Guidelines for Spectra-Based Principal Component Analysis of TOF-SIMS Data

Scope and summary

This code allows you to perform principal component analysis (PCA) on time-of-flight secondary ion mass spectrometry (ToF-SIMS) spectra. The main information that it gives you are:

(i) Which peaks of your mass interval list explains most of the differences between your samples (loadings);

(ii) If different samples can be grouped together thanks to similar characteristics (score plots).

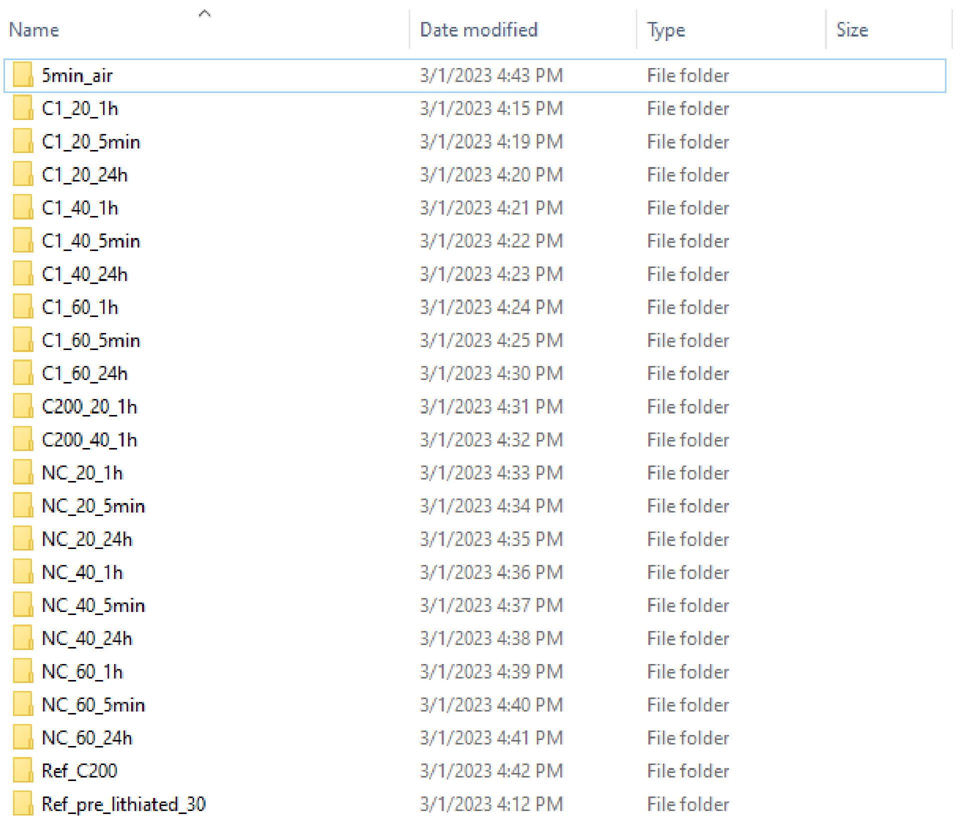
IF you are new to PCA, I advise you to watch this rapid YT video (very accessible but also accurate): https://youtu.be/FgakZw6K1QQ

The utilization of the code requires exporting your data and your mass interval list (the same one for all the samples) in a readable format from SurfaceLab and organizing it in the given subfolders structure ("Data Exporting and Data Structure" section), entering some input in an Excel located in the same folder of the code ("Code's Inputs" section) and run the code.

When the code star it will print a message (that will be closed automatically after 5 seconds) informing you that the program started, and a second message (that will not close automatically) will appear as soon as the analysis is finished.

Data Exporting and Data Structure

Data should be structured in the subfolder structure shown in Figure 1.

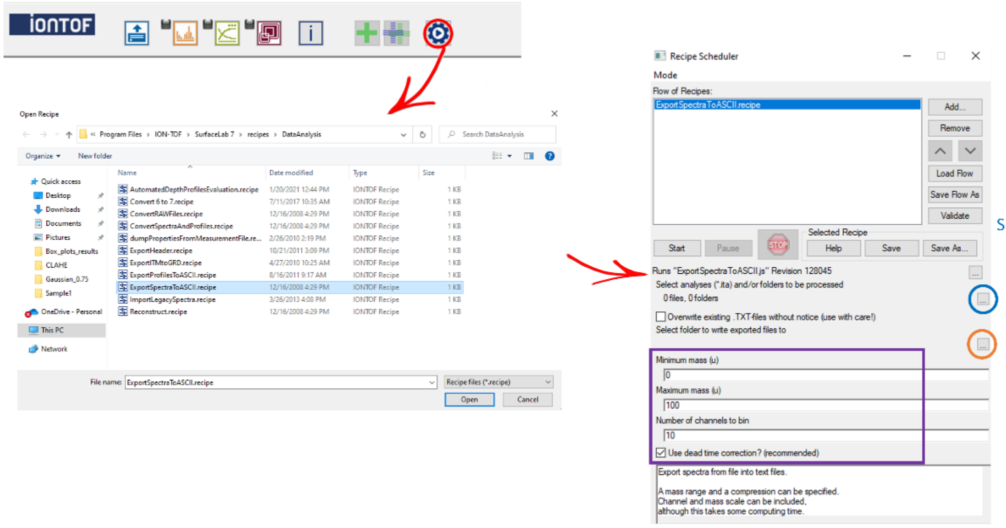


Name	Date modified	Type	Size
5min_air	3/1/2023 4:43 PM	File folder	
C1_20_1h	3/1/2023 4:15 PM	File folder	
C1_20_5min	3/1/2023 4:19 PM	File folder	
C1_20_24h	3/1/2023 4:20 PM	File folder	
C1_40_1h	3/1/2023 4:21 PM	File folder	
C1_40_5min	3/1/2023 4:22 PM	File folder	
C1_40_24h	3/1/2023 4:23 PM	File folder	
C1_60_1h	3/1/2023 4:24 PM	File folder	
C1_60_5min	3/1/2023 4:25 PM	File folder	
C1_60_24h	3/1/2023 4:30 PM	File folder	
C200_20_1h	3/1/2023 4:31 PM	File folder	
C200_40_1h	3/1/2023 4:32 PM	File folder	
NC_20_1h	3/1/2023 4:33 PM	File folder	
NC_20_5min	3/1/2023 4:34 PM	File folder	
NC_20_24h	3/1/2023 4:35 PM	File folder	
NC_40_1h	3/1/2023 4:36 PM	File folder	
NC_40_5min	3/1/2023 4:37 PM	File folder	
NC_40_24h	3/1/2023 4:38 PM	File folder	
NC_60_1h	3/1/2023 4:39 PM	File folder	
NC_60_5min	3/1/2023 4:40 PM	File folder	
NC_60_24h	3/1/2023 4:41 PM	File folder	
Ref_C200	3/1/2023 4:42 PM	File folder	
Ref_pre_lithiated_30	3/1/2023 4:12 PM	File folder	

Figure 1. Schematics of the subfolder organization required for this code to work. Each folder represents one sample, and inside each folder there should be reported all the spectra (≥1) measured for that sample in a readable format (Figure 2).

The procedure needed to export the spectra in a readable format is schematized in Figure 2. Through this procedure, you can export (txt format) at once more spectra at once.

In addition to the spectra, the mass interval list you want to use should also be exported in a readable format. For doing this, you can open the “Peak list” widget (In Surface Lab) and then click on “Export peak list as/Ascii”.



Open Recipe

« Program Files › ION-TOF › SurfaceLab 7 › recipes › DataAnalysis

Name	Date modified	Type	Size
AutomatedDepthProfilesEvaluation.recipe	1/20/2021 12:44 PM	IONTOF Recipe	1 KB
Convert 6 to 7.recipe	7/11/2017 10:35 AM	IONTOF Recipe	1 KB
ConvertRAWFiles.recipe	12/16/2008 4:29 PM	IONTOF Recipe	1 KB
ConvertSpectraAndProfiles.recipe	12/16/2008 4:29 PM	IONTOF Recipe	1 KB
dumpPropertiesFromMeasurementFile.re...	2/26/2010 2:19 PM	IONTOF Recipe	1 KB
ExportHeader.recipe	10/21/2011 3:09 PM	IONTOF Recipe	1 KB
ExportITMtoGRD.recipe	4/27/2010 10:25 AM	IONTOF Recipe	1 KB
ExportProfilesToASCII.recipe	8/16/2011 9:17 AM	IONTOF Recipe	1 KB
ExportSpectraToASCII.recipe	12/16/2008 4:29 PM	IONTOF Recipe	1 KB
ImportLegacySpectra.recipe	3/26/2013 4:08 PM	IONTOF Recipe	1 KB
Reconstruct.recipe	12/16/2008 4:29 PM	IONTOF Recipe	1 KB

File name: ExportSpectraToASCII.recipe

Recipe files (*.recipe)

Recipe Scheduler

Flow of Recipes:

ExportSpectraToASCII.recipe

Runs "ExportSpectraToASCII.js" Revision 128045

Select analyses (*.ita) and/or folders to be processed

0 files, 0 folders

☐ Overwrite existing .TXT-files without notice (use with care!)

Select folder to write exported files to

Minimum mass (u)

0

Maximum mass (u)

100

Number of channels to bin

10

☑ Use dead time correction? (recommended)

Export spectra from file into text files.

A mass range and a compression can be specified.
Channel and mass scale can be included,
although this takes some computing time.

Select all the files to be exported (all the measurements done for one sample)

Select the folder where the exported (.txt format) files are saved

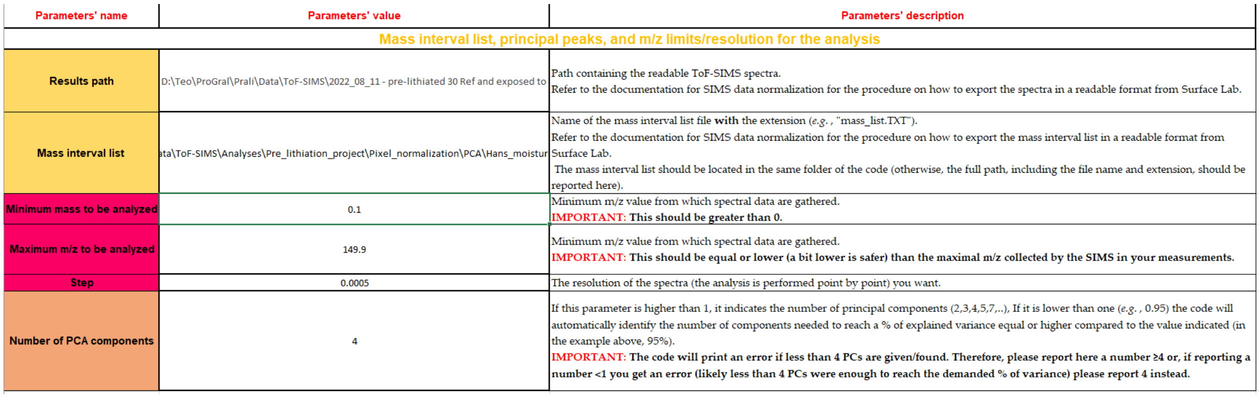
Parameters

Figure 2. Schematics of the procedure to be followed in SurfaceLab to export ToF-SIMS spectra in a readable format. Through this procedure, more spectra can be extracted at once. It is advised to use a number of channels to bin equal to 1 (greater resolution) and to use the dead time correction (advised by ION TOF).

Code's Inputs

The code inputs need to be reported in the associated Excel file, which should be located in the *same* folder containing the code file.

An example of parameters (image-based code example) is reported in Figure 3 (and in the "Inputs_ToF - PCA - Example.xlsx" file).



Parameters' name	Parameters' value	Parameters' description
Mass interval list, principal peaks, and m/z limits/resolution for the analysis		
Results path	D:\Teo\ProGral\Prali\Data\ToF-SIMS\2022_08_11 - pre-lithiated 30 Ref and exposed to	Path containing the readable ToF-SIMS spectra. Refer to the documentation for SIMS data normalization for the procedure on how to export the spectra in a readable format from Surface Lab.
Mass interval list	ata\ToF-SIMS\Analyses\Pre_lithiation_project\Pixel_normalization\PCA\Hans_moistur	Name of the mass interval list file **with** the extension (*e.g.*, "mass_list.TXT"). Refer to the documentation for SIMS data normalization for the procedure on how to export the mass interval list in a readable format from Surface Lab. The mass interval list should be located in the same folder of the code (otherwise, the full path, including the file name and extension, should be reported here).
Minimum mass to be analyzed	0.1	Minimum m/z value from which spectral data are gathered. IMPORTANT: **This should be greater than 0.**
Maximum m/z to be analyzed	149.9	Minimum m/z value from which spectral data are gathered. IMPORTANT: **This should be equal or lower (a bit lower is safer) than the maximal m/z collected by the SIMS in your measurements.**
Step	0.0005	The resolution of the spectra (the analysis is performed point by point) you want.
Number of PCA components	4	If this parameter is higher than 1, it indicates the number of principal components (2,3,4,5,7,..), If it is lower than one (*e.g.*, 0.95) the code will automatically identify the number of components needed to reach a % of explained variance equal or higher compared to the value indicated (in the example above, 95%). IMPORTANT: **The code will print an error if less than 4 PCs are given/found. Therefore, please report here a number ≥4 or, if reporting a number <1 you get an error (likely less than 4 PCs were enough to reach the demanded % of variance) please report 4 instead.**

Figure 3. Example of the Excel input file being already filled.

In general, each parameter is associated with a devoted and complete explanation in the Excel file (Inputs_ToF - PCA.xlsx), which however is also summarized below. All the parameters of this code are mandatory

Result path: Here you should indicate the complete path to the root folder (Figure 1).

Mass interval list: Here report the location (path) + name (including the file extension, *e.g.*, .txt) of your mass interval list. If the mass interval list is located in the same folder of the code you are using, reporting the name with its extension (without the path) is enough.

Minimum and maximum m/z to be analyzed: This are the minimum and maximum m/z value used for the analysis. The minimum value should be greater than 0 (*e.g.*, 0.1) and the maximum value should be slightly lower than the maximum *m/z* you extracted from your spectra (Figure 2). For example, if you extracted your spectra from an *m/z* of 0 to an *m/z* of 150, possible values to reported in the excel (for minimum and maximum) would be 0.1 and 149.9, respectively.

Step: This parameter determines the desired spectra resolution to be used for the PCA analysis. Smaller values mean longer running time but increased accuracy. A value I used (low to enhance accuracy as much as possible) is 0.0005 *m/z*, but you can play with this parameter to see the effect on the final results and on the running time.

Number of PCA components: This parameter determines the number of principal components (PCs) to be calculated. You can report values both ≥1 or <1. In the former case, the value will be read as the minimum % of variance that should be explained (the number of PCs will be defined as the minimal one needed to explain that % of variance or more). In the latter case, the number identifies the number of PCs to be calculated. **IMPORTANT:** The code will print an error if less than 4 PCs are given/found. Therefore, please report here a number ≥4 or, if reporting a number <1 you get an error (likely less than 4 PCs were enough to reach the demanded % of variance) please report 4 (or more) instead.

Running the code for the first time

This code has functionalities that at the moment cannot be included in the executable, therefore it should be run directly the code – but don't worry, it is really easy.

I advise you to use the Jupyter Notebook version of the code, as this will allow you to re-launch (if needed) only some part of the code and not all of it (saving time) and because the Jupyter version of the code offers some interactive plotting to check the goodness of each processing step (cropping and alignment, light enhancement, segmentation).

The easiest path for using Jupyter Notebook is probably installing Anaconda (containing Jupyter Notebook, among other things):

https://www.anaconda.com/products/distribution

After that, you can launch the Anaconda Navigator (just search it after the installation) and from there you can run Jupyter Notebook.

The libraries to be installed are the ones reported in the very top of the code ("import X" or "from X import Y"). It is very easy to install them (you can also google, library by library, how to install them, but typically it works for all in the same way):

https://youtu.be/Yr_ihLKq_yY

If you want to understand a bit better and learn more about Python (that can really be handy, so I definitely advise it!), you can look at the following videos series (focused on image analysis, but the first videos are very general on python):

https://youtu.be/7uE6hypji0o

Results

The PCA analysis allows to obtain two important information:

(i) The loadings: higher the loading higher the importance of this peak to explain the differences between your samples (if this does not sound clear to you, please look at https://youtu.be/FgakZw6K1QQ). An example of loading plot (bottom left) is reported in Figure 4. You will get graphs reporting the loading plots for all the peaks of your mass interval list (if you have many, this graph will be basically unreadable), as for the most important ones only (for PC_1, PC_2, PC_3, and PC_4) and excel files for the first PCs with the loading value for each peak;

(ii) Scores plot, allowing to visualize most of variance in your data in a small dimensionality space (one or a few 2D plots), allowing to identify groups (i.e., samples showing similar patterns/results) and discriminate among the different groups (if this does not sound clear to you, please look at https://youtu.be/FgakZw6K1QQ).. It will be up to you afterward to understand the underlying reasons for which certain samples are part of a group and others are part of another group. For this, the loading plots (to identify the most important peaks) and got back to the original spectra focusing on those peaks could be a good approach. An example of loading plot (bottom left) is reported in Figure 5. You will get score plots PC$_1$ *Vs* PC$_2$, PC$_1$ *Vs* PC$_3$, PC$_1$ *Vs* PC$_3$, PC$_2$ *Vs* PC$_3$, PC$_2$ *Vs* PC$_4$ and PC$_3$ *Vs* PC$_4$.

(iii) Another important information is the % of variance (*i.e.*, the differences between your samples) is described by each PC. For this, a devoted graph is created by the code (reported on the right side of both Figure 4 and 5.

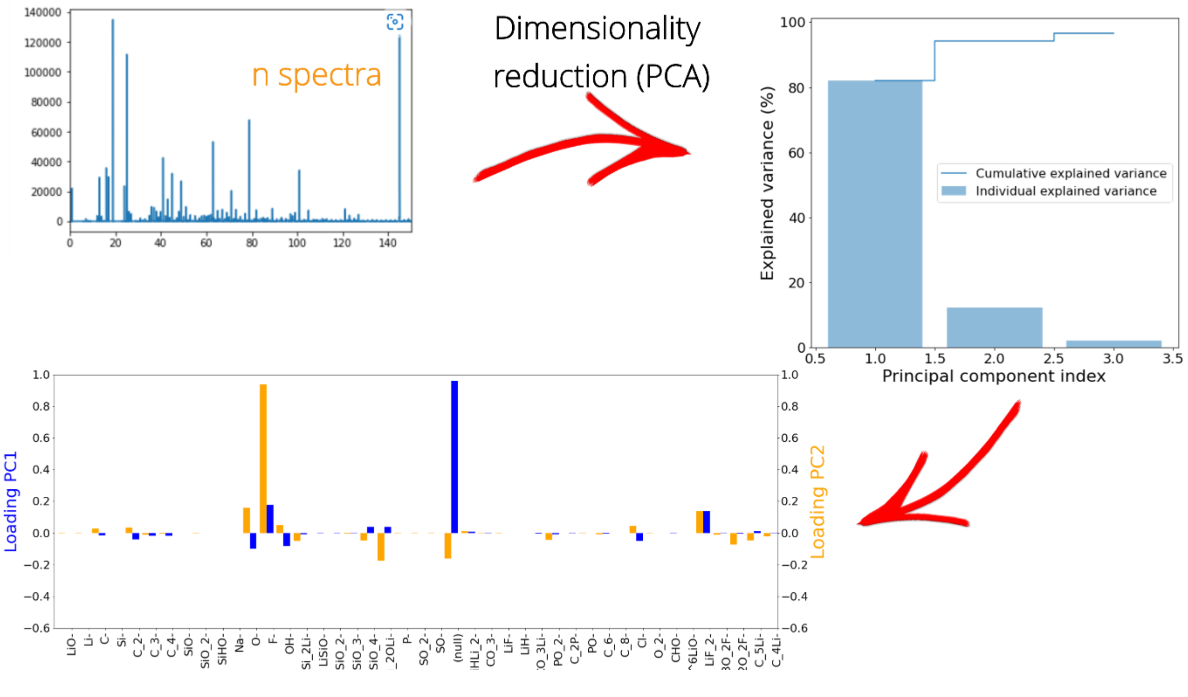


Figure 4. PCA procedure and example of results (loading plot for PC1 and PC2).

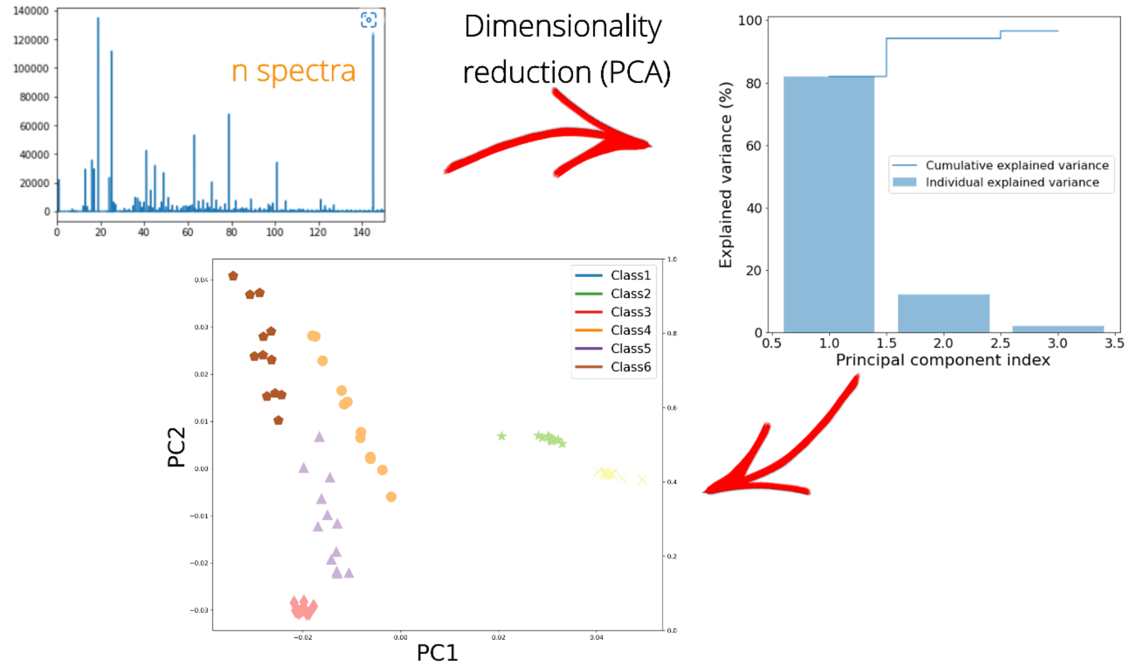


Figure 5. PCA procedure and example of results (score plot PC1 and PC2). For the score plot, please read “class i” (with i=1,2,3,4,5,6) represent name of your sample (the one you gave to the associated folder – Figure 1).

Contact for problems/doubts

If you have any problem using the code, feel free to contact me on my personal email: teo.lombardo3@gmail.com.

You can also reach me on LinkedIn or Twitter.