PCA-GCA tutorial for model selection

PCA-GCA Separates common and distinctive components in multiple data blocks, using a combination of PCA and GCA

Reference: Smilde AK, Måge I, Næs T, et al. Common and Distinct Components in Data Fusion. J Chemom. 2017;In press. doi:10.1002/cem.2900.

The data blocks are in the "saisir" format. The "saisir" data set structure contains three fields:

.d: The data matrix, size NxP

.i Sample names (character array, N rows)

.v Variable names (character array, P rows)

See http://www.chimiometrie.fr/saisir_webpage.html for more information about the saisir toolbox.

Set default options by running

```
options=PCAGCA(X)
```

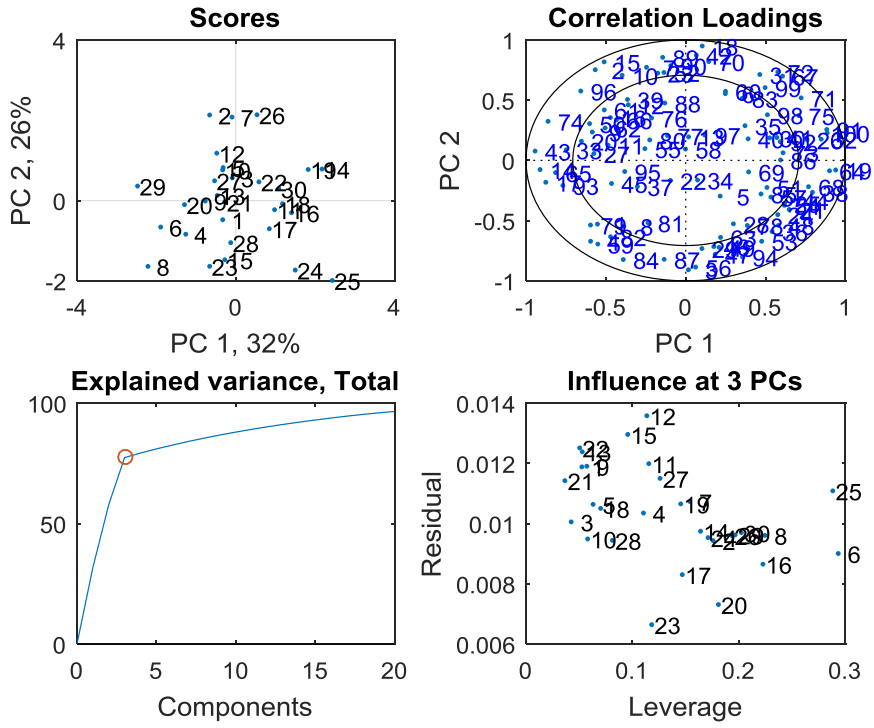
where X is a cell array of data blocks (saisir structures)

Modify the fields of the options structure if necessary (see description of fields in function documentation.

Fit the model by running

```
model=PCAGCA(X,options)
```

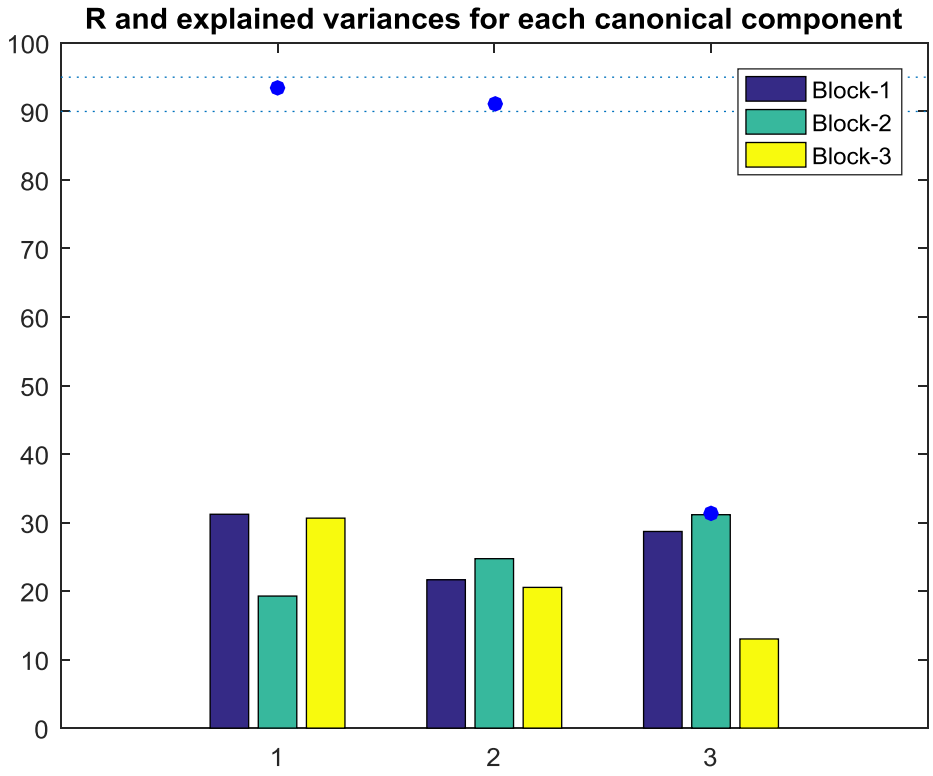
If the 'autoselect' option is set to zero, a number of plots will pop up for model selection.
First, you need to set the number of PCA components for each block separately. PCA plots like this will pop up, and you need to input the number of components in the command window:




```
Number of components in block 1: 3
```

This is repeated for all blocks. The numbers of components can also be specified in the options field ‘nCompsLocalPCA’, in which case these plots will not pop up.

Then, the numbers of common components is selected by evaluating these plots:



Where the bars represent the explained variance in each block, and the dots represent the correlation coefficient (x100). A component must have high correlation and also explain a significant amount of variance in each bloc, in order to be defined as common. In this case, the first two components satisfy these criteria. The third component explain a fair amount of variance in each block, but the correlation coefficient is low (0.31). The third component can therefore not be regarded as common.

Input your choice in the command window:

```
Common components, block 123, keep which components (vector)? [1 2]
```

This decides the number of global common components (common for all three blocks). Similar plots will pop up for the number of local common components. If none of the components are common, type '0' or [].

Then, a results table is displayed:

Explained variance per data block

	Block1	Block2	Block3	Correlation
C123_1	31.2	19.3	30.6	0.94
C123_2	21.6	24.7	20.5	0.91
C12_1	28.7	31.2		0.91
D3_1			26.1	
Total	81.6	75.1	77.3	

This table summarizes how much variance each component explained in each block. ‘C123’ that the component is common across block 1, 2 and 3, ‘C12’ that the component is common only across block 1 and 2. ‘D1’ means that the component is distinct for block 1, etc.