PLSvar_sel - Tutorial

This tutorial shows how to run the variable selection procedure and how to plot the selected variables as shown in the paper. Variable selection is based on bootstrapped-VIP scores calculated from different PLS models.

Main functions: [X_sel]=PLSvar_sel(X,Y1,Y2,MaxFac,Type,btnr,btmd)
PLOTvar_sel(X_two_way,Y1_two_way,Ytime,nr)

PLSvar_sel provides five PLS models with different combinations of bilinear/trilinear X and group/time response dummy Y. Based on the type of model chosen, the X should be construct into two way or three way accordingly. PLOTvar_sel plots temporal profiles for selected variables.

Variable Selection Function

[X_sel]=PLSvar_sel(X,Y1,Y2,MaxFac,Type,btnr,btmd)

INPUT:

X Array of independent variables, which is a two way or three way matrix.
Mode 1, Subject, S = number of subjects;
Mode 2, Metabolite, J= number of metabolites;
Mode 3, Time, T = number of time points.

Y1 Array of dependent variables representing group information.
E.g, Samples from intervention and control group are labelled with 1 and -1 respectively.

Y2 Array of dependent variables representing time response information.
E.g, Samples from response and non-response class are labelled with 10 and 1 respectively.

When the same dataset is constructed to adapt to different types of model, the sizes of X, Y1 and Y2 should be as follows:

Model Type	Size		
	X	Y1	Y2
1	ST×J	ST×1	ST×1
2	ST×J	ST×1	ST×1
3	ST×J	ST×1	ST×1
4	S×T×J	S×1	1×T
5	S×T×J	S×1	1×T

During the running of the function, X will be preprocessed (autoscaling) and Y1 and Y2 will be constructed into a new dummy Y for further modelling.

MaxFac Maximal number of components.
The optimal number of components will be decided based on a single cross validation.

Type Type of model to use
1 bilinear X and group dummy Y
2 bilinear X and time-response dummy Y
3 bilinear X and group×time-response dummy Y
4 trilinear X and group dummy Y
5 trilinear X and group/time-response dummy Y

btnr Number of bootstrap datasets.

btmd Resampling method for bootstrap
1 Balanced resampling
2 Balanced resampling within individual groups

PLS models with an optimal number of latent variables are built on each bootstrap subset and the Variable Importance in Projection (VIP) is calculated for each variable. For each variable, the mean (VIP*) and standard deviation (σ_{VIP}) of the btnr VIP values is obtained. The variable is selected if the lower-bound of the one standard deviation error bar is above 1 (i.e., VIP*-σ_{VIP}>1).

OUTPUT:

X_sel.data	Array of selected variables
X_sel.index	Index of selected variables
X_sel.vip	Mean VIP for each variable from bootstrapping

Selected variables are sorted with mean VIP in descending order.

In the datasets folder, a simulated dataset and a real dataset are provided as examples. These datasets are constructed in both two way and three way based on the requirements of the function PLSvar_sel. They can be tested directly without any transformation.

Plot Selected Variables

PLOTvar_sel(X_two_way,Y1,Ytime,nr)

INPUT:
X_two_way Array of independent variables, which is a two-way matrix.
Y1_two_way Array of dependent variables representing group information.
E.g, Samples from intervention and control group are labelled with 1 and -1 respectively.
Ytime Array of dependent variables representing time.
nr The index of selected variable resulted from X_sel.index. Input one number at a time.

OUTPUT:
Figure Temporal profiles for selected variables

Examples

The simulated dataset is used as an example here. The original dataset is X_two_way (64 samples × 3000 variables). The group and time information of the samples are stored in Y1_two_way and Ytime, respectively. SubID provides the subject ID for each sample and VarID shows the index of the variables.

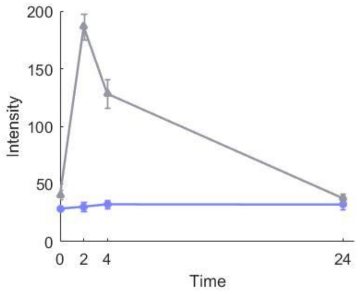
To apply the Type 3 PLS model with 200 times bootstrapping and 1 as VIP threshold, the command lines are shown as follows:

```
X=X_two_way;
Y1=Y1_two_way;
Y2=Y2_two_way;
MaxFac=5;
Type=3;
btnr=200;
btmd=1;
VIPt=1;
[X_sel]=PLSvar_sel(X,Y1,Y2,MaxFac,Type,btnr,btmd,VIPt);
```

The selected variables are stored in X_sel. To plot the temporal profiles of the 5th selected variable, use the following commands:

```
PLOTvar_sel(X_two_way,Y1_two_way,Ytime,X_sel.index(5));
```

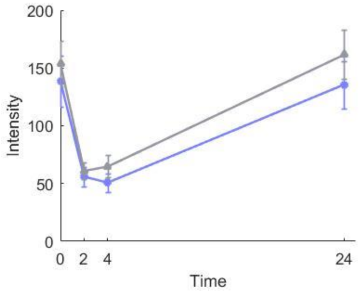
The figure is shown as below:



A non-selected variable is plotted as follows:

```
VarID_non_sel=setdiff(VarID,X_sel.index);
PLOTvar_sel(X_two_way,Y1_two_way,Ytime,VarID_non_sel(1));
```

The figure is shown as below:



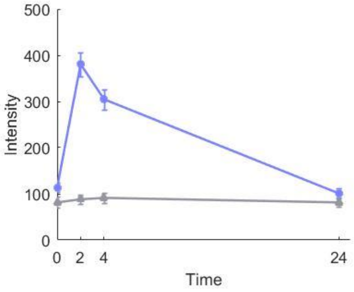
When applying Type 4 or 5 PLS models, the original dataset needs to be transformed into three way. The sequence of the variables should not be changed since when plotting the temporal profiles of the selected variables, the original two way data would be used. The following command lines shows the application of the Type 5 PLS model with 300 times bootstrapping and 1 as VIP threshold, the command lines are shown as follows:

```
X=X_three_way;
Y1=Y1_three_way;
Y2=Y2_three_way;
MaxFac=5;
Type=5;
btnr=300;
btmd=1;
VIPt=1;
[X_sel]=PLSvar_sel(X,Y1,Y2,MaxFac,Type,btnr,btmd,VIPt);
```

The temporal profiles of the 10th selected variable are plotted using the following commands:

```
PLOTvar_sel(X_two_way,Y1_two_way,Ytime,X_sel.index(10));
```

The figure is shown as below:


Intensity
Time