



**HAL**  
open science

## Variable selection by exploiting correlation

Quentin Grimonprez, Alain Celisse, Guillemette Marot

► **To cite this version:**

Quentin Grimonprez, Alain Celisse, Guillemette Marot. Variable selection by exploiting correlation. XXVIIIth International Biometric Conference, Jul 2016, Victoria, Canada. . hal-04031191

**HAL Id: hal-04031191**

**<https://hal.science/hal-04031191v1>**

Submitted on 15 Mar 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Variable selection by exploiting correlation

## Framework

### Notations

- $X \in \mathcal{M}_{n,p}(\mathbb{R})$  matrix containing the  $p$  covariates for  $n$  samples
- $y \in \mathbb{R}^n$  response variable,  $\epsilon$  random noise
- $\beta^* \in \mathbb{R}^p$  containing  $k$  non-zero elements. Note  $S^* = \{j | \beta_j^* \neq 0\}$  the support of  $\beta^*$ .

**Model:**  $y = X\beta^* + \epsilon$

### Objective

Find the  $k$  relevant covariates.

## Background

Variable selection in the presence of correlated covariates.

### Lasso [5]:

- Performs variable selection
- Consistency selection under some assumptions
- Problems in presence of correlation

### Group-Lasso [6]:

- Performs group selection
- Requires one predefined partition of variables (the number of groups must be chosen before)

## Proposed Method

### 1) Hierarchical Clustering [3]

Perform hierarchical clustering with  $\frac{n}{2}$  samples.  
Result: set of partitions with different numbers of groups.

### 2) Overlap Group-Lasso [2]

Perform a Group-lasso with **the full set of partitions** and the  $\frac{n}{2}$  remaining samples.  
Contrary to the usual Group-Lasso, it enables to **select groups from different partitions**.

### 3) Choice of optimal groups

For a value of  $\lambda$ , some selected groups can be included in other selected ones. To select **non-overlapping** groups, apply a hierarchical testing procedure with Family-Wise Error Rate (FWER) control.

### 4) Choice of $\lambda$

Choose the  $\lambda$  value for which **the number of rejections is maximal**.

## Overlap Group-Lasso

Given  $G = \{\mathcal{G}_1, \dots, \mathcal{G}_S\}$  a set of different partitions of  $\{1, \dots, p\}$ ,  $\bar{X} = [X_{\mathcal{G}_1}, \dots, X_{\mathcal{G}_S}]$ .

$$\hat{\beta}_\lambda^G = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \|y - \bar{X}\beta\|_2^2 + \lambda \sum_{s=1}^S \rho_s \sum_{g \in \mathcal{G}_s} w_g \|\beta_g\|_2 \right\}$$

with  $\lambda \geq 0$  a regularization parameter,  $w_g$  the weight associated with  $g$  and  $\rho_s$  the weight associated with the quality of partition  $\mathcal{G}_s$ .

## Choice of parameter $\rho_s$

- $h_s$ : the criterion value at which 2 groups join
- $l_s = h_{s-1} - h_s$ : difference of criteria between two successive levels

**Highest jump rule:** A large value of  $l_s$  indicates the aggregation of distant groups to form the partition  $\mathcal{G}_{s-1}$ . In this case,  $\mathcal{G}_s$  is a better choice than  $\mathcal{G}_{s-1}$ .

### Choice of $\rho_s$

$$\rho_s = \frac{1}{\sqrt{l_s}}$$

## Hierarchical clustering

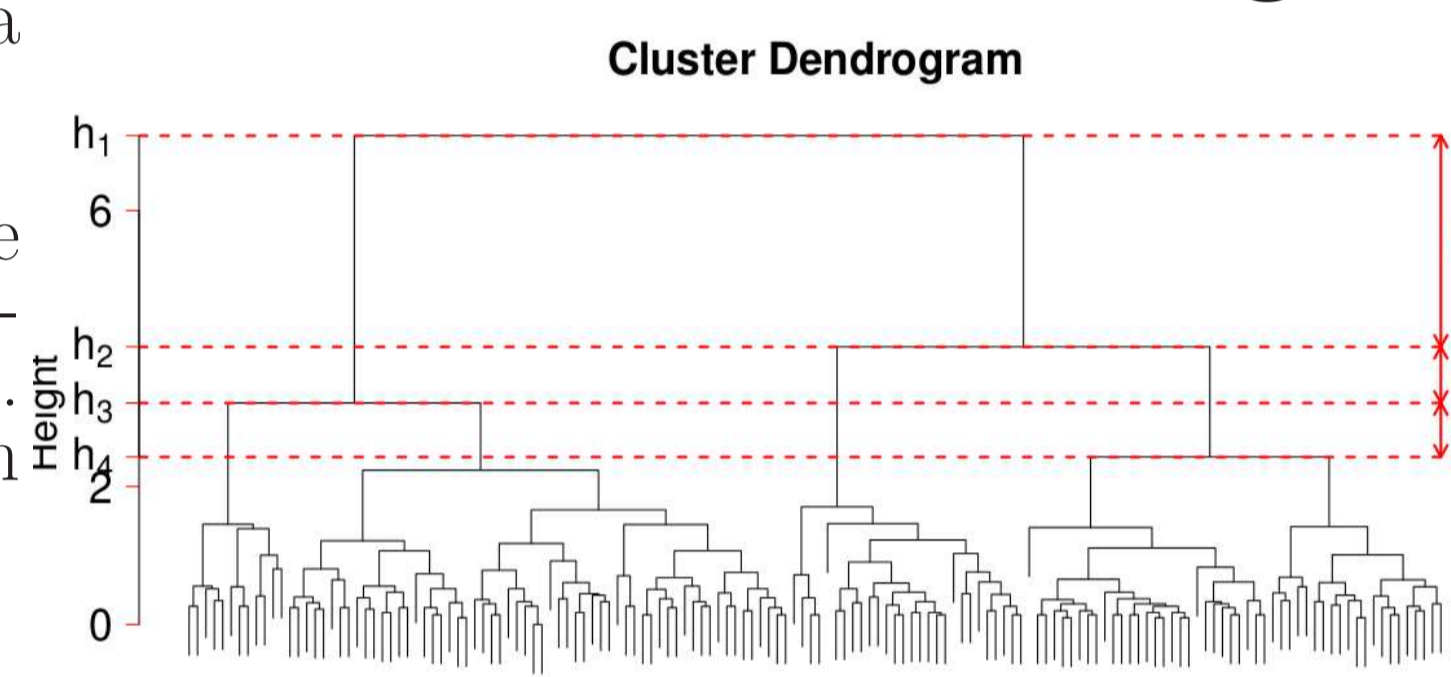


Figure 1: Dendrogram of Hierarchical Clustering performed on the iris data.

## Optimal groups

Let  $S_\lambda$  the selected groups for a specific value of  $\lambda$ . For each value of  $\lambda$ :

1. For each  $g \in S_\lambda$ , compute the principal component  $\bar{X}_g$  of  $X_g$  to summarize the groups.
2. For groups forming a hierarchy, apply a hierarchical testing procedure [4] for controlling FWER.
3. For other groups, apply the Bonferroni multiple testing procedure.
4. Keep groups with an adjusted p-value under a chosen threshold.

## Application (I)

### Design

- $X_1, \dots, X_n \sim \mathcal{N}(0, \Sigma_\rho)$
- $\Sigma_\rho$  covariance matrix with blocks structure

### Results

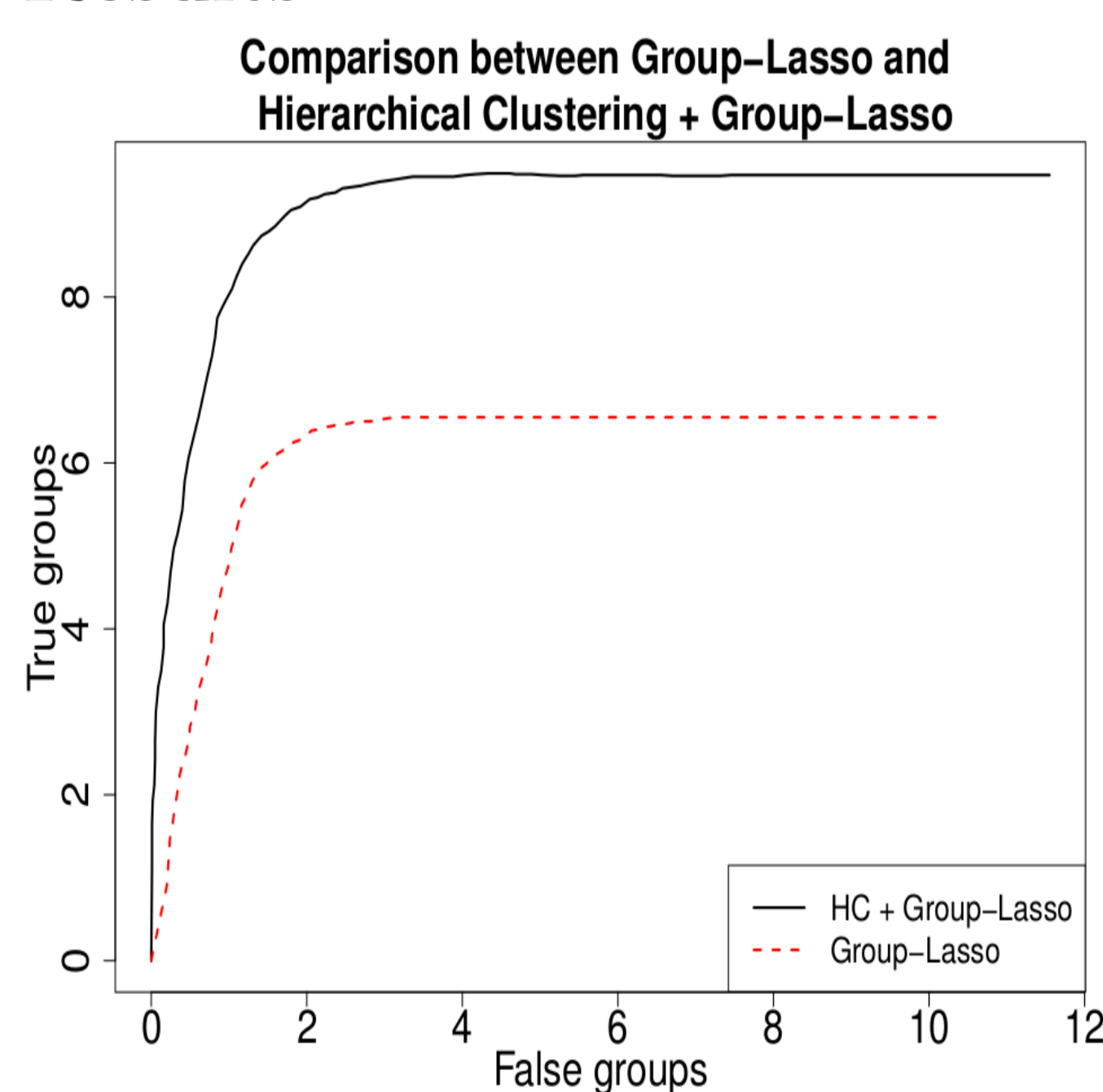


Figure 2: Number of selected true groups with regards to the number of false groups selected. In black, the proposed method (hierarchical clustering + group-lasso), in red, group-lasso on the best partition. The curve is the average of 100 trials.

Our method provides a **better solution path** than group-lasso on the best partition. For an equivalent number of true groups selected, our method selects less false groups.

- $\beta^*$  containing  $K$  non-zero elements, each in a different block
- $y = X\beta^* + \epsilon$  with  $\epsilon$  a gaussian noise

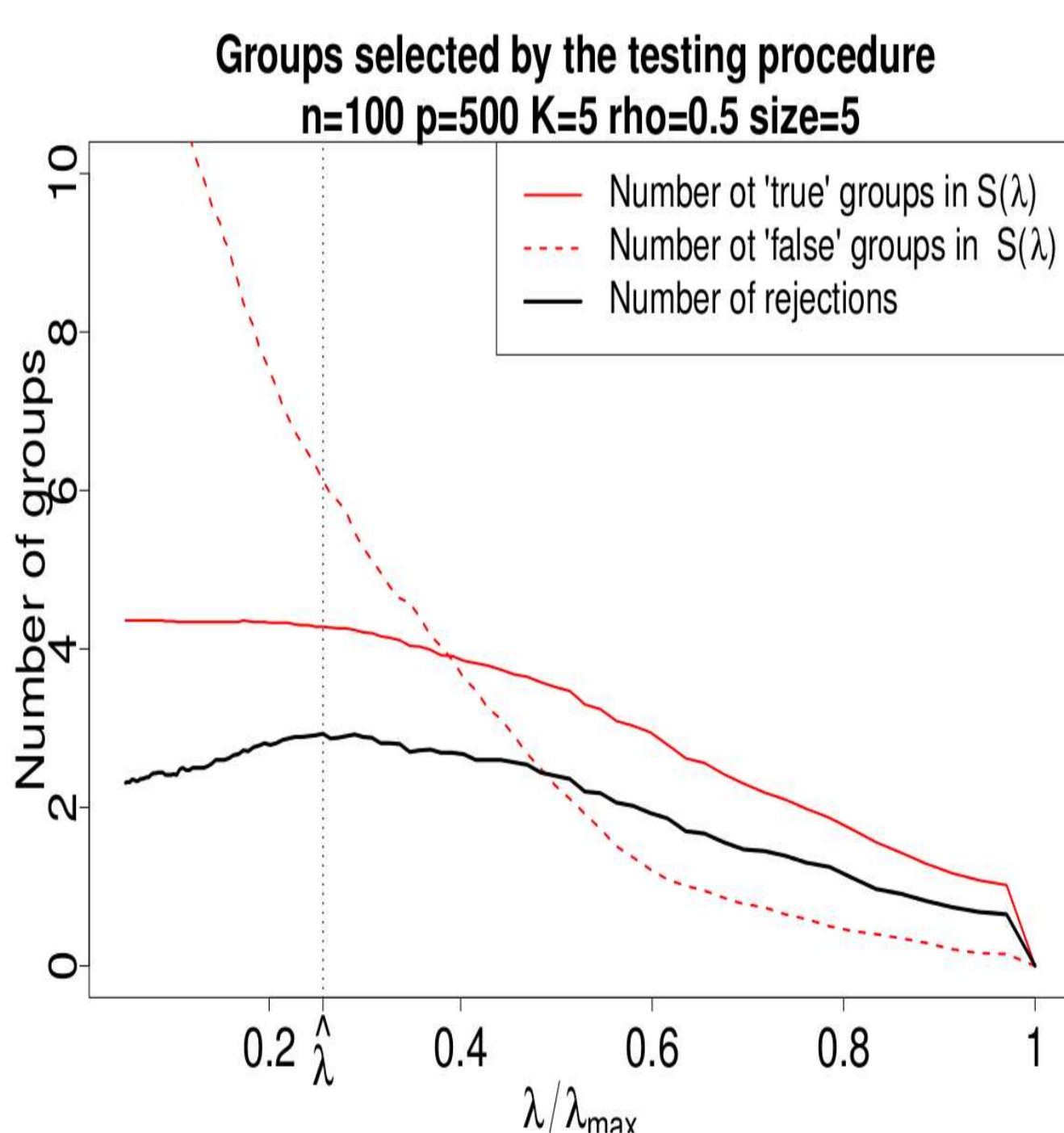


Figure 3: Selected groups after the testing procedure. The dotted vertical line corresponds to the  $\lambda$  value with the maximal number of rejections.

**Selected groups for  $\lambda = \hat{\lambda}$ :**

	True groups	False groups
Before testing	4.28	6.12
After testing	2.84	0.09

## Application (II)

### Data & design [1]

Data set about riboflavin (vitamin B2) production by bacillus subtilis. The covariates are measurements of the logarithmic expression level of  $p = 1000$  genes for  $n = 71$  samples.

Response variable is generated as follows:  $y = X\beta^* + \epsilon$ . The support  $S^*$  is chosen as a randomly selected variable  $i$  and the 9 covariates with the highest absolute correlation to this variable.

### Results

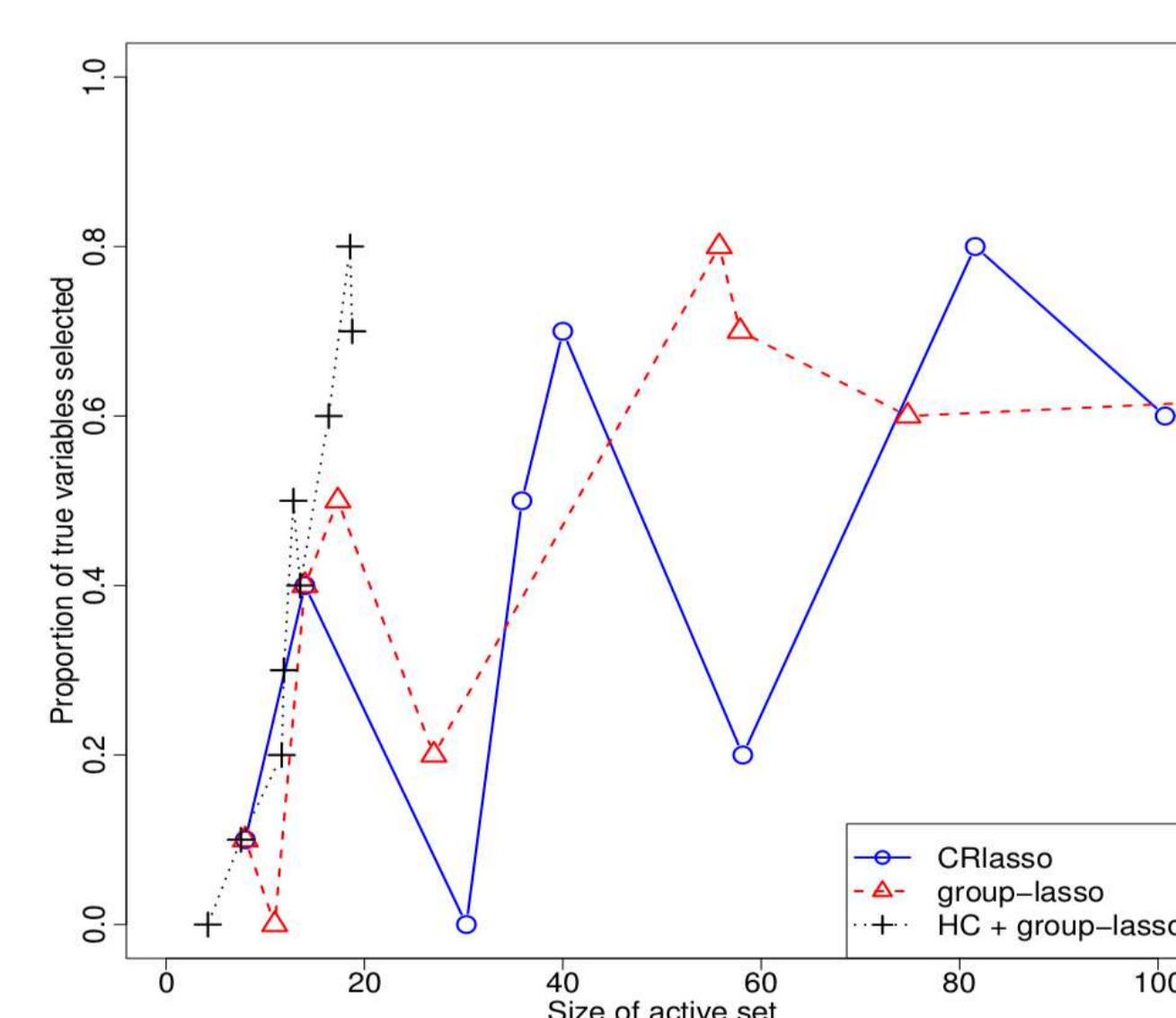


Figure 4: Plot of the frequency of true variables selected versus the size of the selected active set  $S$ . In black, the proposed method (HC + group-lasso), in red, group-lasso on the best partition, in blue the Cluster Representative Lasso. The curve is the average of 100 trials.

For group-lasso and Cluster Representative Lasso [1], the partition used is the one selected by the highest jump rule. A correlation-based distance is used for generating the dendrogram. It results in a partition with a few groups and a very large cluster.

Methods used in [1] with this partition achieve a good rate of true variable selection with a larger number of variables.

The possibility to use different partitions enables our method to select more efficiently.

## Conclusion & Perspectives

- New method combining Group-lasso and Hierarchical Clustering
- Choice of optimal  $\lambda$  and optimal groups
- Kernel Hierarchical Clustering
- Supervised clustering methods

## Bibliography

- [1] Bühlmann, P. et al. (2013). *Correlated variables in regression: clustering and sparse estimation*. Journal of Statistical Planning and Inference 143, pp. 1835-3871.
- [2] Jacob, L. et al. (2009). "Group Lasso with Overlap and Graph Lasso". Proceedings of the 26th Annual International Conference on Machine Learning, ICML 09. ACM, pp. 433-440.
- [3] Jain, A. K. et al. (1999). *Data Clustering: A Review*. ACM Comput. Surv. 31.3, pp. 264-323.

- [4] Meinshausen, N. (2008). *Hierarchical testing of variable importance*. Biometrika 95.2, pp. 265-278.
- [5] Tibshirani, R. (1994). *Regression Shrinkage and Selection Via the Lasso*. Journal of the Royal Statistical Society, Series B 58, pp. 267-288.
- [6] Yuan, M. et al. (2006). *Model selection and estimation in regression with grouped variables*. Journal of the Royal Statistical Society, Series B 68, pp. 49-67.