R package `plsdepot`
SIMPLS

Gaston Sanchez
www.gastonsanchez.com/plsdepot

1 Introduction

SIMPLS is a technique proposed by Sijmen de Jong (1993) as an alternative algorithm for PLS regression. SIMPLS is an acronym for *Straightforward Implementation of a statistically inspired Modification of the PLS method*. It is a method for analyzing the relationships between two data tables (i.e. two blocks of variables). One block consists of predictors, usually denoted by X. The second block consists fo responses, usually denoted as Y. As with most multivariate methods for analyzing two blocks of data, the basic idea behind SIMPLS is to look for components $t_h = Xa_h$ and $u_h = Yb_h$ in such a way that:

1. they explain well their own block of variables, and
2. they are as much correlated as possible with each other.

under the following conditions

- normalized coefficients: $\|a_h\| = 1$ and $\|b_h\| = 1$
- orthogonal components: $t_h'(t_1, \ldots, t_{h-1}) = 0$

In this case, the attention is focused on the t_h components, which are expected to be good representants of X, but also to be good predictors of Y.

Doing some matrix algebra, it turns out that the solution of SIMPLS is the one that maximizes

$$\max_{\|b_h\|=1} cov(Xa_h, Yb_h) = \sqrt{\sum_{k=1}^{q} cov^2(y_k, Xa_h)} \tag{1}$$

2 Data `linnerud`

For this demo we are going to use the data set `linnerud` that already comes in `plsdepot`. This data contains 6 variables measured on 20 individuals. The variables can be grouped in two blocks. One block X for three physical measurements, and another block Y for exercise outputs.

```r
# load the package
library(plsdepot)

# load the data
data(linnerud)

# let's take a peek
head(linnerud)
```

```
##   Weight Waist Pulse Pulls Squats Jumps
## 1    191    36    50     5    162    60
## 2    189    37    52     2    110    60
## 3    193    38    58    12    101   101
## 4    162    35    62    12    105    37
## 5    189    35    46    13    155    58
## 6    182    36    56     4    101    42
```

3 Function `simpls()`

The function `simpls()` has 3 arguments: `X`, `Y`, and `scaled`. `X`, as you may guess, is the data containing the predictors. This can be either a matrix or a data frame. `Y` is the data containg the responses, which can also be either a matrix or a data frame. `scaled` specifies whether to standardize the data (`TRUE` by default). Let's apply `simpls()` on `linnerud`.

```r
# apply simpls
my_sim = simpls(linnerud[, 1:3], linnerud[, 4:6])

# what's in my_sim?
my_sim
```

```
##
## SIMPLS
## ----------------------------------------
## $x.scores    X-scores (T-components)
## $x.wgs       X-weights
## $y.wgs       Y-weights
## $cor.xt      X,T correlations
## $cor.yt      Y,T correlations
## $R2X         explained variance of X by T
## $R2Y         explained variance of Y by T
## ----------------------------------------
##
```

What you get in `my_sim` is an object of class `"simpls"`, and everytime you print an object of such class you get a display with the list of results.

3.1 PLS components

The first two elements in the list are `$x.scores` and `$x.wgs` which contains the extracted PLS components, and its associated weights (i.e. coefficients).

```
# check scores T
head(round(my_sim$x.scores, 3))

##       t1     t2
## 1 -0.643 -0.583
## 2 -0.770 -0.164
## 3 -0.907  0.513
## 4  0.688  0.670
## 5 -0.487 -1.116
## 6 -0.229  0.071

# weights
my_sim$x.wgs

##             t1      t2
## Weight -0.5899 -0.3635
## Waist  -0.7713  0.6902
## Pulse   0.2389  0.6257
```

3.2 Correlations

Remember that our main interest is focused on the components t_h. These guys are not only supposed to summarize the information in X, but also they are supposed to be well predictors of Y. So let's check the correlations between X and T.

```
# correlations X-T
my_sim$cor.xt

##             t1     t2
## Weight -0.9476 0.0128
## Waist  -0.9620 0.2349
## Pulse   0.5108 0.7901
```

The first component t_1 is capturing well enough the information of `Weight` and `Waist`. In contrast, t_2 is the one that better summarizes `Pulse`.

Now let's see how well the components t_h are correlated with the exercise measurements Y.

```
# correlations Y-T
my_sim$cor.yt

##            t1       t2
## Pulls  0.4862 -0.22285
## Squats 0.5921 -0.19266
## Jumps  0.2035 -0.04306
```

3.3 Explained Variance

Besides the correlation among the data blocks and the extracted components, another important result is the proportion of explained variance. Here the purpose is to assess how well the components explain each block of variables.

```
# explained variance of X by T
my_sim$R2X

##            t1     t2
## Weight 0.8980 0.8982
## Waist  0.9255 0.9806
## Pulse  0.2609 0.8852

# explained variance of Y by T
my_sim$R2Y

##            t1      t2
## Pulls  0.2363 0.28601
## Squats 0.3506 0.38771
## Jumps  0.0414 0.04325
```

4 Plotting `"simpls"` objects

An accessory function is the `plot()` method that allows us to get some graphics of the basic results. Basically, we can plot either the variables and the observations on the specified components. The variables are plotted inside a circle of correlations. In turn, the observations are plotted using a scatter-plot.

4.1 Plotting variables

The default output when using `plot()` is a graphic showing the correlations of the variables with the first two components. This plot can be regarded as a radar. The closer a variable appears on the perimiter of the circle, the better it is represented. In addition, if two variables are highly correlated, they will appear near each other. If two variables are negatively correlated, they will tend to appear in opposite extremes. If two variables are uncorrelated, they will be orothogonal to each other.

```
# default plot
plot(my_sim)
```

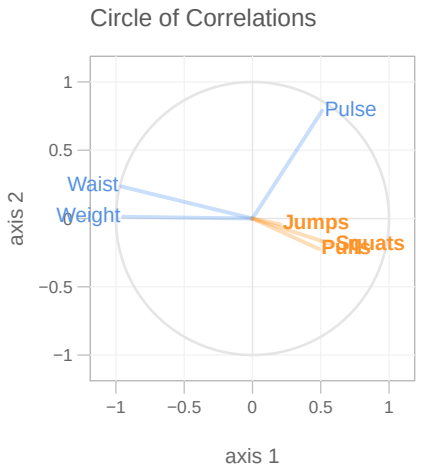

Figure 1: Circle of correlations (axes 1-2)

4.2 Plotting observations

The alternative output when using `plot()` is to show a scatter-plot of the observations on the specified components.

```
# default plot
plot(my_sim, what = "observations", show.names = TRUE)
```

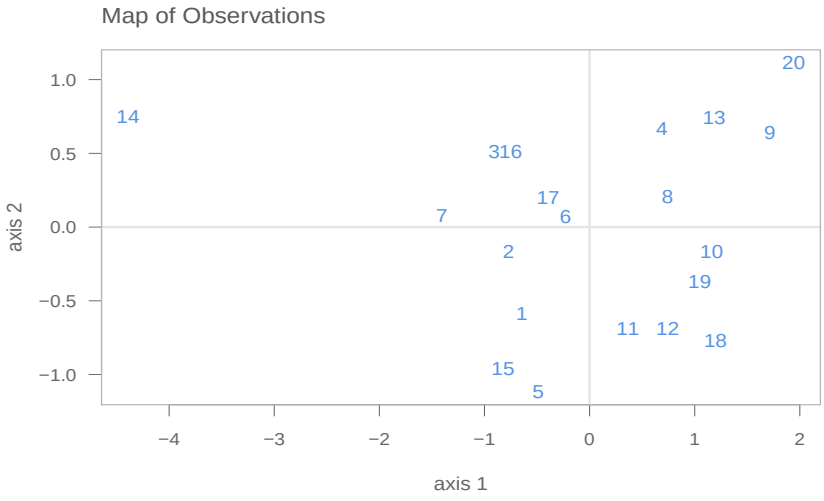

Figure 2: Plot of observations (comps 1-2)

References

de Jong S. (1993) SIMPLS: An Alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18: 251–263.
Tenenhaus M. (1998) *La Regression PLS. Theorie et Pratique*. Paris: Editions TECHNIP.