Variable selection with Random Forests

Robin Genuer,
based on joint work with
Jean-Michel Poggi and Christine Tuleau-Malot

December 8, 2016
Master MAS-MSS, High dimensional data analysis

Variable selection

We distinguish two different objectives :

1. to select all important variables, even with high redundancy, for interpretation purpose
2. to find a sufficient parsimonious set of important variables for prediction

Our aim is to build an automatic procedure, which fulfills these two objectives

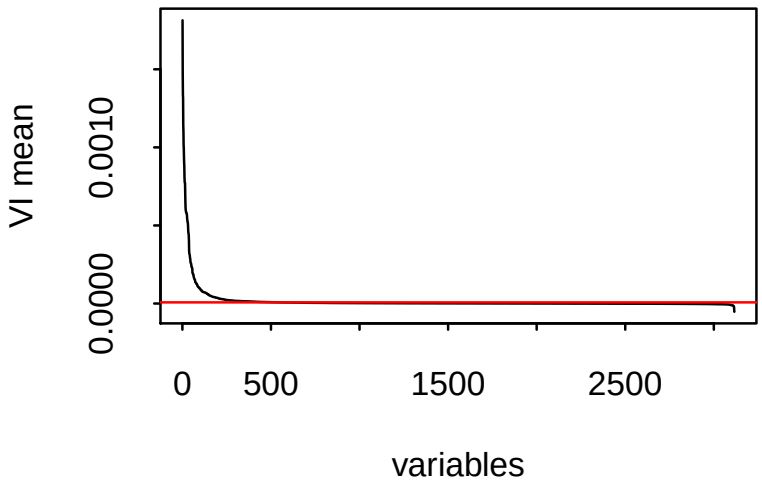
We use the `VSURF` package (Variable Selection Using Random Forests).

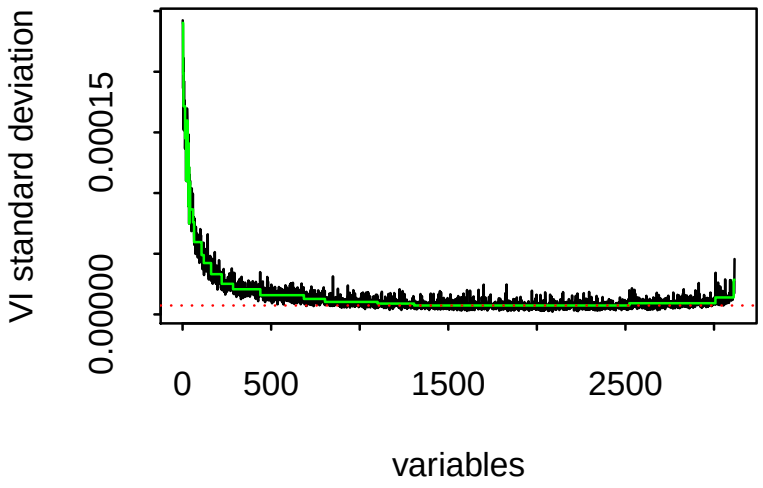
```r
library(VSURF)
library(mixOmics)
data(liver.toxicity)
x <- liver.toxicity$gene
y <- liver.toxicity$clinic$ALB.g.dL.
n <- nrow(x)
p <- ncol(x)
toxi_VSURF <- VSURF(x, y, parallel = TRUE, ncores = 40,
    clusterType = "FORK")
```

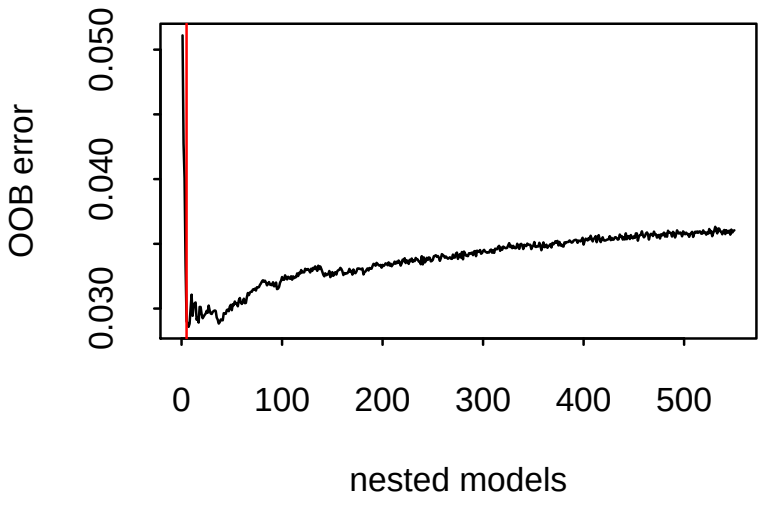
```r
summary(toxi_VSURF)
```

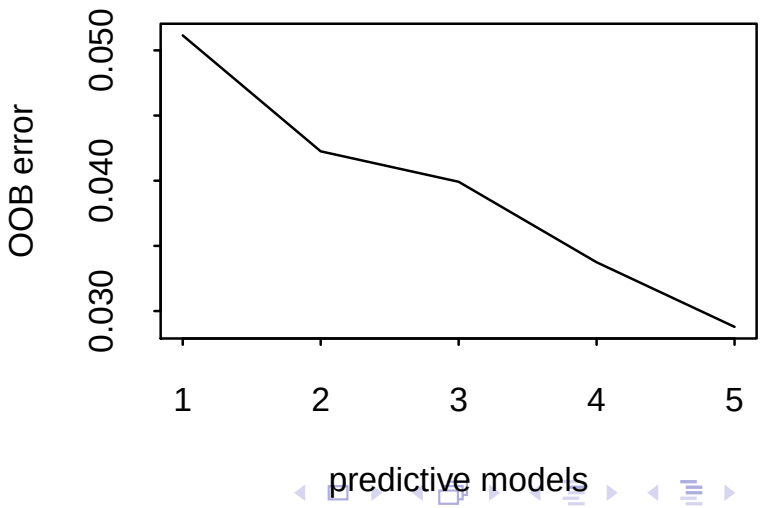
```
##
##  VSURF computation time: 5.9 mins
##
##  VSURF selected:
##  550 variables at thresholding step (in 1.1 mins)
##  5 variables at interpretation step (in 4.8 mins)
##  5 variables at prediction step (in 3.4 secs)
##
##  VSURF ran in parallel on a FORK cluster and used 40 cores
```

```
plot(toxi_VSURF)
```







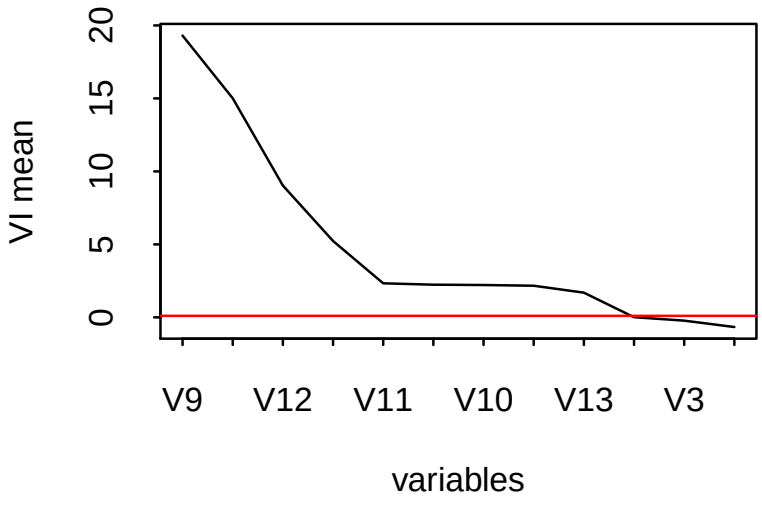


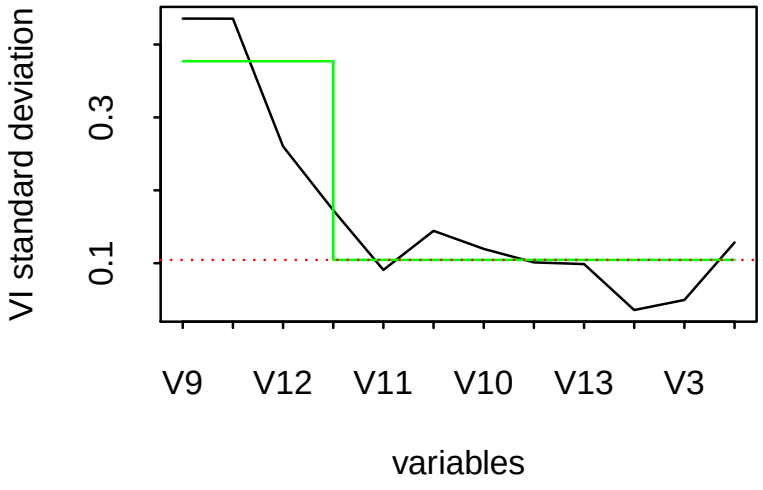
```
library(VSURF)
library(mlbench)
data(Ozone)
```

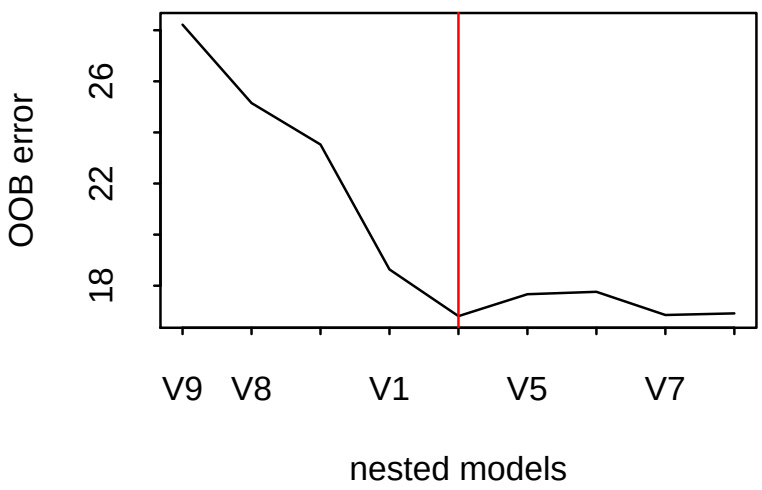
```
ozone_VSURF <- VSURF(V4 ~ ., data = Ozone, na.action = na.omit)
```

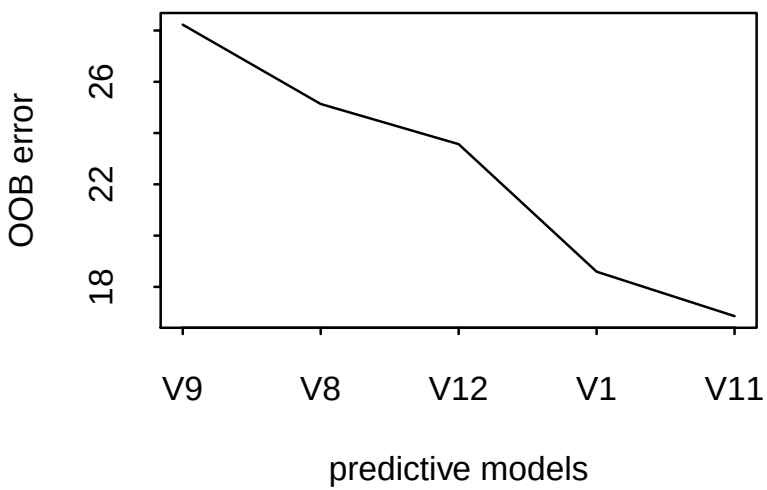
```
summary(ozone_VSURF)

##
##  VSURF computation time: 1.7 mins
##
##  VSURF selected:
##  9 variables at thresholding step (in 1.1 mins)
##  5 variables at interpretation step (in 26.3 secs)
##  5 variables at prediction step (in 12.4 secs)
```

```
plot(ozone_VSURF, var.names = TRUE)
```









Références

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

Díaz-Uriarte R., Alvarez de Andrés S. *Gene Selection and classification of microarray data using random forest*. BMC Bioinformatics (2006)

Genuer R., Poggi J.-M. and Tuleau-Malot C. *VSURF : An R Package for Variable Selection Using Random Forests*. The R Journal (2015)