MATLAB code for multiblock regression methods:

SO-PLS (Sequential Orthogonalized PLS)
PO-PLS (Parallel Orthogonalized PLS)

Ingrid Måge

Nofima AS, Osloveien 1, NO-1430 Ås

e-mail: ingrid.mage@nofima.no

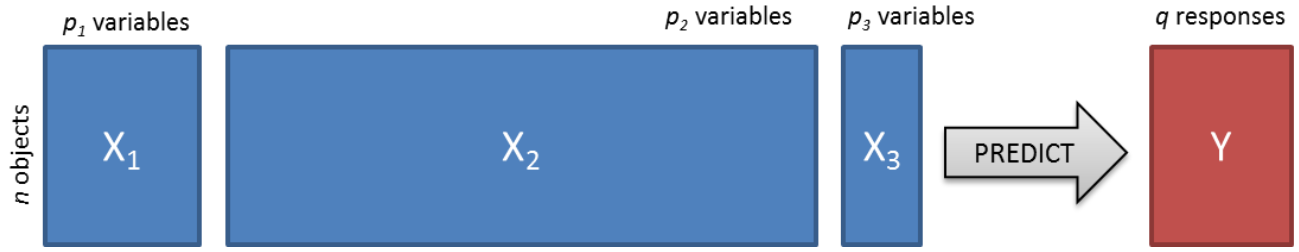

Figure 1. Scematic example of a multiblock regression problem with 3 predictor blocks and q response variables. The three predictor blocks have p_1, p_2 *and* p_3 variables each.

SO-PLS

SO-PLS extracts information sequentially from each X-block. This method is useful when there is a natural order of blocks, and one wants to focus on what additional information each block contains. For this method, the order of the X-blocks is crucial. For detailed description of SO-PLS, see references 1-4.

PO-PLS

PO-PLS extracts information from the X-blocks in a parallel way, and the focus is on identifying common/overlapping and unique information between the X-blocks. This method is useful when all the blocks are of equal interest, and there is no natural order. The method starts by extracting components that are common between blocks, i.e. information that is present in more than one block. When all the common information is extracted, what is left in each block is called unique information. See references 4-6 for more information on PO-PLS.

Matlab implementation

The MATLAB code is a mix of in-house routines and the free, open-source "SAISIR" software package [7].

Each data block is organized in a SAISIR structure array with fields:

- d: Data matrix of dimension n x p. No missing values are allowed
- i: object names, character array with n rows
- v: variable names, character array with p rows

The main functions are *POSO_PLS, crossval_POSO_PLS* and *plot_POSO_PLS*. All other functions in the toolbox are called by these, so it is only necessary to use these three functions. Examples of how to run the functions are found in the script "EXAMPLE_SCRIPT".

Example

```
%Data blocks are organized in structures X1, X2, X3 and Y

X={X1 X2 X3};                          %three predictor blocks

options=POSO_PLS(X,Y)                  % makes default options
options.preprocX{1}='autoscale';       %the first block is autoscaled
options.autoselect=1;                  % sets automatic selection of components

model=POSO_PLS(X,Y,options)            % fits the model
model=crossval_POSO_PLS(model,'r20')% cross-validates the model
plotoptions=plot_POSO_PLS(model)     % make default plot options
plot_POSO_PLS(model,plotoptions)     % plots model results
```

POSO_PLS(X,Y,options)

```
options=POSO_PLS(X,Y)
model=POSO_PLS(X,Y,options)
```

X is a cell array of length 1 x nBlocks, where each cell contains a predictor block (as a SAISIR structure)

Y is the data block with responses (as a SAISIR structure)

options is a structure array with fields:

- cvi: type of cross-validation. vector with indices or 'loo'/'r10'/r'20' (leave-one-out/random 10/random 20)
- autoselect: 0/1 automatic selection of components
- preprocX: cell array (length nBlocks) with preprocessing method for each X block. Either 'mean center' or 'autoscale'
- preprocY: preprocessing of Y. Either 'mean center' or 'autoscale'
- Amax: Vector (1 x nBlocks). Maximum number of components for each block.
- blockCombinations: cell array defining relevant combinations of blocks. default is all combinations.
- nCompsLocalPls: cell array (length blockCombinations) with number of components in local PLS models within each block combination
- nCompsForEachContribution: cell array (length blockCombinations) defining number of components from each blockCombination

model is a struct with fields:

- X: Input X data
- Y: Input Y data
- options: struct with fields described above. Not the same as input, the options are modified during modelling

- Scores: cell array with scores from each blockCombination
- Loadings: cell array with loadings from each blockCombination
- fittedY
- beta: coefficients from final prediction model (based on Scores)
- forprediction. Various parameters needed to predict new samples
- expVarX: cell array (blockCombinations x nBlocks) with cummulative explained X variances
- expVarY: cell array (blockCombinations x nY) with explained Y variances
- RMSEC: root mean squared error of calibration
- ANOVA: CV-anova results (this field is added after running 'crossval_POSO_PLS')
- cvres: CV results (this field is added after running 'crossval_POSO_PLS')

crossval_POSO_PLS

```
model=crossval_POSO_PLS(model,cvi,wb)
```

model is a PO-PLS or SO-PLS model fitted by "POSO_PLS"
cvi defines the cross-validation segments. Can either be:

- 'loo': leave-one-out
- 'r10': ten random segments
- 'r20': twenty random segments
- A vector (length n) with indices defining in which segment each object belongs

wb is optional, and can be used to suppress the waitbar. wb=0 means no waitbar

plot_POSO_PLS

```
plotoptions=plot_POSO_PLS(model)
plot_POSO_PLS(model,plotoptions)
```

model is a PO-PLS or SO-PLS model fitted by "POSO_PLS"

plotoptions is a structure array with fields:

- plots: 'all' (default),'expVarPie','Predplot' or 'ScoresLoadings'
- spec: =1 if data are spectra, loadings will be plotted as curves
- PCs: 1x2 vector defining which components to plot. Default is [1 2]
- Samplegroups: numeric vector or character array defining sample groups
- Variablegroups: cell array containing numeric array or character array

References

1. Jørgensen, Kjetil; Mevik, Bjørn-Helge; Næs, Tormod. Combining designed experiments with several blocks of spectroscopic data. *Chemometrics and Intelligent Laboratory Systems* 2007; Volum 88.(2) s. 154-166
2. Jørgensen, Kjetil; Næs, Tormod. The use of LS-PLS for improved understanding, monitoring and prediction of cheese processing. *Chemometrics and Intelligent Laboratory Systems* 2008; Volum 93.(1) s. 11-19
3. Næs, Tormod; Tomic, Oliver; Mevik, Bjørn-Helge; Martens, Harald. Path modelling by sequential PLS regression. *Journal of Chemometrics* 2011; Volum 25.(1) s. 28-40
4. Næs, Tormod; Måge, Ingrid; Segtnan, Vegard. Incorporating interactions in multi-block sequential and orthogonalised partial least squares regression. *Journal of Chemometrics* 2011; Volum 25.(11) s. 601-609
5. Måge, Ingrid; Mevik, Bjørn-Helge; Næs, Tormod. Regression models with process variables and parallel blocks of raw material measurements. *Journal of Chemometrics* 2008; Volum 22. s. 443-456
6. Måge, Ingrid; Menichelli, Elena; Næs, Tormod. Preference mapping by PO-PLS: Separating common and unique information in several data blocks. *Food Quality and Preference* 2012; Volum 24.(1) s. 8-16
7. Dominique Bertrand, Christophe Cordella, 2011. SAISIR package . Free toolbox for chemometrics in the Matlab, Octave or Scilab environments. Available at http://www.chimiometrie.fr/saisir_webpage.html