Combining clustering of variables and random forests for high-dimensional supervised classification

M. Chavent[1,2] R. Genuer[1] J. Saracco[1,2]

[1] Bordeaux University, [2] INRIA Bordeaux Sud-Ouest

August 28

COMPSTAT 2012

Introduction

Main goal:

- dimension reduction for high-dimensional data
- supervised classification framework

Proposed methodology:

1. clustering of variables (using *ClustOfVar*)
2. selection of the most important synthetic variables (using *VSURF* a variable selection procedure based on random forests)

Main idea:

eliminate redundancy before applying a variable selection method

Framework

$\mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ i.i.d. random vectors with same distribution as (X, Y)

$X \in \mathbb{R}^p$: (quantitative) variables

$Y \in \{-1, +1\}$: (dichotomous) response

Typically, $n << p$

Examples: gene expression data, fMRI data...

ClustOfVar (Chavent et.al. 2012)

Generality:

- lumps together strongly related variables into clusters of variables
- computes synthetic variables associated to each cluster
- useful for case studies and dimension reduction

Tools:

- homogeneity criterion based on squared correlations (hence, synthetic variable = 1st principal component of PCA applied to the cluster variables)
- hierarchical clustering algorithm (Ward like)
- k-means type partitioning algorithm

VSURF (Genuer et.al. 2010)

VSURF = Variable Selection Using Random Forests

Based on Random Forests (Breiman 2001)

- nonparametric statistical learning method
- agregation of a collection of classification trees
- trees constructed on bootstrap samples with randomly drawn variables

Interesting outputs of RF

- Out-Of-Bag (OOB) error: prediction error estimation
- Variable Importance (VI) score: helps to determine which variables explain the most the response

$\Rightarrow$ VSURF uses both OOB error and VI for variable selection

Toys data (Weston et.al. 2003)

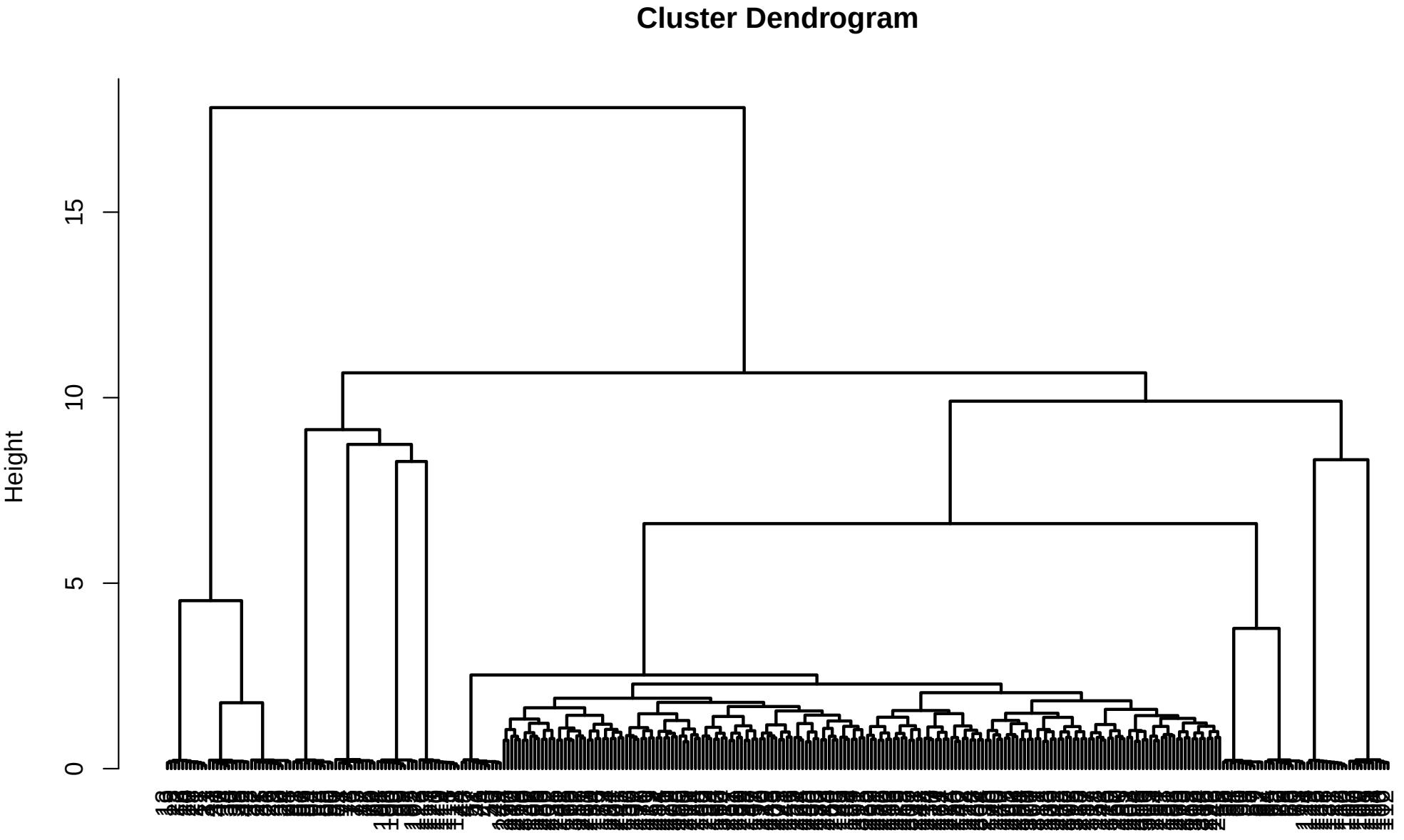
Original dataset: 6 true variables and 194 noise variables

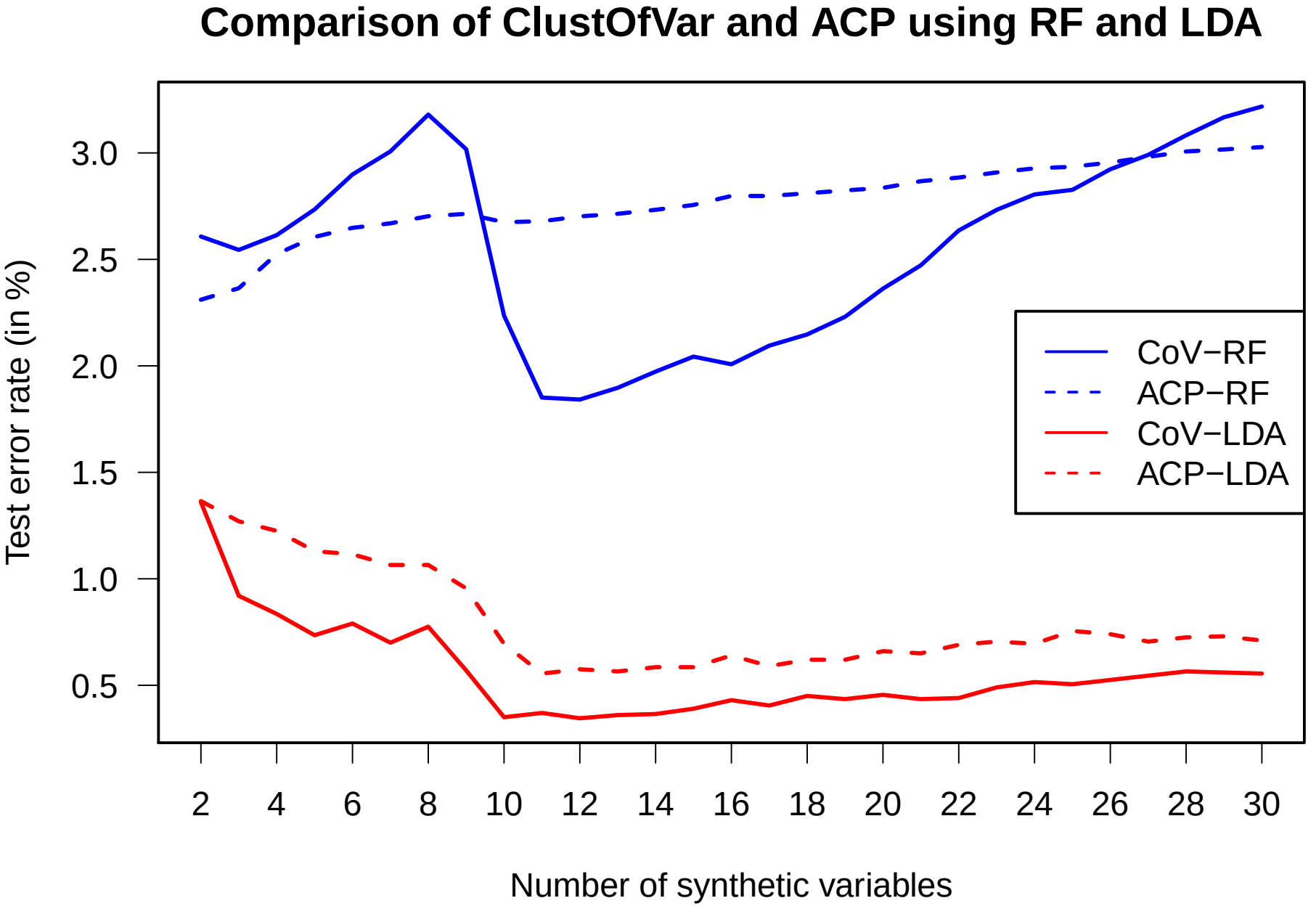
- two independent groups of 3 true variables, related with Y
- within each group, true variables are mediumly correlated with each other
- noise variables, independent with Y

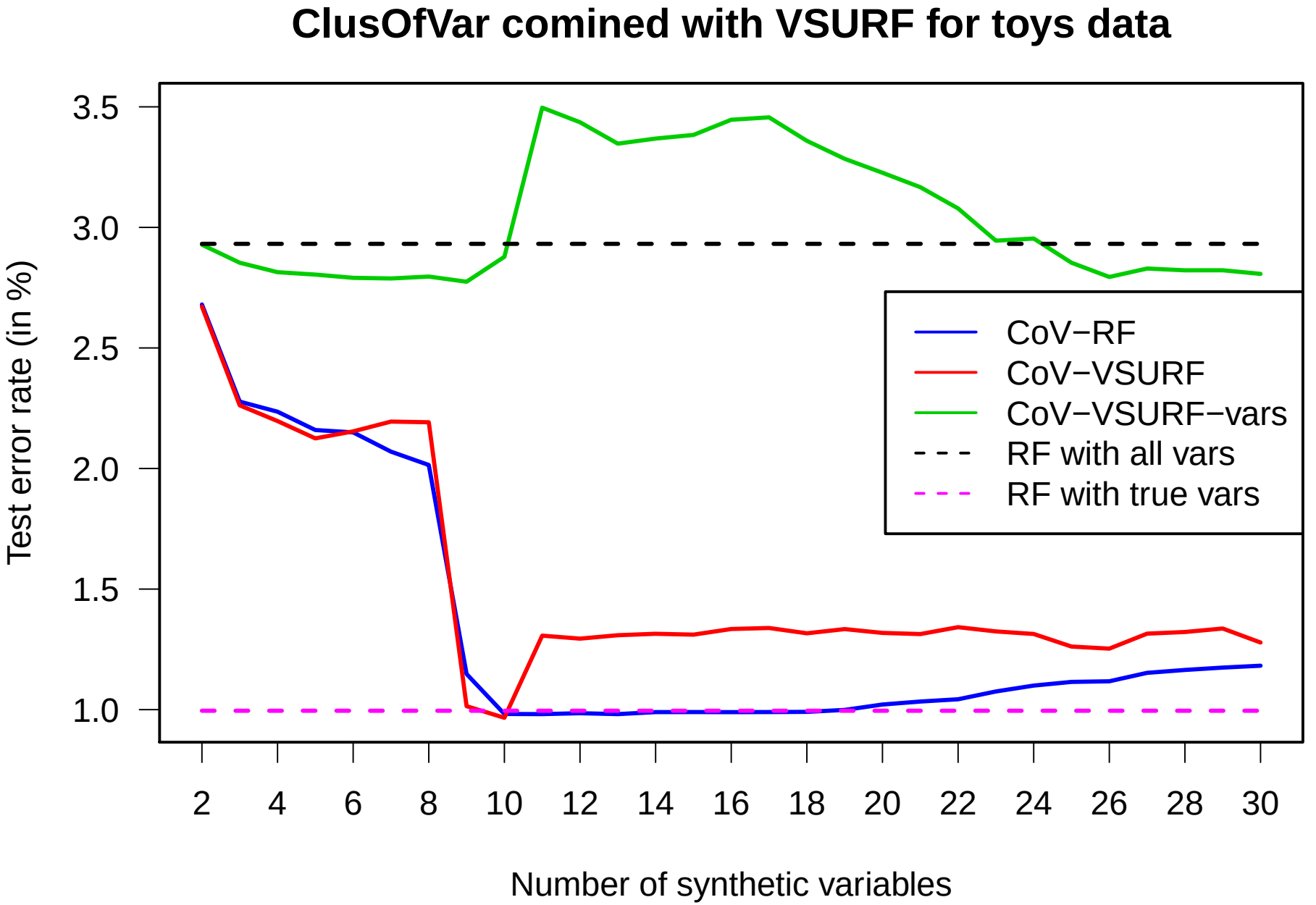
Modified dataset:
for each 6 true variables and 6 noise variables: addition of 10 highly correlated variables

$\Rightarrow$ 12 groups of 11 highly correlated variables, and 188 noise variables

$$n = 200 \quad p = 320$$


Cluster Dendrogram
Height
0
5
10
15


Comparison of ClustOfVar and ACP using RF and LDA
Test error rate (in %)
3.0
2.5
2.0
1.5
1.0
0.5
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
Number of synthetic variables
CoV−RF
ACP−RF
CoV−LDA
ACP−LDA


ClusOfVar comined with VSURF for toys data
Test error rate (in %)
3.5
3.0
2.5
2.0
1.5
1.0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
Number of synthetic variables
CoV−RF
CoV−VSURF
CoV−VSURF−vars
RF with all vars
RF with true vars

Prostate data (Stephenson et.al. 2005)

- Gene expression data in a cancer study
- 79 treated patients:
  - 37 recurrent primary prostate tumor
  - 42 non-recurrent
- 7884 gene expressions

$$n = 79 \quad p = 7884$$

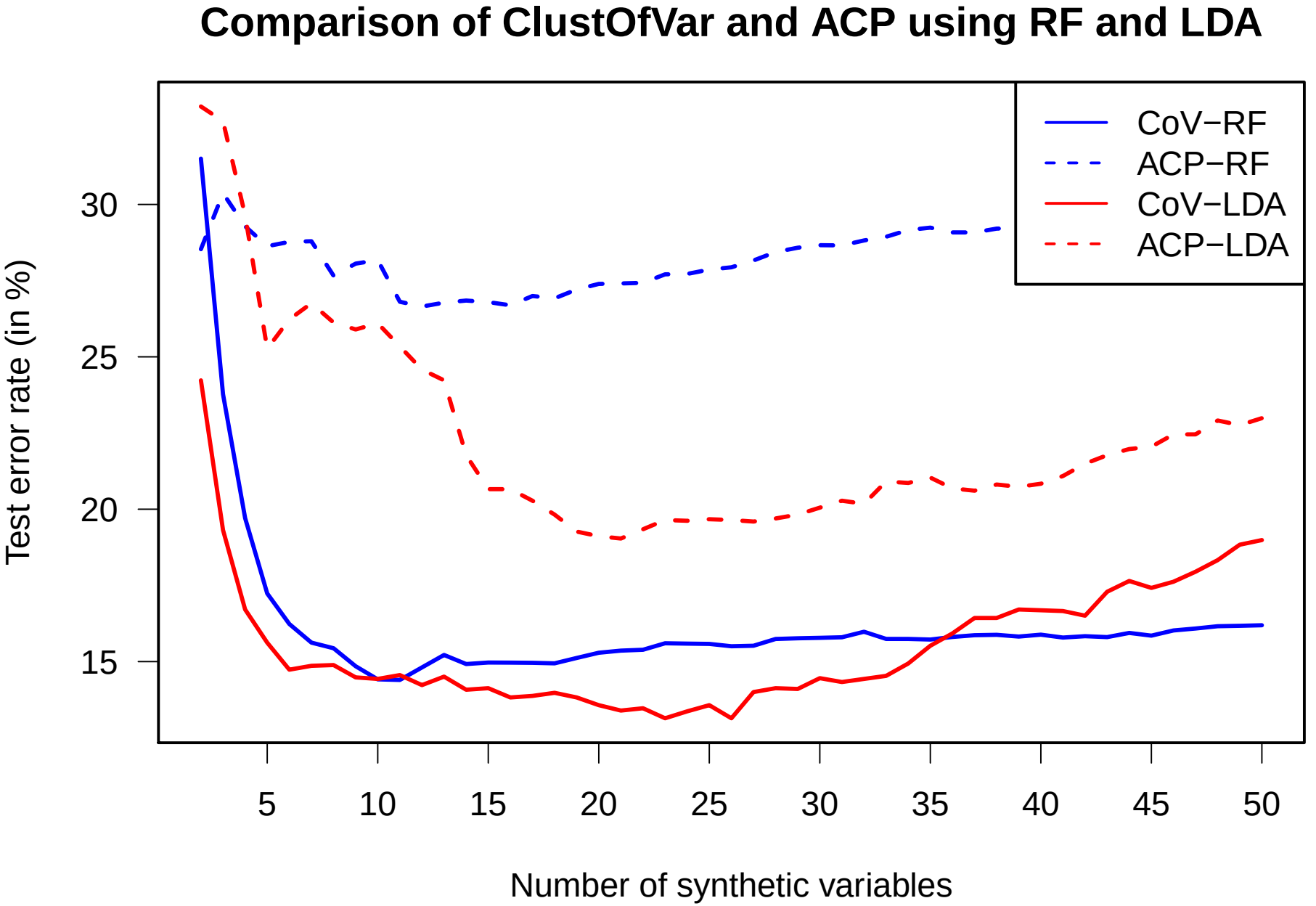
Prostate data (Stephenson et.al. 2005)

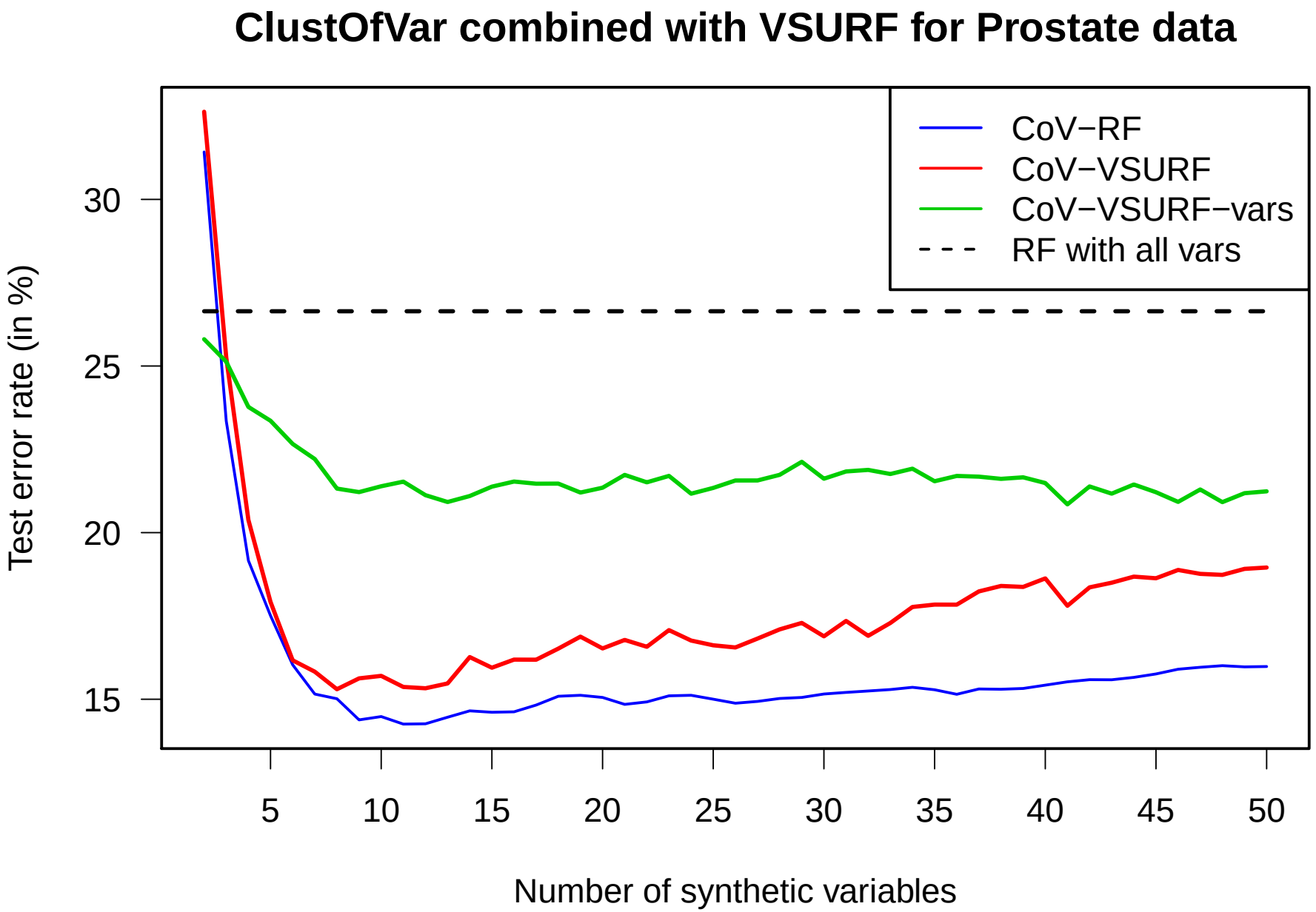
- Gene expression data in a cancer study
- 79 treated patients:
  - 37 recurrent primary prostate tumor
  - 42 non-recurrent
- 7884 gene expressions

$$n = 79 \quad p = 7884$$

Clustering of the 7884 variables:

1. kmeansvar with 120 clusters of variables
2. hclustvar on the 120 corresponding synthetic variables


Comparison of ClustOfVar and ACP using RF and LDA
CoV−RF
ACP−RF
CoV−LDA
ACP−LDA
Test error rate (in %)
30
25
20
15
5
10
15
20
25
30
35
40
45
50
Number of synthetic variables


ClustOfVar combined with VSURF for Prostate data
CoV−RF
CoV−VSURF
CoV−VSURF−vars
RF with all vars
Test error rate (in %)
30
25
20
15
5
10
15
20
25
30
35
40
45
50
Number of synthetic variables

Discussion

Results:

- Choice of the number of synthetic variables using OOB error rates $\Rightarrow$ 13 synthetic variables for Prostate
- VSURF then selects 4 synthetic variables corresponding to 516 original variables (gene expression)
  $\Rightarrow$ we discard 93.5% of the variables.

Remarks:

- Eliminate redundancy before variable selection seems to bring some benefits for prediction.
- R packages : `ClustOfVar` (available), `VSURF` (in construction)

Perspective:

- 516 variables may still be too large.
  $\Rightarrow$ Select variables within each selected cluster ?
  Select variables directly among the 516 variables ?

Références

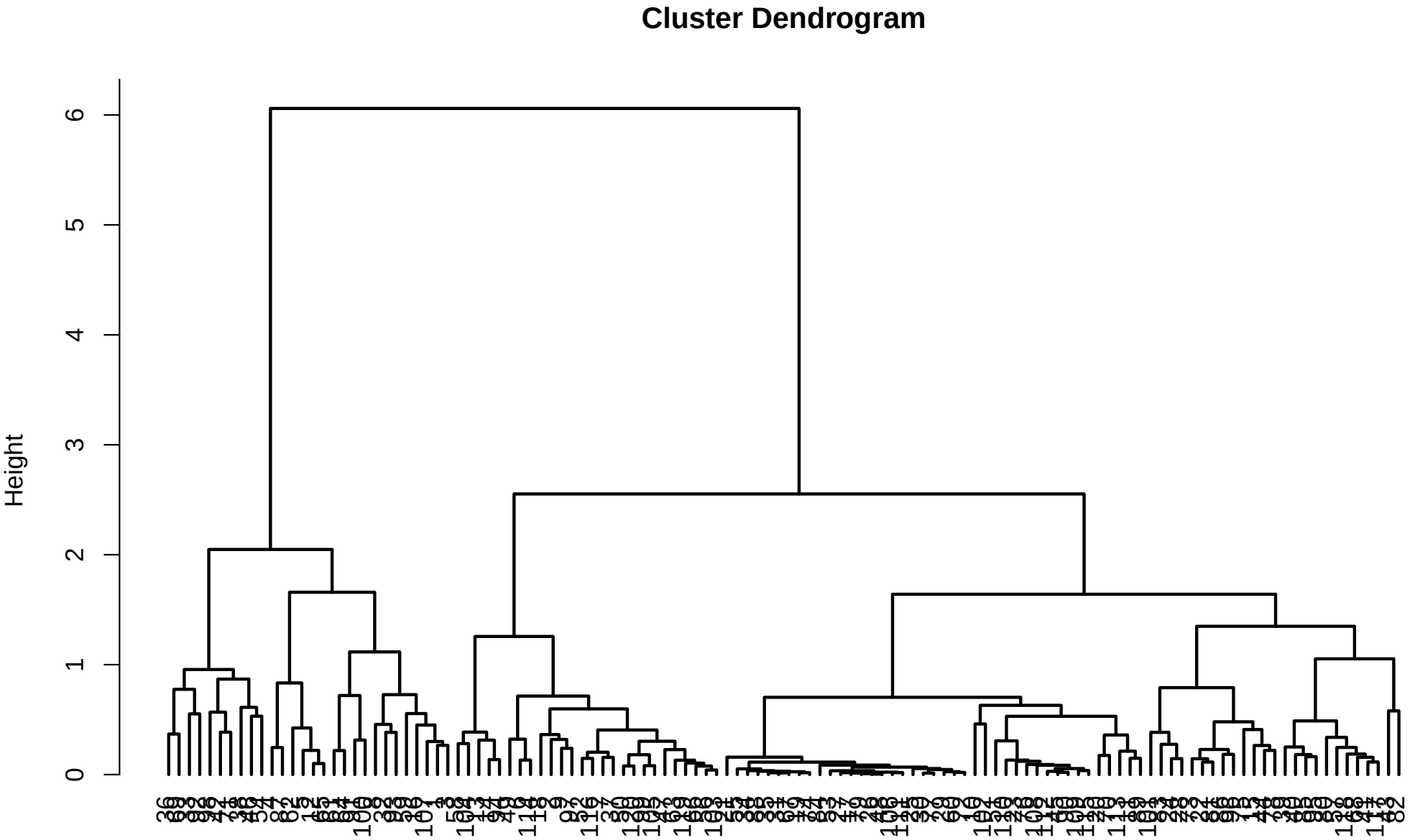

Breiman, L. *Random Forests*. Machine Learning (2001)

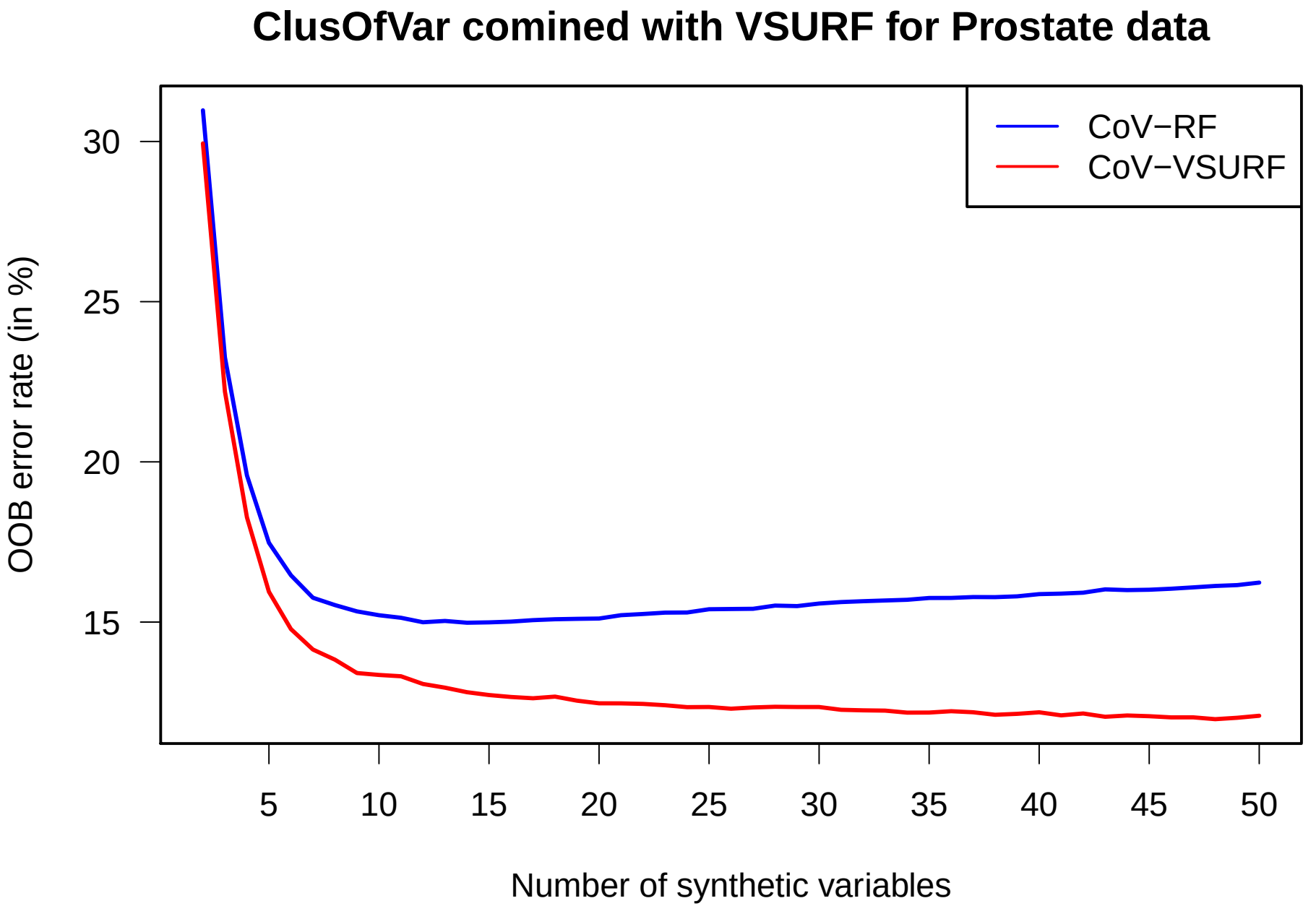

Chavent M., Kuentz-Simonet V., Liquet B. and Saracco J. *ClustOfVar: An R Package for the Clustering of Variables*. Journal of Statistical Software (to appear)

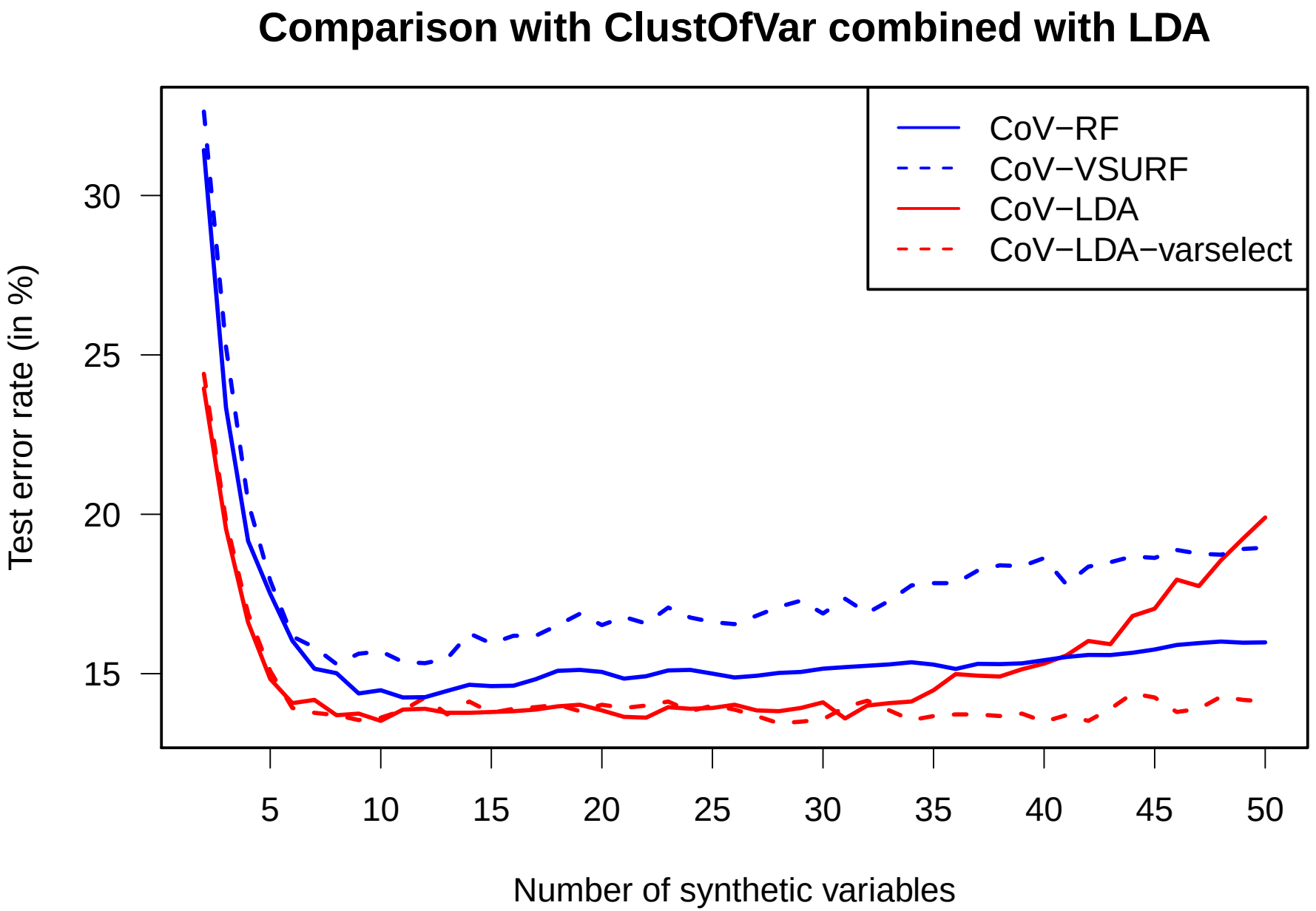

Genuer R., Poggi J.-M. and Tuleau C. *Variable selection using random forests*. Pattern Recognition Letters (2010)

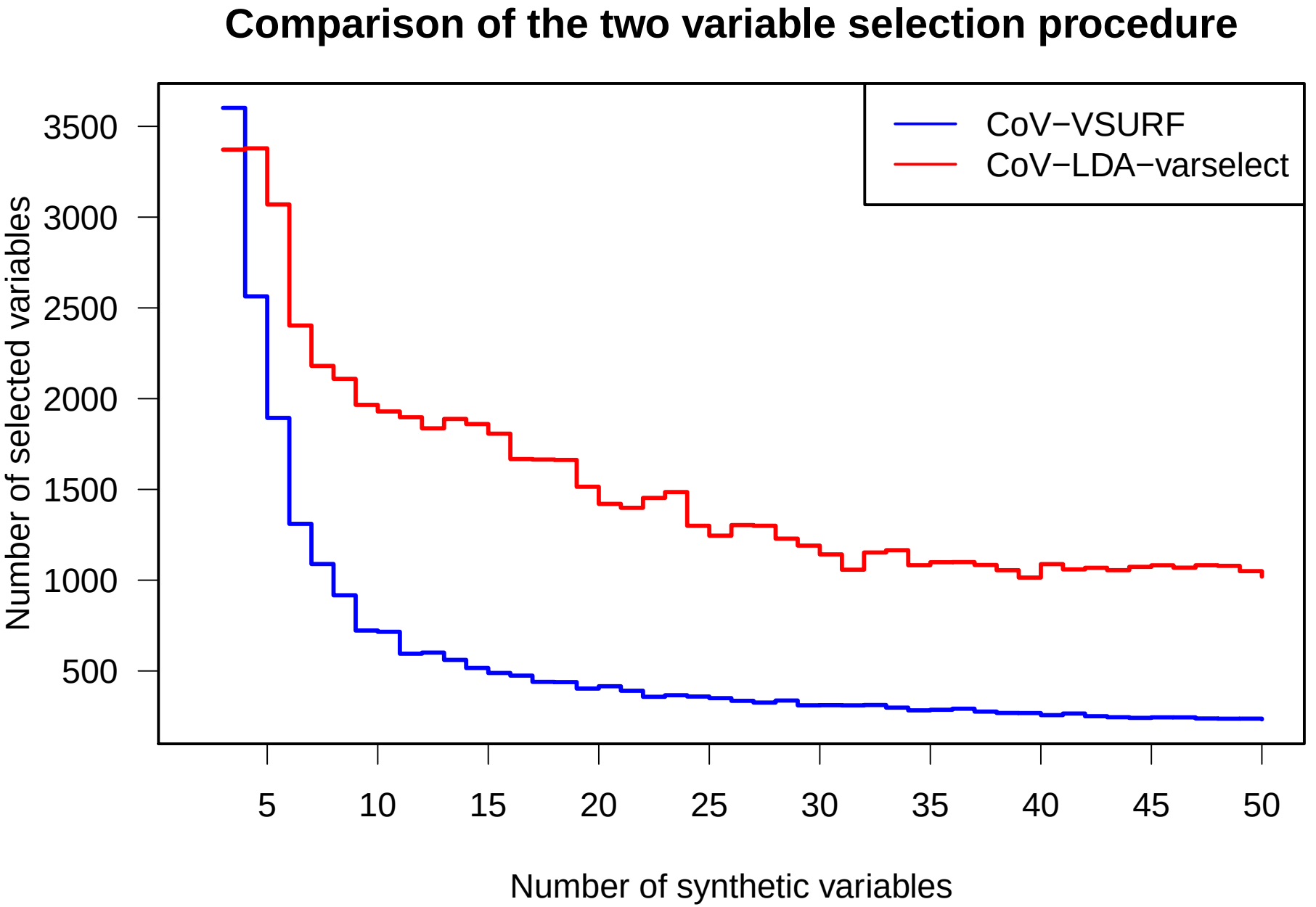

Stephenson,A. et al. *Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy*. Cancer (2005)


Weston, J., Elisseff, A., Schoelkopf, B., Tipping, M. *Use of the zero norm with linear models and kernel methods*. Journal of Machine Learning Research (2003)


Cluster Dendrogram
Height
0
1
2
3
4
5
6


ClusOfVar comined with VSURF for Prostate data
CoV−RF
CoV−VSURF
OOB error rate (in %)
30
25
20
15
5
10
15
20
25
30
35
40
45
50
Number of synthetic variables


Comparison with ClustOfVar combined with LDA
CoV−RF
CoV−VSURF
CoV−LDA
CoV−LDA−varselect
Test error rate (in %)
30
25
20
15
5
10
15
20
25
30
35
40
45
50
Number of synthetic variables


Comparison of the two variable selection procedure
CoV−VSURF
CoV−LDA−varselect
Number of selected variables
3500
3000
2500
2000
1500
1000
500
5
10
15
20
25
30
35
40
45
50
Number of synthetic variables

ClustOfVar (Chavent et.al. 2012)

Homogeneity measure $\mathcal{H}$ of a partition $P = (C_1, \ldots, C_K)$:

$$\mathcal{H}(P) = \sum_{k=1}^{K} H(C_k)$$

with

$$H(C_k) = \sum_{X^j \in C_k} r^2(X^j, \mathbf{c}_k)$$

$\mathbf{c}_k$ is a synthetic variable

r^2 is squared correlation

ClustOfVar (Chavent et.al. 2012)

Homogeneity measure $\mathcal{H}$ of a partition $P = (C_1, \ldots, C_K)$:

$$\mathcal{H}(P) = \sum_{k=1}^{K} H(C_k)$$

with

$$H(C_k) = \sum_{X^j \in C_k} r^2(X^j, \mathbf{c}_k)$$

$\mathbf{c}_k$ is a synthetic variable

r^2 is squared correlation

Synthetic variable of a cluster:

$$\mathbf{c}_k = \arg\max_{\mathbf{u} \in \mathbb{R}^n} \left\{ \sum_{X^j \in C_k} r^2(X^j, \mathbf{u}) \right\}$$

$\Rightarrow \mathbf{c}_k =$ 1st principal component of PCA applied to the cluster

Classification tree

Tree: piecewise constant predictor obtained by dyadic recursive partitioning of $\mathbb{R}^p$

At each step of the partitioning process, we seek for the "best" split of the data from $\mathcal{L}_n$

Example : CART, Breiman et.al. (1984).

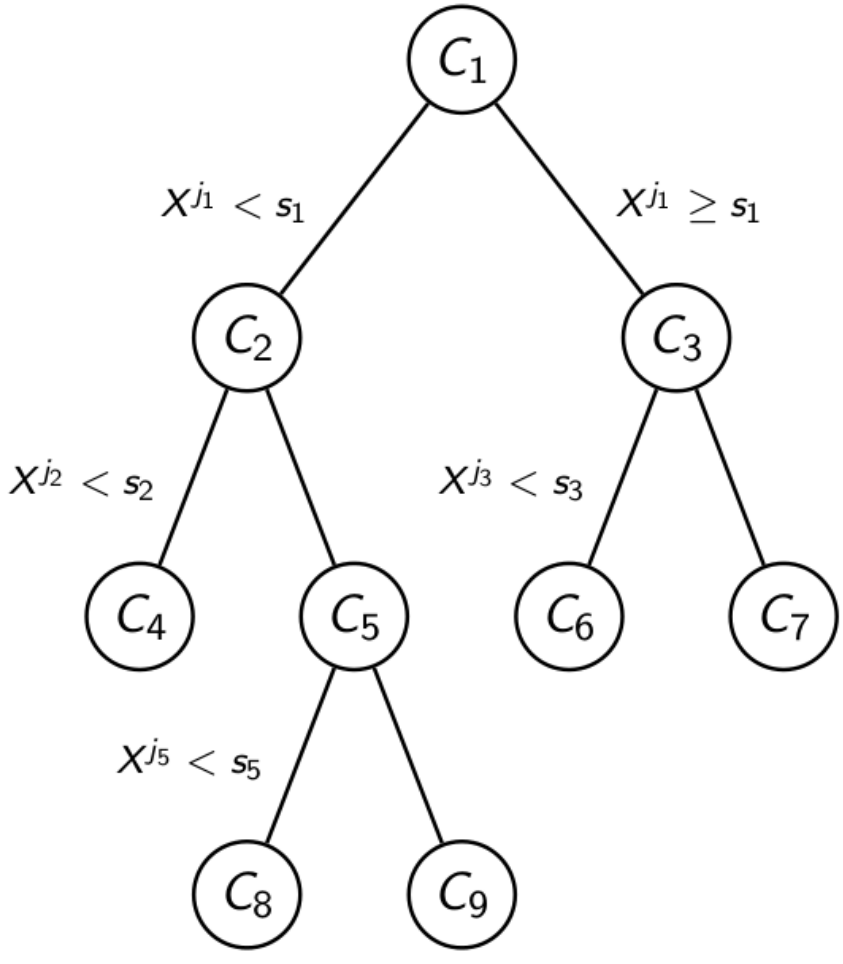


Figure: Classification tree

Random Forests