MultiVarSel: Package for variable selection in high-dimensional general linear models

M. Perrot-Dockès, C. Lévy-Leduc and J. Chiquet

April 4, 2017

This vignette explains how to use the package MultiVarSel which is dedicated to the variable selection in high-dimensional general linear models by taking into account the dependence that may exist between the columns of the observations matrix. For further details on the methodology we refer the reader to [1].

After having installed the package in R, the package has to be loaded by using the following instruction:

```
> library(MultiVarSel)
```

In the following, we shall explain how to analyze the `copals_camera` dataset provided within the package.

To load this data set, type

```
> data("copals_camera")
> dim(copals_camera)
```

```
[1]   62 1022
```

```
> #### We limit ourselves to the following data
> copals = copals_camera[copals_camera$Include == 1, -1]
```

We extract the data matrices

```
> Y  <- as.matrix(copals[, -(1:2)])
> X1 <- copals[,   1]
> X2 <- copals[,   2]
```

We remove individuals with class 1155 and 1551 which are isolated

```
> rm <- which(X1 %in% c("1155","1551"))
> Y <- Y[-rm, ]
> X1 <- X1[-rm]; X1 <- factor(as.character(X1))
> X2 <- X2[-rm]; X2 <- factor(as.character(X2))
```

According to the following table, the problem is in fact a simple one-way MANOVA

```
> table(X1,X2)
```

```
      X2
X1     Class 0 East West
  1960      13    0    0
  1967       0    9    8
```

```
> ## -> X1 is useless => We have a one-way MANOVA model with 3 levels
```

We build the design matrix

```
> X <- model.matrix(lm(Y ~ X2 + 0))
> p <- ncol(X)
> n=nrow(X)
> n
```

```
[1] 30
```

```
> q=dim(Y)[2]
> q
```

```
[1] 1019
```

We scale the $\boldsymbol{Y}$ matrix

```
> Yscaled=scale(Y)
> Y=Yscaled
```

In the following, in order to speed up the computations, we only focus on the first 200 columns of $\boldsymbol{Y}$

```
> Y=Y[,1:200]
```

The residuals are defined as follows:

```
> residuals=lm(as.matrix(Y)~X-1)$residuals
```

We apply the whitening test to this residuals matrix in order to know if it is useful to whiten the observations or not

```
> pvalue=whitening_test(residuals)
> pvalue
```

```
[1] 5.676735e-229
```

Whitening is useful since the p-value is smaller than 0.05.

In order to select the type of dependence that is the most adapted to the data we apply the `whitening_choice` function

```
> result=whitening_choice(residuals,c("AR1","nonparam","ARMA"),pAR=1,qMA=1)
> result
```

```
          Pvalue       Decision
AR1            0 NO WHITE NOISE
nonparam   0.992    WHITE NOISE
ARMA 1 1   0.653    WHITE NOISE
```

The non parametric choice has the highest p-value. We select this dependence to model the residuals.

We compute the square root of the inverse of the estimator of the covariance matrix of each row of the residuals matrix using the non parametric modelling as follows:

```
>  square_root_inv_hat_Sigma=whitening(residuals,"nonparam",pAR=1,qMA=0)
```

We then applied the variable selection technique. Here, in order to provide an example having a low computational burden, we only applied the stability selection with 100 replications. We suggest to the reader to take at least 500 replicates to have a robust result.

```
>   Frequencies=variable_selection(Y,X,square_root_inv_hat_Sigma,
+                                  nb_repli=100,parallel=FALSE)
```

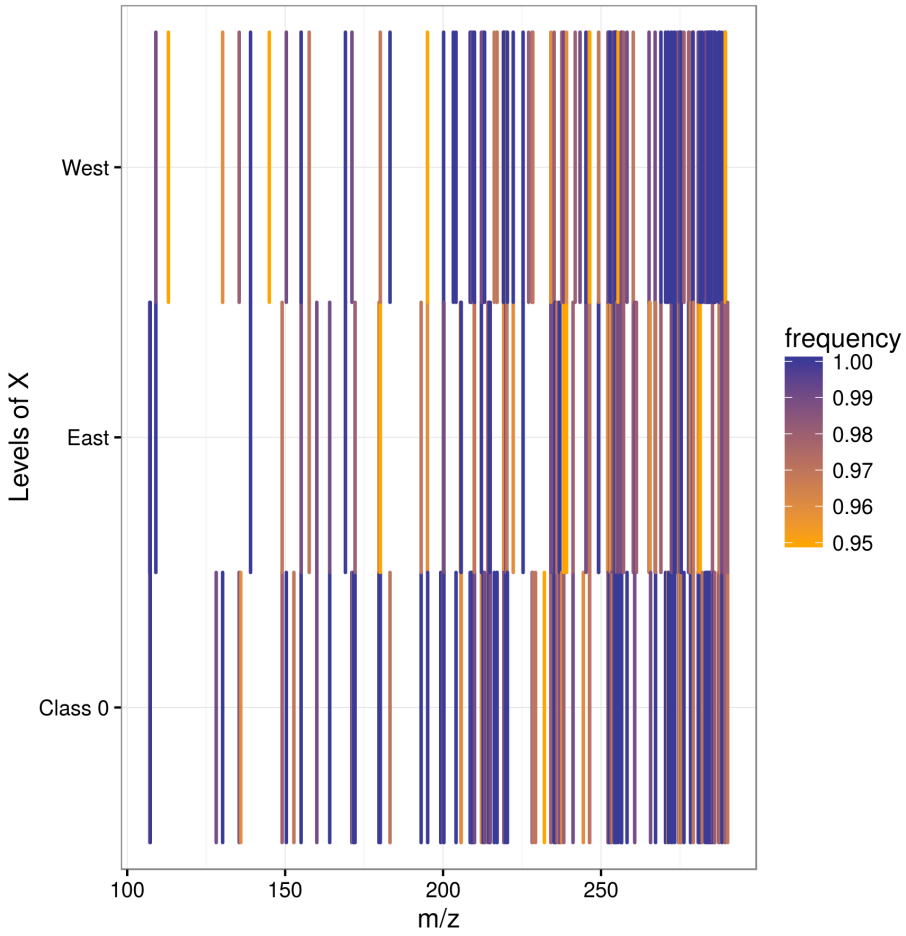
Parallel computing is also supported by the function `variable_selection`. To make it work, users must download the package doMC which is not available on Windows platforms (it is on others).

```
> require(doMC)
> registerDoMC(cores=4)
> Freqs=variable_selection(Y,X,square_root_inv_hat_Sigma,
+                      nb_repli=10,parallel=TRUE,nb.cores=4)
```

This function provides the selection frequencies of the variables for the different levels of the qualitative variable.

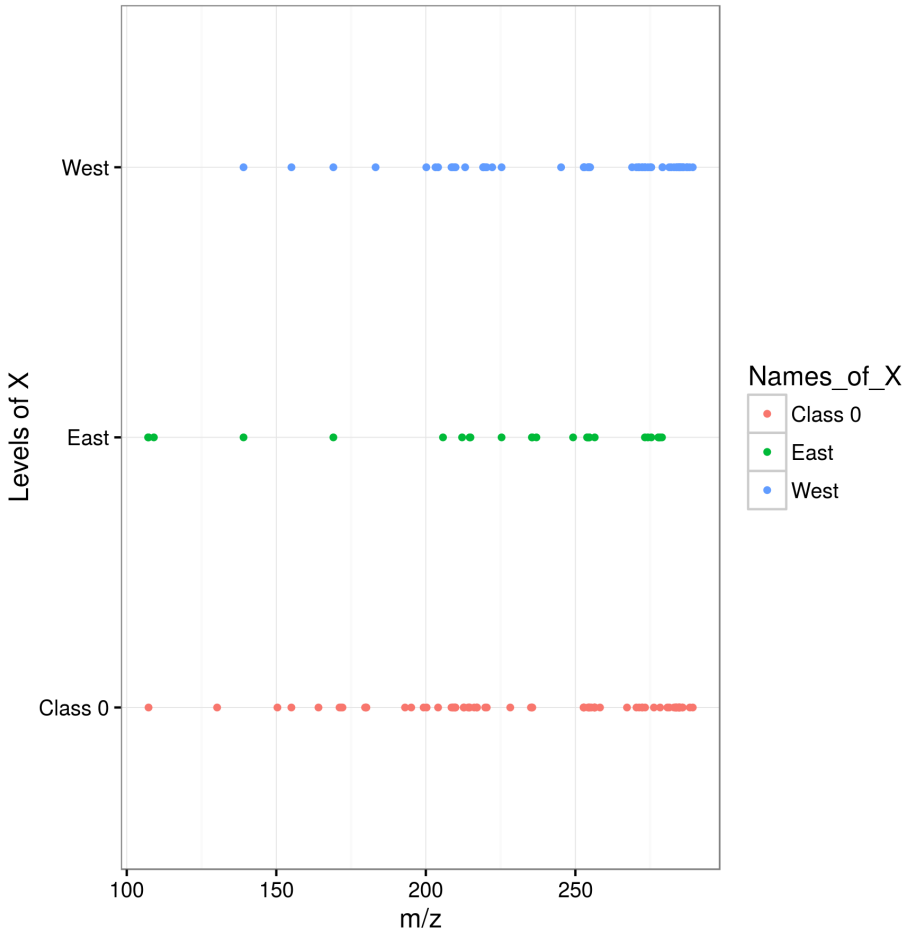
To display the positions of the metabolites that are selected with a frequency larger than 95%, the following code can be used.

```
> colnames(Frequencies)<-c('Names_of_Y','Names_of_X','frequency')
> # Here we can consider the names of Y as numerical since they correspond
> # to the ratio m/z of the metabolites.
> Frequencies$Names_of_X<-sub('X2','',Frequencies$Names_of_X)
> Frequencies$Names_of_Y<-as.numeric(gsub('X','',gsub('\\.1$','',Frequencies$Names_of_Y)))
> p<-ggplot(data=Frequencies[Frequencies$frequency>=0.95,],
+           aes(x=Names_of_Y,y=Names_of_X,color=frequency))+
+           geom_tile(size=0.75)+scale_color_gradient2(midpoint=0.95,mid ='orange')+
+           theme_bw()+ylab('Levels of X')+xlab('m/z')
> p
```



To avoid false postive we only consider the variables that are always selected (with a frequency equal to one)

```
> p<-ggplot(data=Frequencies[Frequencies$frequency==1,],
+           aes(x=Names_of_Y,y=Names_of_X,color=Names_of_X))+
+           geom_point(size=1)+theme_bw()+ylab('Levels of X')+xlab('m/z')
> p
```



Since the number of replications that we have chosen here is very low, the result is not very sparse. For results with larger number of replications we refer the reader to [1].

Hereafter, we also provide some information about the R session

```
> sessionInfo()

R version 3.3.3 (2017-03-06)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.5 LTS

locale:
 [1] LC_CTYPE=fr_FR.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=fr_FR.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=fr_FR.UTF-8    LC_MESSAGES=fr_FR.UTF-8
 [7] LC_PAPER=fr_FR.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=fr_FR.UTF-8 LC_IDENTIFICATION=C
```

```
attached base packages:
[1] stats     graphics  grDevices utils     datasets
[6] methods   base

other attached packages:
[1] MultiVarSel_1.0 ggplot2_2.1.0   glmnet_2.0-5
[4] foreach_1.4.3   Matrix_1.2-7.1

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.7      lattice_0.20-35  codetools_0.2-15
 [4] digest_0.6.10    grid_3.3.3       plyr_1.8.4
 [7] gtable_0.2.0     scales_0.4.0     labeling_0.3
[10] iterators_1.0.8  tools_3.3.3      munsell_0.4.3
[13] parallel_3.3.3   colorspace_1.2-7
```

References

[1] M. Perrot-Dockes et al. "A multivariate variable selection approach for analyzing LC-MS metabolomics data", arXiv:1704.00076