temDM

temDM MSA advanced version

Pavel Potapov

February 21, 2022

Use this manual in combination with the guide for the temDM MSA basic version. The present manual considers only the issues NOT COVERED in the previous manual.

Contents

1 Installation

Installation of the advanced version is not more difficult than that for the basic version of **temDM MSA**.

You should place the **temDM MSA.gtk** plugin into a plugins folder of DigitalMicrograph. There are typically several such folders. As the advanced version of **temDM MSA** is of restricted access, you are recommended to install it in the user-specific folder available only for you in a given PC. If import of Velox XEDS spectrum-images is desired, you should also place there an appropriated HDF5-reading plugin. These open-source plugins were compiled by Tore Niermann, TU Berlin :
hdf5_GMS2X_amd64.dll for the 64-bit system,
hdf5_GMS2X_x86.dll for the 32-bit system.

The script **find plugins folders.s** included in the distribution package will help you to localize the desired folders. Open **find plugins folders.s** in DigitalMicrograph and run it by pressing execute or by pressing ENTER with holding the CNTR key. Read the list of available plugins folders. The first folder in the list is typically the user-specific folder and therefore most appropriated for placing the temDM plugins. Read this folder path and drop **temDM MSA.gtk** there. If you cannot localize the user-specific plugins folder, then use any available plugins folder.

Some folders can be hidden in Windows. If you do not see all folders, make them visible in the file explorer of Windows 10:View tab - click hidden items checkbox. In some network-based systems, **find plugins folders.s** could fail to localize the user-specific DigitalMicrograph plugins folder. You might try to find them manually checking for the paths like
`.../users/user/AppData/Local`
`/VirtualStore/Program Files (x86)`
`/Gatan/DigitalMicrograph/Plugins`.

After you localize the appropriate plugins folder you should

- drop **temDM MSA.gtk** and **hdf5_GMS2X_amd64.dll** into the chosen plugins folder.
- restart DigitalMicrograph.

If you are using the most recent versions of GMS (Gatan Microscopy Suite), you might get the warning message "You have incompatible plugin *hdf5_GMS2X_amd64.dll*" during the start of DigitalMicrograph. That is because this dll was compiled for the older versions of GMS. In fact, the plugin works with any version. This message is just over-security and you can safely ignore it. Alternatively you can get rid of this boring message by removing *hdf5_GMS2X_amd64.dll* from the plugin folder. However you then would not be able to import Velox spectrum-images. The problem with the warning message will be fixed in the nearest future.

IMPORTANT: if you have already installed the basic version of ***temDM MSA****, you should OVERWRITE it with the plugin of the advanced version. This tip is also applicable for updating the version.* All versions of temDM MSA have the same name to avoid confusion with loading ambiguous commands. If you have several versions, it is recommended to keep plugins in individual folders with meaningful names like
`temDM MSA GMS3 advanced version 2_XX`.

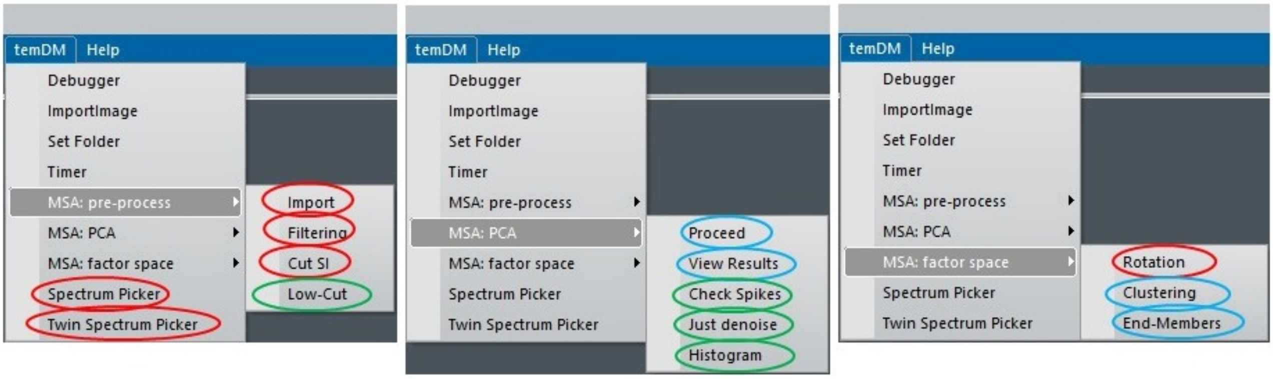
During the first start of the advanced version, the program will request you to input your individual installation code. You have to do it only once for a given PC.

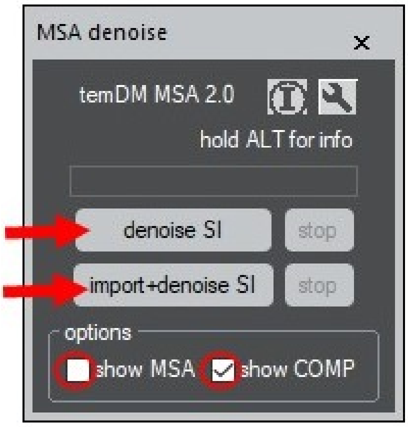
After successful installation, check the MSA menu in your DigitalMicrograph (see figure at the top of the next page). This resembles the menu for the basic version although some extra items appear. The new tools are outlined with green ellipses in the figure. The tools with the extended functionality are outlined in blue, and the tools that are exactly same as in the basic version are outlined in red.

2 Just denoise, no theory

This tool is designed for those who currently have no time or fun to learn all aspects of Multivariate Statistical Analysis (MSA) but just wish to denoise their spectrum-images without thinking much how it works exactly. You can test it with example **EELScube.dm3** included in the distribution package.

- Open the just denoise tool by choosing temDM - MSA: PCA -Just denoise,
- Having **EELScube.dm3** data cube in front press denoise SI. Wait until the denoised data cube appears. If you feel something goes wrong, you can interrupt the process at any moment by pressing stop.





The number of the PCA components used in reconstruction is determined automatically. At the end, you get the denoised spectrum-image. If you checked the show COMP box, you will also see an image called **EELScube COMP.dm3**. This consists of your denoised data cube in the *compressed* form. When saved, it takes quite little space in the disc. For decompression consult the Manual for the basic version, end of section “Basic PCA treatment”. You might also check the show MSA box, then the similar image but requiring much more memory image will appear. This could be only needed if you plan to do a lot of MSA treatment afterward.

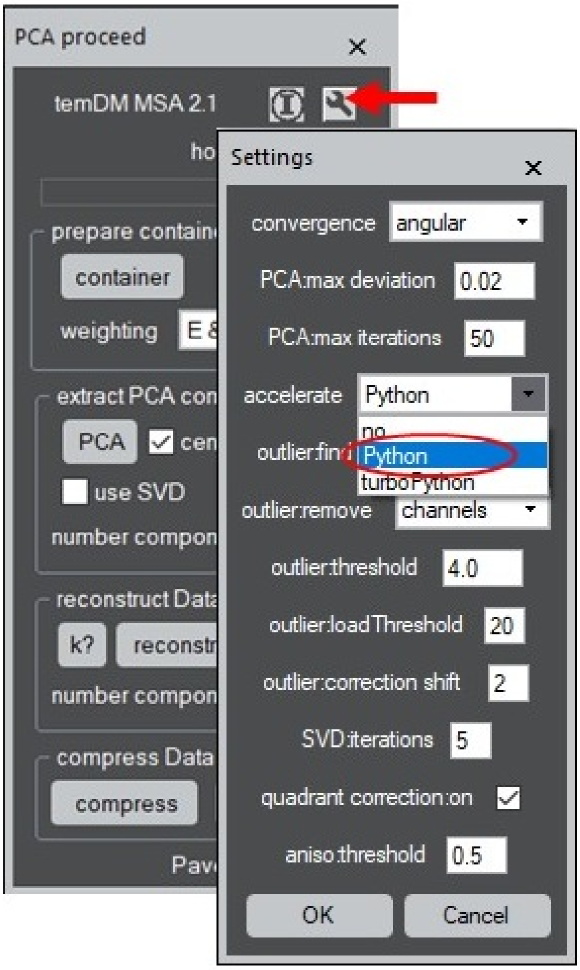
That’s all. You do not need necessarily to know how denoising works, just press a knob.

Even more: you can *import* an XEDS spectrum-image collected by Velox or Bruker *and then denoise* it in the non-stop way by the singe-click.

- press import+denoise SI.

Enjoy the fully automated processing flow! However, if this denoising works nicely for your spectrum-images, you are advised to learn a bit more from this manual and from the previous manual for the basic version. The automatic treatment is not always optimal, you probably could improve it.

3 Boost processing



convergence angular

PCA:max deviation 0.02

PCA:max iterations 50

accelerate Python

outlier:fin

outlier:remove channels

outlier:threshold 4.0

outlier:loadThreshold 20

outlier:correction shift 2

SVD:iterations 5

quadrant correction:on ☑

aniso:threshold 0.5

The most time consuming operation in the MSA analysis is extraction of principal components. Advanced version optimizes this process through the appended Python miniconda package. The speed is increased in approximately 3 times. Additionally one can choose the turbo mode, which boosts the speed up to 5 times. The acceleration mode can be chosen in the settings of the Proceed tool.

Turbo mode is the fastest although it implies a certain restriction on the extraction of principal components. In particular, the score outliers cannot be in-situ removed.

However, the *Gatan Microscopy Suite version higher than 3.4.0 is to be installed* in order to enjoy this speed boost. The "Python Support" should be enabled during the installation.

Except of principal components extraction, numerous other processing steps, e.g. in pre-treatment of EDX spectrum-images, are boosted.

4 Treatment of outliers

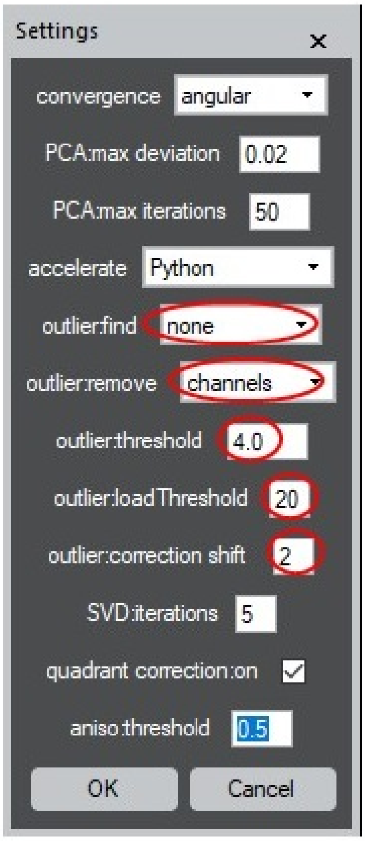
STEM data cubes nowadays consist of a huge number of data points. If few of them, for whatever reasons, are accidentally located far away from the main data distribution you have a problem. This is a typical case for EELS data where few of channels eventually show sharp peaks due to X-ray spikes. These outliers (they are sometimes named spikes) might be also observed in EDX data.

Even a singular outlier can make your PCA analysis completely wrong. Most cases when people “are disappointed with PCA...” arise exactly from this issue.

The advanced version of **temDM MSA** offers powerful algorithms for removing such outliers in the course of the PCA decomposition.

- Open the PCA proceed tool by choosing temDM - MSA: PCA - Proceed . Click the spanner icon. The Settings window will appear.
- In the drop-down menu outlier find choose the method for finding outliers : in score or in loading.

Finding outliers in score (recommended) means that at each iteration, the program removes the points deviating too strongly from the main distribution. The threshold value is indicated in the outlier:threshold field and is expressed in the fractions of the standard deviation (*sigma*) for a given component. The *sigma* threshold of *4.0* is default, but this is something you can align for specific datasets.


Settings

convergence: angular

PCA:max deviation: 0.02

PCA:max iterations: 50

accelerate: Python

outlier:find: none

outlier:remove: channels

outlier:threshold: 4.0

outlier:loadThreshold: 20

outlier:correction shift: 2

SVD:iterations: 5

quadrant correction:on ☑

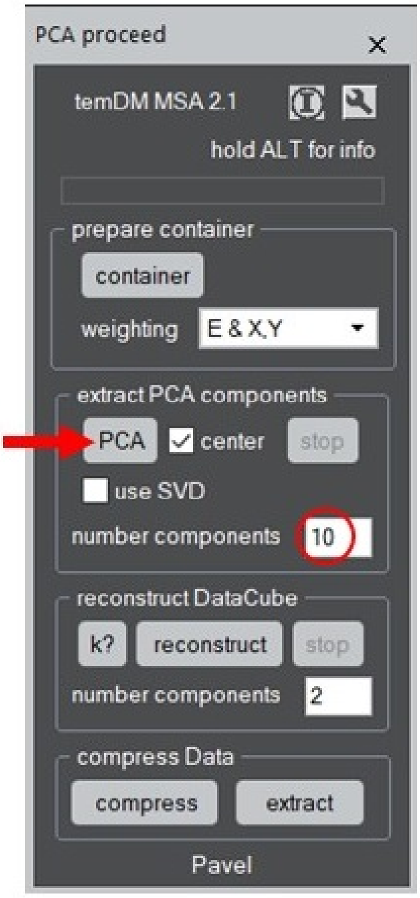
aniso:threshold: 0.5

The in score method of finding outliers is most natural for the PCA and delivers usually the best results. However, you may also choose finding outliers in loading. Then, “too sharp” features in the obtained loadings will be detected. How? The obtained loading is checked for the sharp variations. If the derivative somewhere exceeds a certain threshold, this channel is marked as an outlier. The threshold value is indicated in the outlier:LoadThreshold field and is expressed in the fractions of the standard deviation of the loading derivative for a given component.

The latter method is more or less a classical way of finding outliers in EELS data. The in score method is however more general and works nicely for EDX data as well.

Finding outliers must be followed by their removal. In the drop-down menu outlier remove, you find two options for correcting outliers: pixels or channels.

Consider first the channel correction. An outlier at a given channel will be corrected by shifting the neighboring channel at the place of an outlier. Both left and right neighbors will be shifted at the target place and averaged. It can happen that shifting the only one channel is insufficient. In particular, the X-ray spikes in EELS data usually extend over several channels according the camera point-spread function. You may control the shift value in the outlier correction shift field.


PCA proceed

temDM MSA 2.1

hold ALT for info

prepare container

weighting E & X,Y

extract PCA components

☑ center

□ use SVD

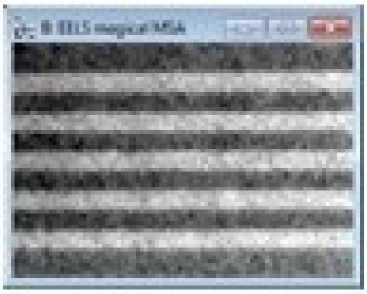
number components 10

reconstruct DataCube

number components 2

compress Data

Pavel



The pixel correction works in the exactly same way with the only difference that the neighboring pixels are shifted. If you think of your data as of a matrix with channels in the horizontal direction and pixels in the vertical one, then the channel correction corresponds to the horizontal shift of channels and the pixel correction corresponds to the vertical shift of pixels in the vicinity of outliers.

The outlier correction shift of **1** is usually sufficient for pixel correction while this should be increased for the channel correction in the case of EELS data.

Now you probably wish to check if correction for outlier makes any difference or not. This requires treatment of a real-life EELS data. In contrast to simulated ones, they *always* consist of some outliers.

- Open **MagiCal EELS** example included in the distribution package.
- Prepare **Magical EELS MSA.dm3** image by pressing container.
- Set first none in the outlier find drop-down menu of settings.
- Run PCA with extracting, say, 10 components.

You get the loadings of 10 components, the screeplot and the scatter plot. You may close the scatter plot but *do not close* the loadings and the screeplot. They will be needed for comparison with the next results. Just drag it somewhere away at your working place or minimize to bottom.

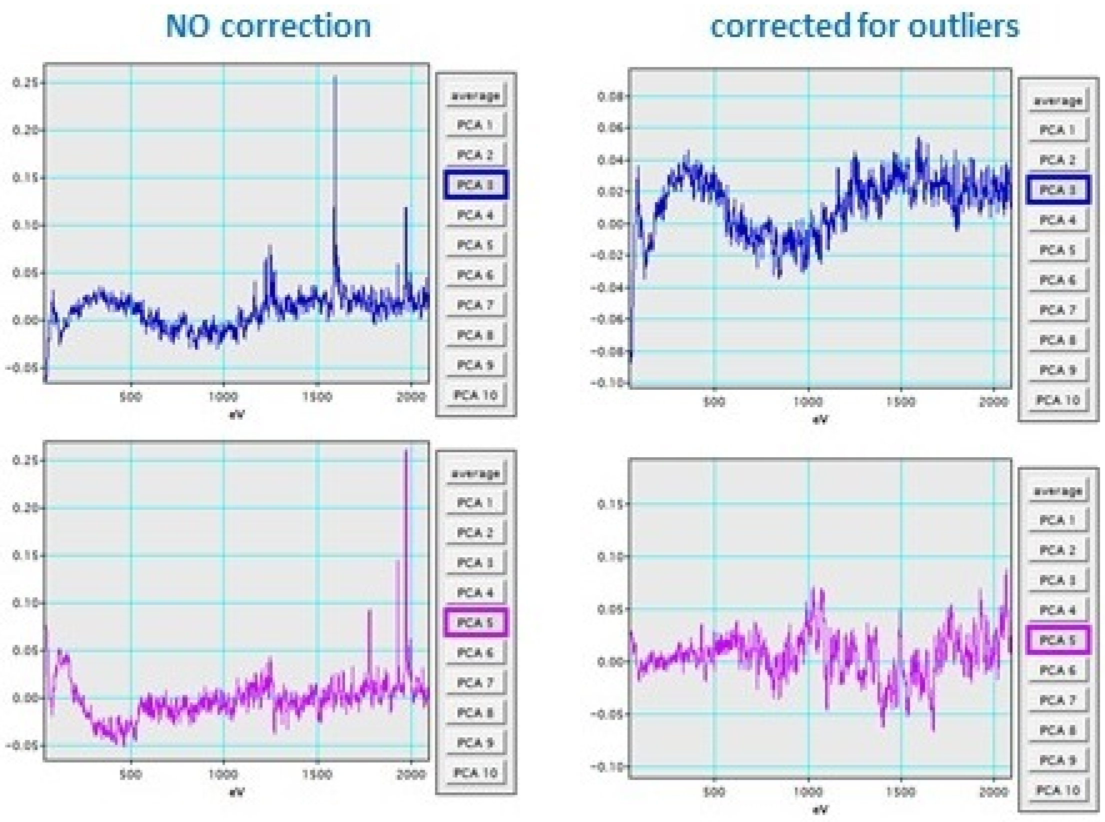
Now do the same treatment *with* correction for outliers .

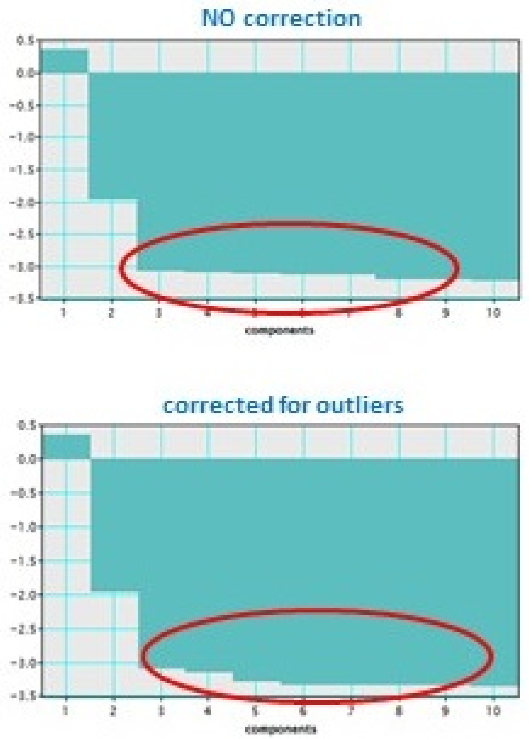
- Set in score in the outlier find drop-down menu of settings.
- Set channels in the outlier remove drop-down menu of settings.
- Set outlier correction shift **2**.
- Run PCA with extracting 10 components.

Now compare the resulted loadings and the scree plot with those extracted previously.

First two PCA components are identical but the components of the higher index are different. Look at the scree plots and you discover that the variance is noticeably higher for the no-outlier treatment. This is however the wrong values - few occasional outliers strongly affect the observable variance and this has little to do with the meaningful variations.

If you think that the error is insignificant, look at the loadings. In the 3rd component, you clearly see the traces of X-ray spikes in one case while they are sealed in another one. More dramatic: the loading of the 5th component is not only contaminated with spikes but is overall wrong when compared with the outliers-sealed treatment. This is because PCA is not able to find the right solution if spikes are comparable with the meaningful data variation

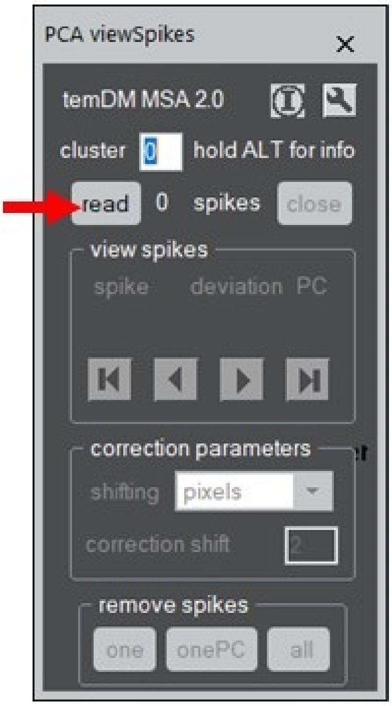




If you still think that correction for outliers is a minor stuff, continue to explore your real-life data examples. You will soon realize that outliers can easily distort even the first major PCA component and make your PCA work senseless.

You might also consult [1], section 4.2 to see how outliers erode your “true” components.

5 Permanent removal of outliers



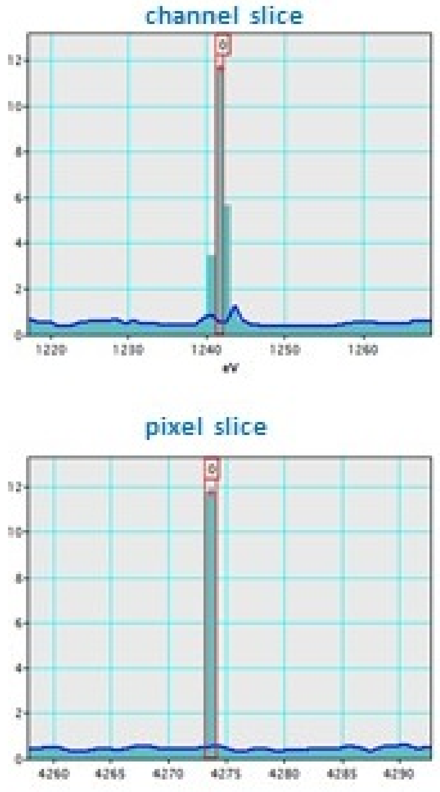
cluster 0

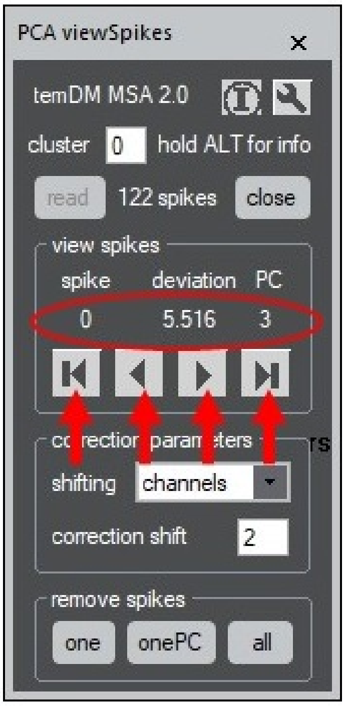
0 spikes

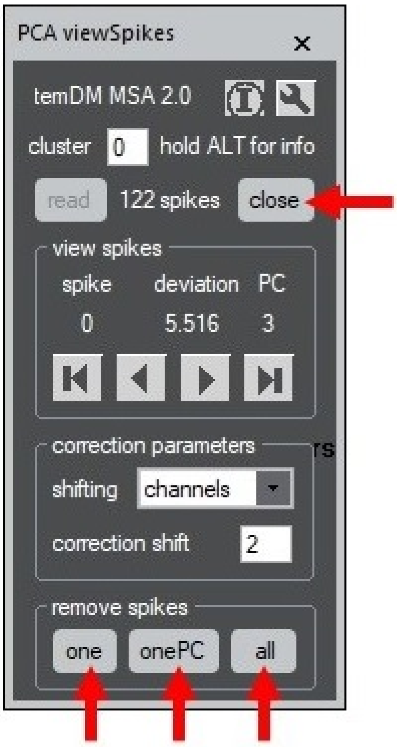
shifting pixels

correction shift 2

In the course of PCA, **temDM MSA** removes out-







liers from the copy of the data matrix. This is implemented in order to minimize the irreversible actions with your original data. Thus, you can play with parameters of the outliers treatment and run PCA again and again - every time outliers will be searched in and removed from the fresh, original data.

However, if you are sure in your treatment, you may remove the outliers permanently.

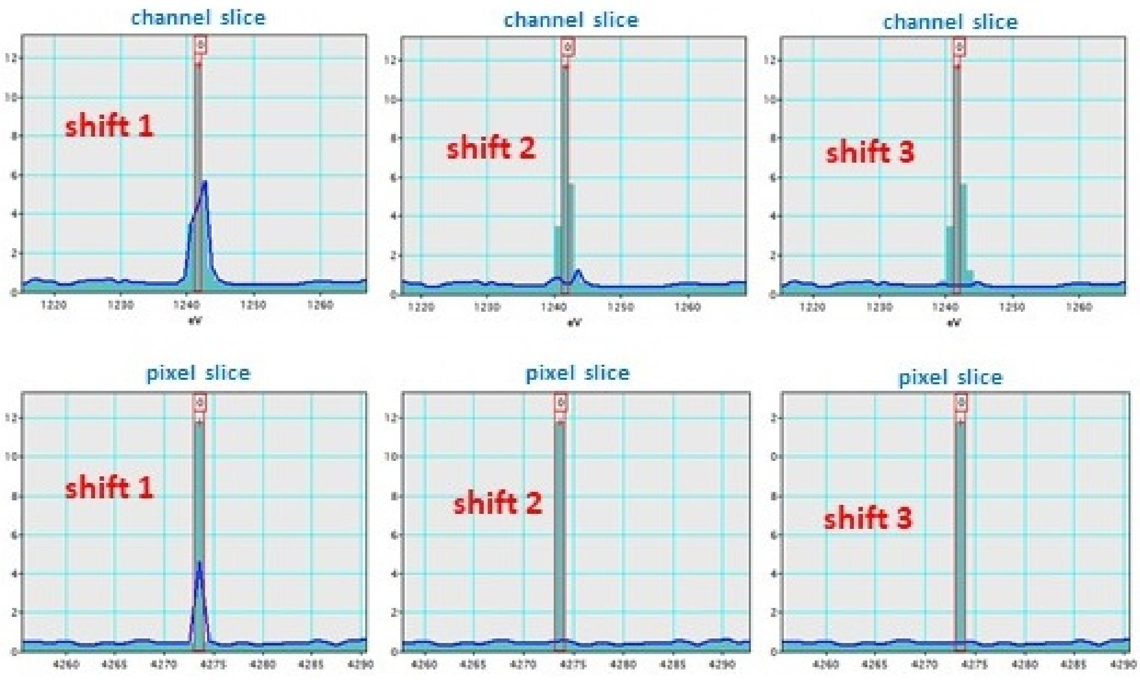
- Open the viewSpikes tool by choosing temDM - MSA: PCA - Check Spikes.
- Having the **Magical EELS MSA.dm3** image in front, click read.

Two displays will appear: one showing the pixel slice near the outlier, another showing the energy slice near the outlier.

Adjust the horizontal and vertical scales in both slides to see clearly your outlier, which is marked by red. The filled profile shows the original data and the blue curve is the suggested correction.

Note that you have the direct access to the correction parameters in this tool. You can play with them and see immediately the result in a given outliers. The global correction parameters are however not changed. To set your global correction parameters click the spanner icon.

It is instructive to learn how your correction parameters affect the results. Play with shifting pixels or channels and the correction shift value. The blue correction curve will change live when changing correction shift (if it is not, press Enter or click on the slice display to set the new number in action). The figure at the top of the next page shows a typical effect of the shift value



on sealing the outliers.

That was the first detected outlier. There are however many more outliers.

- Scroll through all detected outliers by clicking play and playback buttons.

Also you can quickly jump to the outliers detected in the other PCA components by pressing the corresponding most-left or most-right scroll buttons. Note that some basic information about a given outlier is displayed in the tool: the spike's index, the deviation from the mean, the PCA component where it was detected.

Up to now you were only inspecting the outliers not changing it. You can anytime finish your inspection by pressing the close button. Data are unchanged.

On the other hand, if you are sure that a given outlier is well corrected, you can press the button one at the bottom of the tool. Then, this outlier will be *permanently* removed from the dataset. You can also click the onePCA button to remove *all* outliers detected in the currently displayed PCA component.

And, most ultimately, you can press the all button. Then all the detected outliers will be removed from the dataset at once.

Now you carefully sorted out and removed the outliers detected during your previous PCA procedure. You might run PCA again - this time without correction for outliers.

- Set none in the outlier find option in the Settings of the PCA proceed tool. Run PCA again.

Hopefully the results would be more accurate because you manually corrected the outliers in the most optimal way.

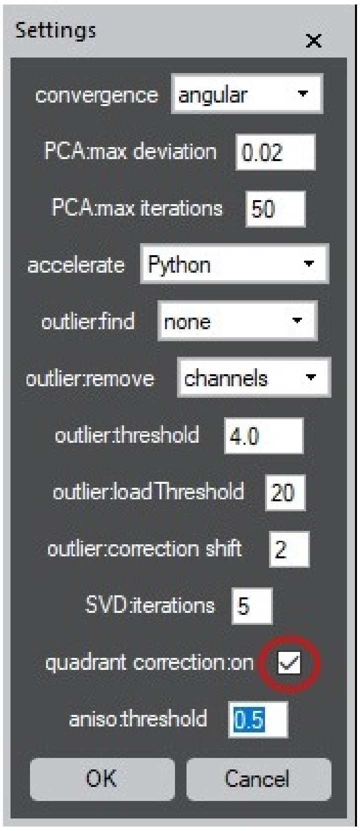
6 Correcting quadrant gain of GIF camera

Tridiem and Quantum GIF cameras sometimes suffer of unequal gain among the camera quadrants. This creates an offset in the EELS spectrum around pixel 1024 (middle of a 2K camera). Such artifacts might show up in the loadings of minor PCA components. More importanly: spontaneous variation of the quadrant gain can distort the PCA analysis ! The advanced version of **temDM MSA** employs a special algorithm to correct for that.

- Open the PCA proceed tool by choosing temDM - MSA: PCA - Proceed. Click the spanner icon.

- Check quadrant correction:on box.

When this option is activated, the PCA loadings are automatically corrected to compensate for the quadrant unequal gain.


Settings

convergence: angular

PCA:max deviation: 0.02

PCA:max iterations: 50

accelerate: Python

outlier:find: none

outlier:remove: channels

outlier:threshold: 4.0

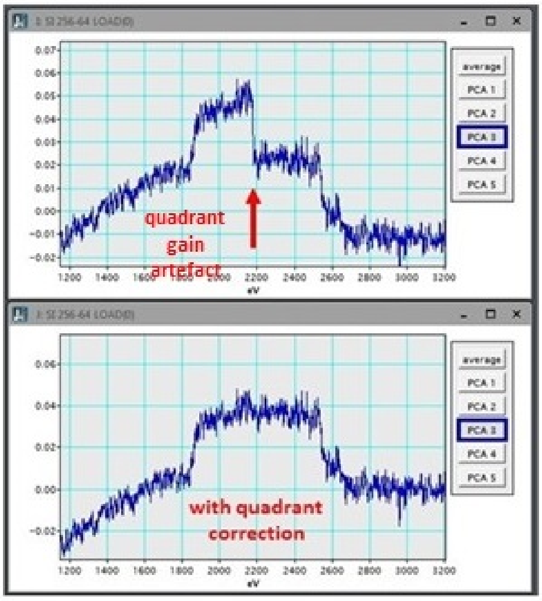
outlier:loadThreshold: 20

outlier:correction shift: 2

SVD:iterations: 5

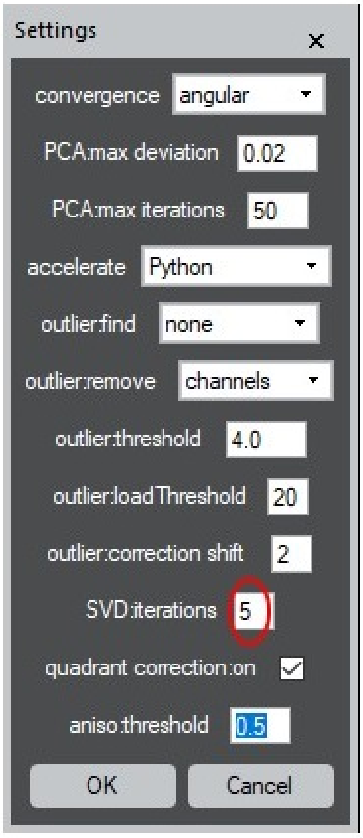
quadrant correction:on: ☑

aniso:threshold: 0.5



7 Finding PCA components with SVD

The default method for finding PCA components in **temDM MSA** is the NIPALS algorithm that extracts components sequentially every time capturing a component with the highest variance. This allows to find sufficiently accurately 10-20 components within a reasonably short time. Usually, 10-20 components are enough to pick up all your meaningful data variations.


Settings

convergence: angular

PCA:max deviation: 0.02

PCA:max iterations: 50

accelerate: Python

outlier:find: none

outlier:remove: channels

outlier:threshold: 4.0

outlier:loadThreshold: 20

outlier:correction shift: 2

SVD:iterations: 5

quadrant correction:on: ☑

aniso:threshold: 0.5

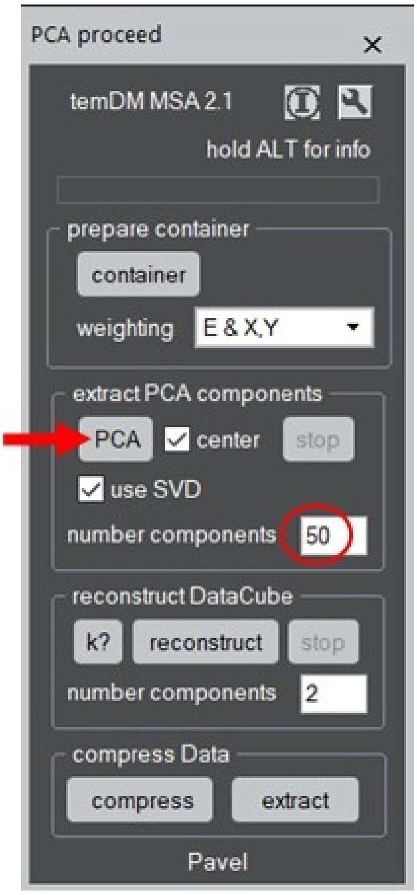
However, sometimes you need a larger number of principal components, for instance, to investigate how the noise variance behaves with increasing the index of a component. Doing that with the NIPALS algorithm would be unreasonably time consuming.

The component's loadings are actually the singular vectors of a given data matrix, thus, it is possible to extract them all at once by the Singular Value Decomposition (SVD).

The advanced version of **temDM MSA** includes the standard SVD algorithm that extracts all principal components available in a given dataset (for instance, if there are 2048 energy channels in your spectrum-image, you can extract 2048 principal components). Note that you should install the version of Gatan Microscopy Suite higher than 3.4.0 to enjoy this option. The algorithm is quite fast but the accuracy of the extracted components might be worse than that in the case of the NIPALS algorithm.

As a compromise between the full SVD and NIPALS, the plugin suggests also the original in-house developed algorithm for the truncated SVD. This algorithm extracts the specified number of largest principal components (typically 50-100)while you can tune their accuracy by playing with the number of iterations. The default number (5) provides the high accuracy while still keeping the calculation time in the reasonable range.

- Open the PCA proceed tool by choosing temDM - MSA: PCA - Proceed.
- Open the simulated example **EELScube.dm3** and prepare the MSA image **EELScube MSA.dm3** by clicking container if it was not prepared yet.
- Click the box use SVD.
- Press PCA.
- Choose between *truncated* and *complete* SVD. The former will proceed longer and extract less components but their precision will be higher.



PCA proceed ×

temDM MSA 2.1

hold ALT for info

prepare container

container

weighting E & X,Y

extract PCA components

PCA ☑ center stop

☑ use SVD

number components 50

reconstruct DataCube

k? reconstruct stop

number components 2

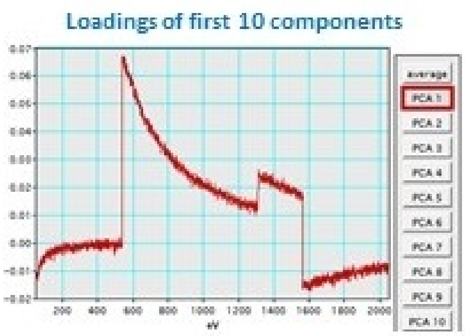
compress Data

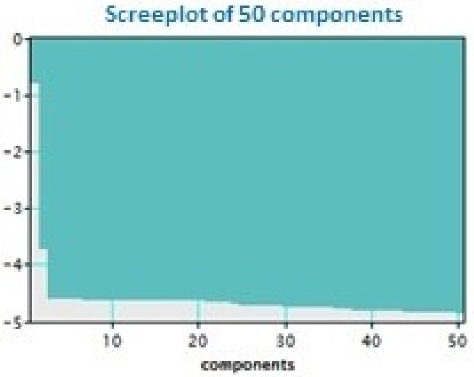
compress extract

Pavel

A good tip about usage of the NIPALS vs SVD algorithms:

If you are unsure how many components you need, start with NIPALS. You can extract only few major components at the beginning and then gradually increase the number of components until you get a clear idea about your data set. There is no need to overwrite the previously found components.





If you know in advance that you need many components (to characterize the noise in your data, for instance) - run SVD.

At the end, you see the scree plot, the scatterplot and the loadings similar to what was obtained with NIPALS, but now there are 50 PCA components extracted with SVD.

From the table below you get an idea about the time consumed for different algorithms. This was obtained by treating a 100x100x2000 spectrum-image in GMS3.4.0 using an i5-3470 processor.

Table 1 – Comparative performance of the NIPALS and SVD algorithms

NIPALS	50 components	220 sec
NIPALS boosted	50 components	102 sec
truncated SVD	50 components	106 sec
truncated SVD	100 components	245 sec
complete SVD	2000 components	35 sec

What should be stressed: *the SVD algorithms DO NOT take care about outliers.*

Because of that, the best strategy is to find the first 5-10 PCA components by the NIPALS algorithm with activating the outliers treatment. All pronounced outliers will be hopefully removed at this stage. Then, you might run truncated or complete SVD for the higher number of components *without overwriting* the first results. The program will take all available (NIPALS-found) components, remove all NIPALS-detected outliers and continue to extract further PCA components with SVD. The **temDM MSA** program allows to extract PCA components incrementally, without overwriting the previously found components. This nice feature is readily applied to the combination of the NIPALS and SVD algorithms. But you should keep in mind that the precision of these algorithms depends on different parameters, thus a slight mismatch at the border NIPALS-found and SVD-found components might be observed.

You know how to adjust the precision of NIPALS (see the manual of basic version, p.22 bottom). Learn how to tune the precision of SVD:

- Click the spanner icon to open Settings of proceed PCA tool.

The parameters PCA max deviation and PCA max iterations would affect the precision of PCA in the similar way as that in NIPALS. However, the precision might slightly differ quantitatively because they are different algorithms. If you need to adjust the precision of neighboring PCA components obtained by NIPALS and SVD, just play a bit with the parameters.

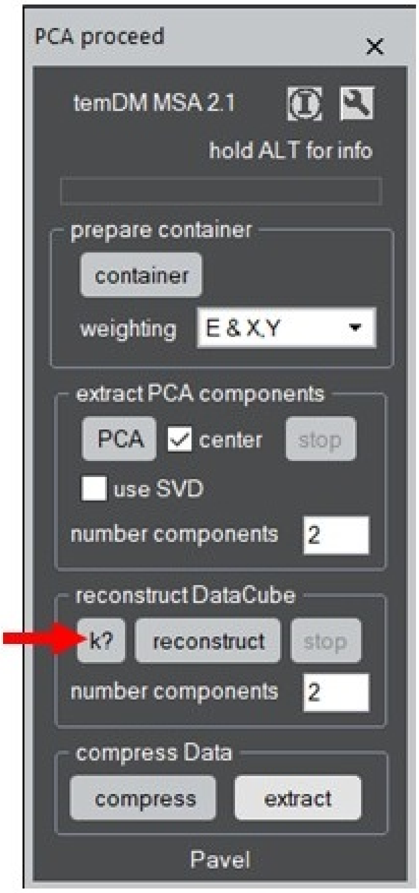
Additionally, there is one extra parameter affecting the precision of SVD - SVD iterations. You might guess that the higher number of iterations delivers the better precision at expense of the higher consumed time.

8 Adviser

This feature will help you to determine the right number of PCA components for reconstruction. Usually people look at the scree plot (dependence of the variance on the component index) trying to guess where the noise region starts. It might work but you should be aware that this is a *very subjective* style of treatment. To stay on the objective basis, **temDM MSA** uses a couple of in-house developed algorithms.

To learn how it works, open your previously prepared example **EELScube MSA.dm3**. Suppose you have already extracted the sufficient number of PCA components and are now deciding how many of them will be used in reconstruction.

- Open the PCA proceed tool by choosing temDM - MSA:PCA - Proceed.



weighting E & X,Y

☑ center

☐ use SVD

number components 2

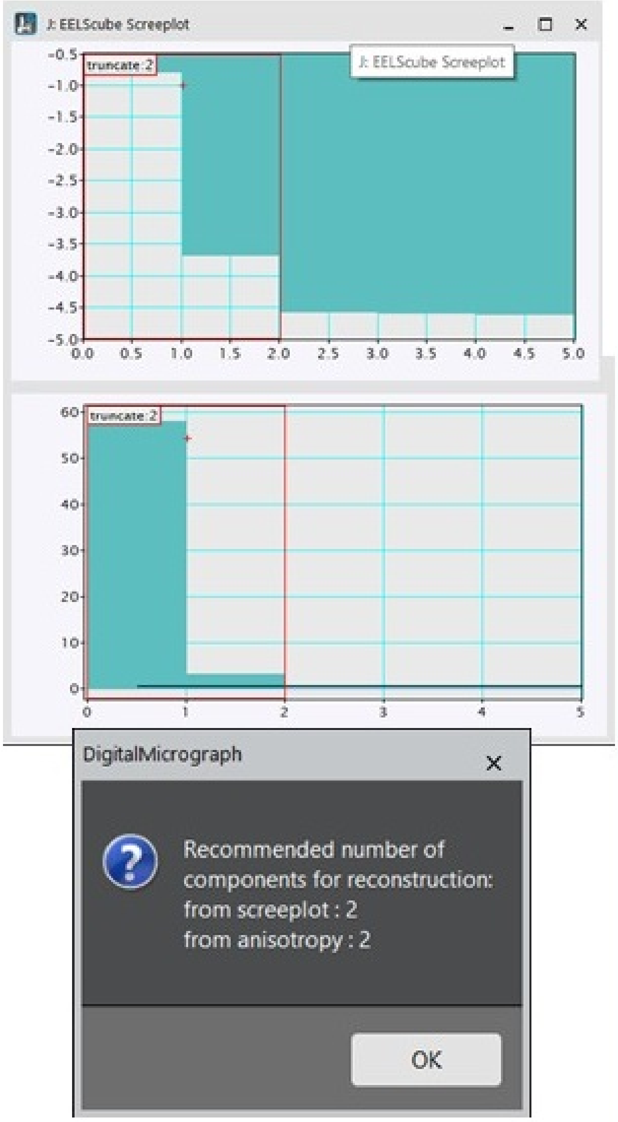
number components 2

- Click the knob **k ?** in the reconstruct data cube box. Then, you get an advice based on two methods.

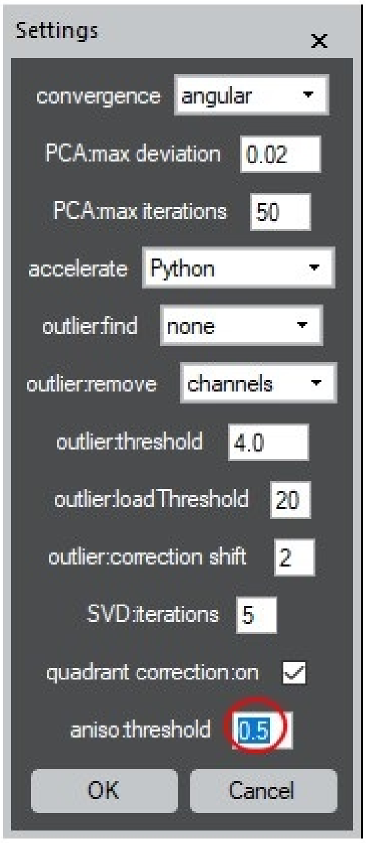
The first method is the analysis of a screeplot by the original algorithm, which localizes up the inflection point between the meaningful (high variance) and noise (low variance) components. The second method employs the analysis of anisotropy plots. What is that?

Noise might have different nature - Poissonian, Gaussianian, mixed; but there is a property that should be conserved in any dataset. The uncorrelated noise must be isotropic in all directions of the factor space. To isolate the noise region, **temDM MSA** calculates the anisotropy of the consequent couples of PCA scatter plots and catches the border component where the anisotropy decays below a certain predefined threshold. All components with the anisotropy value below this threshold will be considered as isotropic, i.e. representing pure noise. The advisor in **temDM MSA** uses the Bayesian approach to find out the most probable border between meaningful (anisotropic) and noise (isotropic) components.

After adviser proceeds, you get the recommended number of the truncated principal components evalu-



Recommended number of components for reconstruction:
from screeplot : 2
from anisotropy : 2



convergence: angular
PCA:max deviation: 0.02
PCA:max iterations: 50
accelerate: Python
outlier:find: none
outlier:remove: channels
outlier:threshold: 4.0
outlier:loadThreshold: 20
outlier:correction shift: 2
SVD:iterations: 5
quadrant correction:on ☑
aniso:threshold: 0.5

ated by the two methods. Also, the screeplot and the anisotropy plot are displayed for your information.

The anisotropy plot can be also generated using the view results tool.

- Open the view results tool by choosing temDM - MSA:PCA - View Results.
- Click the knob anisotropy while having the **EELScube MSA.dm3** image in front.

More about this method can be learned from [2, 3]

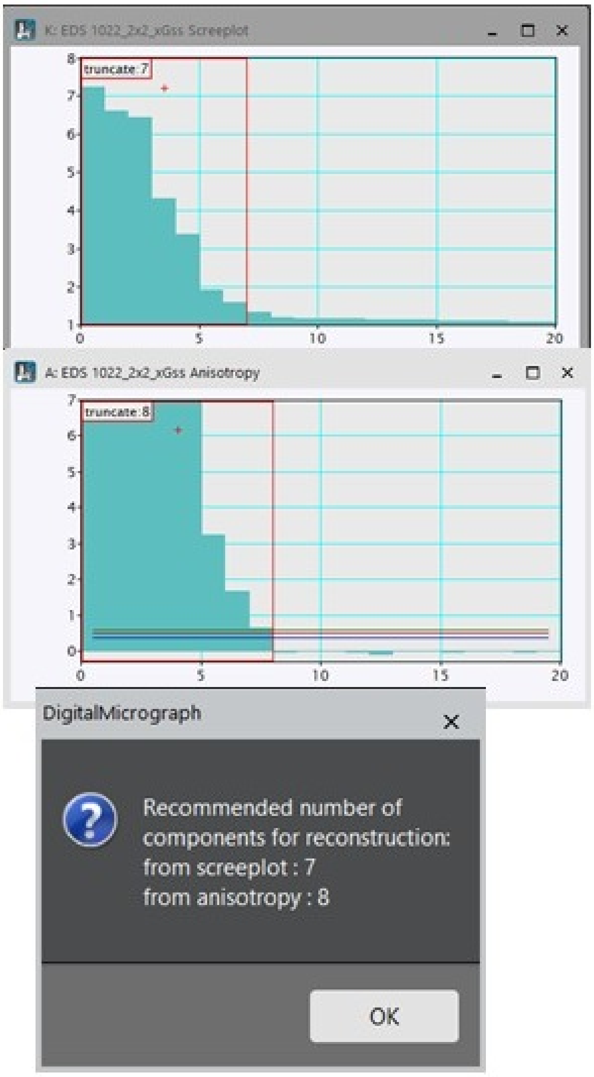
Well, the **EELScube MSA.dm3** dataset was a simplified synthetic example. In fact, you did not need the adviser to realize that this was a two-componential object. The real objects might be more complicated. The next example shows the situation where the screeplot and anisotropy methods advice slightly different numbers of components to truncate (7 vs 8). The final choice must be yours.

You can access the settings for the anisotropy method by clicking the spanner icon. The tuned parameter is:aniso:threshold. All scatter plots with the anisotropy below this value will be considered as *isotropic*. The reasonable number for the threshold is **0.5-1.0**. The lower is this parameters, more marginal components will be accounted for. But it is not recommended to set it below 0.5 because then the noise can be accidentally classified as a meaningful variation. The tuning parameters for screeplot auto-truncation are not recommended to change, thus they are in the more deeply hidden tags.

9 Setting of “Just denoise”

You are now well prepared to review the settings of the just denoise tool. Click the spanner icon at the right-upper corner of the tool and see what you can tune.

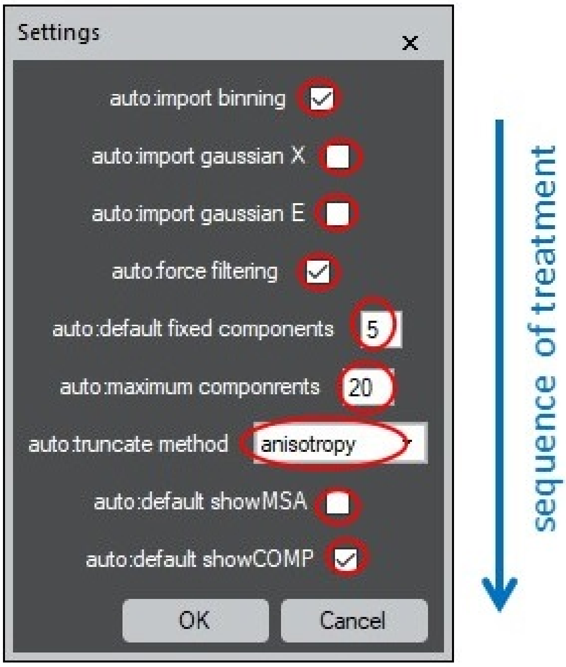
The list of setting reflects the typical sequence of treatment. The first three parameters tell how to treat imported XEDS data.
import binning: check this box if you want to bin an imported data cube. The default binning values from


K: EDS 1022_2x2_xGss Screeplot
truncate: 7
8
7
6
5
4
3
2
1
0
5
10
15
20
A: EDS 1022_2x2_xGss Anisotropy
truncate: 8
7
6
5
4
3
2
1
0
0
5
10
15
20

DigitalMicrograph

Recommended number of components for reconstruction:
from screeplot : 7
from anisotropy : 8

OK



Settings

auto:import binning ☑
auto:import gaussian X □
auto:import gaussian E □
auto:force filtering ☑
auto:default fixed components 5
auto:maximum componrents 20
auto:truncate method anisotropy
auto:default showMSA □
auto:default showCOMP ☑

OK Cancel

the settings of the filter tool will be used.

import gaussian X: check this box if you want to apply Gaussian smoothing in spatial directions. The default *sigma* value from the settings of the filter tool will be used.
import gaussian E: check this box if you want to apply Gaussian smoothing in the energy direction. The default *sigma* value from the settings of the filter tool will be used.

It should be stressed that in the case of very sparse data, the program might *force* binning or smoothing even if you unclicked these boxes. This is necessary because the optimal amount of the components cannot be determined in sparse datasets otherwise. If you ultimately want to avoid binning or smoothing, proceed the treatment flow manually.

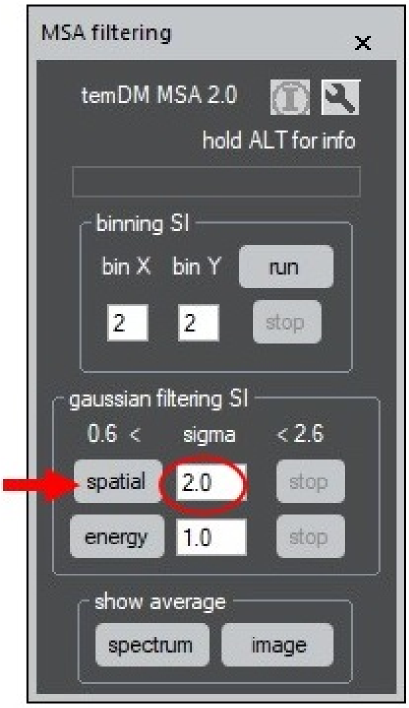
The next parameters relate with finding PCA components and reconstructing denoised data.
truncate method: For the moment, the only anisotropy method is used to automatically determine the optimal number of components for reconstruction. However, you can change it for fixed, then the fixed number of components will be always used. This makes sense if you treat a lot of very similar datasets.
fixed components: the number of components used for reconstruction if the fixed truncate method is chosen.
maximum components: This is the maximal number of the PCA components extracted by the NIPALS method. If for any reasons, the automatic truncation method fails, this number of components will be also used for reconstruction.

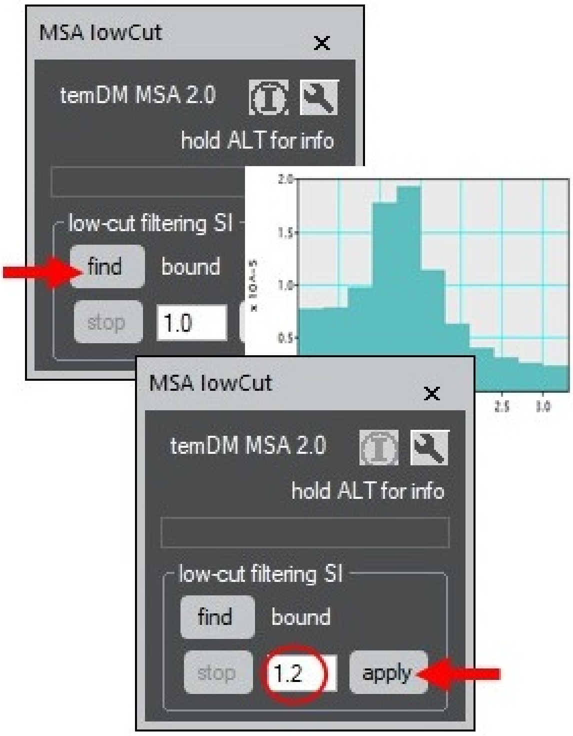
The last two check boxes control appearance of images consisting of all the PCA results like loadings, scores et cet. in the full or compact form. If you are interested in the denoised data cube only, unclick these boxes.
default show MSA: check this box if you wish to see the **XXX MSA.dm3** image.
default show COMP: check this box if you wish to see the compressed **XXX COMP.dm3** image.

10 Low-cut filter

This tool is useful for special spectrum-images only. Namely, i) this must be an XEDS data cube, ii) low-cut filtering makes sense for very noisy data cubes only.

Lets learn how it works with an example **GaAs-110** included in the distribution package. This is a learning example representing extremely noisy data. If you are interested in an idea behind the extraction of elemental maps from such hopeless sets, please consult [4]. Here there is just a straight sequence of the steps:





- Open the filtering tool by choosing temDM - MSA: pre-process - Filtering.

- Having the raw data cube **GaAs-110.dm3** in front, click spatial. Use the default sigma of **1.0**. The resulted **GaAs-110_hGss.dm3** data cube represents the Gaussian-smeared data.

- Open the low-cut tool by choosing temDM - MSA: pre-process - low-Cut.

- Having **GaAs-110_hGss.dm3** in front, click find.

  A graph appears showing the dependence of some characteristic -low-bound - (see the reference above) on the low-bond value. This operation might last a bit. You may stop it anytime by pressing stop . The optimal low-bound corresponds to the maximum on the graph. In our example, this is **1.2** which is automatically inserted in the corresponding field.

- Press apply. The low-bound value of 1.2 will be applied to cube **GaAs-110_hGss.dm3** and the filtered data cube **GaAs-110_hGss_LB.dm3** will appear.

Why so much pre-processing? This smoothing is an obligatory treatment for noisy XEDS data [3], otherwise the forthcoming PCA would fail.

You see that we have filtered the raw data twice - first with the Gaussian smoothing and second with the low-cut filter, which is an analogue of a top-hat filter cutting off some signal of low intensity.

The Gaussian filter is quite safe - in 90% cases it improves the EDX data cubes while keeping the reasonable spatial resolution.

The low-cut filter is somehow more invasive. It is recommended to use in the rare case when the other means deliver only noise.

Important: the Low-cut filter may be applied ONLY after Gaussian filtering!

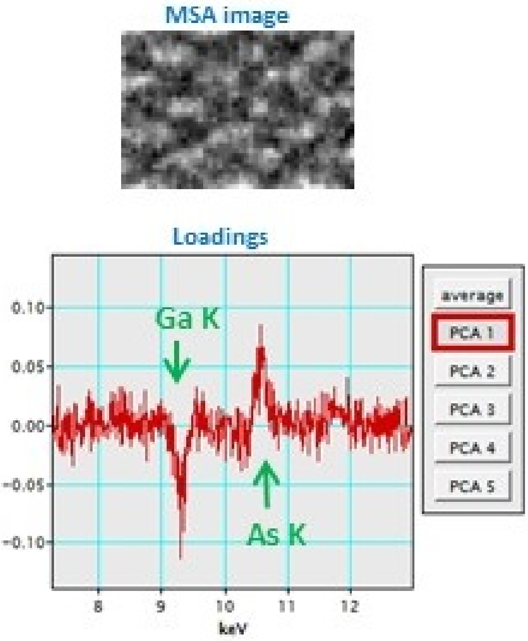
Now you can make PCA on the filtered cube.

- Open the PCA proceed tool by choosing temDM - MSA: PCA - Proceed.

- Having **GaAs-110_hGss.dm3** front-most, press container button. The MSA container will be prepared.

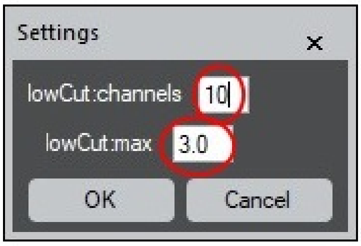
- Choose 5 PCA components in the PCA proceed tool. Having **GaAs-110_hGss MSA.dm3** in front, press PCA .

The resulted PCA loadings show the anticorrelated *Ga K* and *As K* peaks in the 1st component. This is sufficient for building the reasonable Ga and As elemental maps.



Quite complicated treatment ? Yes, but this is a very noisy data. The other treatment strategies would most probably deliver nothing but white noise in this case.

Through the spanner icon you have access to some settings of the low-cut tool.


Settings

lowCut:channels 10

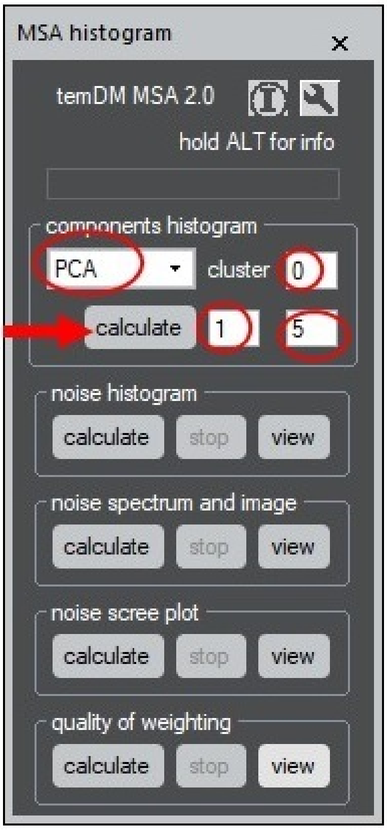
lowCut:max 3.0

Those are:
find low-bound: channels - the number of channels for searching the optimal low-bond value or, say, the resolution of search. 10 channels are usually quite sufficient.
find low-bond: max - the maximal value for the searched low-bound values. It is measured in the fractions of the average signal.

11 Histogram

This tool offers you a plenty of options for statistical characterization of the obtained PCA results. Test it with the example **EELScube** included in the distribution package.

- Open the histogram tool by choosing temDM - MSA: PCA - Histogram in the DigitalMicrograph menu.
- Having **EELScube MSA.dm3** in front, calculate the histogram of the retrieved PCA components by clicking calculate in the components histogram box.


MSA histogram

temDM MSA 2.0

hold ALT for info

components histogram

PCA cluster 0

calculate 1 5

noise histogram

noise spectrum and image

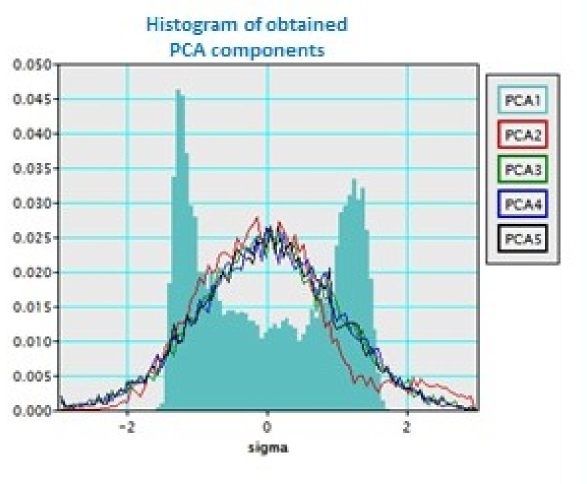
noise scree plot

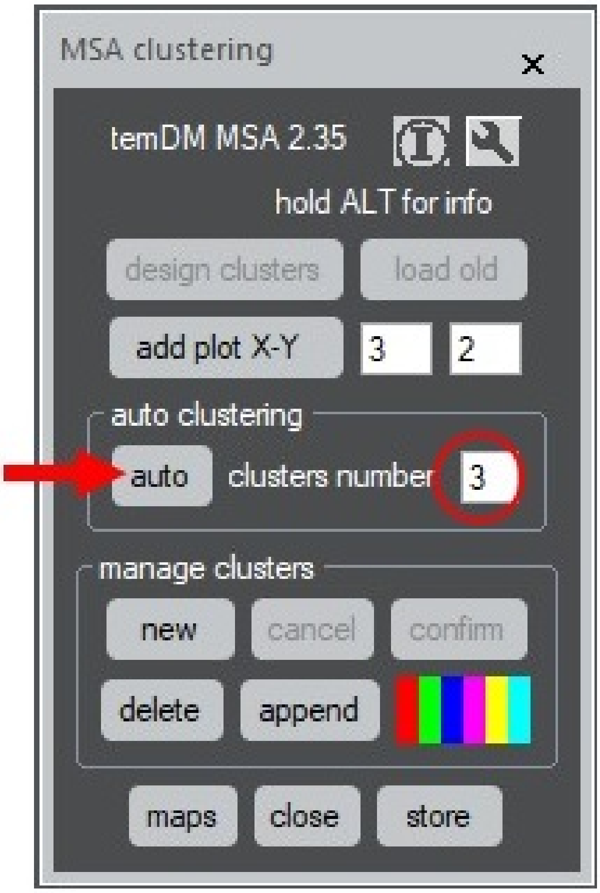
quality of weighting

You can choose the range of the components by changing the numbers in the corresponding fields. In case you clustered your data and proceeded PCA in the clusters, you can treat the clustered components by indicating the desired cluster's number instead of cluster "0" (as before, 'cluster "0" means NO clustering). If you performed the rotation of the PCA components, you may calculate histograms of the rotated components. Just choose ICA, varimax or free in the drop down menu.

You get the histograms of all desired components. The X-axis displays the obtained values normalized to the standard deviation (*sigma*) of the data distribution in each component. The Y-axis shows the frequency of observation. The total area under the distribution is normalized to **1**.

You can also calculate the histogram of the noise (the stuff that remains after you subtracted all available PCA components is assumed to be a noise).

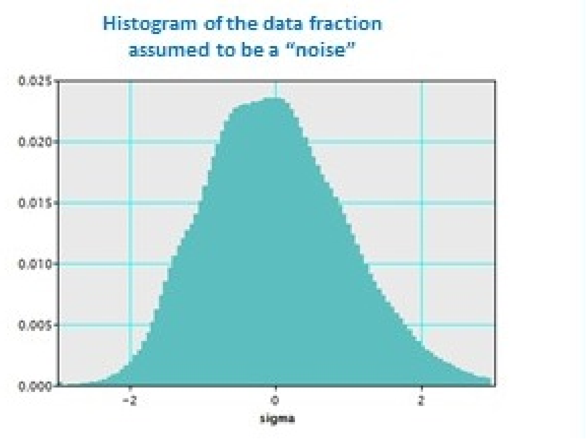




- Having **EELScube MSA.dm3** in front, click calculate in the noise histogram box.

This calculation requires the reconstruction of the data from all the available PCA components, thus it might take some time. You may anytime halt the calculation by pressing stop.

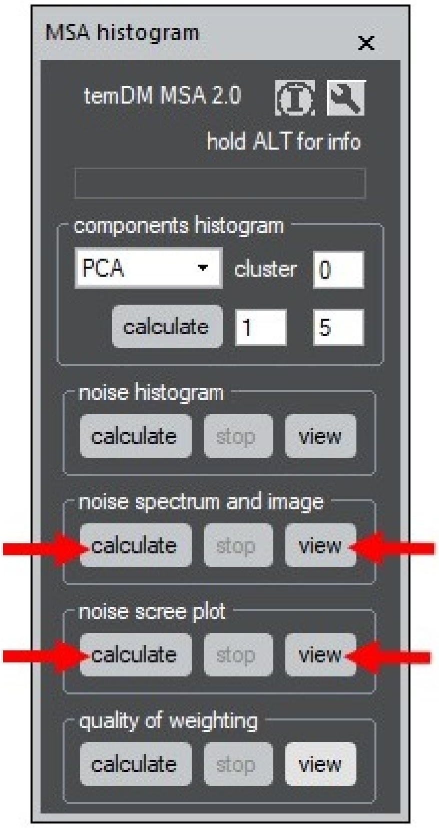
Since you have calculated it once, the histogram is stored in **EELScube MSA.dm3**. If you want to see it again, no need to spend time repeating the calculations.



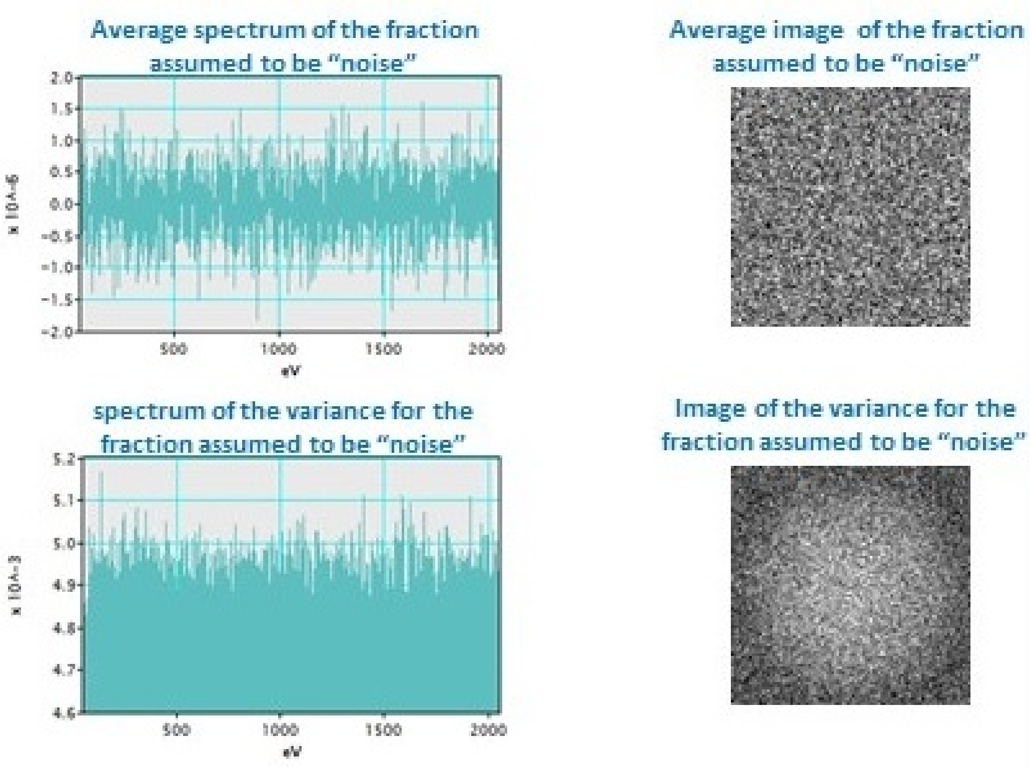
- just click the view button. The histogram appears immediately.

It is very important to evaluate carefully the data fraction remained after extraction of all available PCA components. It is assumed to be a noise, but is it really? Maybe you did not reach the real noise level yet? Maybe you must extract more PCA components ?

The histogram tool delivers not only a histogram but more information about your noise fraction helping you to understand whether there is still some useful information buried in the "noise".



- Having **EELScube MSA.dm3** in front, click cal-



culate in the noise spectrum & image box.

You get the datacube consisting of residuals after your reconstruction. Also, the following spectra will be displayed:
the average spectrum of your residuals ,
the average image of your residuals,
the spectrum of the variance of the "noise",
the image of the variance of the "noise".

You can anytime halt the calculation by pressing stop.

Since you have calculated these images once, they are stored in **EELScube MSA.dm3**. When you want to see them again, no need to spend time repeating the calculations. Just click the view button. The images appear immediately. However, the datacube of residuals is not stored in **EELScube MSA.dm3** as it would take too much memory. To see this datacube you have to run calculate again.

Wish to know even more about your "noise" ?

What would happen in case you continue to extract the PCA components within the noise domain? How a scree plot (the variance of the components) would look like if you extract ALL theoretically possible PCA components from your data ? The upper limit equals the number of energy channels and is 2000 for the case of **EELScube** dataset. Practically, you cannot extract so many components but you can evaluate how such a scree plot should look like provided that your noise fraction is indeed random noise. This was predicted in E.R. Malinowski "Theory of the distribution of error eigenvalues resulting from PCA with application to spectroscopic data" J. Chemometrics 1 (1987) 33-40.

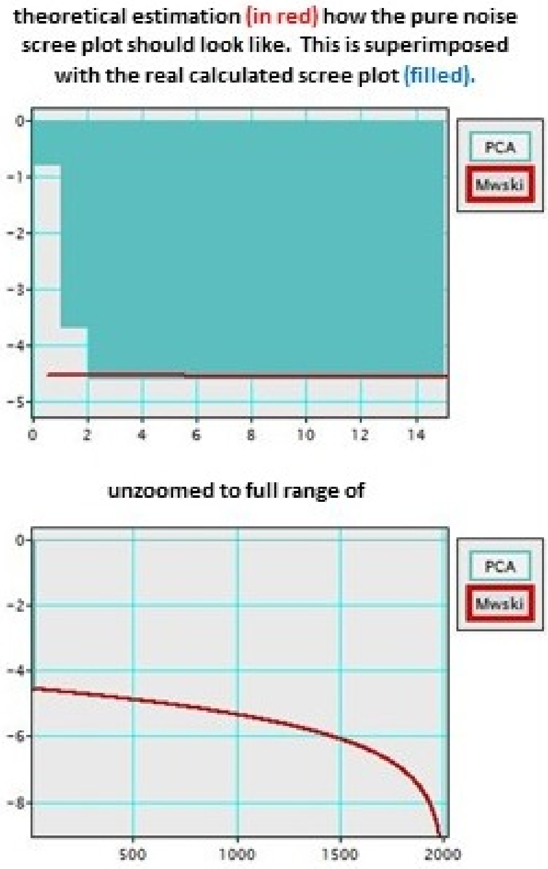
- Having **EELScube MSA.dm3** in front, click calculate in the noise scree plot box.

You get the predicted scree plot beyond the number of components you actually calculated. To see the full range, make mouse right-click and choose "home display" or just drag the horizontal scale by keeping CNTR key pressed. As usual, you can anytime halt the calculation by pressing stop.

Ideally, your estimated noise scree plot (red line) should meet the actually calculated one (filled profile). This is the case of **EELScube MSA.dm3** because it was a synthetic data. Real experimental scree plots might deviate from those predicted for the pure noise variance. In the other words, your noise fraction will be never a 100% noise. Some fraction of meaningful variations will be eventually buried in the "noise". If you did a good job in tuning your PCA treatment, this part is hopefully small.

Since you have calculated such a noise scree plot once, you can load it immediately.

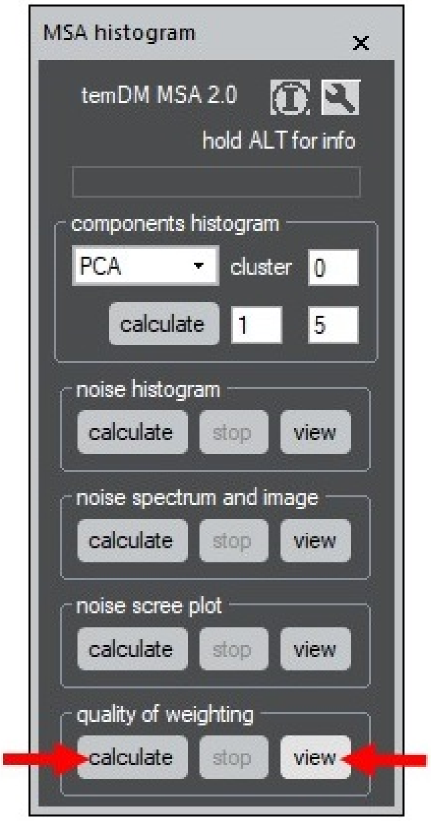
- Just click view button and the theoretical noise



theoretical estimation (in red) how the pure noise scree plot should look like. This is superimposed with the real calculated scree plot (filled).

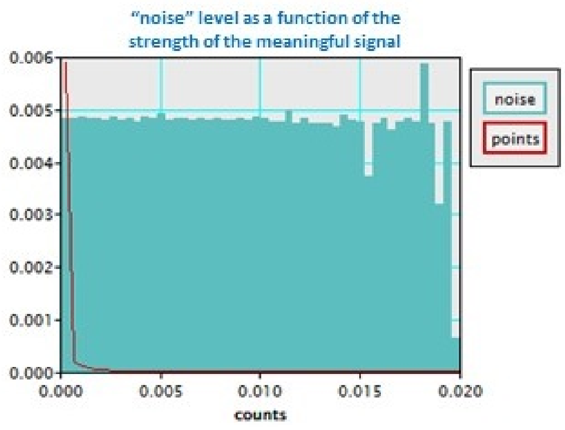
unzoomed to full range of

scree plot appears.



You can also evaluate *a posteriori* how accurate your weighting was when you generated the MSA container. Just to remind - the weighting procedure is aimed to equalize the Poissonian noise over all pixels and energy channels. Now you can check how homoscedastic your “noise” appears at the end.

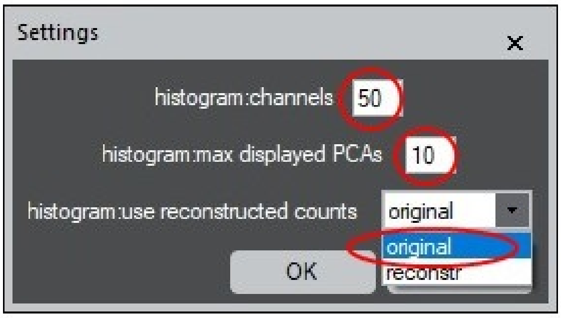
- Having **EELScube MSA.dm3** in front, click calculate in the quality of weighting box.



“noise” level as a function of the strength of the meaningful signal

Now you see the level of your noise as a function of the local value of signal (counts). In the case of **EELScube MSA.dm3**, the noise level is indeed quite same for all signal counts. The number of data points giving rise to the signal is indicated by the overlaid red curve. The resulted graph may look a bit noisy at the very right of the plot where too little points are available.

Again - you can anytime halt the calculation by pressing stop. Since you have calculated such a graph once, you can load it immediately afterward. Just click the view button.



histogram:channels 50

histogram:max displayed PCAs 10

histogram:use reconstructed counts original

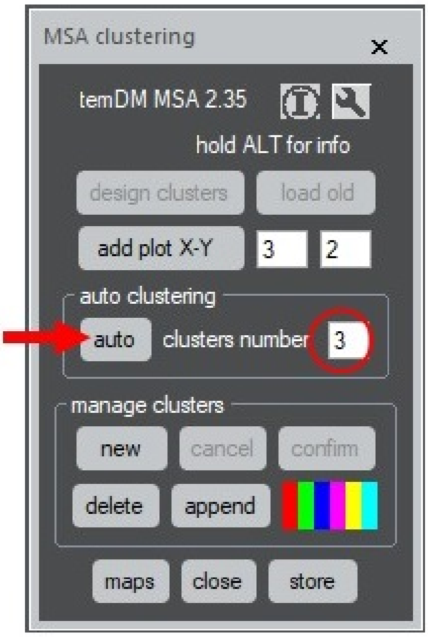
By clicking the spanner icon you get access to some parameters of the histogram tool.

Those are:
histogram channels: Number of channels used to build a histogram. This is used for components and noise histogram as well as for the graphs created in the qual-

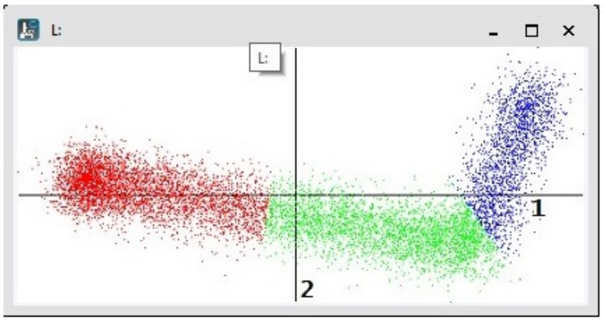
ity of weighting box.
max displayed PCAs: Maximal number of PCA components overlaid in the component's histogram display. Keep it reasonably small, otherwise quick display manipulation us difficult.
QofW based on: The evaluation of Quality-of-Weighting requires knowledge of the "true" signal. You can choose how to estimate the level of the meaningful signal. The default parameter is "original", which means that the raw counts (signal + noise) are assumed to be a measure for the strength of the meaningful signal (signal only). For very noisy data, this might not work. Then evaluate the strength of the signal from the reconstructed data.

12 Automatic Clustering

The section "Clustering" of the manual for the basic version describes how to design clusters in your dataset. The advanced version offers you an extra possibility for clustering - automatic breaking data to the desired numbers of clusters. For the moment, the simplest (although most popular) K-means clustering method is used.



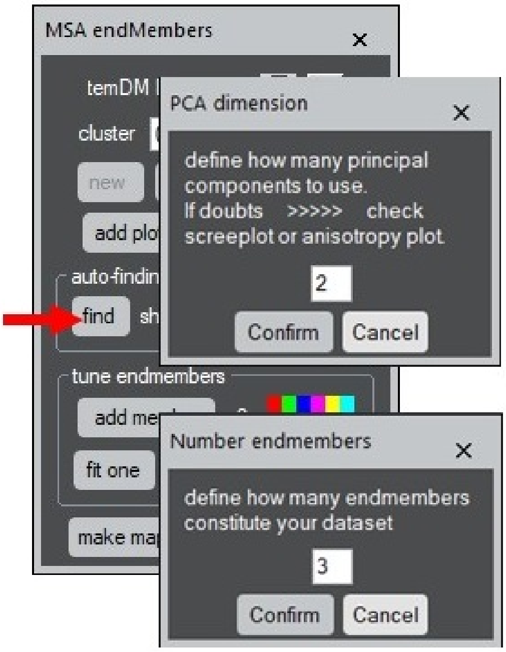
- Open the clustering tool by choosing temDM - MSA: factor space - Clustering in the DigitalMicrograph menu.
- Having **EELScube MSA.dm3** in front, press design clusters.
- Define the desired number of clusters, for instance **3**, in the corresponding field of the auto-clustering box. Press the auto button.



You see that the data are automatically clustered according to the k-means criteria.

Important: you can further correct the resulted clusters manually as described in the manual for the basic version. You can delete any of them or manually add another one if you want.

13 Automatic finding of endmembers



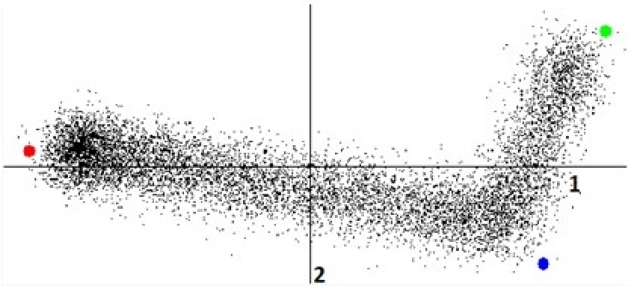
define how many principal components to use. If doubts >>>>> check screeplot or anisotropy plot

define how many endmembers constitute your dataset

Perhaps, the most powerful feature of the advanced version is automatic finding endmembers. The section "Endmembers" of the manual for the basic version describes how to create and tune endmembers manually. That works reasonably well for datasets having not more than 3 meaningful principal components after the PCA decomposition. If the dimensionality is higher, the problem of finding endmembers becomes more tricky and the automatic procedure is needed.

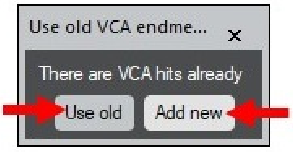
- Open the endMembers tool by choosing temDM - MSA: factor space - End-members in the DigitalMicrograph menu.
- Having **EELScube MSA.dm3** in front, press new.
- Press the find button in the auto-finding endmembers box.
- You will be prompt to confirm the number of the meaningful principal components in your data. The number **2** will be suggested. Press Confirm.
- After a while you are prompt to confirm the number of endmembers constituting (to your opinion) your dataset. By default this is the number of meaningful principal components plus one. In the given example, this is **3**. Press Confirm.

The three automatically found endmembers will be displayed on your scatter plot(s) and their spectra will be shown accordingly. They look a bit "too extreme". This is because we used a simplest algorithm - repeatable Vertex Component Analysis (VCA) that is searching for points maximally extended from each other in the factor space. "Repeatable" means that the VCA procedure is performed many times to improve the representability of results. The final endmembers are found by clustering all available points. However, this algorithm does not account for the noise in data. More advanced Bayesian algorithm will be considered in the next section.



It is very important to specify how many principal components are used. This defines the dimensionaly of the space where endmembers are searched for. Too many dimensions might make evaluation time consuming and inaccurate. Too little dimensions will leave the part of your data variation unattended. The problem of the right truncation of the principal components was already discussed in section "Adviser". You might consult with it. If you have performed already the denoising reconstruction, the last applied number of the truncated PCA components will be used as a suggestion.

Then try to press the find button again. After confirming the number of principal components, the program realizes that some potential endmember points (we call it VCA hits) were already collected by the VCA procedure. The system is asking you whether you want to hit another points or simply add new points to the old set. Choose add new and notice that the final endmembers are slightly changed because a bigger statistical set is now used.


Use old VCA endme...

There are VCA hits already

You might argue that it is probably easier to set these 3 endmembers manually in this case. True. But that was just a learning example. How would you treat the case with e.g. 8 endmembers in the 7-dimensional space ? The automatic algorithm is needed.

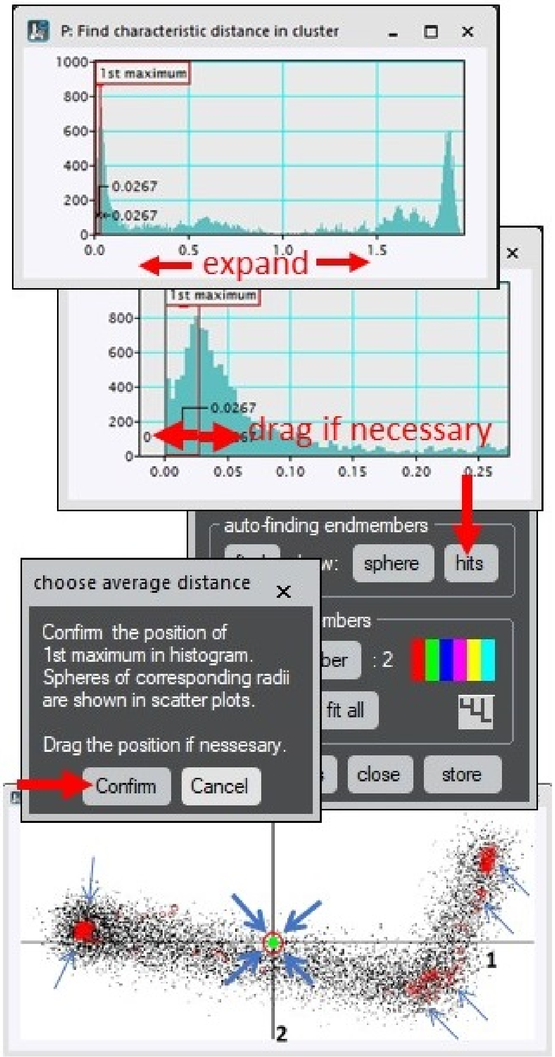
14 More advanced automatic algorithms

Now lets try to explore more complicated algorithms described in [5].

- Go to the setting through the spanner-icon. Set the "hits method" to Bayes VCA and "cluster method" to mean-shift.
- Press the find button in the auto-finding endmembers box.
- Confirm the number of principal components (**2**).
- Wait until some histogram is displayed in the upper-left corner of the DigitalMicrograph workspace. Press Confirm in the dialog window.
- After a while, a graph of potential endmembers is displayed at the same place. The **3** endmembers will be selected but you can change it. Do not do that for the moment, just press Fix.

Your endmembers are now much more realistic. The spreading of data due to noise is accounted for.

As you see, the algorithm requires a certain user interaction. In principle, all decisions can be made automatically. However, the user confirmation is implemented in order to minimize the risk of mistreatment.



Confirm the position of 1st maximum in histogram. Spheres of corresponding radii are shown in scatter plots.

Drag the position if nessesary.

It is always better to have some control on what is going on, is not it ? What should be stressed: while waiting for your confirmation, the program does not block any other actions with **EELScube MSA.dm3**. You may load the View Results tool and generate a screeplot or an anisotropy plot in order to decide on the PCA dimension. You can also press the k ? button in the Proceed tool and get an advice. The dialog window will be still waiting for your decision till you press OK. Lets consider in more detail what program is doing.

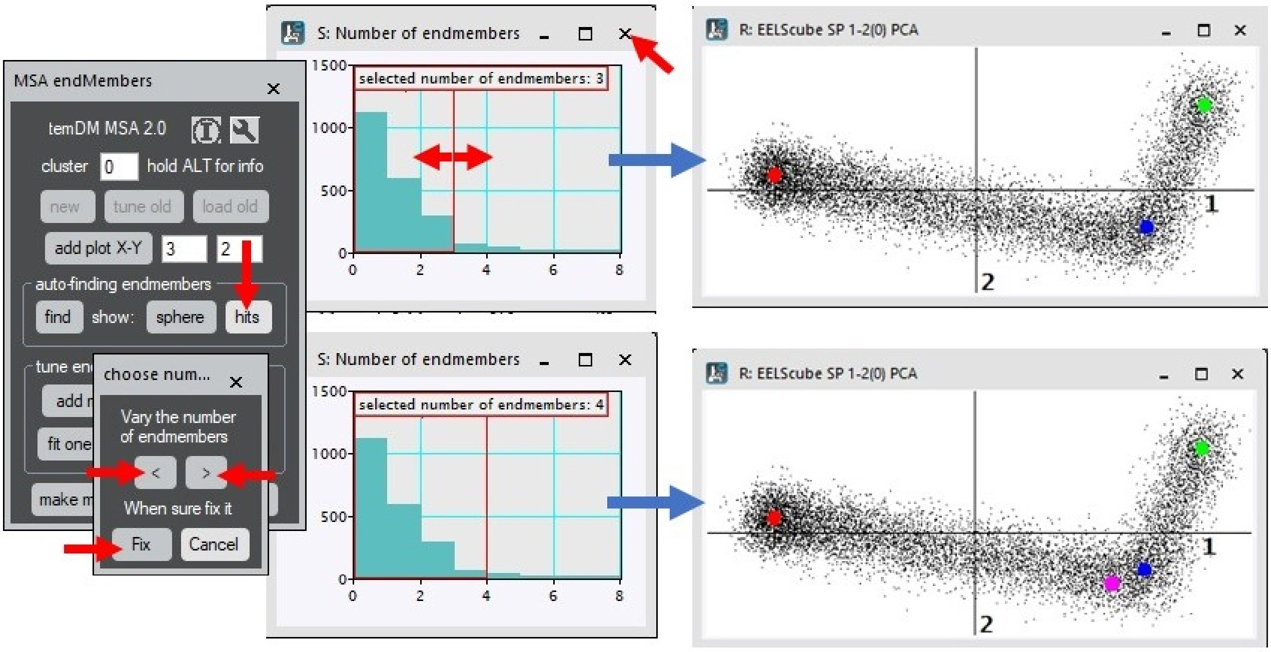
Confirm the the position of the 1st maximum in histogram, or in the other words, the size of clustering sphere At this step, the program evaluates if the VCA hits are clustered around certain positions in the factors space. The algorithm requires to define a certain sphere, within which the nearest VCA hits consolidate. For that purpose, the histogram of mutual distances among all VCA hits is calculated where the position of the first maximum indicates the average distance among the VCA hits within a cluster. The histogram is displayed in the upper-left corner of the DigitalMicrograph workspace. In principle, the first maximum is found automatically but you should check whether it is really the *first, not other* maximum. You might expand the scale of the histogram to see the region of the first maximum better.

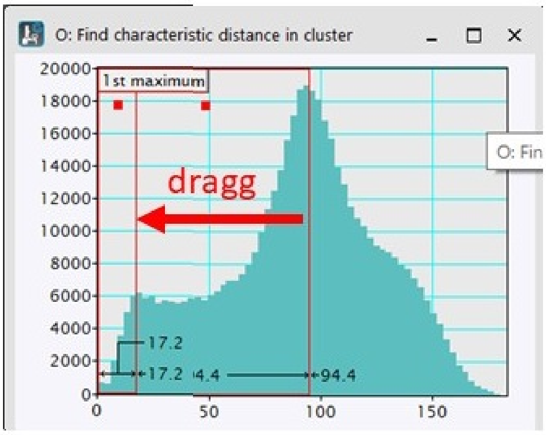
In few cases the program might be wrong with the position of the first maximum. Then drag manually the red selection line to the right position. The figure below illustrates such possible case. When you are dragging the red line, the sphere of the corresponding radius will be live displayed in the center of the scatter plot. That would help you to decide whether the clustering sphere is reasonable or not.

When judging about the size of the clustering sphere, it is worth to display temporarily *all* your VCA hit points. Click the knob hits. That would plot all VCA hit points generated by the system so far. Then the possible clusters are visualized and you can adjust the clustering sphere accordingly. The second click on hits will make the VCA hits invisible again. Do that because such numerous points can disturb evaluation of the scatter plot at the next step of treatment.

Choose the number of endmembers. At the last step, a graph of potential endmembers is displayed with the red rectangular denoting the selected endmembers. This graph shows the *probability* that a given endmember is actually relevant. The endmembers are ordered according their relevance - most probable are at the left of the graph, less probable are at the right. As default, the (*number of PCA components + 1*) is selected. This is a theoretical value, however, in each particular case, it might deviate from the theory. The endmembers are supposed to reflect some real latent factors and nobody except of you can judge if a given endmember makes a physical sense or just represents some artifact. Try to increase or decrease the number of endmembers by clicking > or < and see what happens. You can also drag the selecting rectangular to the left or to the right directly in the graph. The new endmembers points will live appear or disappear in the scatter plots. Decide what you consider to be a reasonable configuration of the endmember points describing all variations in your dataset. You might also check the corresponding spectra of the endmembers. As soon as you are confident on how many endmembers to choose, press the Fix button.

Sounds complicated ? Well, if you think there are only three endmembers in your data, it is easier to set them manually in the relevant scatter plot. However, you might face much more complicated cases. The ex-





ample of 8 endmembers in the 7-dimensional space is shown in this page. We cannot display 7 dimensions, we only show few selected two-dimensional projections of such a complicated object. Understanding the geometry of data distribution in 7-dimensions is quite tricky. Probably, it is better to rely on the automatic algorithm like implemented in **temDM MSA** and described here.

15 Semi-automatic finding of endmembers

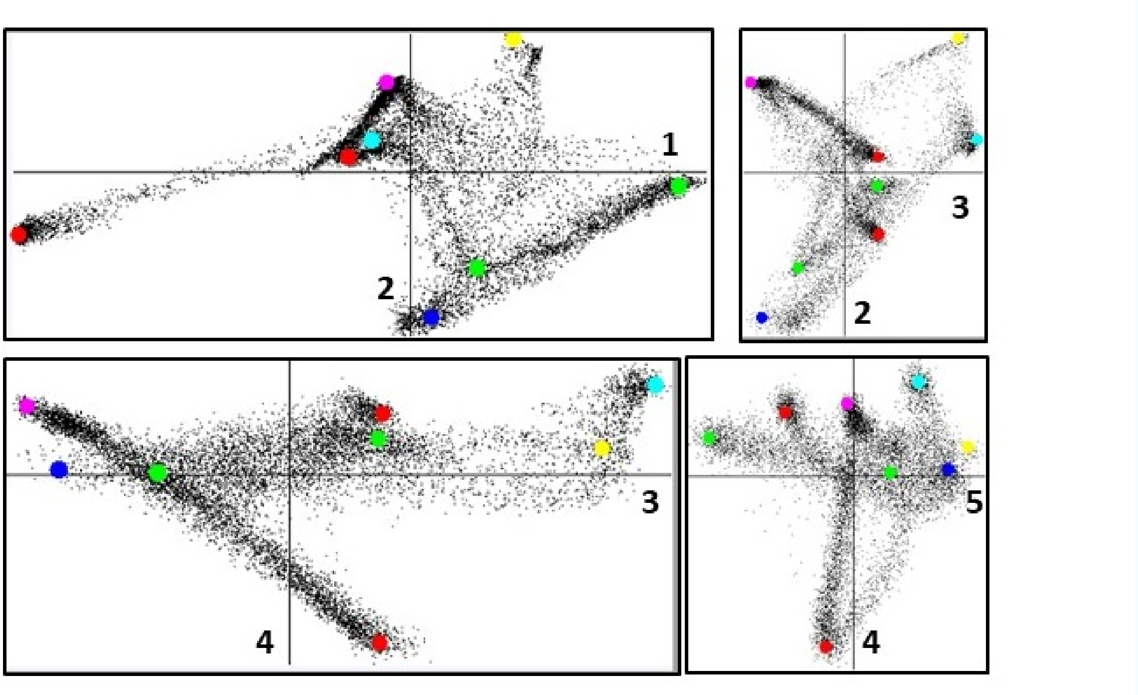
What is important about the automatically found endmembers - *you can further tune them manually* as described in the manual for the basic version. Your final target is to extract the latent factor with a clear physical meaning, is not it ? The automatic procedures are good but only humans may finally judge what does make a physical sense.

It is not easy to navigate in the space with the dimensionality greater than 3. The Endmembers tool offers the semi-automatic means to tune endmembers in the invisible dimensions. Suppose that you set some endmember point (say endmember 1 in **EELScube MSA.dm3** example) in the scatter plot **1-2** but you are not sure about its position in the *other principal components coordinates*. It is not crucial for the given two-dimensional example but could be extremely challenging for the case of higher dimensions. Just to play with this option:

- Generate an extra scatter plot by entering **3** in the relevant field and pressing add Plot X-Y
- Intentionally shift the marker of endmember 1 (red) to some crazy position far from the main data distribution trend.
- Select the marker of endmember 1 (red) in the **1-2** scatter plot, press the fit one button in the tune endmembers box. The red marker in the **2-3** scatter plot (and all other potential plots even not visible) will jump to the area of the maximal density of the data points.

Well, the optimal position of the endmember is not always situated in the regions of highest data points density, but in 90 percent cases it does, thus this button is useful.

You can also try to optimize the positions of all your



endmembers at once.

- Again, displace the endmember markers to the crazy positions in the **3-2** scatter plot.
- Click several times the fit all button in the tune endmembers box. You will notice that all markers tend to align along a certain line in the **3-2** scatter plot and its orientation will slowly approach the direction of the 2nd PCA principal axis. That is the right result because the 3rd principal component (and any other components with the higher indexes) is just noise in this dataset.

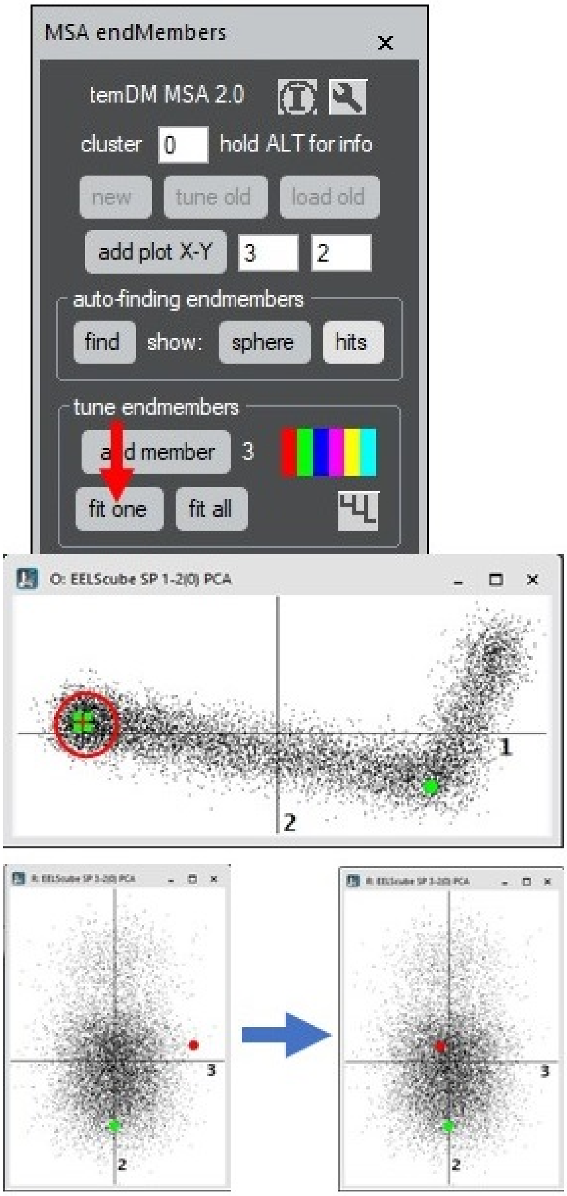
The latter procedure uses the so-called Alternative Least-Square (ALS) fit. This is quite different from the automatic algorithm described above. ALS results depend on the initial conditions, thus you can try apply first the automatic VCA procedure with the button find and then ALS with fit all. Just explore different approaches and track the positions of your endmembers in the different scatter plots. You will definitely have a lot of fun exploring your data in the n-dimensional factor space !

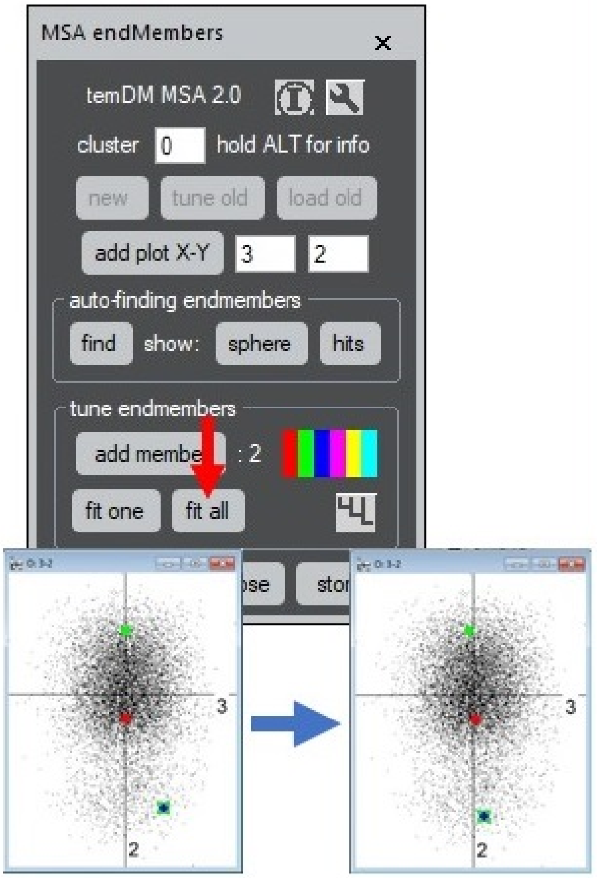
There are a couple of new parameters in the end-Members settings related to the automatic and semi-automatic finding of endmembers. You can acces that via the "spanner" icon:

lambda: Regularization parameter in Alternate Least Square (ALS) fit.
min points: This is needed for the fit one procedure. The algorithm must capture at least *this* number of data points to find the maximal density area in the factor space.
auto:VCAhits: The algorithm of automatic finding endmembers uses Vertex Component Analysis (VCA) in its repeatable form. This parameters defines the number of the VCA attempts (hits) to determine the most probable endmembers. The higher number generates more statistically significant results but also leads to the longer calculation time. See [5] for details.
auto:hits method: You can choose between the simple plain VCA method described in the previous section or more advanced Bayes VCA. The latter accounts for the noise in the data distribution but requires longer computation time and possible user interaction.
auto:cluster method: The repeatable VCA procedure generates a number of potential endmember points that must be then clustered to evaluate the most probable location of the real endmembers. The K-means clustering method is fast but requires to define in advance the number of endmembers. The more sophisticated mean-shift method sorts the potential endmembers according their probability and allows a user to judge finally on how many endmembers are there.

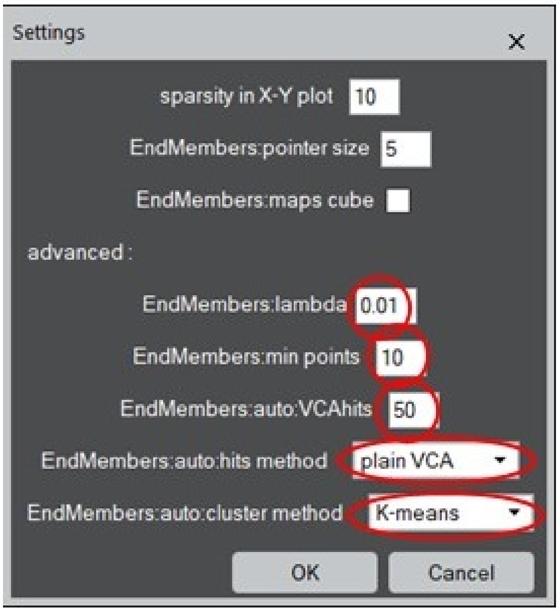
References

[1] P. Potapov. Why principal component analysis of STEM spectrum images results in abstract, uninterpretable loadings? *Ultramicroscopy*, 160:197–212, 2016.





„end members“ tool


Settings

sparsity in X-Y plot 10

EndMembers:pointer size 5

EndMembers:maps cube □

advanced :

EndMembers:lambda 0.01

EndMembers:min points 10

EndMembers:auto:VCAhits 50

EndMembers:auto:hits method plain VCA

EndMembers:auto:cluster method K-means

[2] P. Potapov, P. Longo, and A. Lubk. A novel method for automatic determination of the number of meaningful components in the the PCA analysis of spectrum-images. *Microsc. Microanal. Proceedings*, 24 S.1:572–573, 2018.

[3] P. Potapov and A. Lubk. Optimal principal component analysis of STEM XEDS spectrum images. *Advanced Structural and Chemical Imaging*, 5:4, 2019.

[4] P. Potapov, P. Longo, and E. Okunishi. Enhancement of noisy EDX HRSTEM spectrum-images by combination of filtering and PCA. *Micron*, 96:29–37, 2017.

[5] P. Potapov and A. Lubk. Extraction of physically meaningful endmembers from STEM spectrum-images combining geometrical and statistical approaches. *Micron*, 145:103068, 2021.

[6] P. Potapov. On the loss of information in PCA of spectrum-images. *Ultramicroscopy*, 182:191–194, 2017.