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
A simple vision of the immune system:
Or how Astérix and Obélix can kick-off the Romans

Figure 1: DC Th dialogue

Experimental set up:
Ordralphabetix and Cétautomatix are two!

Figure 2: DC Th dialogue
Dataset description:

- **X**: \( n \times p \) design matrix: the DC signals

- **Y**: \( n \times q \) response matrix: the Th responses

Question: Which variables influence the responses?

Approach:

- Variable selection in

\[
Y = XB + E,
\]

where

- **B**: \( p \times q \) sparse coefficients matrix
- **E**: \( n \times q \) error matrix with

\[
\forall i \in \{1, \ldots, n\}, (E_{i,1}, \ldots, E_{i,q}) \sim \mathcal{N}(0, \Sigma_q)
\]

- We take the dependence into account by estimating \( \Sigma_q \).
Dataset description:

- **X**: $n \times p$ design matrix : the DC signals
- **Y**: $n \times q$ response matrix : the Th responses

**Question**: Which variables influence the responses?

**Approach**:

Variable selection in

$$Y = XB + E,$$

where

- **B**: $p \times q$ sparse coefficients matrix
- **E**: $n \times q$ error matrix with

$$\forall i \in \{1, \ldots, n\}, (E_{i,1}, \ldots, E_{i,q}) \sim \mathcal{N}(0, \Sigma_q)$$

We take the dependence into account by estimating $\Sigma_q$. 
Differents penalties: for different point of view

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

\[
\hat{b}_L = \text{Argmin}_b \left\{ \| y - Xb \|_2^2 + \lambda \| b \|_1 \right\},
\]

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

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

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

\[
\hat{b}_F = \text{Argmin}_b \| y - Xb \|_2^2 + \left\{ \lambda_1 \sum_{(i,j) \in G} |b_i - b_j| + \lambda_2 \| b \|_1 \right\},
\]
Differents penalties : for different point of view

- **Lasso**: select variables without taking into account potential links.
  \[ \hat{b}_L = \text{Argmin}_b \left\{ ||y - Xb||_2^2 + \lambda ||b||_1 \right\}, \]

- **Group-Lasso**: select a group of variables.
  \[ \hat{b}_G = \text{Argmin}_{b_1, \ldots, b_L} \left\{ ||y - \sum_{1 \leq \ell \leq L} X(\ell) b(\ell)||_2^2 + \lambda \sum_{1 \leq \ell \leq L} \sqrt{p_\ell} ||b_\ell||_2 \right\}, \]

- **Fused-Lasso**: influence a group of variables to have the same coefficient.
  \[ \hat{b}_F = \text{Argmin}_b \left\{ ||y - Xb||_2^2 + \left\{ \lambda_1 \sum_{(i,j) \in G} |b_i - b_j| + \lambda_2 ||b||_1 \right\} \right\}, \]
VariSel for one model type

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

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

Regularization Path

value of the coefficients

Lambda

IL12p70 on IFNg
TNFa on IL3
TNFa on IFNg
IL12p70 on IL3
IL10 on IFNg
IL10 on IL3

group

Cetau
Ordra

Lambda

1e-04 1e-03 1e-02 1e-01
Different modelling strategy

\texttt{compar\_path(mods = list(mod, m2, m3, m4))}

![Regularization Path Graph](image)
Different modelling strategy

\texttt{compar\_path(mods = list(mod,m2,m3,m4))}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{regularization_path.png}
\caption{Regularization Path}
\end{figure}

- IL12p70 on IFN\(\gamma\)
- TNF\(\alpha\) on IL3
- TNF\(\alpha\) on IFN\(\gamma\)
- IL12p70 on IL3
- IL10 on IFN\(\gamma\)
- IL10 on IL3

\begin{itemize}
  \item \texttt{Cetau}
  \item \texttt{Ordra}
\end{itemize}
ct <- `compar_type`( Y = TResp, regressors = DC_sign, 
group = dc, 
types = c("group_multi_regr", "group_multi_both", 
"fused_multi_regr", "fused_multi_both", "lasso_multi" ), times = 10)
bm <- get_best_models(ct, criterion = "MSE_boot")
plot_md(bm)
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 associating any of them

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