VariSel: An R package to perform variable selection in the multivariate linear model

Or how a Gallic village stays irreductible

Marie Perrot-Dockès, Julien Chiquet

UseR2019

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

INRA
SCIENCE & IMPACT

A simple vision of the immune system :

Or how Astérix and Obélix can kick-off the Romans

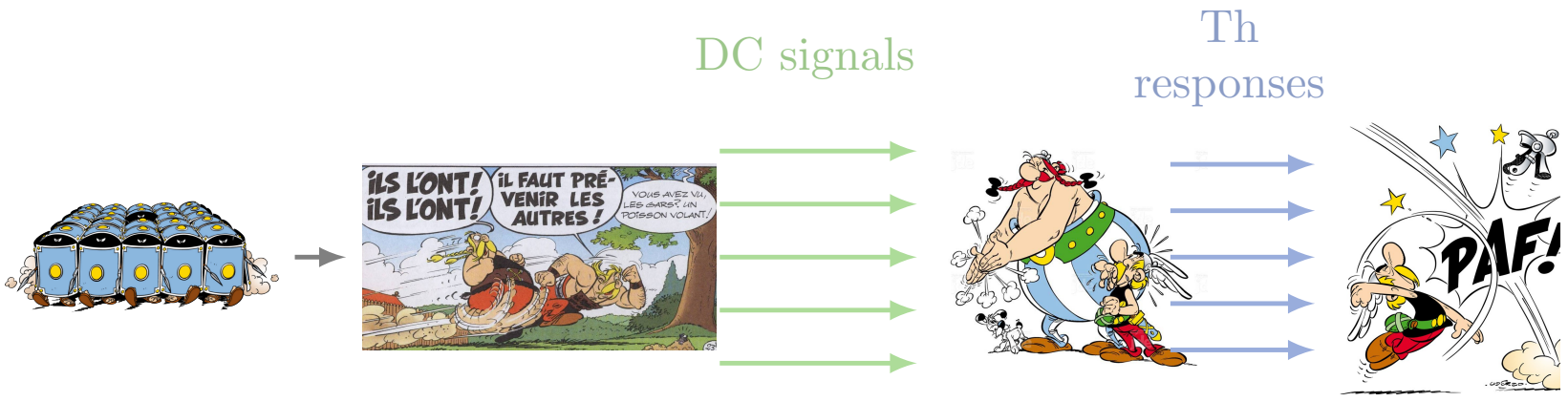


Figure 1: DC Th dialogue

Grandclaudon, M., Perrot-Dockès, M, Trichot, C et al. *A Quantitative Multivariate Model of Human Dendritic Cell-T Helper Cell Communication (March 15, 2019)*. Available at http://dx.doi.org/10.2139/ssrn.3353217

Experimental set up :

Ordralphabetix and Cétautomatix are two!

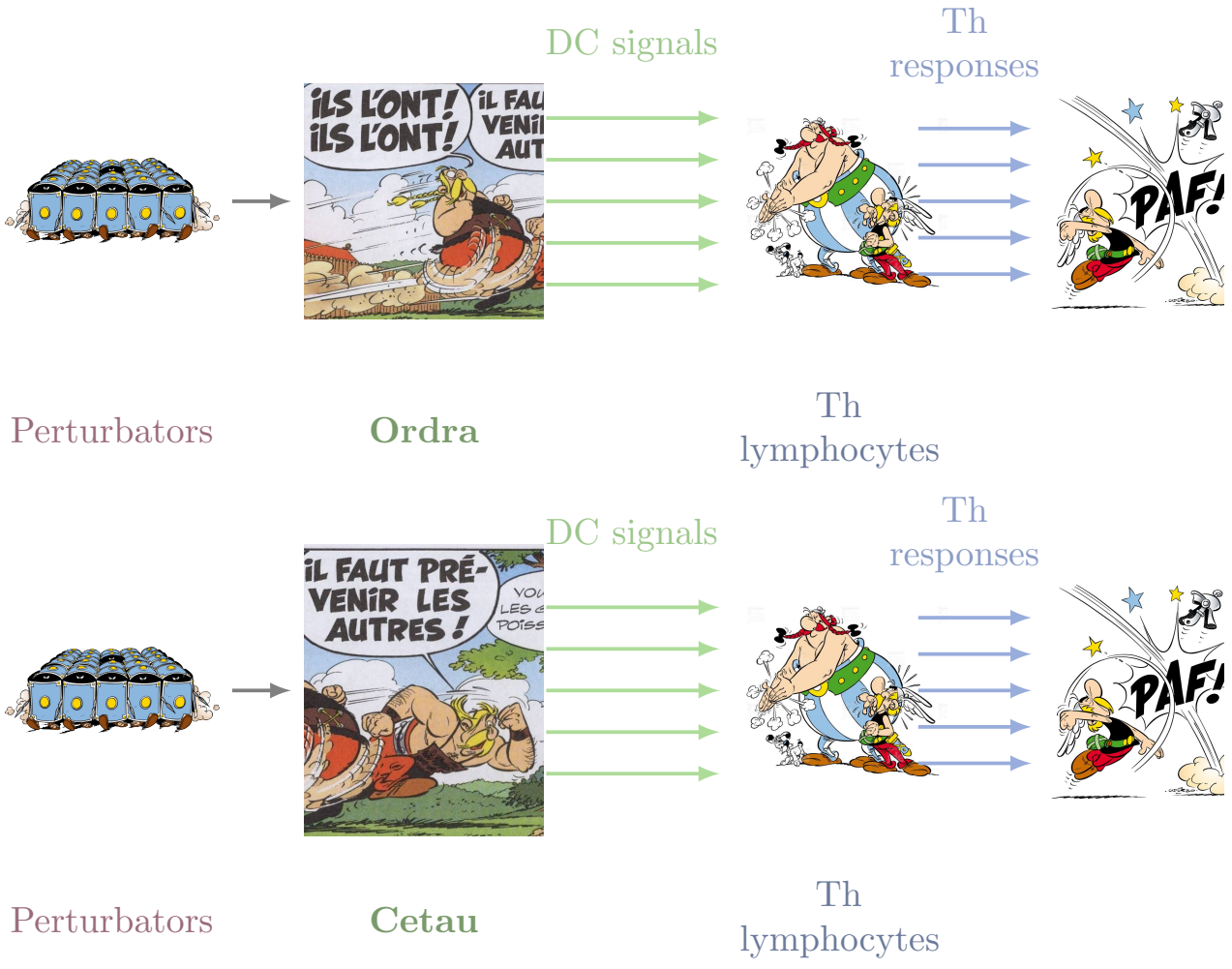


Figure 2: DC Th dialogue

Statistical modelling
Understand the Gallic language

- **Dataset description:**
  - $\boldsymbol{X}$: $n \times p$ design matrix : the DC signals 
  - $\boldsymbol{Y}$: $n \times q$ response matrix : the Th responses 
- **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 $\boldsymbol{\Sigma}_q$.

Statistical modelling
Understand the Gallic language

- **Dataset description:**
  - $\boldsymbol{X}$: $n \times p$ design matrix : the DC signals 
  - $\boldsymbol{Y}$: $n \times q$ response matrix : the Th responses 
- **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.*

$$\widehat{b}_L = \text{Argmin}_b \left\{ ||y - \boldsymbol{X}b||_2^2 + \lambda ||b||_1 \right\},$$

- **Group-Lasso** : *select a group of variables.*

$$\widehat{b}_G = \text{Argmin}_{b_1,\ldots,b_L} \left\{ ||y - \sum_{1 \leq \ell \leq L} \boldsymbol{X}_{(\ell)} b_{(\ell)}||_2^2 + \lambda \sum_{1 \leq \ell \leq L} \sqrt{p_\ell} ||b_\ell||_2 \right\},$$

- **Fused-Lasso** : *influence a group of variables to have the same coefficient.*

$$\widehat{b}_F = \text{Argmin}_b ||y - \boldsymbol{X}b||_2^2 + \left\{ \lambda_1 \sum_{(i,j) \in \mathcal{G}} |b_i - b_j| + \lambda_2 ||b||_1 \right\},$$

Differents penalties : for different point of view

- **Lasso** : *select variables without taking into account potential links.*

$$\widehat{b}_L = \text{Argmin}_b \left\{ ||y - \boldsymbol{X}b||_2^2 + \lambda ||b||_1 \right\},$$

- **Group-Lasso** : *select a group of variables.*

$$\widehat{b}_G = \text{Argmin}_{b_1,\ldots,b_L} \left\{ ||y - \sum_{1 \leq \ell \leq L} \boldsymbol{X}_{(\ell)} b_{(\ell)}||_2^2 + \lambda \sum_{1 \leq \ell \leq L} \sqrt{p_\ell} ||b_\ell||_2 \right\},$$

- **Fused-Lasso** : *influence a group of variables to have the same coefficient.*

$$\widehat{b}_F = \text{Argmin}_b ||y - \boldsymbol{X}b||_2^2 + \left\{ \lambda_1 \sum_{(i,j) \in \mathcal{G}} |b_i - b_j| + \lambda_2 ||b||_1 \right\},$$

VariSel for one model type

```
mod <-  train_VariSel( Y = T_resp,
                       regressors = DC_sign,
                       group = dc,
                       type ="group_multi_regr")
```

```
X <-  model.matrix(~DC_sign:dc -1)
mod <-  train_VariSel( Y = T_resp,
                       X = X,
                       sepx = ":",
                       type ="group_multi_regr")
```

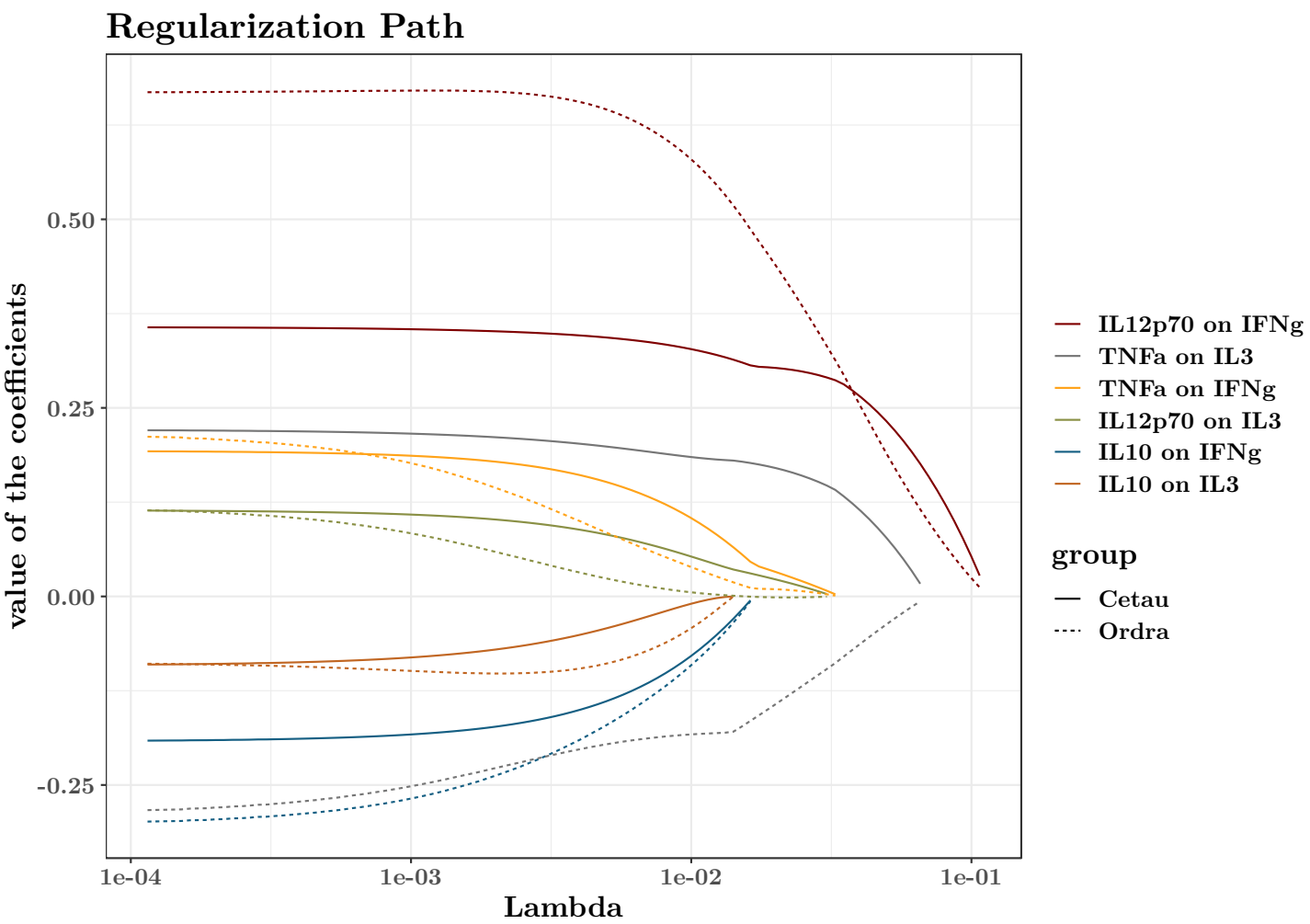
VariSel for one model type

```
mod <-  train_VariSel( Y = T_resp,
                       regressors = DC_sign,
                       group = dc,
                       type ="group_multi_regr")
```

```
X <-  model.matrix(~DC_sign:dc -1)
mod <-  train_VariSel( Y = T_resp,
                       X = X,
                       sepx = ":",
                       type ="group_multi_regr")
```

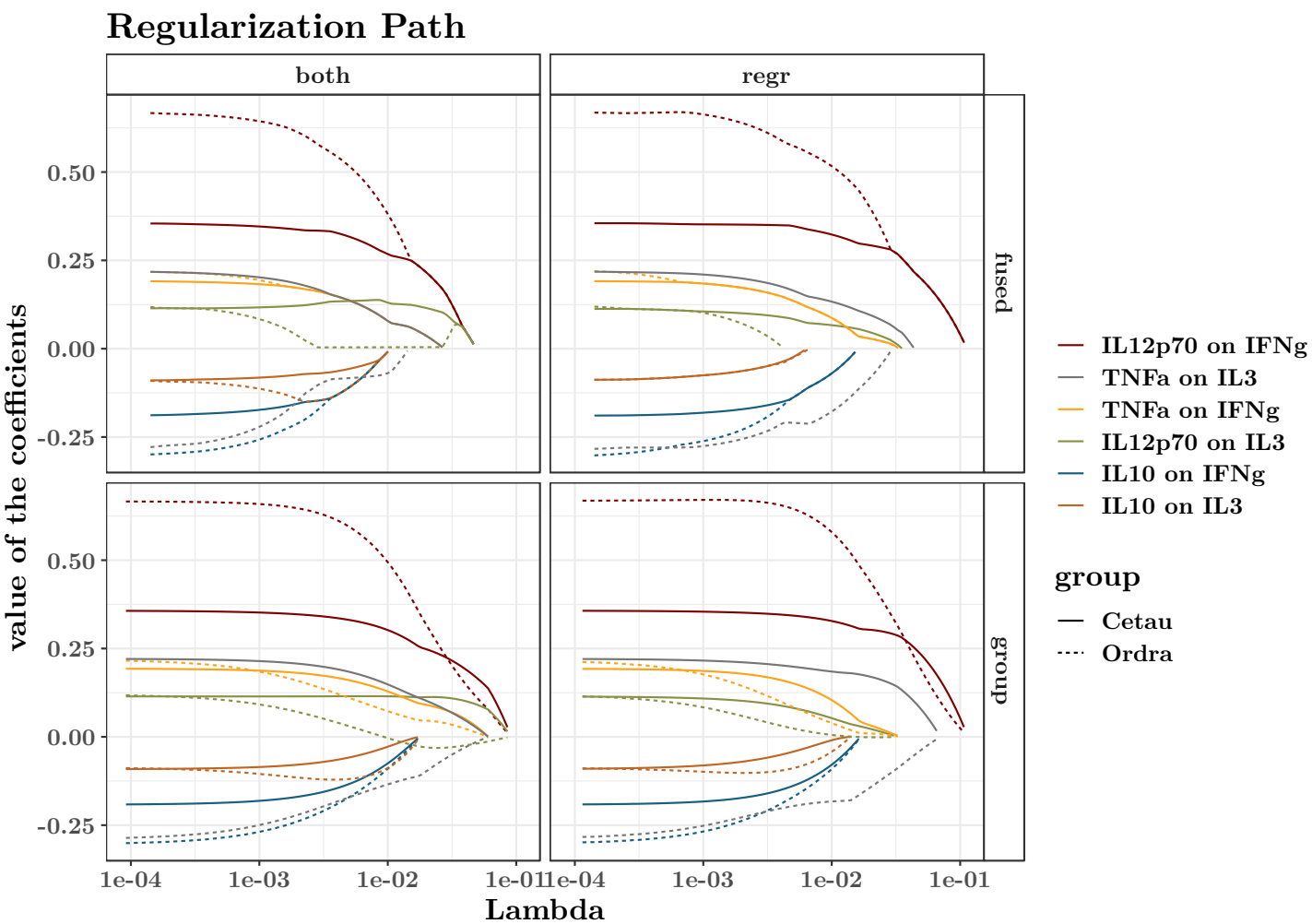
VariSel for one model type : Outcome

```
plot(mod)
```



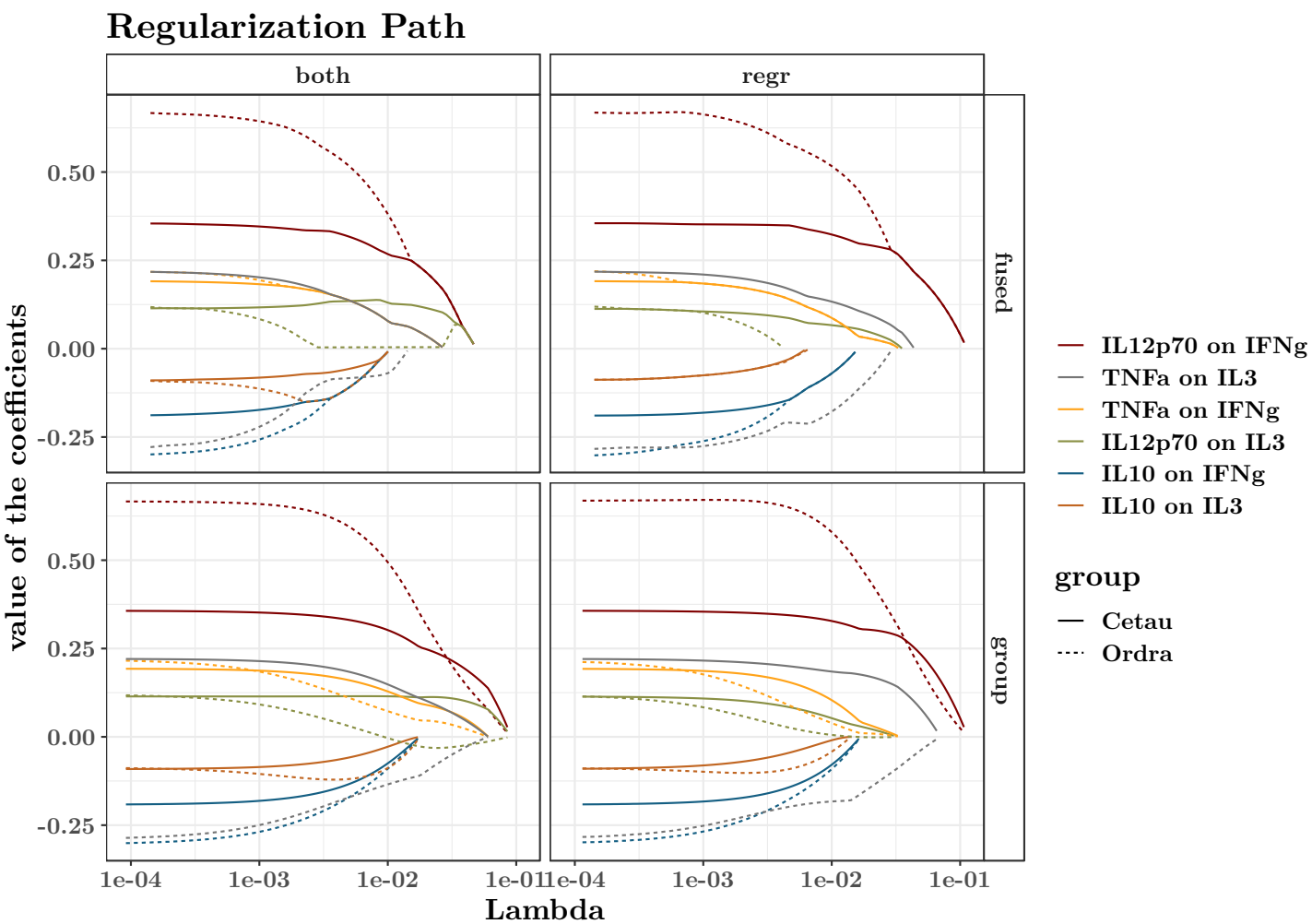
Different modelling strategy

```
compar_path(mods = list(mod,m2,m3,m4))
```



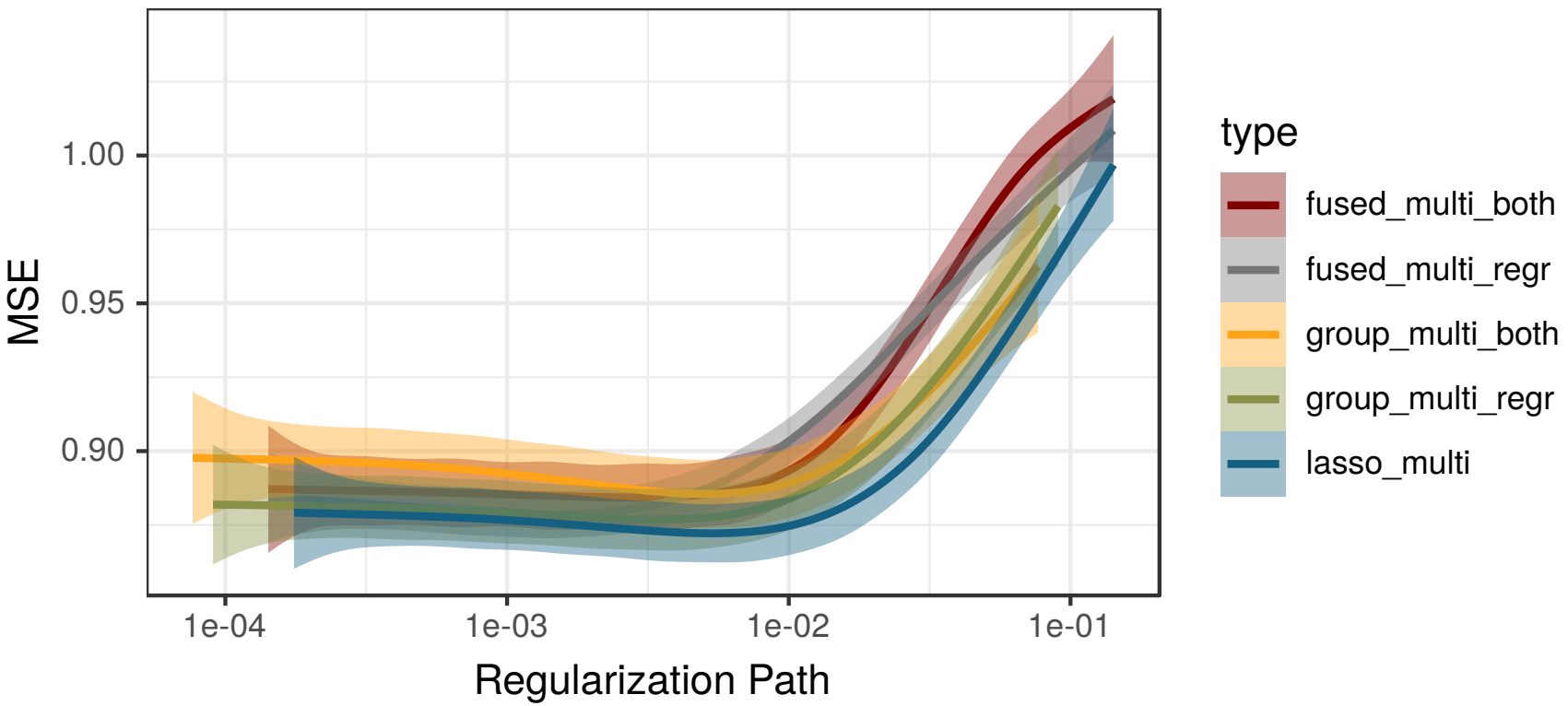
Different modelling strategy

```
compar_path(mods = list(mod,m2,m3,m4))
```



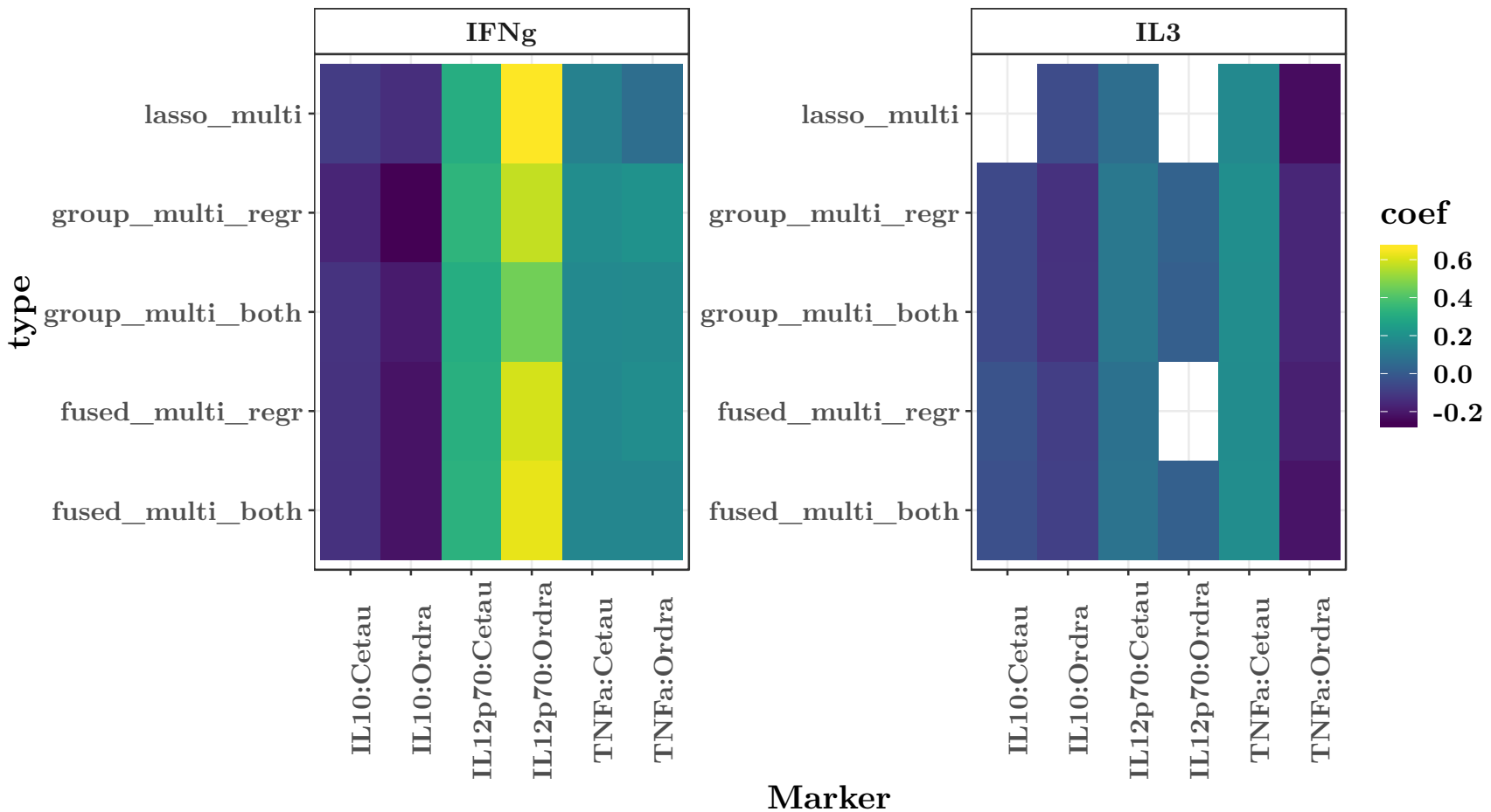
Models selection

```
ct <- compar_type( Y = T_resp, regressors = DC_sign,
                   group = dc,
 types = c("group_multi_regr" , "group_multi_both" ,
  "fused_multi_regr", "fused_multi_both",
  "lasso_multi" ), times = 10)
```



Best models representation

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



Conclusion

This is an R package to perform variable selection in multivariate linear models. It can

- Associate explicative variables
- Associate responses
- Associate both explicative variables and responses
- Let all variables 'free', without assoiciating any of them

Come and see the vignette!
https://github.com/Marie-PerrotDockes/VariSel