SK-Ana: Spectro-Kinetic matrices Analysis

P. Pernot

(2020-09-09)

Contents

Introduction **1**

Workflow **2**

Modules reference **2**

`Project` module . . . 2

`New Project` tab . . . 2

`Open` and `Save` tabs . . . 3

`Data Selection` module . . . 3

`Selection` tab . . . 4

`Baseline` tab . . . 4

`Wavl Mask` tab . . . 5

`Delay Mask` tab . . . 5

`SVD` module . . . 5

`ALS` module . . . 6

Controls . . . 6

Outputs . . . 7

`Kinet` module . . . 8

Controls . . . 8

Outputs . . . 10

`Downloads` module . . . 11

`About` module . . . 11

Introduction

The `SK-Ana` graphical interface is organized in sequential order of project management:

- `Project`: define a project's name and load the data

- `Data Selection`: define the data subset to be treated
- `SVD`: perform Singular Values Decomposition analysis
- `ALS`: perform Alternated Least-Squares decomposition
- `Kinet`: constrain the analysis by a kinetic model
- `Downloads`: download saved the results and/or a report
- `About`: information about the code

A good introduction to the methods can be found in the article

> C. Ruckebusch, M. Sliwa, P. Pernot, A. d. Juan and R. Tauler (2012) "Comprehensive data analysis of femtosecond transient absorption spectra: A review". *J. Photochem. Photobiol. C* **13**:1–27. (http://dx.doi.org/10.1016/j.jphotochemrev.2011.10.002)

Workflow

A typical workflow consists in the sequence:

```
`Project` > (`Data Selection` > `SVD` > `ALS`) > `Downloads`
```

where the sequence between parentheses is iterated until *satisfecit*. Technically, the `SVD` step could be avoided for an ALS analysis, but it provides a lot of useful information and should not be overlooked.

To perform a SAS or DAS analysis, the sequence would be

```
`Project` > `Data Selection` > `SVD` > `Kinet` > `Downloads`
```

here again, the `SVD` step helps to decide the number of species to include in the chemical scheme.

Modules reference

`Project` module

Project definition and data input.

`New Project` tab

- **`Project Name`**: choose a name. If not, a name will be generated from the datafiles selected below.
- **`Predefined File Formats`**: a few data file formats have been predefined from the datafiles of different experiments. Choose the one corresponding to your data. For fine tuning, select 'Other...' which will open a new panel.

	Header	Separator	Decimal	Data structure
CSV	FALSE	','	'.'	wxd
ELYSE	FALSE	'\t'	'.'	wxd
Fluo	FALSE	';'	'.'	wxd
Streak	TRUE	','	'.'	wxd

  - `Header`: does the first line contain column headers ?
  - `Separator`: symbol used to separate the columns
  - `Decimal`: character used in the file for decimal points
  - `Data structure`:
    * `wxd`: wavelength in columns; delays in lines
    * `dxw`: delays in columns; wavelengths in lines

- **`Load-time compression factors`**: the data can be averaged by blocks at load time to save processing time and reduce noise.
  - `Delay` width (in pixels) of the block in delay dimension
  - `Wavl` width (in pixels) of the block in wavelength dimension

- **`Select data file(s)`**: select one or several files to be analyzed. Selecting the files will create new items in the right panel:
  - a success message '**Data Loaded !**'
  - an active table with a description of the file(s). When several files have been loaded:
    * it is possible to use the table to select a subset or to reorder them.
    * a menu appears with processing options:
      · `Average`: average the selected files
      · `Tile Wavl`: assemble the matrices in the wavelength dimension
      · `Tile Delay` (default): assemble the matrices in the delay dimension. In this case, the delay coordinate is replaced by an index.
      · press on **`Do It!`** to process the data
  - a summary of the processed matrix
  - `Save Matrix`: to save the processed matrix in a .csv file
  - a vignette of the processed matrix

- **`Post-process compression factor`**: the block averaging is performed *after* the data files are assembled.

`Open` and `Save` tabs

These are placeholders. The functionalities are not active.

`Data Selection` module

This module enables to fine tune a subset of data to be analyzed, mainly to remove over-noisy area or artefacts (signal rise, Cherenkov. . .).

Important: you have to visit this tab to activate/enable the the analysis tabs (`SVD`, `ALS`. . .).

The left panel contains four tabs:

- `Selection`: to select the external limits of the treated area
- `Baseline`: to define the areas used to correct the signal baseline
- `Wavl Mask`: to define masks on the wavelength axis (wavelengths ranges not to be analyzed)

- `Delay Mask`: to define masks on the delay axis, typically to mask signal artefacts

and a set of buttons to `Reset`, `Save` and `Load` selections.

Warning: the `Save` and `Load` operations are experimental, meaning unstable. Presently, any difference in the loaded matrices prevents the reuse of saved selections.

The right panel contains two tabs with graphical representations that show the modifications due to the actions in the left panel:

- the `Data` tab represents the data matrix and averaged cuts along both axes.
  - The data matrix is zoomable (`click and drag` + `dble click`; another `dble click` cancels the zoom).
  - The cuts position are marked by violet dashed lines and they are controlled by two sliders (`Reference wavl` and `Reference delay`). Each cuts data can be saved to disk by clicking on the corresponding `Save` buttons.
  - Masks are represented by grayed areas.
- the `Cuts` tab provides the usual stacked lines representation, for the spectra (left) and kinetic traces (right), where you can choose the cut frequency `Cut freq.` for each axis. Both figures are zoomable and the corresponding data can be saved to disk.

`Selection` tab

This tab contains three sliders:

- `OD Range`: select the Optical Density range to improve visualization, notably when there are spikes. **This has no impact on data selection.**
- `Wavelength Range`: selects the min and max wavelengths of the matrix
- `Delay Range`: selects the min and max delays of the matrix

`Baseline` tab

A Baseline mask is used to correct the baseline in a data matrix, by delaywise averaging the masked values to zero.

When several matrices have been delay-tiled, each baseline correction is applied to the data between this mask and the next one (or the end if none is present).

Note: The data covered by the Baseline masks are *not* excluded from the data analysis.

The initial tab contains two elements:

- a `Nb of masks` input where you can select the desired number of masks.
- a `Auto` button, which attempts to generate and locate the adequate number of masks.

For each mask, a slider is created enabling to define its min and max positions. The masks are represented by salmon transparent areas on the matrix and cuts figures.

Tip: Zooming on the data matrix is helpful to define precise limits.

`Wavl Mask` tab

The wavl masks are intended to exclude wavelength-delimited area(s) from data analysis, typically over-noisy areas or laser wavelengths.

The initial tab contains two elements:

- a `Nb of masks` input where you can select the desired number of masks.
- a `Auto` button, which attempts to generate and locate the adequate number of masks.

For each mask, a slider is created enabling to define its min and max positions.

`Delay Mask` tab

The delay masks are intended to exclude delay-delimited area(s) from data analysis, typically baseline areas and artefacts (Cherenkov).

The initial tab contains two elements:

- a `Nb of masks` input where you can select the desired number of masks.
- a `Auto` button, which attempts to generate and locate the adequate number of masks.

For each mask, a slider is created enabling to define its min and max positions.

`SVD` module

This module provides the Singular Values Decomposition of the selected data. The main utility of SVD is to inform us on the complexity of the data matrix.

For more details about the method, see SVD in Wikipedia.

The left panel contains two control inputs:

- `SVD parameters`: the `Dimension` selector enables to select the number of singular values used to build figures in the `Data vs. Model`, `Residuals` and `Contributions` tabs of the right panel.
- `Glitch removal in kinetics` enables to remove spikes in the data from the visualization of singular vectors in tab `Vectors`
  - `Level` is the index of the target **delay** vector from which the spike is to be removed.
  - the `Clean` button removes the spike. The code masks the point with the largest absolute value in the selected vector.
  - the `Cancel` button cancels the last spike removal

The right panel contains a set of tabs covering different aspects of the results:

- `Singular Values` contains two figures
  - the spectrum of singular values (golden dots, dotted blue line) and a baseline of noise estimated from the largest singular values (violet dashed lines). The number of species that can be identified in the data is given by the index of the smallest singular value standing out of the noise.

  - the lack-of-fit spectrum, which gives the percentage of the signal that is not represented depending on the number of singular vectors used in the signal recomposition. One can appreciate on this graph how adding a new species improves the model. For large indexes, on gets the noise-to-signal ratio.

  Rq: except for ideal data matrices, there is always an ambiguity of plus or minus one species (at best) on the cutting level from both figures. One gets rather a clear indication on the largest decomposition that would *not* be acceptable.

- `Vectors` presents the wavelength-wise and delay-wise singular vectors. The idea here is to discard vectors that contain pure noise. Here again, the step from signal to noise is often not clearcut.

  Spikes in the data matrix can create artificial signal, and one can remove the spikes in the *decay-wise* vectors by using the `Glitch removal in kinetics` tool in the left panel.

- `Data vs. Model` shows the SVD data recomposition and the original matrix side-by-side. The recomposition is driven by the `Dimension` parameter entered in the left panel.

  This for illustration, but in order to appreciate the effects of `Dimension` on the quality of the model, it is better to focus on the nex tab: `Residuals`

- `Residuals` shows the difference between the data matrix and its reconstruction from SVD vectors, controlled by `Dimension` in the left panel.

- `Contributions` shows the individual components of the SVD reconstruction.

- `Statistics` provides a table of results with the singular values, the lack-of-fit and the standard deviation of the residuals, versus the number of singular vectors.

`ALS` module

The Alternated Least Squares (ALS) algorithms factorizes the data matrix given a number of species and some constraints (*e.g.*, positivity of kinetics...).

The left panel contains the controls, and the right panel displays the outputs.

Controls

- `Dimension` is the number of species. This should be consistent with the results of the SVD analysis.
- `Max # Iter.` is the maximal number of iterations allowed before stopping the optimizer.
- `Log convergence threshold` controls the logarithm of the stopping convergence threshold of the optimizer.
- The `Run` button starts the ALS optimization.

Several tabs enable to fine tune the ALS analysis:

- `Options` tab
  - **Initialization** enables to select a starting point
    * `|SVD|` takes the absolute values of the SVD vectors
    * `NMF` takes the solution of a Non-negative Matrix Factorization algorithm
    * `Sequential` performs a series of ALS optimizations with increasing dimension
    * `Restart` enables to restart from a previous run. It does not work if `Dimension` is changed.

  - `Use SVD-filtered matrix` uses the noise filtering ability of thee SVD reconstruction. The matrix is computed with the dimension specified in the `SVD` tab.
  - `Opt S first` start by optimizing the spectra vectors, instead of the kinetics vectors by default.

- `S const.` tab: constraints on the spectra vectors
  - `S > 0`: positivity constraint
  - `S unimodal`: unimodality constraint
  - `Normalize`: normalize the spectra:
    * `SUM(S) = 1` normalizes the area of the spectra. The default is to normalize the intensities.
  - `Smooth`: a smoothing factor to get less noisy spectra, used as the `span` parameter in the `loess` function.
  - `External Spectrum Shape` opens a new control enabling to read a `.csv` file with spectra to be constrained.

    By default, the spectra are used as such (hard constraint) which is often too strong and results in poor solutions. Activating `Soft constraint` enables to input a weight for the similarity constraint in the loss function of the ALS. The `logWeight` slider enables to tune this weight.

- `C const.` tab: constraints on the kinetics vectors
  - `C > 0`: positivity constraint
  - `Closure`: imposes the conservation of matter by normalizing the sum of the kinetics to 1 (at each delay).

    Warning: This works only for unimolecular processes and could be in conflict with normalization constrains on the spectra.
  - `Presence matrix` enables to specify the occurrence of individual species in different experiments when datasets have been delay-tiled. By default the matrix is filled with ones (1). Put 0 where a species is not expected to occur. **When finished, press `Done`.**

Outputs

- `Alternated Least Squares` tab

  This shows the convergence message of the ALS code.

  Tip: For a succesful fit, the lack-of-fit should be versy close to the lack-of-fit statistics of a SVD with the same dimension.

- `Diagnostics` tab

  This gives access to several results:

  - `Data vs. Model` whichh compares side by side the best fit model to the data matrix
  - `Residuals`which shows the residuals map and an histogram of the residuals compared to the histogram of the data.
  - `SVD of residuals` which provides the Singular Values Decomposition of the residuals matrix.In the ideal case, all vectors should be featureless and appear as pure noise. A normal Q-Q plot is provided to assess the normality of the residuals distribution.

- `Kinetics and Spectra` tab

  This tab provides zoomable plots of the spectra and the associated kinetics, identified by color code. The data can be saved to disk.

- `Contributions` tab

  This tab shows the contribution matrix of each species with its weight.

- `Ambiguity` tab

  In most cases, the ALS decomposition is not unique and subject to rotational ambiguity. If one transforms/combines the spectra, the inverse transformation applied to the kinetics will leave their combination unchanged. The range of eligible transformations is limited by the varoius constraints on the ALS solutions (positivity...).

  The algorithm performs a brute force exploration of transformation matrices and might require a long time to finish. It returns a subset of valid spectra and kinetics, from which one can appreciate the level of ambiguity and/or search for better behaved solutions than the ones returned by the ALS.

  Several controls are available:

  - `Pick 2 or 3 vectors`: according to the dimension of the ALS decomposition, one has to choose a set of vectors. This might be the full set for dimensions 2 and 3, but the algorithm does not allow explorations of more than three-vectors transformations.
  - `Relative positivity threshold`: because of the noise in the data, one has to enable some level on non-positivity in the transformed vectors. The slider enables to pick a level.
  - `Exploration step`: the step amplitude for the exploration of the transformation matrix elements. The smaller the better and more accurate, but very small steps might incur very long calculations.
  - `Start`: click to start process when all other parameters have been chosen.
  - `Stop`: early stop of the process (it works sometimes...)
  - `Save`: save the subset of transformed spectra and kinetics to disk

`Kinet` module

This module enables to introduce an explicit chemical scheme, leading to a hybrid model, as described in Ruckebusch (2012). The spectra are optimized by least-squares for each value of the kinetic parameters, which are optimized by a non-linear algorithm.

The optimizer is based on a Bayesian statistical model where one maximizes the posterior probability density function (pdf) of the kinetic parameters, given the data and model.

The left panel contains the controls, and the right panel displays the outputs.

Controls

- `Model` enables to describe the chemical scheme. It contains two groups of tabs. At the top, `Type`, `Load` and `Save` manage the model and model files. At the bottom, `Scheme`, `Rates`, `Conc.` and `Eps.` display the parameters values and enable to edit some of them.

  - `Type` to enter manually the model in the text box. It must contain the reaction scheme and initial values for the reaction rates, the initial concentrations of species and their maximum extinction coefficients

    * **Reactions** are typed one per line, with two segments, separated by a semi-column (;). The first segment contains the reaction, for instance `A + B -> C + D`, where A, B are the reactants, '->' the reaction symbol and C, D the products. The second segment contains the initial value for the reaction rate and its uncertainty factor, separated by a slash '/'. For instance, `1e8 / 1.2`, meaning a rate constant of 1e8 (units should be consistent with your data) and a multiplicative uncertainty factor of 1.2, which corresponds to a relative uncertainty of about 20%.
      The full line for this example is thus `A + B -> C + D ; 1e8 / 1.2`
      Note(s)

- to fix a rate constant, its uncertainty factor should be 1.
- lines starting with a sharp (#) are treated as comments
- the reaction scheme is the same for all the experiments when data have been delay-tiled.

* **Extinction coefficients** are typed in the form `eps_X = value / Feps`, where `X` is the name of a species declared in the scheme, `value` is the value and `Feps` the uncertainty factor on this value.
  For instance `eps_A = 0.001 / 3`
  Note(s)
  - the extinction coefficient of all species in the recation scheme is 0 by default. The extinction coefficients of 'visible' species have to be declared.
* **Initial concentrations** are decalred with a format similar to the extinction coefficients `c0_X_i = value / Fc0`, where `X` is a species, `i` is the index of an experiment (should be 1 for single experiments) and `Fc0` is an uncertainty factor.
  For instance `c0_A_1 = 1 / 1`, meaning that the initial concentration of `A` is fixed to 1 in the first (or single) experiment.
  Note(s)
  - all initial concentrations are 0 by default.
* **Examples**

```
# 3-species DAS
A -> 0 ; 1 / 3
B -> 0 ; 0.5 / 3
C -> 0 ; 0.001 / 3

eps_A = 0.001 / 3
eps_B = 0.001 / 3
eps_C = 0.001 / 3

c0_A_1 = 1 / 1
c0_B_1 = 1 / 1
c0_C_1 = 1 / 1
```

```
# Transformation of A to C with a blind intermediate
A -> B ; 1 / 3
B -> C ; 0.5 / 3
eps_A = 0.001 / 3
eps_C = 0.001 / 3
c0_A_1 = 1 / 1
```

* Click `Done` to process the model

– `Load` to load an existing model file

– `Save` to save the typed model file (this does not save the optimized values)

– `Scheme` displays the reaction scheme after a model is defined (`Type > Done` or`Load`). It is not editable.

– `Rates` displays the reaction rates and their uncertainty factors, which are editable.

– `Conc.` displays the initial concentrations of all species, with one tab per experiment. All values are editable.

– `Eps.` displays the extinction coefficients of all species. All values are editable.
  Note(s)
  * values modified in the `Rates`, `Conc.` or `Eps.` tabs affect only the initial values for the optimizer. They are not taken into account when saving a model file in `Save`.

- `Run` contains the controls for the optimizer. By default, the optimizer performs a local search around the initial values, in a box defined by the parameters uncertainty factors. A global optimization can be performed by using multiple random starting points within this uncertainty box.
  - `Global Optimization Iterations` controls the number of iterations of the global optimizer (default 0)
  - `Global Population Factor` controls the number of random starting points. Active only if the number of lobal iterations is not zero.
  - `Log Convergence Threshold` enables to tweak the convergence threshold os the optimizer.
  - `S>0` defines the positivity constraint on spectra. It should be unchecked for DAS analysis.
  - `Smooth`: a smoothing factor to get less noisy spectra, used as the `span` parameter in the `loess` function.
  - `Weighted data` controls the use of a weighted least-squares criterion (experimental...)
  - `Restart` controls the initialization of the optimizer with the last results (shoul not be checked at first run)
  - `Run` launches the optimized.

  Note(s)

  - It is recommended to make a final run with `Restart` on to check that the oprimized value is stable
  - For a gloal optimization, it is recommended to run several small sets of iterations with `Restart` on to ensure a faster convergence, rather than a single large set of iterations. This avoids a premature trapping of the walkers in a local minimum.

Outputs

- `Best Params` tab

  It displays the results of the optimization (notably the final Lack-of-fit, which can be analyzed in `Diagnostics`) and signals possible problems with the solution, for instance when a value is at a limit of the initial uncertainty box. Such problems can be visualized in the `Identifiability` tab.

- `Trace` tab

  It shows the output and messages of the optimizer.

- `Identifiability` tab

  It presents two tabs to appreciate the identification of the parameters:

  - `Densities` plots the marginal densities of the Laplace approximation of the posterior pdf (salmon) in comparison to the prior densities defined by the parameters uncertainty factors (blue). If a parameter is well identified, its posterior density should be more concentrated than its prior density. Also, the posterior density should not be concentrated at a limit of the uncertainty box.
  - `Sample` displays histograms and pairs scatterplots for parameters samples drawn from the Laplace approximation of the posterior pdf. The upper panel shows the correlation coefficients between the parameters.

- `Diagnostics` tab

  This tab contains a series of tabs

  - `Lack-of-fit` presents the level of LOF reached by the solution, compared to the level for a SVD decomposition with the same number of species.
  - `Integ. kinet.` presents the wavelength-integrated optimized matrix compared to the data.

  - `Data vs. Model` whichh compares side by side the best fit model to the data matrix
  - `Residuals`which shows the residuals map and an histogram of the residuals compared to the histogram of the data.
  - `SVD of residuals` which provides the Singular Values Decomposition of the residuals matrix. In the ideal case, all vectors should be featureless and appear as pure noise. A normal Q-Q plot is provided to assess the normality of the residuals distribution.

- `Kinetics and Spectra` tab

  This tab provides zoomable plots of the spectra and the associated kinetics, identified by color code. The data can be saved to disk.

- `Contributions` tab

  This tab shows the contribution matrix of each species with its weight.

`Downloads` module

Two functionalities in this module:

- `Generate Report`: generates a .html file containing the results of the analysis. You can choose to omit `SVD` or `ALS`. To generate and download the report, **`Ctrl+Click`** on `Download` (if you simply click, this will crash the app and you will loose all your hard work...)
- `Get my files`: generates a .zip archive with all the files you saved to disk in the previous modules. To download the archive, **`Ctrl+Click`** on `Download` (if you simply click, this will crashe the app and you will loose all your hard work...)

  Note: all files starting by you project's name are included in the .zip file. If you do not change the project's name, you might gather older files along...

Remarks

- these downloads should go to your default `Downloads` folder
- if you run SK-Ana locally, the saved files are stored in `SK-Ana/outputdir` and you can simply pick them up from there.

`About` module

Provides miscellaneous information about the code, notably:

- `How to cite` provides you the correct citation to report if/when you use SK-Ana for a publication, namely

  > Pernot, P. (2018) SK-Ana: Analysis of Spectro-Kinetic Data (Version X.X).
  > https://doi.org/10.5281/zenodo.1064370

  where the version number is provided just above the `How to cite` link.

- `code@github` links to the Github page for the source code
- `Bugs report, Features request` guides you to the 'Issues' page of the github deposit, where you can interact with the author.
- `Users Manual` should lead you to the present documentation...