Renishaw plc
Spectroscopy Products Division
New Mills, Wotton-under-Edge,
Gloucestershire GL12 8JR
United Kingdom

Tel +44 (0) 1453 524524
Email raman@renishaw.com

www.renishaw.com

RENISHAW
apply innovation™

TM015 – Multi-file data analysis (multivariate) WiRE™ 5

This document aims to show the WiRE™ 5 user how and when to use multivariate data analysis methods on image data.

Multivariate methods require significant computational power and it is recommended a high specification computer be used for such analysis (multi core, > 2 Gb Ram). Such methods can be applied to large datasets containing up to 50 million spectra.

This document assumes the user is familiar with the basic image creation methods within WIRE 5 and has purchased the multivariate data analysis software module (chemometrics).

Why use multivariate data analysis?

The role of chemometrics in the analysis of Raman spectroscopy data is becoming increasingly important for many different application areas. The main reasons for this are:

- Fast imaging techniques, such as StreamLine™ imaging, produce too much data for manual, non-automatic, methods
- Chemometrics enables the potential qualification and quantification of very complex systems such as biological materials
- Multivariate methods analyse the whole data set simultaneously. This enables spectral information to be more directly correlated with chemical / structural properties.

WiRE 5 contains the following multivariate image creation methods:

- Direct classical least squares component analysis (DCLS)
- Principal component analysis (PCA)
- Multivariate curve resolution – alternating least squares (Empty Modelling™)

Note: Prior to conducting multivariate data analysis, ensure significant cosmic ray features (CRFs) are removed from the dataset (See module TM13). If significant CRFs are present in the dataset, the multivariate model may not accurately determine the components or their distributions.

When do I use multivariate analysis?

When Raman datasets have been collected from complex systems, and contain many spectra, multivariate data analysis is an efficient method for accurately determining the distribution of chemical components.

The different multivariate methods should be used in the following situations:

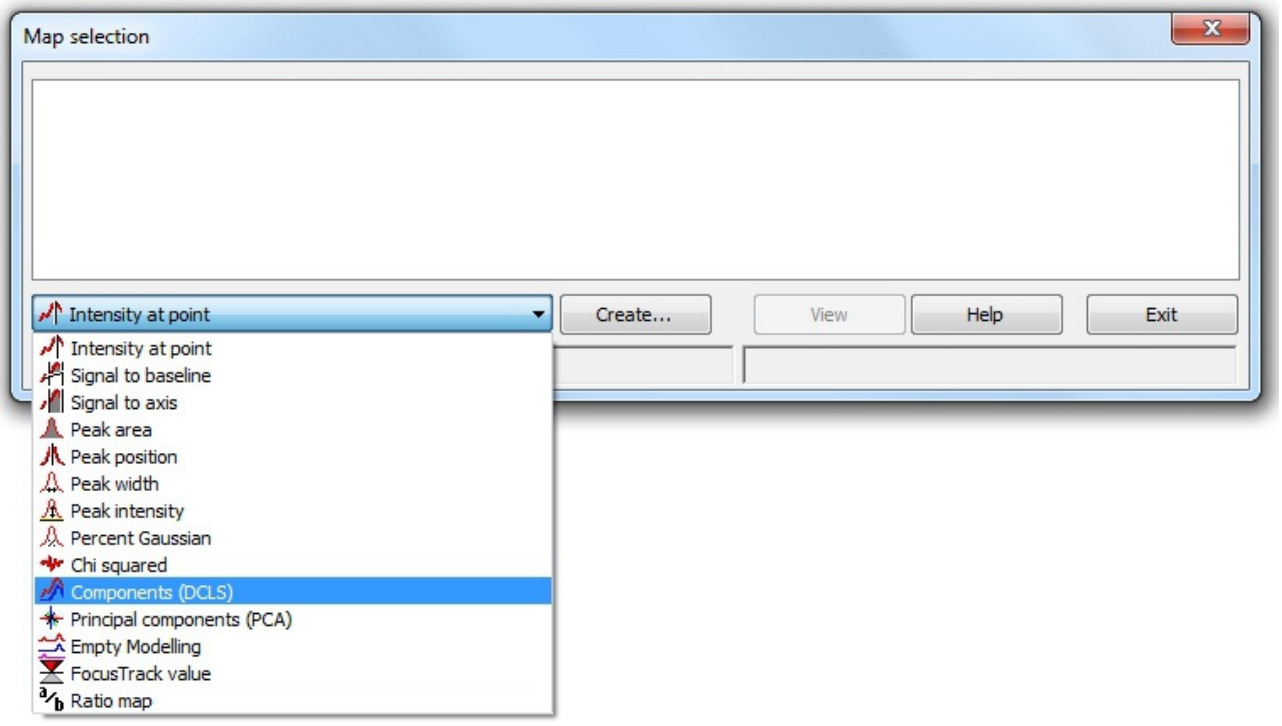
- 'Components (DCLS)' – when reference spectra are available for **all** components in the mixture. Component images and quantitative information (i.e. fraction estimates) are obtained for each reference spectrum.

  If only a **limited** number of reference spectra are available, DCLS can be used to create useful but potentially inaccurate (and non-quantitative) component images.

- 'Principal components (PCA)' – when **no** reference spectra are available and the user wishes to investigate different components within the map data. This method provides abstract component spectra (loadings) which are not usually recognizable as real component spectra, as these are abstract vectors and not constrained to be positive.

- 'Empty Modelling' – when **no** reference spectra are available and the user wishes to gain images and (pure) spectra **representing physically meaningful components**. This method is most successful where the main components are present in near pure form somewhere in the map.

Once data has been collected the create map icon  can be used to define the map type to be created. The multivariate methods are listed after the univariate analysis methods.



Direct classical least squares component analysis (DCLS)

The ‘Components (DCLS)’ method is used:

- when reference spectra are available for **all** components in the mixture. Component images and quantitative information (i.e. fraction estimates) are obtained for each reference spectrum.

- If only a **limited** number of reference spectra are available, DCLS can be used to create useful but potentially inaccurate (and non-quantitative) component images.

Reference spectra can be taken from previously collected separate spectra, or from within the map dataset.

How does the ‘Components (DCLS)’ method work?
The Direct Classical Least Squares (DCLS) algorithm approximates the spectrum at each point in the map by summing together scaled copies of each reference spectrum. The scaling factors are chosen to get the best possible fit to the original map spectrum, and then used to create a map showing the distribution (correlation values) of each reference spectrum.

If there is an unexpected component in the map with no corresponding reference spectrum, the algorithm will not be able to produce a good fit between the actual map spectrum and the scaled reference spectrum, and so the results may be misleading.

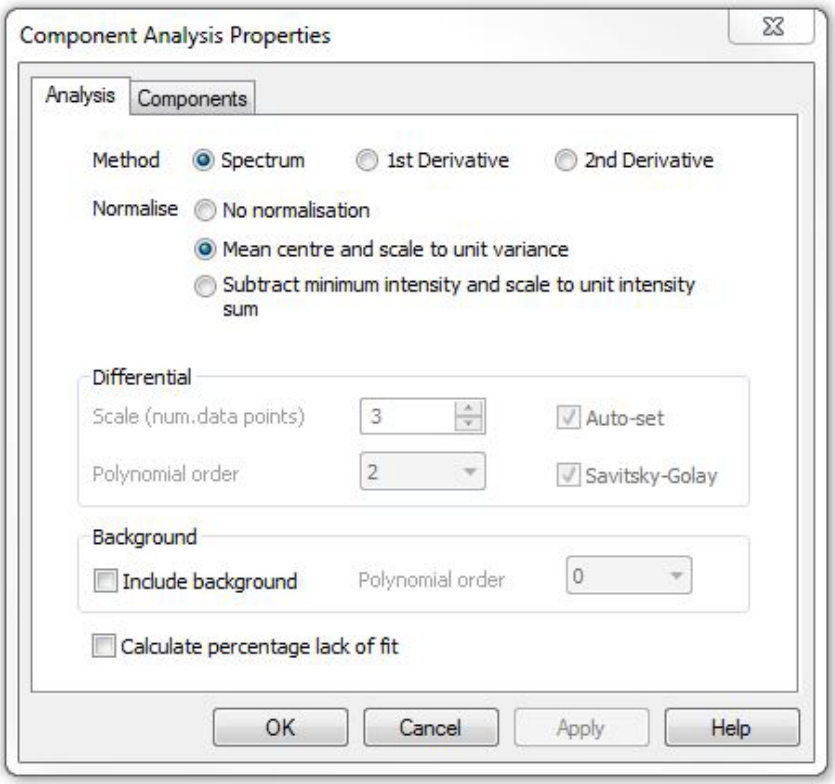
The ‘lack of fit’ option creates a further map, which displays regions of the sample, which deviate most from the optimum DCLS solution (i.e. the spectral information is not accounted for by the reference spectra). The LOF map can be used to identify Raman features representative of unexpected chemical components which have been omitted from the reference spectra.

- **When reference spectra are available for all components**
  DCLS correlation values can be used to estimate **quantitative** information, such as the total fractional component make-up of the map. For accurate fractional analysis, DCLS should be performed using ‘no normalisation’, and the reference spectra must **all** be collected under identical conditions (or scaled appropriately). If the data originates from a non-representative region, the resulting fraction estimate may not be representative of the entire sample.

- **When only some reference spectra are available**
  The DCLS algorithm will not be able to produce a good fit between the actual map spectrum and the scaled reference spectrum, and so the results should be interpreted with care. **This method will not yield any useful quantitative information.**

Using ‘component analysis’ to create component images

1. Ensure Raman imaging data are loaded into WiRE 5 software and select ‘Components (DCLS)’ option from the map selection dialog.



2. Select ‘Method’ type.

   Choose spectrum if a direct comparison is required, especially where absolute intensity values are important (i.e. not normalising).

   Using derivative spectra can be useful if the data contain background offsets and / or slopes. The 1st Derivative option is suitable for removing backgrounds with a consistent shape but variable offset. The 2nd Derivative option is suitable for spectra where the background changes dynamically for the same component (e.g. fluorescence quenching).

3. Select ‘Normalise’ option.

   Choose ‘no normalisation’ if absolute intensity variations are required (important to extract quantitative information i.e. fraction estimates).

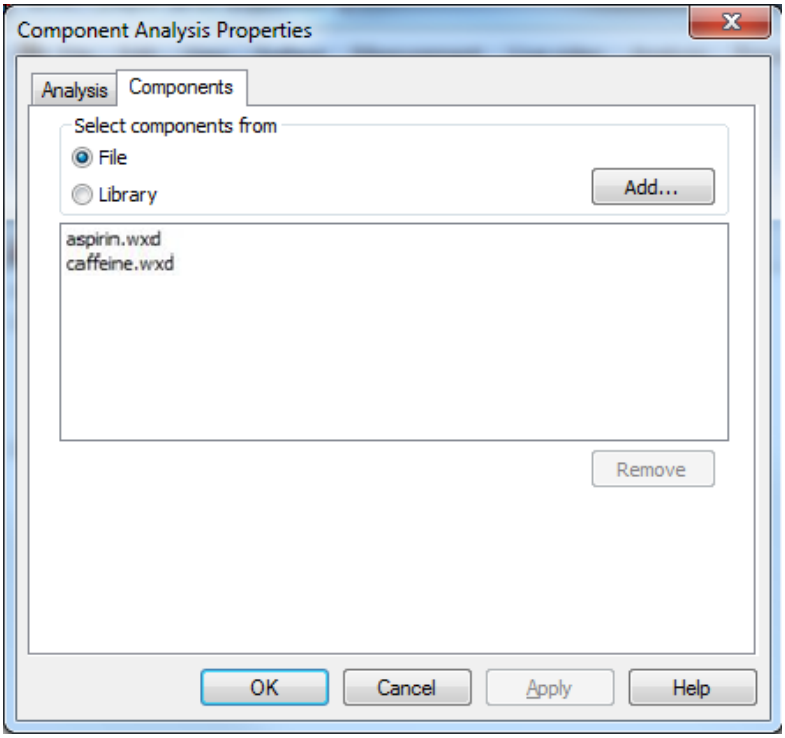
   Choose a desired normalisation method if absolute intensities are not important (Mean centre and scale to unit variance is set as default).

4. Select the ‘Include background’ option (only available for Spectrum method) to reduce the skewing of the DCLS analysis by the contribution of any background variation. This option can also be used if there is a variable background apparent in the map dataset, but not in the component spectra and this difference should be ignored.

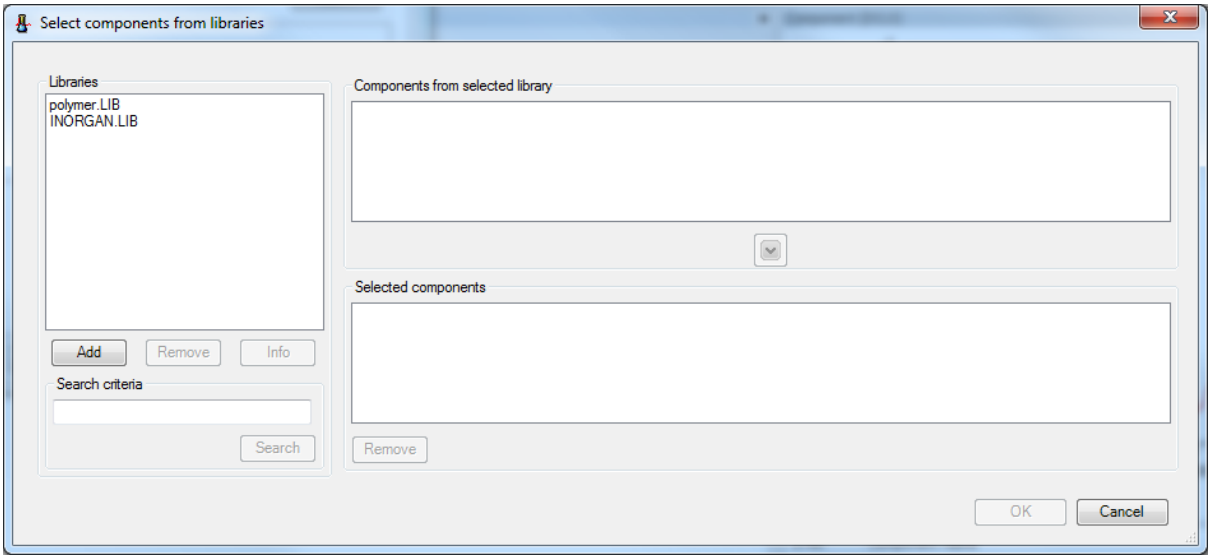
5. The ‘lack of fit’ option creates a further image, which displays regions of the sample, which deviate most from the optimum DCLS solution (i.e. spectral information not accounted for by the reference spectra). This can be used to identify Raman features representative of unexpected components which have been omitted from the reference spectra.

6. Select the ‘Components’ tab and add reference spectra.

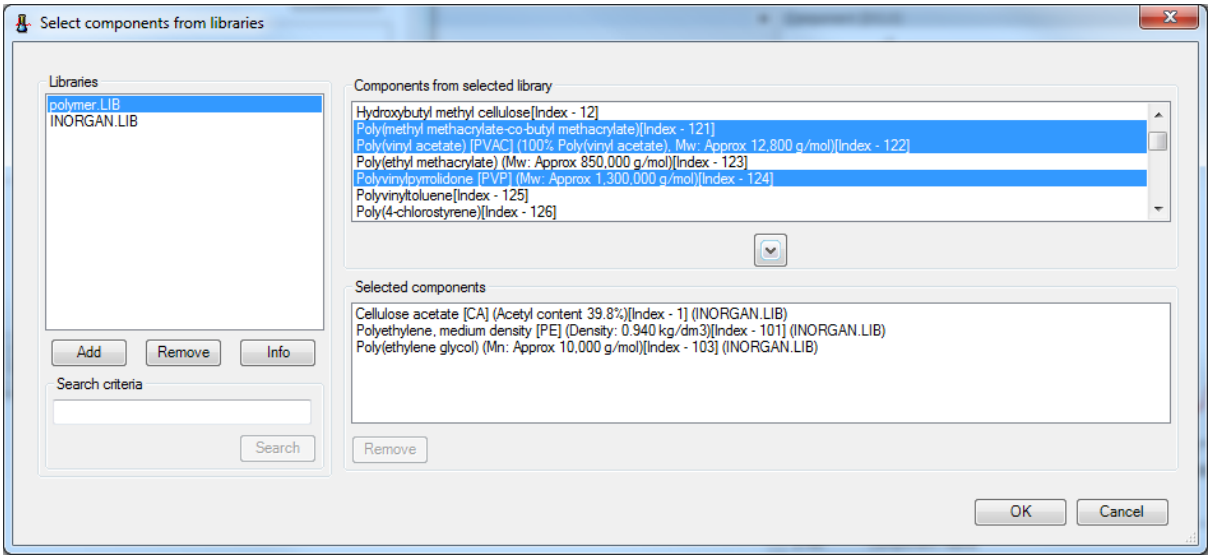
   Reference spectra can be added by:
   - Selecting ‘File’ and navigating to the file containing the reference spectra
   - Loading in from an existing library
   - A combination of reference spectra and library spectra



   To add a component from a library, the user must initially add the desired library or libraries to the ‘Libraries’ section of the ‘Select components from libraries’ window (opened by selecting **Library...Add...**). Once loaded, WiRE will remember the library unless it is removed from the ‘Libraries’ section.



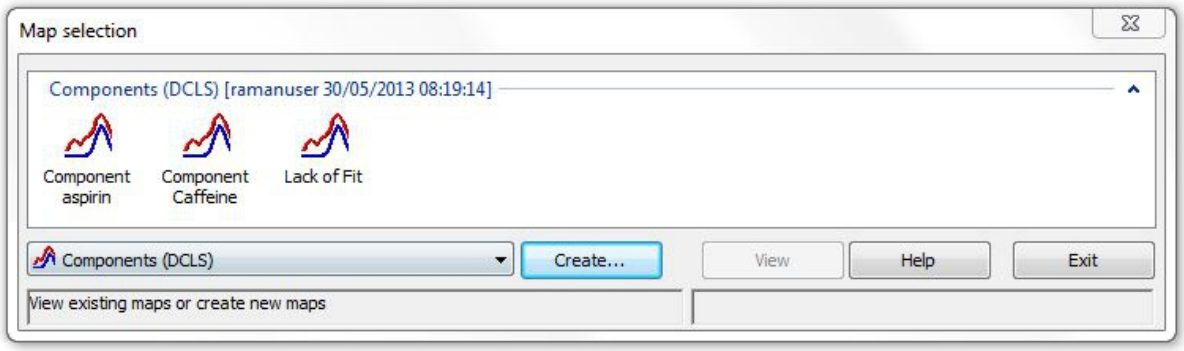
Highlighting a library allows the user to remove, view information on, or search for key words within the selected library. The 'Components from selected library' section lists all reference spectra within the selected library. This list can be reduced by entering key words into the 'Search criteria' section. To add component(s) from the library, click on the desired reference spectra so that they are highlighted. Click the down arrow to add the reference spectra to the 'Selected components' section. Components can be removed from this section by clicking to highlight and selecting 'Remove'.



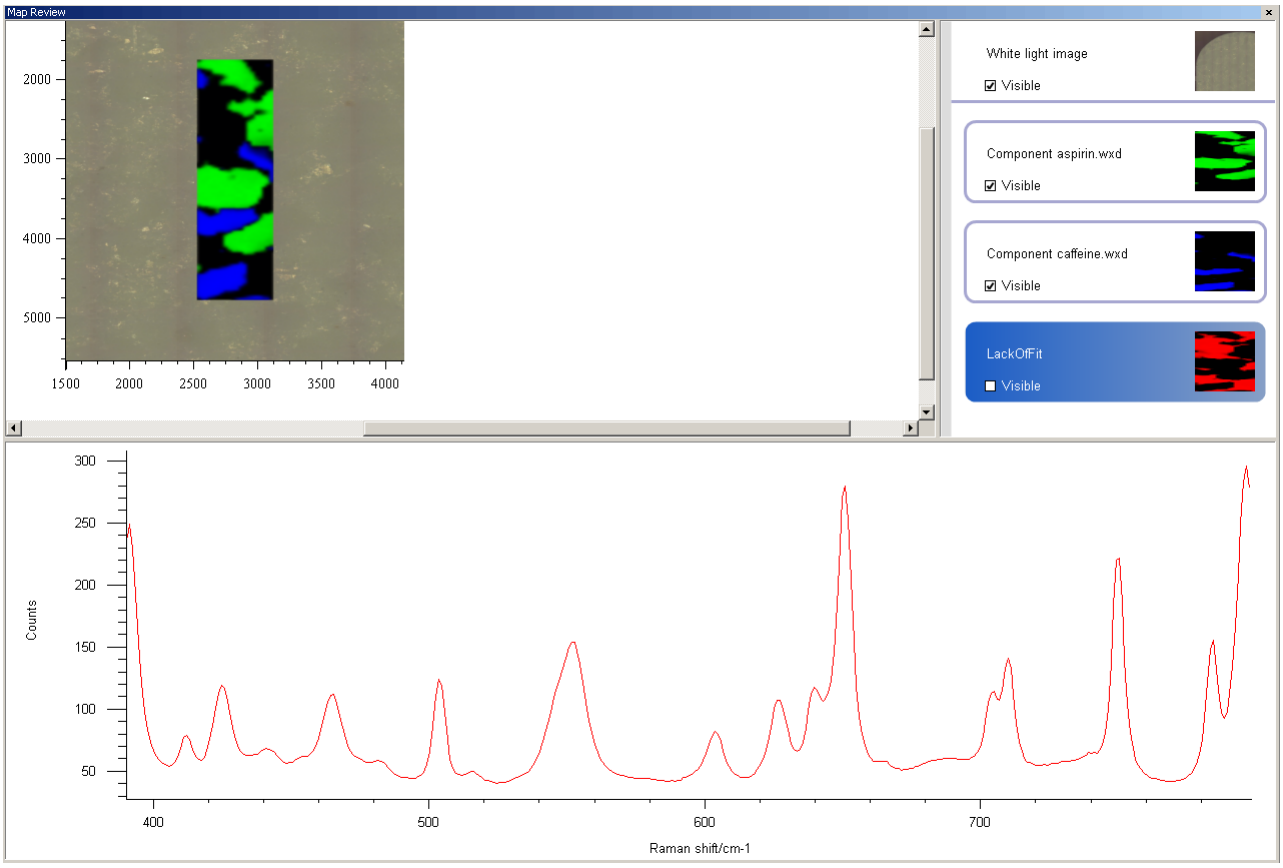
Clicking 'OK' adds the selected reference spectra to the components list.

7. The image is created by selecting 'OK', and can be viewed from the 'map selection' dialogue.

This example shows component images built from ‘aspirin’ and ‘caffeine’ reference spectra only. The ‘paracetamol’ reference is missing. A lack of fit image is also created.



Even though the reference spectra do not describe the whole dataset, Raman images representative of those components are still created. The lack of fit image now represents the third component (i.e. paracetamol).

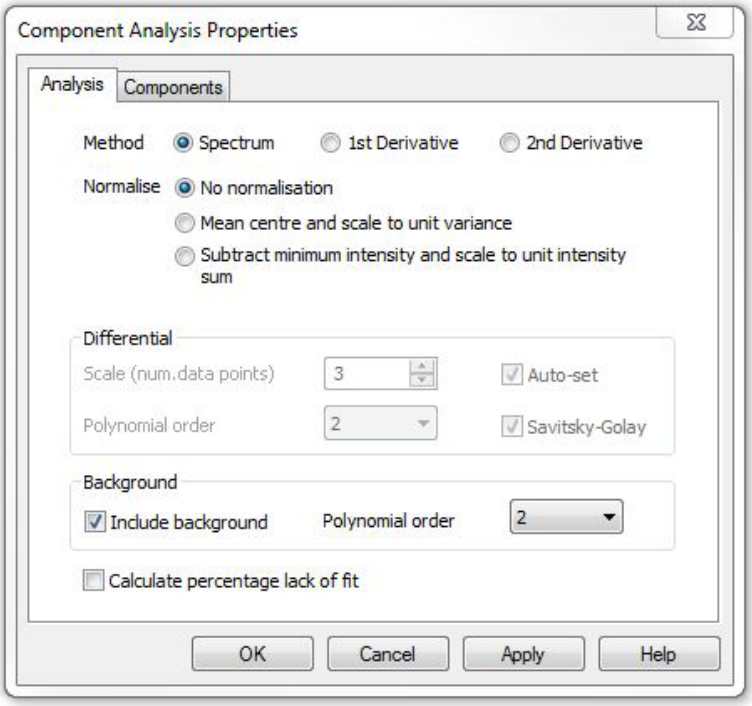


See module TM18 for further information regarding viewing multiple and overlaid images.

Using ‘component analysis’ to create component images and fraction estimate information

Note: To obtain accurate fraction estimates, DCLS should be performed **using ‘no normalisation’** and **all reference spectra must be scaled correctly**. If different experimental conditions (acquisition time, laser power, etc.) were used to collect reference spectra, then a single equivalent condition must be determined for each spectrum by use of suitable data arithmetic (See module TM4).

1. Ensure Raman data is loaded into WiRE 5 software and select ‘Components (DCLS)’ option from the map selection dialogue.



2. Select ‘Spectrum’ as method type.

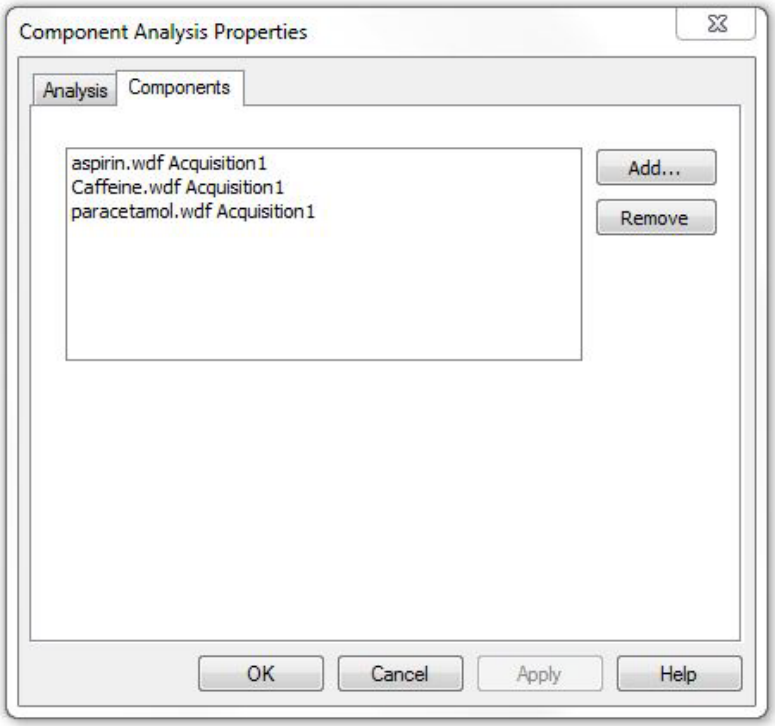
3. Select ‘No normalisation’ option to extract quantitative DCLS maps.

4. Select the background option and set the polynomial order to 2.

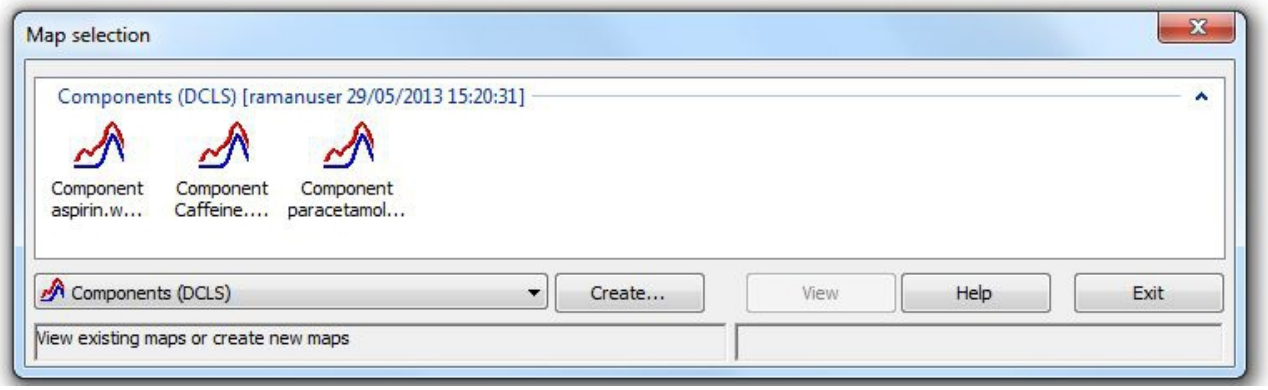
5. The ‘lack of fit’ option creates a further image, which displays regions of the sample, which deviate most from the optimum DCLS solution (i.e. spectral information not accounted for by the reference spectra). This is optional.

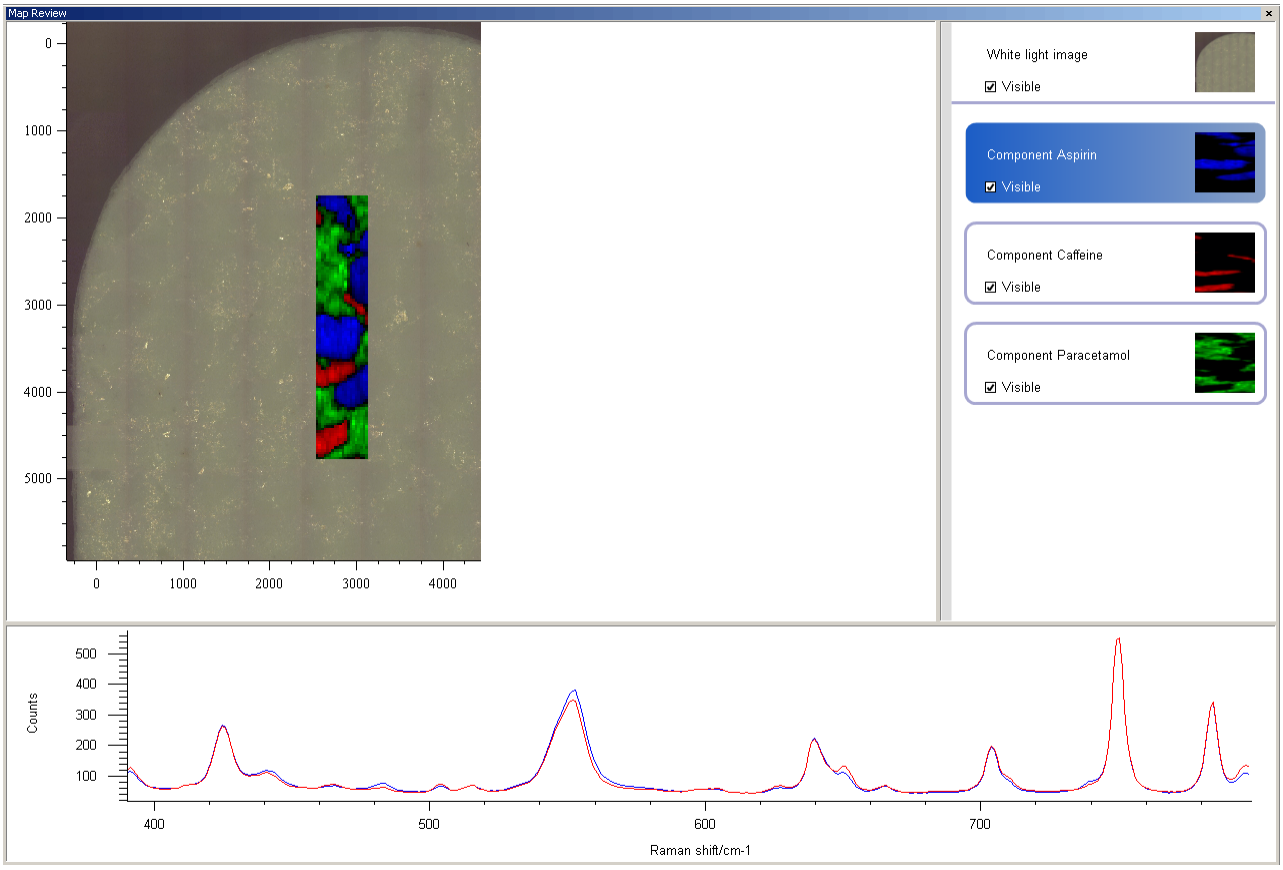
6. Select the ‘Components’ tab and add reference spectra that have been collected under identical conditions (or appropriately scaled).

Navigate to the file containing your reference spectra. Select one or more reference spectra and click Open. The reference components can be renamed by clicking twice on the name in the dialogue.



7. Select ‘OK’ to create the DCLS images, which can be viewed from the ‘map selection’ dialogue.



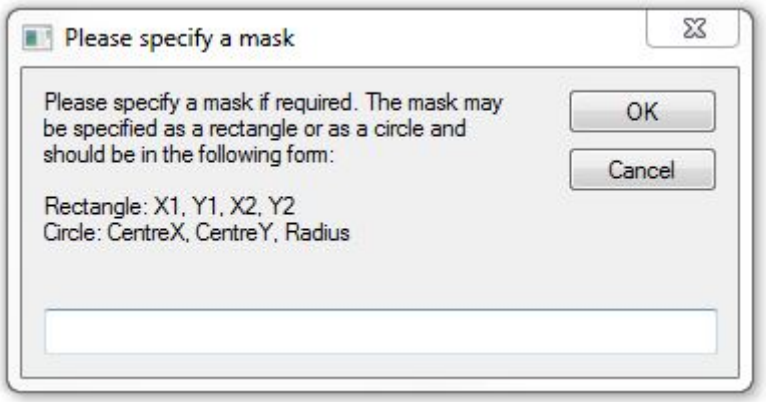


8. Before running the fraction estimate procedure, the active viewer should be a Spectrum viewer displaying a single spectrum from the map dataset. To do this, close the Map review dialog (shown above) by clicking the × in the top right corner.

9. To initialise the 'Fraction estimate' procedure, either select *Procedure > FractionEstimate* or click on the button

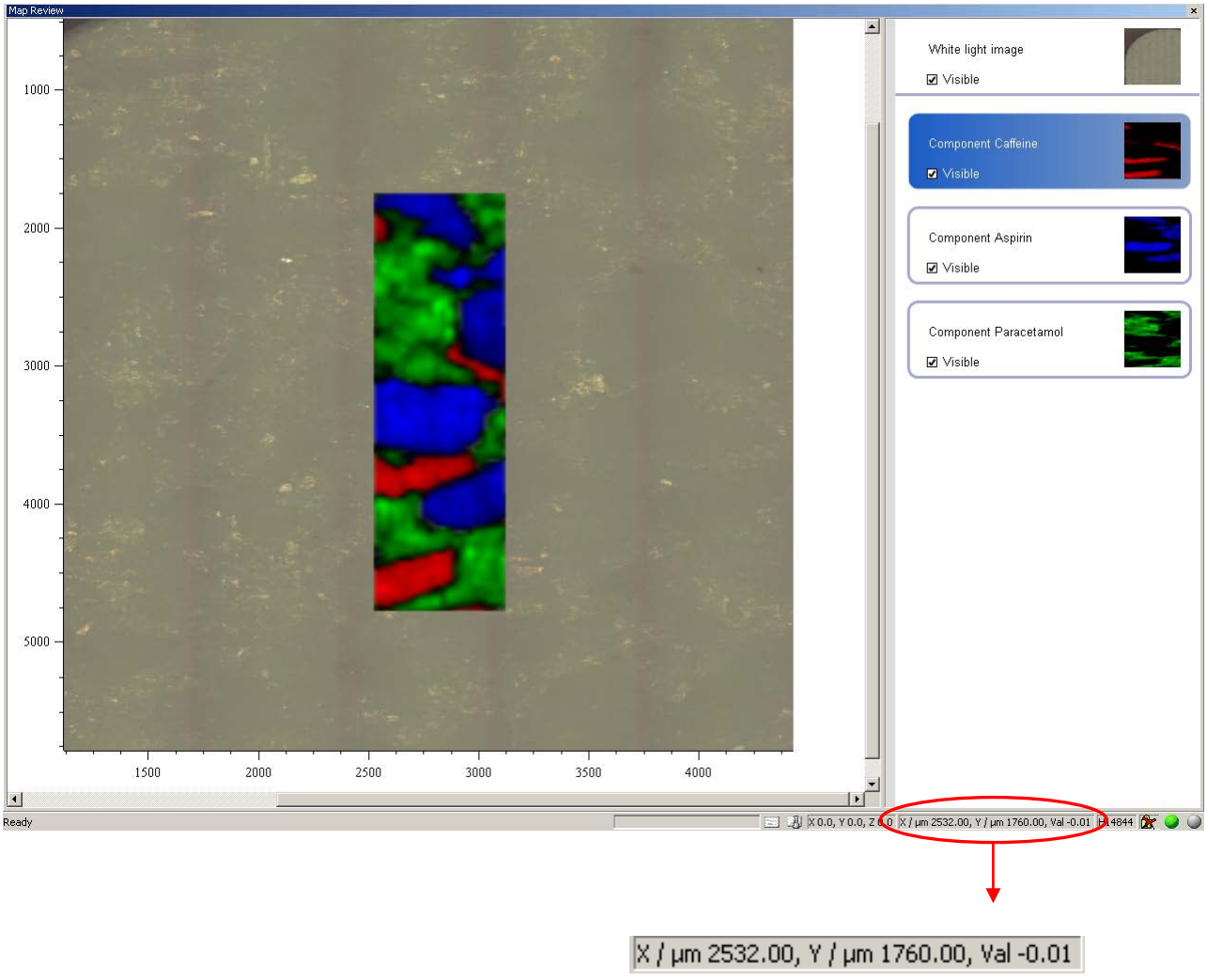
10. When performing 'Fraction estimate' the user has the option to sub-sample the map by specifying a mask.

    If the fraction estimate should include the entire map, simply leave the dialog blank and select 'OK'.

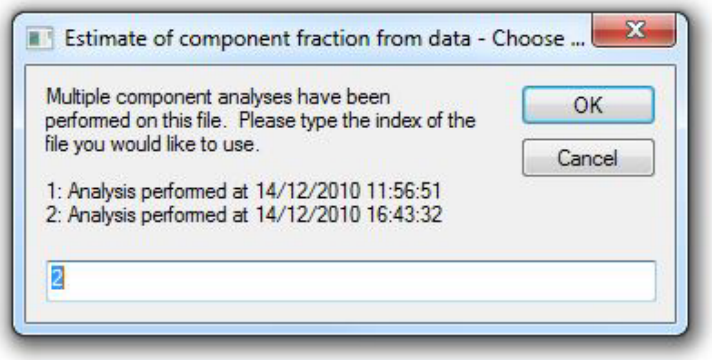


A rectangular mask can be defined using the corner points on the image in micrometres (input 4 numerical values).
A circular mask can be defined by determining the centre XY coordinates and circle radius in micrometres (input 3 numerical values).

The XY coordinates within an image can be determined by clicking on the desired location in the map review. The corresponding XY coordinates can be read off the lower right corner of the main WiRE window. All XY coordinates should be determined before initialising the 'Fraction estimate' procedure.

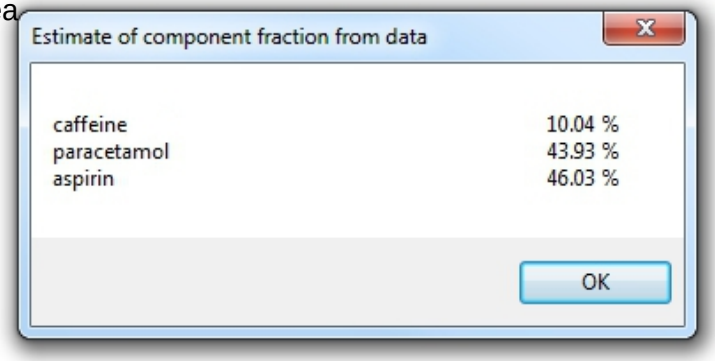


11. Fraction estimates can be determined for each instance of component DCLS analysis. A list of those created is displayed. The user can simply enter the index of the DCLS analysis to be used.

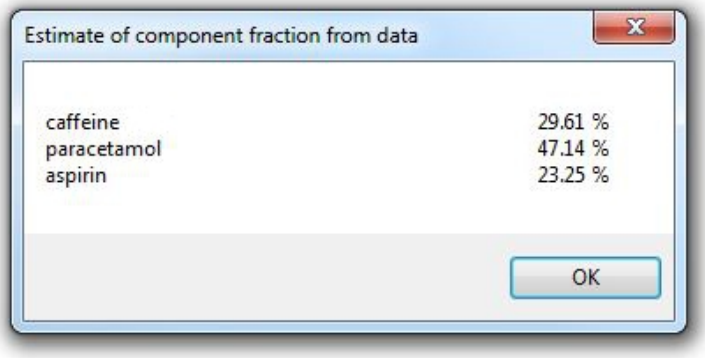


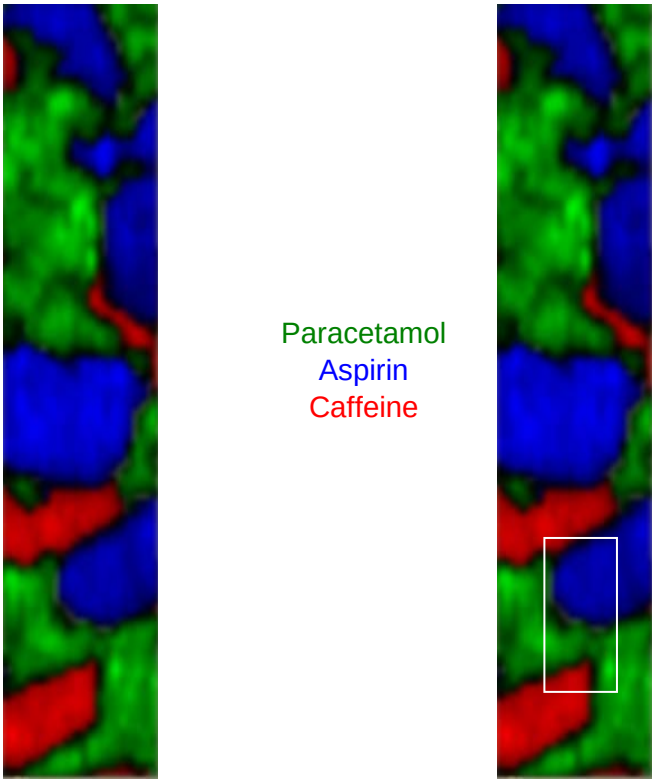
The data fraction (%) results for each component will then be displayed.

Whole image area



Rectangular sub-area





Whole image area

Rectangular sub-area (white box)

Note: The fractions provided are estimates. To optimise the accuracy of these values each step in this guide must be stringently followed.

Principal component analysis (PCA)

When do I use the 'Principal Components Analysis (PCA)' method?

The Principal Components method is used:

- when **no** reference spectra are available and the user wishes to investigate different components within the map data.

This method allows images to be created for each principal component. It is worth noting that a principal component is an abstract mathematical representation of the data in the map, and may have no obvious relationship to the chemical species present within the sample.

How does the 'Principal Components Analysis (PCA)' method work?

PCA rotates the data matrix onto a new co-ordinate system where the greatest variance by any projection of the data comes to lie on the first coordinate (first principal component), the second greatest variance, of that remaining, on the second coordinate (orthogonal to the first), and so on.

$$X = T*V + E$$

This co-ordinate system is decomposed into two matrices that are abstract representations:

- Scores (T) – concentration-like
- Loadings (V) – spectrum-like

PCA is therefore used to show data variance, but the scores and loadings do not necessarily directly represent chemical information. The new matrix also allows the data dimensionality to be conveniently reduced.

Termination methods for ‘Principal components (PCA)’

The maximum number of PCs to be extracted by PCA can be specified by the user (‘Choose number of Principal Components’) or automatically determined (by defining the ‘Minimum loadings auto-correlation’ or the ‘Minimum total variance explained’).

- ‘Minimum loadings auto-correlation’ automatically determines the number of components based on whether the different loadings spectra “look like noise” or not.

  How does ‘Minimum loadings auto-correlation’ work?
  The point-to-point auto-correlation for each Principal Component “loadings” vector is calculated. Loadings containing “real” spectral information will be smoother and therefore have a higher point-to-point auto-correlation value (close to 1). Conversely, loadings that are primarily noise will have a lower auto-correlation value (close to 0). This method requires the spectral resolution to be better than the width of spectral features (i.e. spectrally over-sampled). The default auto-correlation value of 0.7 is a good starting threshold.

  Are there any potential caveats to the use of this method?
  It is possible for a PC loading containing ‘real’ spectral features to have a lower auto-correlation value than a noisy and uninformative PC loading. This might occur when the loadings auto-correlation values have been influenced by cosmic rays or other spectral anomalies.

  How does pre-processing of the dataset affect PCA using the ‘Minimum loadings auto-correlation’ method?
  The ‘minimum loadings auto-correlation’ method is largely independent of the scaling method used.

- ‘Minimum total variance explained’ automatically determines the number of components based on how much variance in the dataset is described by the PCA model.

  How does ‘Minimum total variance explained’ work?
  PCs are ordered by the amount of variance (information in the dataset) that they explain, with PC1 always explaining more than PC2, and so on. Using this option allows the user to decide what percentage of all the variance (information) in the dataset they wish to retain. Even in datasets containing multiple chemical components, PC1 will often explain > 0.9 (90%) of the data. It is recommended that the default value of 0.98 be used to determine major PCs within the data.

  How does pre-processing of the dataset affect PCA using the ‘Minimum total variance explained’ method?
  The threshold ‘minimum total variance’ value corresponds to the percentage of information explained AFTER pre-processing with a scaling method. Therefore, the appropriate value to use will be influenced by the choice of scaling mode (especially the “Mean-centre” option).
  For example: If the variance in a (theoretical) dataset is 80% background, 15% Raman signal, and 5% noise, and no pre-processing was applied, then a threshold variance of 95% [=80+15] would be required to retain all the useful spectral information. However, if an appropriate scaling method was successfully used for background removal, only 75% [=15/(15+5)] of the total variance needs to be retained to explain all the Raman information.

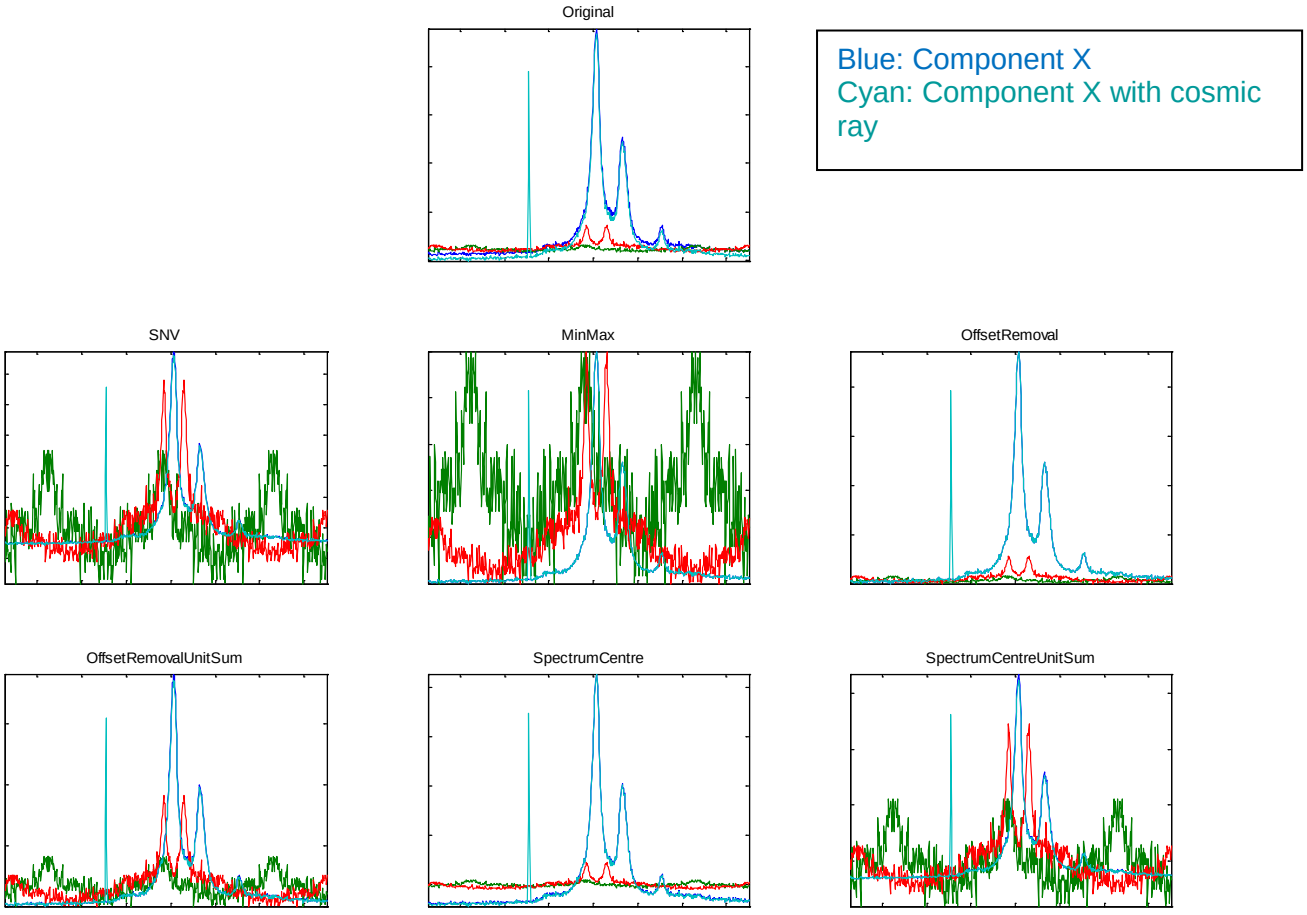
- ‘Choose number of Principal Components’ performs PCA where the maximum number of PCs is user-specified. This method is useful if the user is only interested in a specific number of components.

  How does pre-processing of the dataset affect PCA using the ‘Choose number of Principal Components’ method?
  The appropriate number of PCs required to describe a dataset can be influenced by pre-processing with different scaling methods.
  For example: with no pre-processing, a (theoretical) dataset contains four PCs, of which PC1 and PC2 are mostly background, while PC3 and PC4 contain real spectral information. If an appropriate scaling method was successfully used for background removal, the number of PCs required to describe the dataset would be reduced from four to two.

(Advanced) Pre-processing options for 'Principal components (PCA)'

In WiRE 5, the user can choose to apply pre-processing, which consists of a scaling method and a 'Mean-centre' option (either on or off). To achieve the same results as in WiRE 3.4, select "Spectrum centring + normalization (SNV)" with the 'Mean-centre' option un-checked.

- Scaling
  A number of scaling options can be used to remove any Y-offset, and / or normalise the influence of different components within a dataset. To help illustrate the different scaling options, consider an example of 4 spectra as shown in the following figure:



(Note: With the exception of 'mean-centre,' all scaling options operate on each spectrum individually. Therefore, intensity values used for pre-processing such as the average, offset, minimum or maximum, are calculated individually for each spectrum.)

No scaling: The raw dataset is used, so the PCs will describe all sources of variance between spectral components.
In the example above, map regions containing Component X (blue) will influence the PCA model more than those containing Component Y (red), simply because the signal intensity from Component X is so much stronger.

Spectrum centring + normalization (SNV): Each spectrum is scaled by subtracting the average value and dividing by the standard deviation. As a result, each spectrum is centred around zero on the Y-axis, and has a variance of exactly 1. Each component now has comparable influence on the PCA model. In the example above, Component X and Component Y spectra now have similar intensity.
SNV is useful for extracting PCs when the raw components have very different integrated intensities. However, SNV will also increase the variance due to 'empty' regions containing only noise. Cosmic rays can also influence the scaling in SNV.

Min-Max scaling: Each spectrum is scaled to fit between 0 and 1 on the Y axis. In the example, all component spectra (including empty noise spectra) have the same maximum intensity. Since PCA takes account of all the values in a spectrum, and not just the single point with the maximum intensity, pre-processing using Min-Max could potentially produce very different results compared to SNV.
In the example, the 'empty' noise spectrum is significantly amplified. Cosmic rays that are of lower intensity than the strongest peak have no effect on the scaling.

Offset removal: Each spectrum is offset so that its minimum value on the Y-axis is exactly zero.
Offset removal is useful for removing a background characterised by variable offset but consistent shape, while preserving any variation in signal intensity (and associated quantitative information). Cosmic rays will have no influence on the scaling result, but can still influence the PCA result.

Offset removal + unit sum: Offset removal is applied, and then each spectrum is scaled such that the sum of all points equals 1.
This scaling method is less effective than SNV at normalizing the influence of components with different intensities (compare the ratio of Component X and Component Y in the example above). However, 'offset removal + unit sum' performs better at not increasing the variance due to 'empty' regions when compared to SNV and Min-Max scaling. Cosmic rays will influence the scaling result.

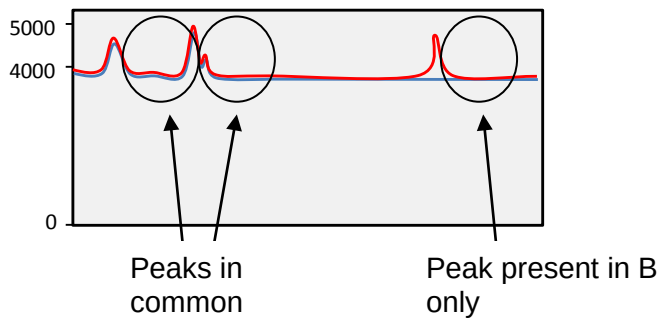
Spectrum centring: Each spectrum is scaled by subtracting the average value.

Spectrum centring + unit sum: Spectrum centring is applied, and then each spectrum is scaled such that the sum of all intensity values equals 1.

- 'Mean-centre' option
  To mean-centre a dataset, the average spectrum is calculated and then subtracted from every spectrum in the dataset. This is applied after any scaling option.

  Select 'mean-centre' if the PCA model should highlight differences between spectra, and if the typical spectrum (i.e. the substrate or background) is not of interest. If mean-centre is used, the 'number of PCs' or 'minimum total variance explained' can probably be reduced.

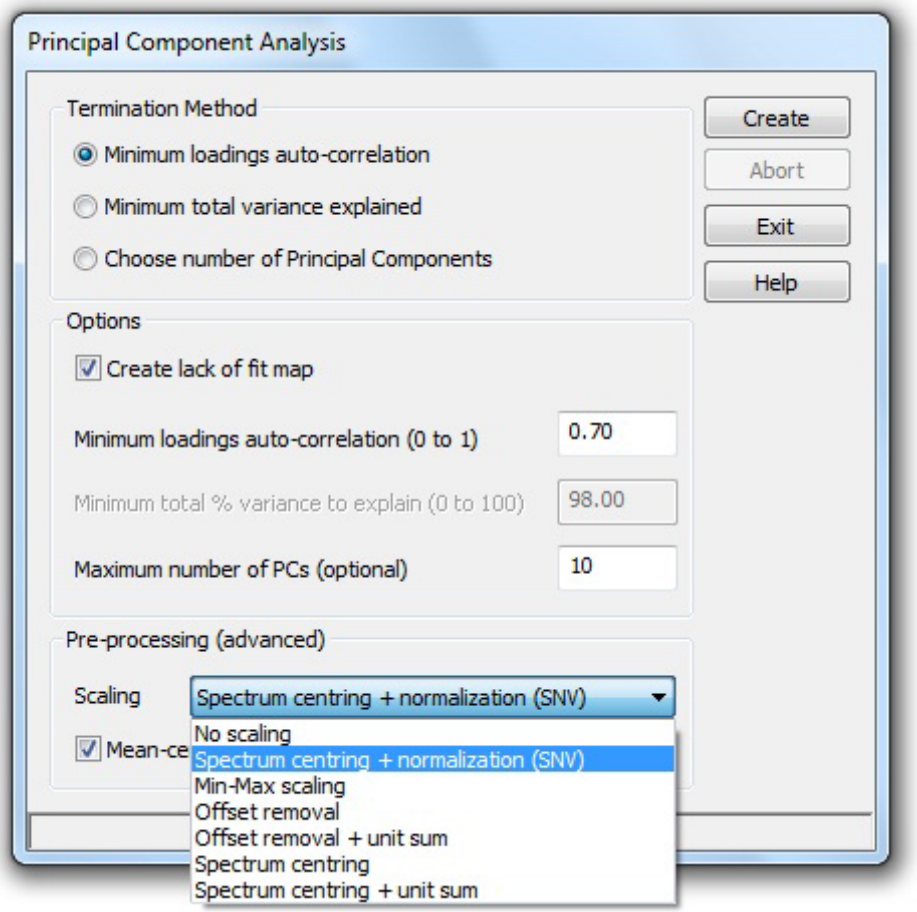
  In many cases, mean-centring has the effect of removing what would otherwise have been in PC1, while promoting PC2 to PC1, and PC3 to PC2, etc.
  For example: A (theoretical) map data set contains 2 types of spectra, A (blue) and B (red), as shown in the figure below:



A and B share two common peaks, and have a similar baseline (Y-offset) around 4000. A and B are differentiated by the presence / absence of a third peak. In the raw dataset, the vast majority of variance is due to the large baseline offset and two common peaks. Only a small proportion of variance is due to the difference between A and B.
Without pre-processing, PC1 will describe the large offset, two common peaks and a diminished version of the third peak. PC2 will describe the third peak, and will be useful for distinguishing A from B. PC3 will be noise.
With 'mean-centre,' the offset and common peaks are subtracted, so PC1 will describe the differences between A and B (mainly the third peak).

Using 'Principal component (PCA)' analysis to create component images

1. Ensure Raman data is loaded into the WiRE 5 software and select 'Principal components (PCA)' option from the map selection dialog.



2. Select a 'Termination method' for specifying the maximum number of Principal Components to be found. The maximum number of PCs can be specified by the user ('Choose number of Principal Components') or automatically determined (by defining the 'Minimum loadings auto-correlation' or the 'Minimum total variance explained').

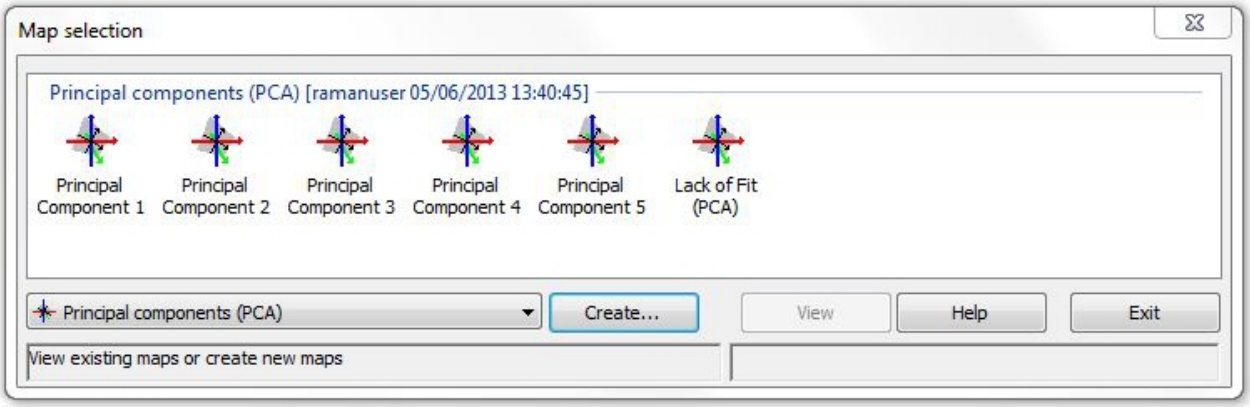
3. Create a ‘Lack of fit’ map, if desired. This creates a further image, which is useful for displaying regions of the sample, which are not explained by the chosen PCA model. If the LOF map reveals any obvious patterns, these can be an indication of further unextracted PCs. If the LOF map reveals only a noise pattern, then all useful spectral information has already been described by other PCs.

4. Select a ‘Pre-processing’ method, which consists of a scaling option and a ‘Mean-centre’ option (either on or off). A number of scaling options can be used to remove any Y-offset, and / or normalise the influence of different components within a map (see Pre-processing options for ‘Principal components (PCA)’ for more detailed information).

   If the resulting PCA component images should highlight only the differences between spectra, do select ‘mean-centre’. If the resulting PCA component images should describe the typical spectrum (i.e. the substrate or background), then do not use mean-centre (see Pre-processing options for ‘Principal components (PCA)’ for more detailed information).

   To achieve the same results as in WiRE 3.4, select “Spectrum centring + normalization (SNV)” with the ‘Mean-centre’ option un-checked.

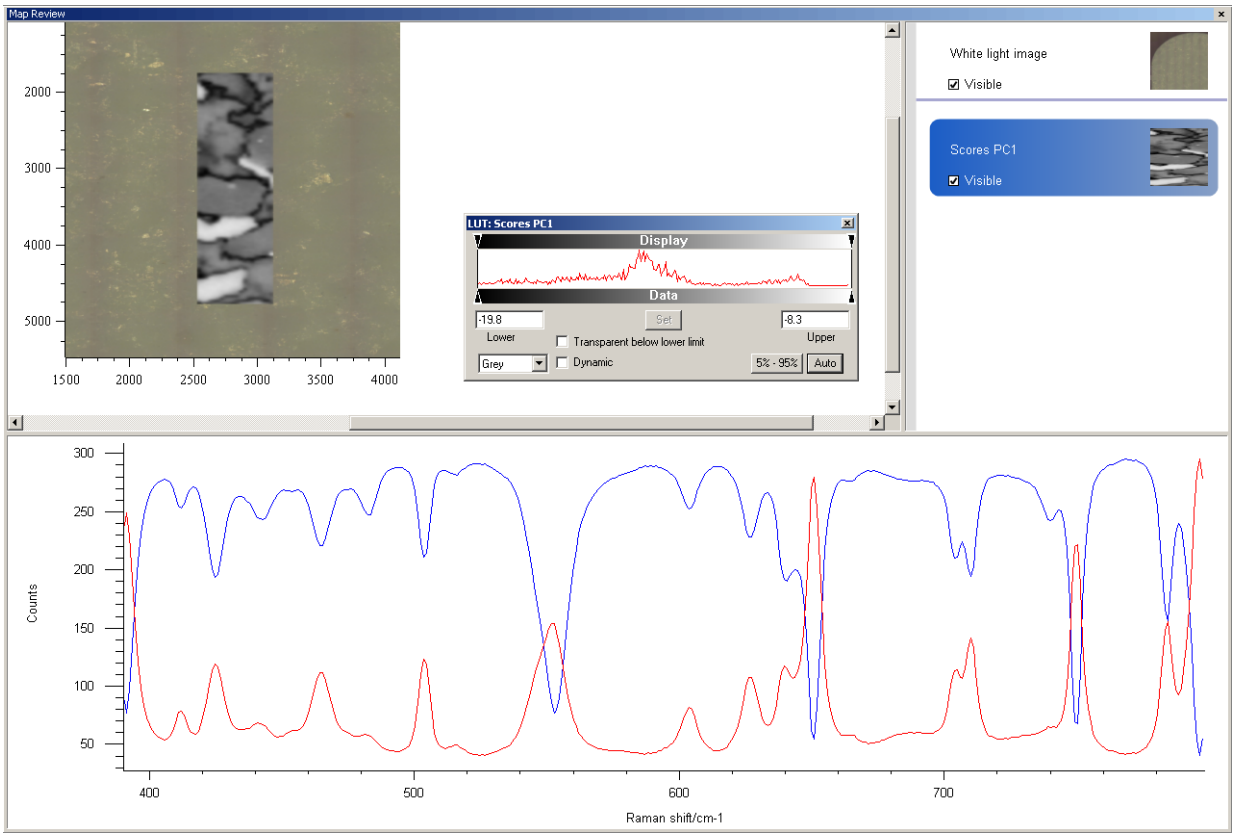
5. Select ‘Create’.



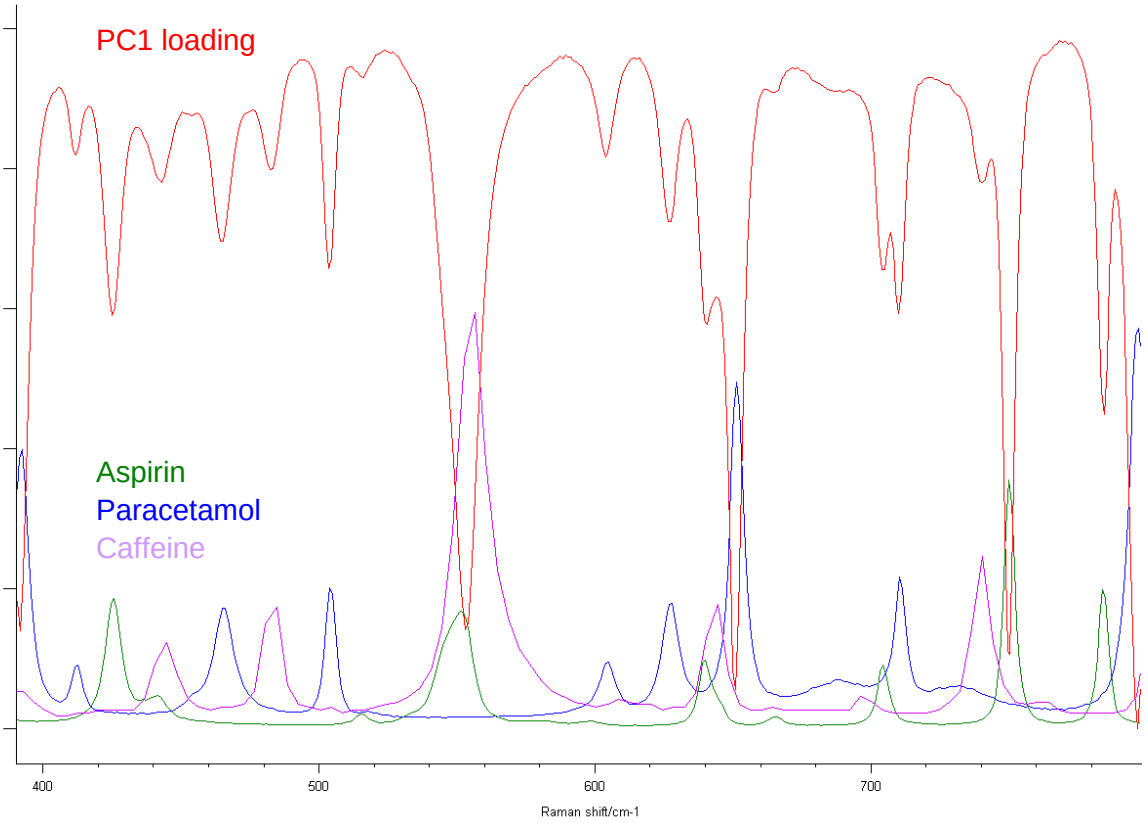
PCA example using ‘Choose number of Principal Components’ method:

PCA was performed using the ‘Choose number of Principal Components’ method. The dataset was pre-processed using ‘spectrum centring + normalization (SNV)’, without mean centring.

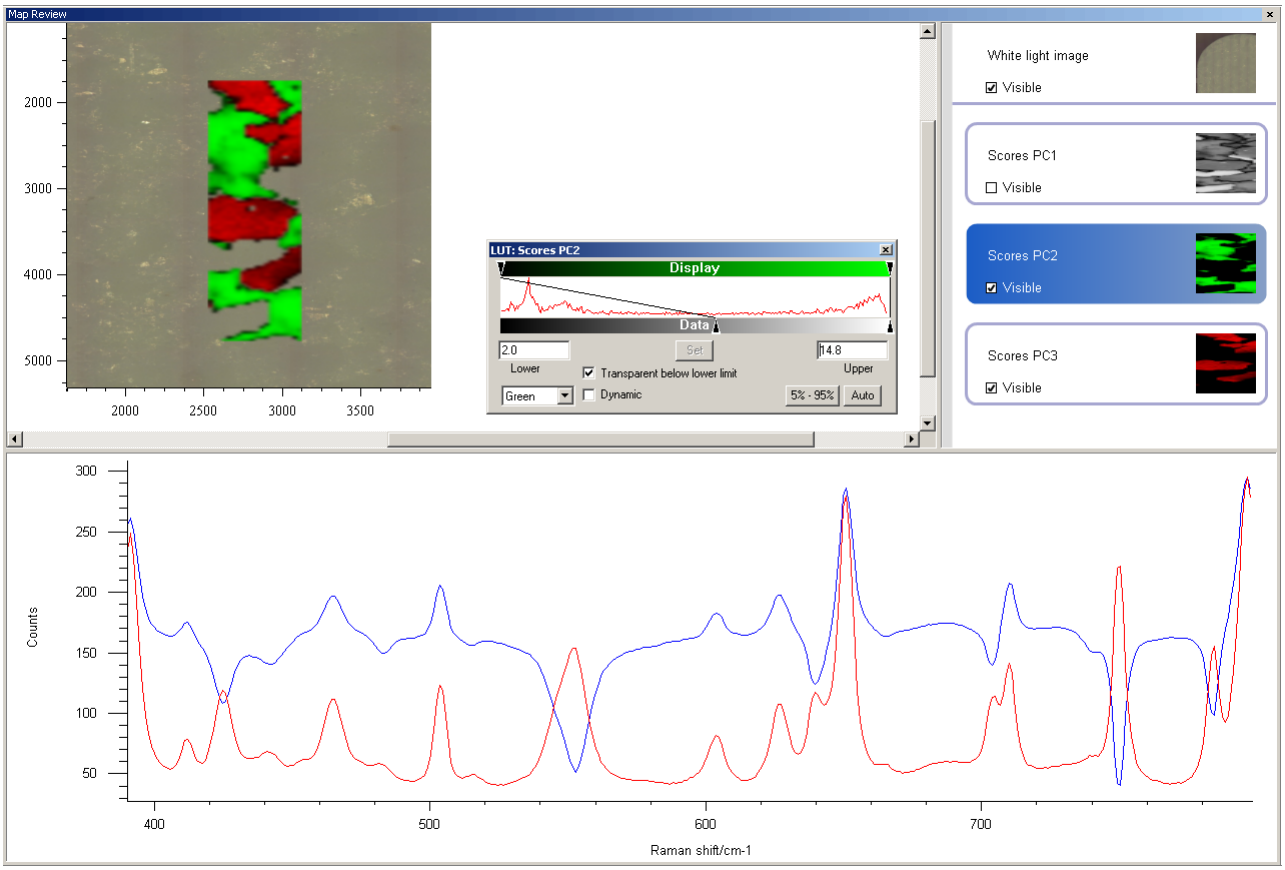
PC1 describes the average spectrum, the loadings in this case will be unidirectional i.e. either positive or negative only, but never mixed. In the example shown the loadings are negative.



The PC1 loading therefore contains a mixture of all the different spectral components which make-up the dataset. Weak spectral components will have a smaller contribution to the average spectrum, PC1.


PC1 loading
Aspirin
Paracetamol
Caffeine
400
500
600
700
Raman shift/cm-1

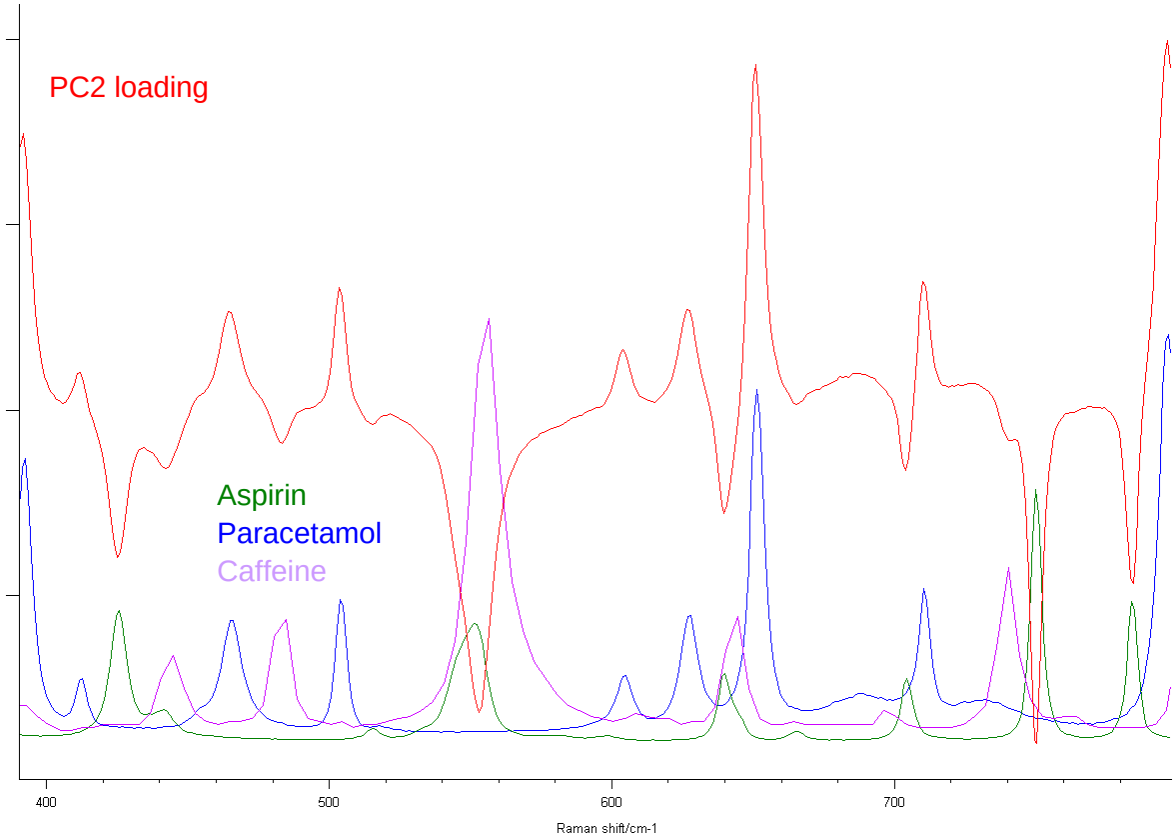
The PC2 and PC3 scores (images) show the regions on the sample which are strongly characterised, either positively or negatively, by the corresponding loadings spectrum.

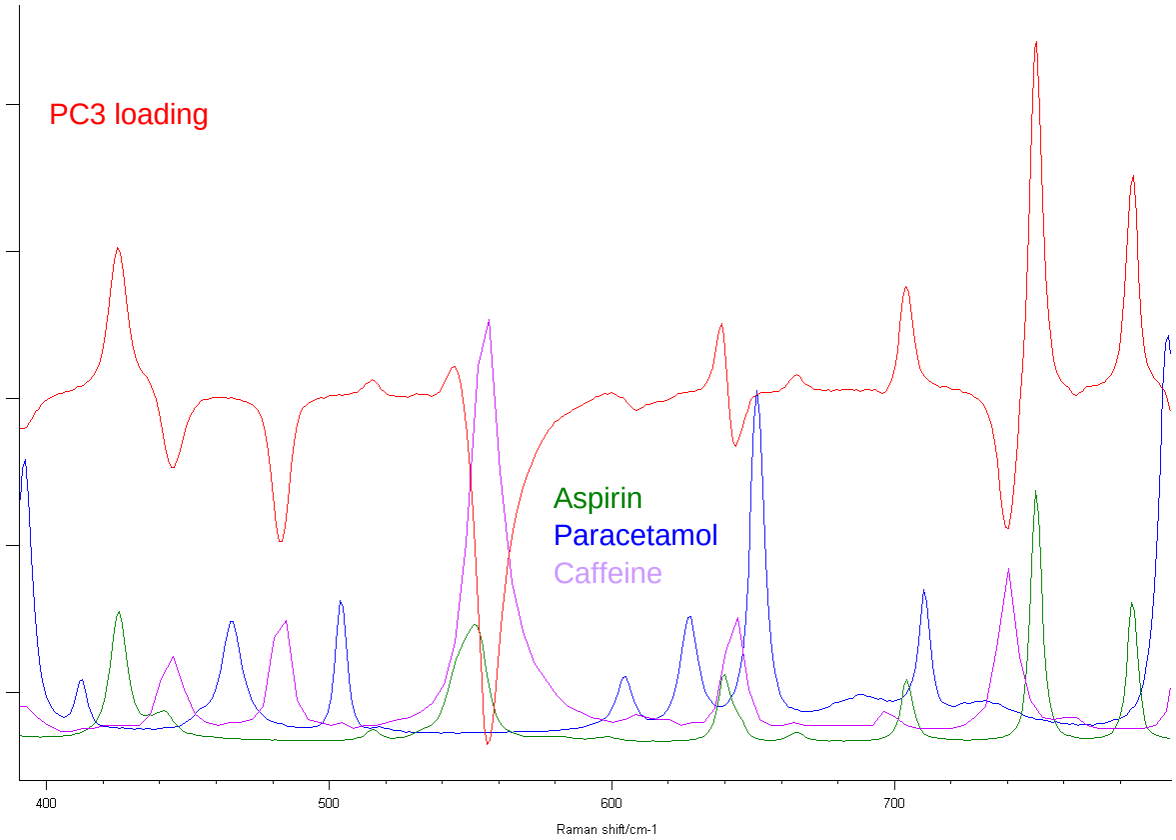


The PC2 loading shows the principal component which describes the greatest remaining variance. The PC2 co-ordinate is orthogonal to PC1 and shows spectral bands relating to PC2 (paracetamol bands) as positive whilst those from other PCs (aspirin and caffeine) are negative.

The PC3 loading shows the principal component which describes the greatest remaining variance after PC1 and PC2. The PC3 co-ordinate is orthogonal to PC1 and PC2 and shows spectral bands relating to PC3 (aspirin bands) as positive whilst those from other PCs remaining, not already described (caffeine only), are negative. Paracetamol information is not observed at all within the PC3 loading as this has been removed in the PC2 loading.

A PC loading that contains positive spectral bands from one species, and negative spectral bands from another species, indicates a strong inverse spatial correlation between the two. In this example, the first three PC loadings demonstrate this behaviour, and indicate that the three main components are present in virtually pure form.


PC2 loading
Aspirin
Paracetamol
Caffeine
400
500
600
700
Raman shift/cm-1


PC3 loading
Aspirin
Paracetamol
Caffeine
400
500
600
700
Raman shift/cm-1

Empty Modelling

When do I use the ‘Empty Modelling’ method?

The ‘Empty Modelling’ method is used:

- when **no** reference spectra are available and the user wishes to extract images and (pure) spectra representing physically meaningful components
- works particularly well when the main components are present in pure form somewhere within the dataset.

How does the ‘Empty Modelling’ method work?

The Renishaw Empty Modelling™ method is a form of multivariate curve resolution – alternating least squares (MCR-ALS) with no requirement for the user to specify initial concentration or spectral estimates. Component spectra and concentration images are iteratively resolved in this technique.

Multivariate curve resolution – alternating least squares (MCR-ALS) methods allow component information to be rotated into physically meaningful components. X, the data matrix, is now decomposed into concentration images (C) and (pure) spectra (S), rather than into abstract vectors:

$$X = C*S^T + E$$

C and S are constrained to positive values and the equation is solved iteratively. This is important as the quality of information will vary depending on:

- purity of spectra in the dataset, particularly of the main component
- specific spectral forms of the components

Termination methods for ‘Empty Modelling’

The maximum number of components to be extracted by Empty Modelling can be specified by the user (‘Choose number of components’) or automatically determined (by defining the ‘Minimum total % variance explained’).

- 'Minimum total % variance explained' continues to add components until a minimum proportion of the variance is described by Empty Modelling.
  A suitable variance threshold can be reduced with appropriate pre-processing of the dataset, but the effect is less pronounced than for PCA with / without mean-centring. To determine major components within a dataset, a typical value of 0.98 is recommended.

- 'Choose number of components' allows the user to perform MCR-ALS where the solution is limited by a maximum number of components and is not constrained by a minimum variance value.

Advanced options for 'Empty Modelling'

Advanced options for 'Empty Modelling' can be used to specify the type of scaling (normalisation) that is used for pre-processing, to turn on/off 'apply closure constraint to estimated concentrations,' and to turn on/off 'accelerate algorithm using sub-sampling for large files.'

- Pre-processing
  No scaling: Empty Modelling is performed on the dataset without any pre-processing.
  This option is suitable where there is minimal background, and intensity (concentration) information is of interest.

  Offset removal: Each spectrum is offset so that its minimum value on the Y-axis is exactly zero.
  This option is suitable for removing any background Y-offset, while retaining any intensity (concentration) information.

  Offset removal + unit sum: Offset removal is applied, and then each spectrum is scaled such that the sum of all points equals 1.
  This option is suitable for removing any background Y-offset, and normalising the influence of different components to obtain a qualitative Empty Modelling result.

- Apply closure constraint to estimated concentrations: This option constrains the sum total of the component intensities (concentrations) at each map pixel to be 100%.

  This closure constraint is unsuitable in cases where:
  - the map contains significant amounts of 'empty' regions (i.e. regions where there is no Raman signal)
  - some components do not produce a Raman signal
  - the sum of the component intensities is expected to vary with concentration (and this is of interest)

  If the pre-processing method is chosen to be either 'no scaling' or 'offset removal' so that intensity (concentration) information is retained, then do not select 'apply closure constraint to estimated concentrations.'

- Accelerate algorithm using sub-sampling for large files: This option enables faster Empty Modelling of large datasets (containing > 15,000 spectra) by estimating pure spectra using a sub-sample of the file, before producing images for the entire dataset.

  This option should normally be enabled when performing Empty Modelling on large datasets. Sub-sampling should only be disabled if a component is represented by less than 5% of spectra within the map.

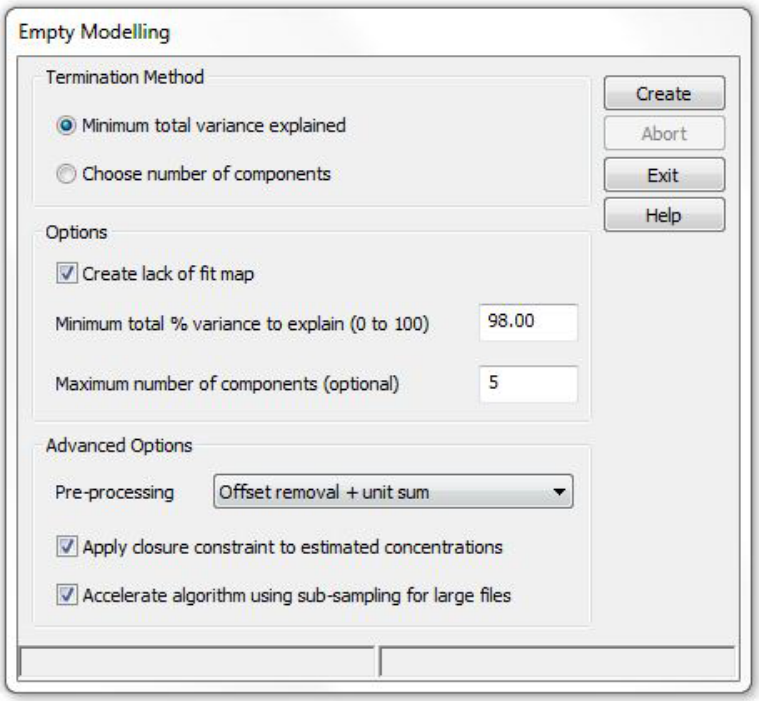
  Using 'Empty Modelling' analysis to create component images

  1. Ensure Raman data is loaded into the WiRE 5 software and select 'Empty Modelling' option from the map selection dialogue.

  2. Select 'Method' type.

     - 'Minimum variance explained' allows the user to set a threshold on how much of the data is described by the MCR-ALS model. It is recommended that a typical value of 0.98 be used to determine major components within the data. This method is useful if the number of distinct components in the map is not known.

     - 'Choose number of components' allows the user to perform MCR-ALS where the solution is limited by the number of components allowed and is not constrained by a minimum variance value. This method is useful if the user already knows with reasonable confidence the number of components they expect to find in the map.



3. Create ‘Lack of fit’ map if desired. This creates a further image which displays regions of the sample which are not explained by the Empty Modelling analysis. If the LOF map reveals any obvious patterns, these can be an indication of further unextracted components. If the LOF map reveals only a noise pattern, then all useful spectral information has already been described by the model.

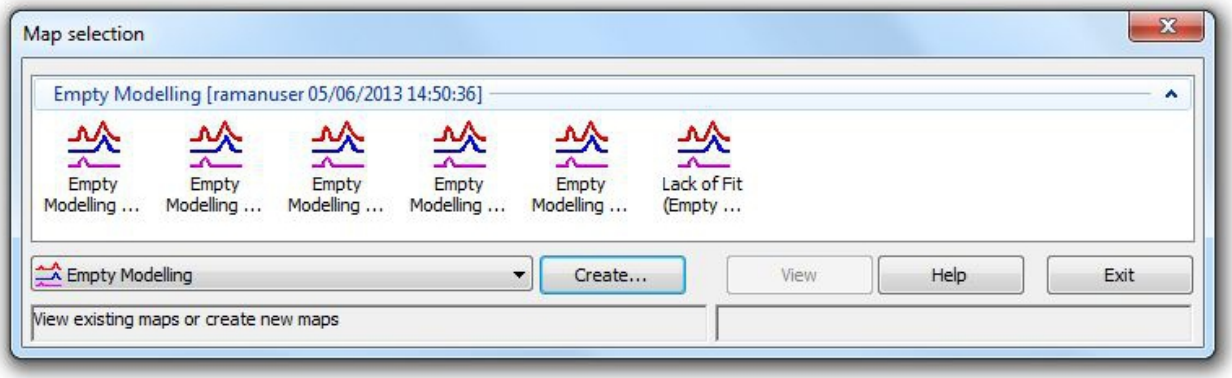
4. Select a pre-processing option, if required. By default, ‘offset removal + unit sum’ is used to normalise the dataset before Empty Modelling is performed.

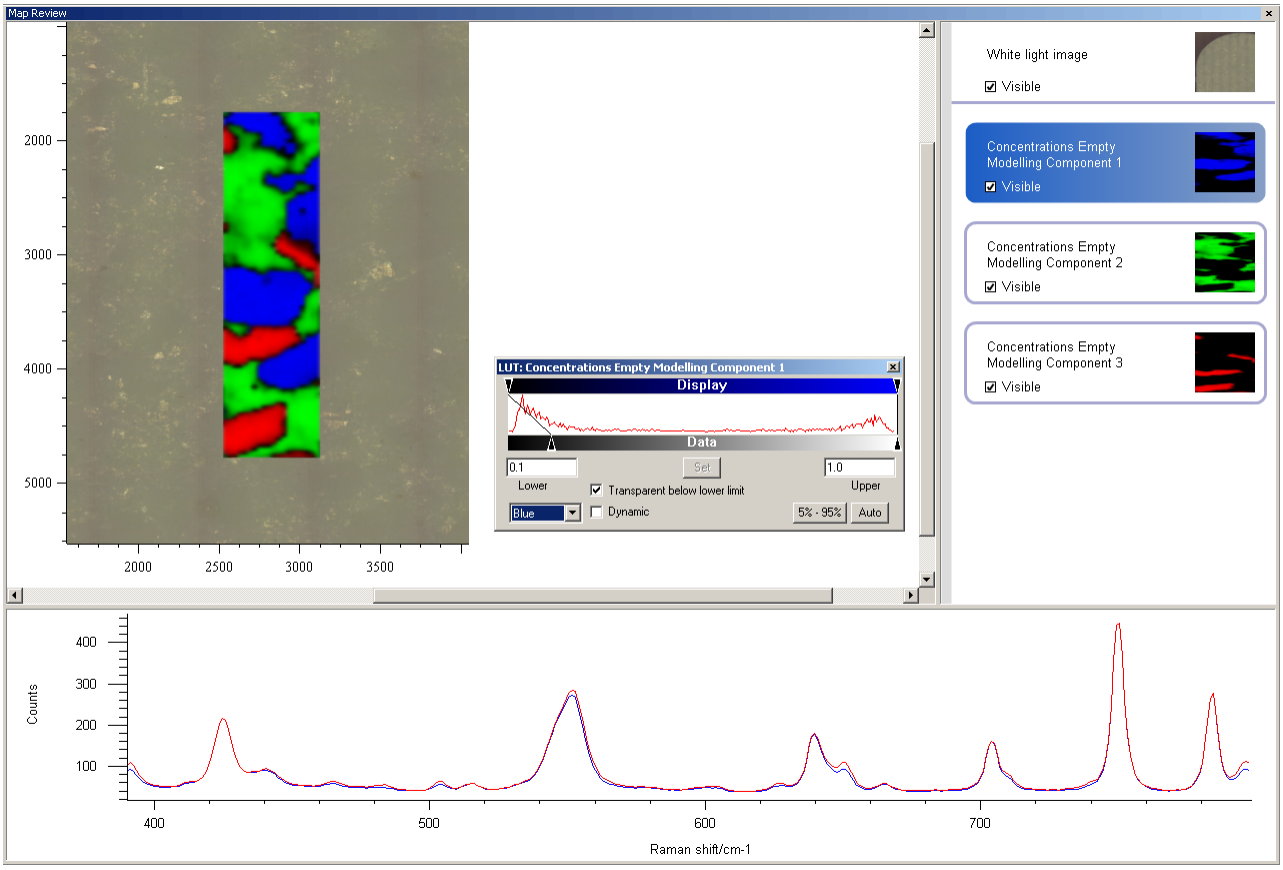
   Select / deselect ‘apply closure constraint to estimated concentrations,’ as required.

   Select / deselect ‘accelerate algorithm using sub-sampling for large files,’ as required.

   (To obtain identical Empty Modelling results to WiRE 3.4, use ‘offset removal + unit sum’ and select ‘apply closure constraint to estimated concentrations.’)

5. Select ‘Create’.





The Empty Modelling images show the location of each chemical species within the image data. This method creates ‘real’ spectra rather than abstract loadings (PCA). As a result of this the component determined by Empty Modelling (blue) and the image spectrum (red) are almost identical.