Cluster toolbox v2.0

Yun Xu

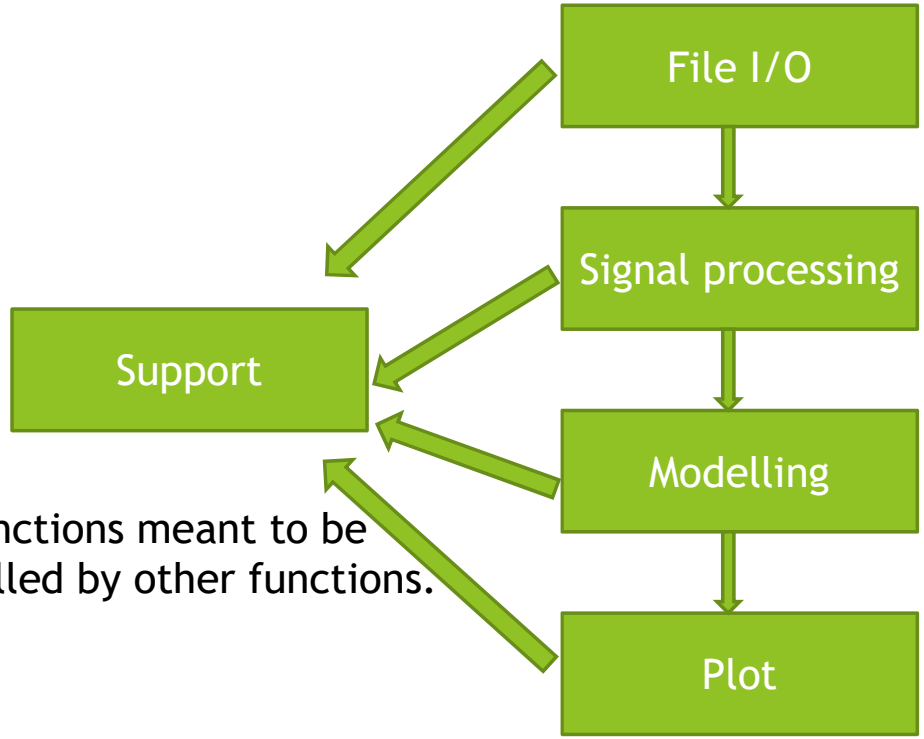
School of chemistry, University of Manchester

18/May/2016

Cluster toolbox is now 20 (1996-2016) years old!

- Some maintenance works are needed
  - File name issues
  - Hidden bugs
  - Better descriptions
- Incorporate new models
  - File I/O
  - signal processing
  - multi-block models
  - PLS and its extensions

Five categories



functions meant to be called by other functions.

netCDF, SPC, ASCII

EMSC, BC, normalization…

PCA, DFA, PLS…

File I/O – newly added

- Thermo-Galactic SPC files
  - **get_spc()**
- MALDI-T.o.F.-MS ASCII files
  - **get_maldi_text()**
- netCDF hyphened mass spectrometry files
  - **get_cdf_tic()** for importing TIC only
  - **get_cdf()** for low resolution MS (binned to unit m/z)
  - **get_cdf_lcms()** for high resolution MS
  - **get_cdf_dims()** for high resolution direct infusion MS

Signal processing

- For baseline correction
  - **baseline_correction()** and **asysm()**
- For vibrational spectroscopy spectra
  - emsc(), CO2corr(), detrendm(), **band_area()**, **band_area2()**
- For GC/MS, LC/MS alignment using on QCs
  - **qc_corr()**
- For general use
  - derivats() for Savizky-Golay derivatives
  - **gaussian_smooth()** for fast signal smoothing
  - **mos()** morphological factors/scores for multivariate signal-to-strength measurement
  - normal(), normalhigh(), normaltot(), scalem(), vecnorm() for various normalisations
  - **dosc()**, **osc()** for orthogonal signal correction

Modelling

- PCA
  - pca() has been renamed to **pca_np()** to avoid confliction with pca() in MATLAB statistics toolbox.
  - **cpca()** and **hpca()** for multi-block PCA.
  - **asca()** for ANOVA-simultaneous component analysis.
  - **msca()** for multi-level simultaneous component analysis.
- DFA
  - dfa() has been updated to prevent giving unrealistic results when the number of classes has been set to high.
  - projpcdf() for projecting test set to the PC-DFA space created by training set.
- Cluster analysis
  - oc_clustering() for hierarchical cluster analysis.

Modelling

- **PLS**
  - **pls()**, **plspred()** and **plspred2()** for basic PLS regression and classification
  - **plsr_boots()** and **plsda_boots()** for PLS-R/DA with built-in bootstrapping validation
  - **opls()**, **opls2()** and **oplsda()** for orthogonal projection to latent structure
    - use **opls()** for single variable prediction (PLS 1)
    - use **opls2()** for multiple variables prediction (PLS 2)
    - use **oplsda()** for classification

Plot functions

- Classical plots
  - plotftir(), plotpyms(), plot_map()
  - plot_pca(), plot_dfa(), p2d_col(), plotnm2(), plotnm3()
- “Fancy” plots
  - **gradientclass_plot()**, **gradientclass_plot2()**, **multiclass_plot()**
- 1-*a* confidence region
  - **error_ellipse()**

How to use it

- Create a folder, e.g. cluster_toolbox and put all the files under it.
- Add this path to MATLAB search path list (“Set Path” button).
- Type “help cluster_toolbox” for the full list of the functions.
- Type “help <function name>” for the help of using each individual function.

To do list

- Demonstration scripts.
- Template scripts for common tasks.
- Improve current functions, e.g. add error ellipses to PC-DFA projection function.
- More functions to be added.

Let me know if you have any suggestions or requests by send your request to yun.xu-2@Manchester.ac.uk