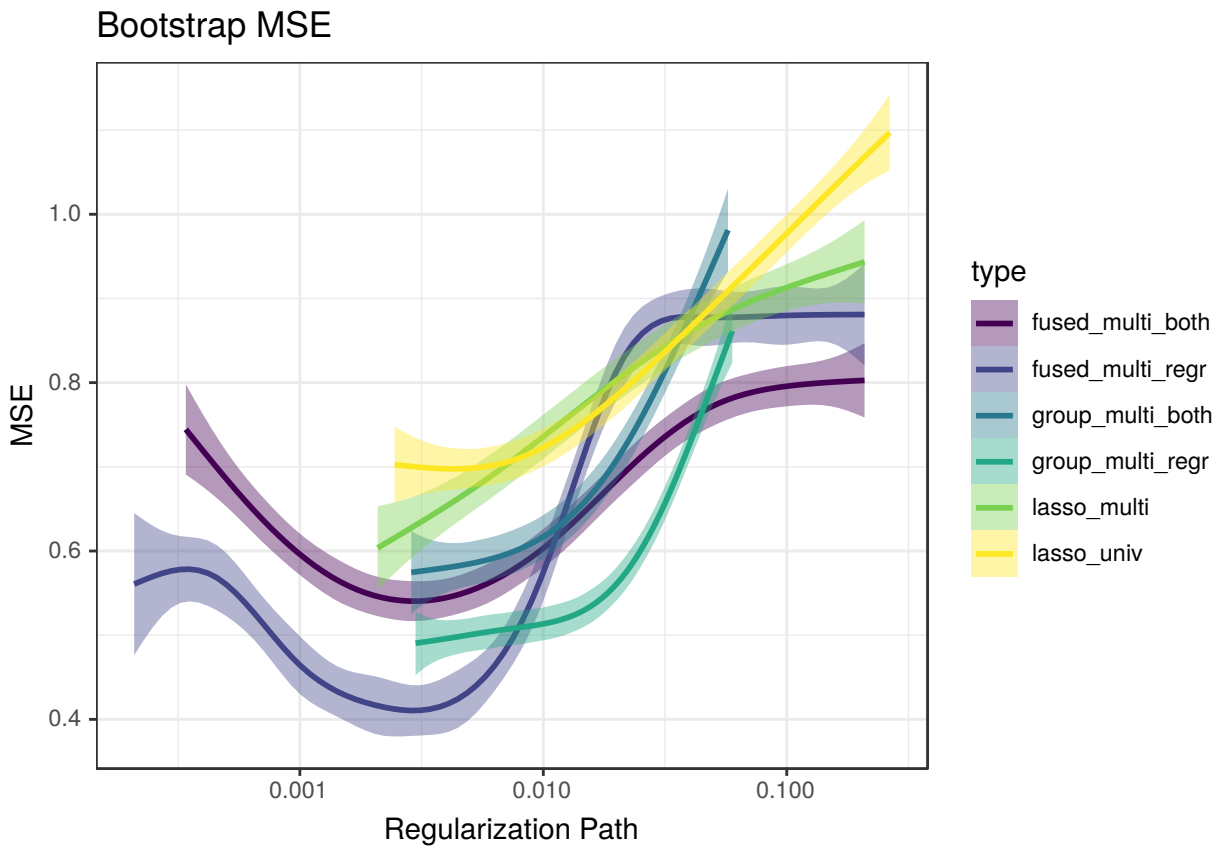



```
mod_lasso_reg <- train_VariSel(Y = Y,
                               regressors = CO2_quant_full_exp,
                               group = as.character(CO2_full$Region),
                               type = "fused_multi_both")
plot(mod_lasso_reg, n =6)
```

gularization Path

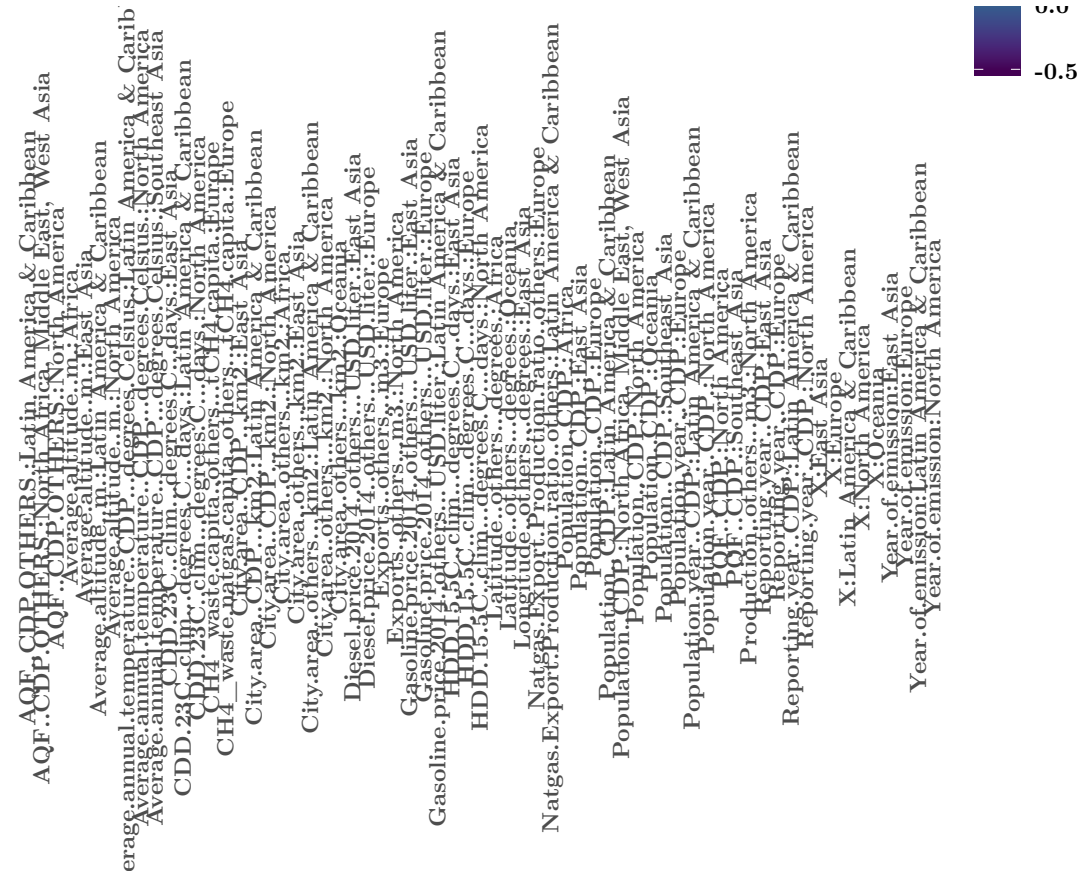



```
set.seed(123)
plan(multiprocess)
ct <- compar_type(Y = Y,
        regressors = CO2_quant_full_exp,
        group = as.character(CO2_full$Region),
        types = c("group_multi_regr" , "group_multi_both" ,
          "fused_multi_regr", "fused_multi_both",
          "lasso_multi", "lasso_univ" ))
plot_ct(ct)
```



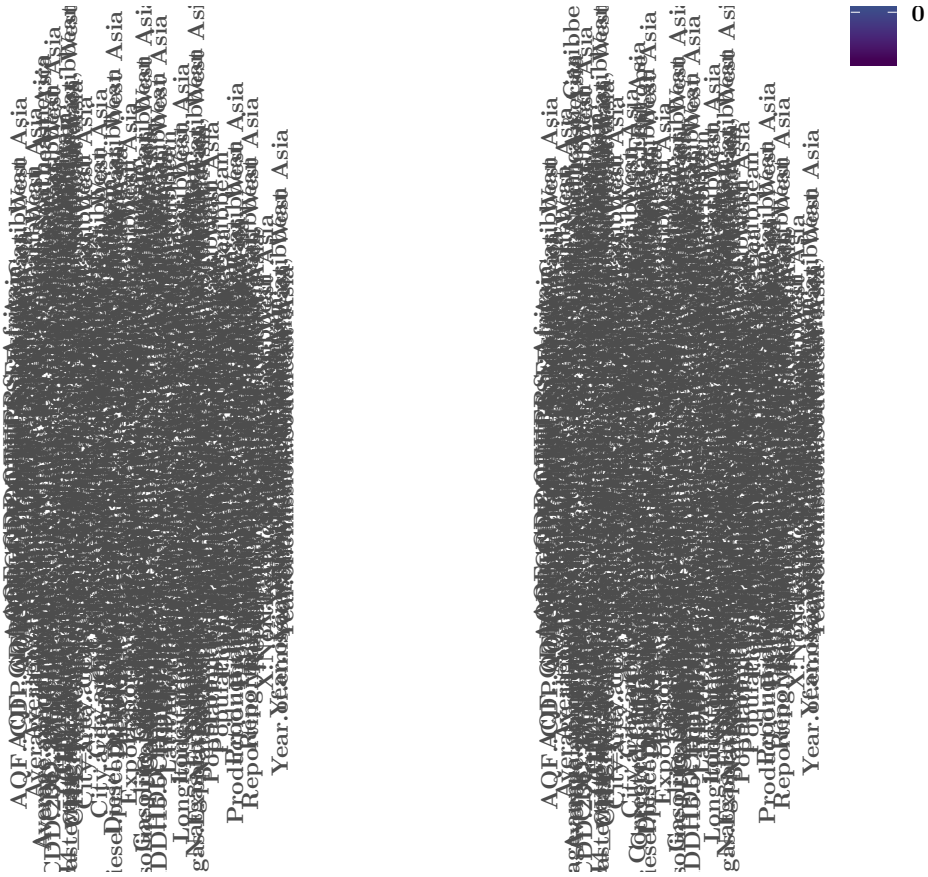

```
bm <- get_best_models(ct, criterion ="MSE_boot")


plot_md(bm, type = "fused_multi_both")
```



Even when influence by the model to have the same coefficient the explicative variables keep different coefficient according to the place

```
plot_md(bm)
```



When there is more reponses than individuals

```
require(MultiVarSel)
dim(MultiVarSel::prot)
```

```
## [1]   9 725
```

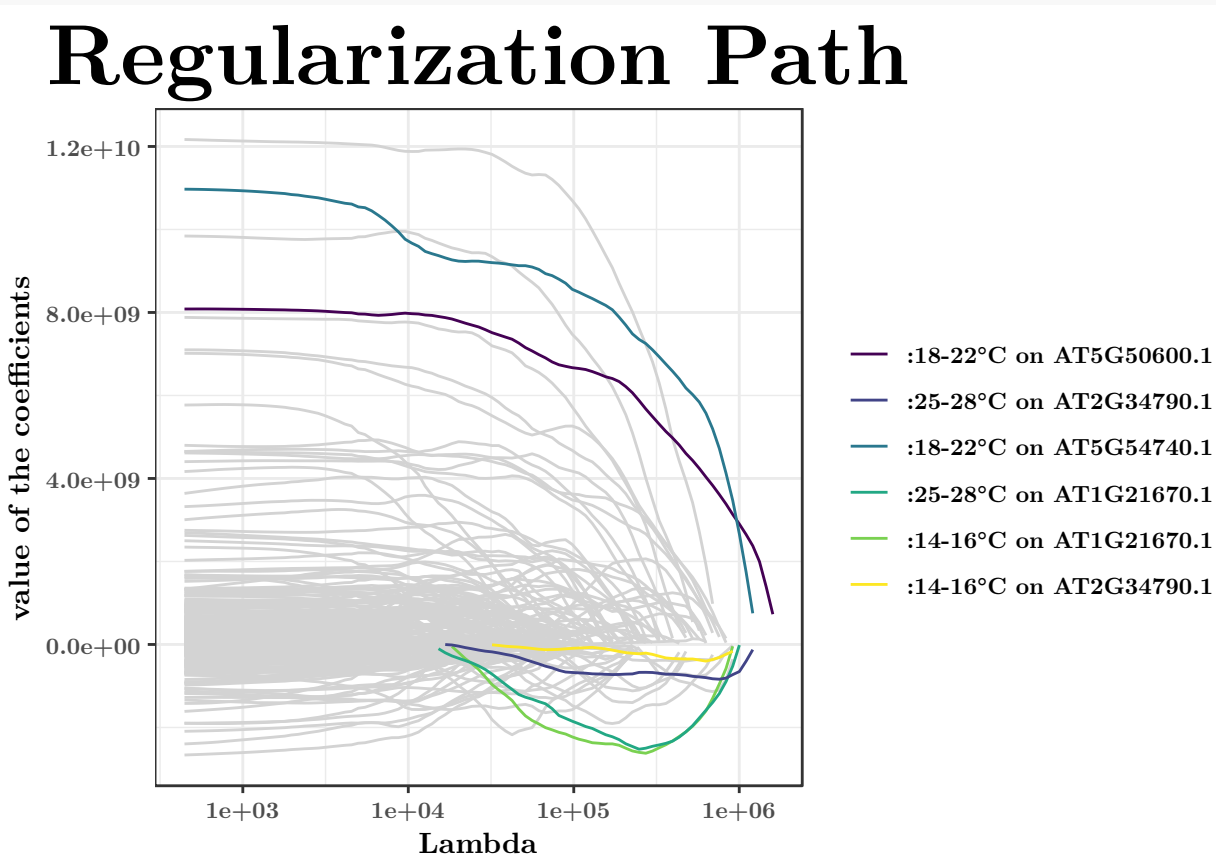
```
kable(head(prot[, 1:10]))
```

temperature	AT1G01470.1	AT1G01900.1	AT1G02700.1	AT1G02780.1	AT1G03030.1	AT1G03880.1	AT1G0389
14-16°C	22866363	383154342	48307337	30263431	5688062	3836522010	7527640
14-16°C	19209303	562809367	48089559	12672606	6224114	4463958133	6229194
14-16°C	24780232	402493402	70132972	11465256	6403075	4981020276	7278106
18-22°C	46092725	284388931	79060256	12062102	9879726	6158403160	7948625
18-22°C	49154336	263711099	85218205	13788091	12741874	9124148585	7642146
18-22°C	36031053	335996708	82237070	8452248	9805074	3253684736	4493027

In this case the number of reponses is the nulbver of proteines (724 for 9 individuals)

```
mod_lasso_reg <- train_VariSel(Y =prot[,-1],
                               regressors = NULL,
                               group = prot[,1],
                               type = "lasso_multi",
                               type_S12_inv = "Block")
```

```
plot(mod_lasso_reg)
```




```
kable(head(coef(mod_lasso_reg)))
```

Lambda	num_lambda	Trait	Marker	value
1595891	2	AT5G50600.1	:18-22°C	729070470
1454116	3	AT5G50600.1	:18-22°C	1393372330
1324937	4	AT5G50600.1	:18-22°C	1998659443
1207233	5	AT2G34790.1	:25-28°C	-125957909
1207233	5	AT5G50600.1	:18-22°C	2377590572
1207233	5	AT5G54740.1	:18-22°C	751013797