# OPEN ACCESS DOCUMENT

Information of the Journal in which the present paper is published:


- NPG, Nature Protocols, 2015, 10, pp. 217-240.
- DOI: dx.doi.org/ 10.1038/nprot.2015.008

# Vibrational spectroscopic image analysis of biological material using multivariate curve resolution – alternating least squares

Judith Felten[a] Hardy Hall[a] Joaquim Jaumot[b] Romà Tauler[b] Anna de Juan[c]
András Gorzsás[d]

[a]Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, Umeå, SWEDEN
[b]IDAEA-CSIC, Barcelona, SPAIN
[c]Chemometrics Group, Department of Analytical Chemistry, Universitat de Barcelona, Barcelona, SPAIN
[d]Department of Chemistry, Umeå University, SE-90187 Umeå, SWEDEN

Correspondence should be addressed to A. G. (andras.gorzsas@chem.umu.se)
Telephone: 00 46 90 786 5918
Fax: 00 46 90 786 76 55

Email addresses:
Judith Felten Judith.felten@slu.se
Hardy Hall Hardy.hall@slu.se
Joaquim Jaumot joaquim.jaumot@idaea.csic.es
Romà Tauler rtaqam@idaea.csic.es
Anna de Juan anna.dejuan@ub.edu
András Gorzsás andras.gorzsas@umu.se

## ABSTRACT

Multivariate data analysis techniques are ideal to decrypt chemical differences between anatomical features or tissue areas in hyperspectral images of biological samples. This protocol provides a user-friendly pipeline and graphical user interface (GUI) for data pre-processing and un-mixing of pixel spectra into their contributing pure components by multivariate curve resolution-alternating least squares (MCR-ALS) analysis. The analysis considers the full spectral profile to identify the chemical compounds and to visualize their distribution across the sample to categorize chemically distinct areas. Results are rapidly achieved (usually less than 30 - 60 min/image) and are easy to interpret and evaluate both in terms of chemistry and biology, making the method generally more powerful than principal component analysis (PCA) or single band intensity heap maps. In addition, chemical and biological evaluation of the results by means of reference matching and segmentation maps (based on k-means clustering) are possible.

## INTRODUCTION

Imaging is an essential tool for biological studies that aim to understand gene function related to developmental processes and phenotypical outputs. By using biochemical, molecular biological and spectroscopic tools to augment imaging, sample anatomy and morphology can be correlated to molecular characteristics. The term hyperspectral imaging is used for the techniques that utilize spatially resolved spectroscopic information to create images. Vibrational microspectroscopic techniques, such as Fourier Transform Infrared (FTIR) and Raman microspectroscopy, are particularly suited for hyperspectral imaging of biological materials, since they are fast, non-invasive, non-destructive and inexpensive. They are also very versatile and provide molecular level information with little to no sample preparation and without staining1. In addition, they can be used without a priori knowledge of sample composition, in contrast to immunolocalization studies of targeted bio-polymers2,3. These advantages make FTIR and Raman microspectroscopy popular in a broad field of sciences, ranging from medicine (e.g. identifying abnormal (cancerous) tissue areas based on their chemical composition4-6) to wood biotechnology (e.g. analyzing the chemical composition of different cell types or cell walls7-9). In plant sciences, in particular, vibrational microspectroscopy contributed to identifying the role of genes involved in the wood biosynthetic machinery10-13 by enabling the mapping of chemical compositional changes in transgenic plants and at different developmental stages. The major types of biopolymers in woody cell walls (cellulose, lignins, hemicelluloses and pectins) have characteristic FTIR / Raman spectroscopic fingerprints8,14-22, making vibrational microspectroscopy suitable for the chemical imaging of wood. Historically, FTIR microspectroscopy has been more commonly used for this purpose23, mainly because of the fluorescence problems encountered in Raman spectroscopy due to lignin. However, with the development of strategies to circumvent fluorescence problems, Raman spectroscopy is rapidly gaining popularity due to its high spatial (confocal and lateral) resolution8,12,14-16. This high spatial resolution enables the chemical analysis of distinct (sub)micron-sized zones or layers within woody cell walls8,14, providing an advantage over standard FTIR microspectroscopy, which usually aims at single cell resolution only7.

### Vibrational microspectroscopic image analysis

The three main steps of gaining chemical information using vibrational microspectroscopic techniques are sample preparation, sample measurement and data analysis. Detailed protocols have been developed for plant biologists regarding sample preparation and spectroscopic measurements7,14,24. However, data analysis often stops at the level of

generating simple band intensity maps (heat maps), or is performed on a case-by-case basis by specialists in the fields of chemometrics and spectroscopy7. As a result, biologists are often disconnected from the data analysis steps, which can refrain them from exploiting vibrational microspectroscopy in their research. Therefore, we present a pipeline (Figure 1), together with a detailed protocol and a software script (Supplementary Figure 1) specifically designed for users with limited background in chemometrics and spectroscopy. This protocol is centered around multivariate curve resolution - alternating least squares (MCR-ALS) analysis streamlined for a well-defined kind of data analysis problem, namely vibrational microspectroscopic image analysis of biological samples. Accordingly, most MCR-ALS parameters are pre-selected, and certain analytical options are entirely omitted (see Table 1 and the ADDITIONAL COMMENTS AND LIMITATIONS section of this protocol) for ease of use and to make it possible for biologists to rapidly and reliably analyze their own data, and to gain a quick overview of a large number of samples.

Consultation with specialists can thus be restricted to the most complicated cases, where the basic application of the algorithm may be insufficient, or to the more advanced steps of the analysis (e.g. comparing the results of different models, spectral band interpretations, etc.).
The protocol integrates the most suitable procedures for advanced vibrational microspectroscopic image analysis into a complete package via an interactive graphical user interface (GUI, freely available as a MATLAB script at www.kbc.umu.se/vibrationaldownload.html). Together with the already existing protocols for sample preparation and recording14,24, it forms a complete vibrational microspectroscopy suite for plant sciences. While we focus on Raman microspectroscopy and plant material, the protocol is not limited to these. It can be directly applied to both Raman and FTIR microspectroscopic images of any kind of biological (including medical) samples (Figure 2), and it covers all the important stages of the analysis in several steps:
1) pre-processing of the spectra;
2) multivariate curve resolution – alternating least squares (MCR-ALS) analysis of the data, focusing on un-mixing the different chemical constituents;
3) evaluation of the results in terms of chemistry by reference spectra matching; and 4) visualization and evaluation of the results in terms of anatomy using pure component maps and segmentation maps. The main goal of the present paper is to provide a step-by-step manual for each of these stages in the PROCEDURE section. In that section, we only provide information that is needed for practical decision-making and troubleshooting.

The underlying principles, theory and other background information regarding all major parts of the analysis are detailed in the BACKGROUND section below.

## BACKGROUND

### Spectra pre-processing

Pre-processing of the recorded raw data is required prior to data analysis in order to remove all variation that is uncorrelated to, and interferes with, the chemical information in the spectra (fluorescence, background or total signal intensity variations, noise, etc). Below, we describe a procedure that accomplishes this task through baseline correction by asymmetric least squares fitting and optional smoothing and area normalization. This procedure assumes that the input data have already been corrected for basic method-related artifacts (e.g. cosmic rays) in accordance with previously described Raman imaging procedures14.

#### *Baseline correction*

The most commonly required pre-processing step is baseline correction. In addition to environmental and instrumental sources (e.g. temperature or source intensity fluctuations, vibrations, etc.), baseline variations can be caused by the inherent optical and physicochemical properties of the sample (edge effects, hot spots, auto-fluorescence, refractive index heterogeneities, etc.). While entirely linear baselines can theoretically exist, they are practically non-existent in real images. This limits the usefulness of simple one-point (offset) or two-point linear baseline corrections, especially in the case of Raman images of plants with contributions from fluorescence. Multi-point linear or polynomial baseline corrections can approximate real baselines better, but can be extremely difficult to perform correctly and reproducibly25. This is especially true in the case of vibrational spectra of biological materials, where broad, overlapping bands cover significant areas of the spectrum and where distinct image regions / pixels can have different, intense and irregularly shaped baselines. These features make it difficult to determine a fixed set of baseline points for polynomial fitting.

The method of choice for baseline calculation of vibrational microspectroscopic images of biological material uses asymmetric least squares (AsLS) fitting (http://www.science.uva.nl/~hboelens/publications/draftpub/Eilers_2005.pdf), originally proposed for chromatograms by Eilers26. AsLS baseline correction is an iterative method based on the use of a Whittaker smoother to fit the baseline to the data. It is fast, flexible, easy to perform and automate and only the selection of two parameters is required. The first parameter is the lambda value, which determines how smoothly (closely) the baseline is fitted to the data. Higher lambda values result in a more linear baseline, while lower lambda values generate baselines that closely follow the curvilinear natural baseline shape. Lambda values should be adjusted so that the corrected spectrum does not contain any broad features from leftover baseline, while at the same time it retains even low intensity bands. Overly high lambda values result in underfitted baselines, where the broad baseline features are not removed. Conversely, overly low lambda values result in over-fitted baselines, where spectral band intensities diminish to such extent that low intensity bands could disappear entirely or suffer an artificial loss of intensity and variation in the band shape (Figure 3). Lambda values should always be adjusted to the working data set and tested on representative spectra of the image, by visually inspecting the fitted baseline and the resulting corrected spectra. The second parameter to be adjusted is the p-value. It gives a different weight to those points in the fitted baseline that have positive residuals (i.e. where bands or spectral features are present) as compared to those points that have negative residuals in the fitted baseline. Since vibrational microspectroscopic images of biological material should not contain negative peaks, p-values should be kept at the minimum value by default. It is important to note that this method is based on local fitting and does not apply any predefined baseline shape. This is why it adapts extremely well to data with irregular baselines, even if each spectrum of the dataset has differently shaped baselines and intensities.

### *Smoothing*

Smoothing is an optional pre-treatment step that needs to be used with caution. It can improve data quality by removing noise, but excessive smoothing causes information loss by decreasing spectral resolution, which is easily noticed by the resulting distortion and the merging of bands (Figure 4). Of the different spectral smoothing methods, Savitzky-Golay filtering27 is the most widely used, because it can improve the signal-to-noise ratio without distorting the signal. The smoothing is controlled by two factors: the polynomial order and the frame size (or number of convolution coefficients). In practical terms, the larger the difference between polynomial order and frame size, the more smoothing is applied. If the difference is 1, no smoothing is produced28. Smoothing parameters must always be visually evaluated from

the graphical output on a case-by-case basis, to ensure that they result in noise reduction without signal distortion.

### *Area Normalization*

Area normalization of the raw image spectra is entirely optional, since it is not required to describe the variation among pixels by multivariate analysis (Supplementary Figure 2). On the contrary, area normalization is generally not recommended in resolution image analysis. However, it can facilitate the interpretation of component distribution maps in cases where significant intensity fluctuations exist (due to the optophysical properties of the sample, such as pixel coverage, sample thickness, refractive index variations, out-of-focus effects, etc.). It ensures that all pixels have the same total intensity, and thus the observed intensity variations in component maps represent (proportional) chemical concentration changes. In short, area normalization is only recommended if significant optophysical contributions are present, seriously hindering data evaluation or interpretation.

## Data analysis

The aim of the data analysis is to identify the chemical composition of the sample in a spatially resolved manner. This often includes the identification of chemically distinct zones and their relation to biological information, such as anatomy, gene/protein expression, hormone gradients, enzyme activity etc. Different approaches can be taken to identify chemically different zones. Selection can simply be based on anatomical features observed in the white light image (cell types, developmental stage, etc). Spectra from the selected areas/pixels can then be exported and the chemical information evaluated by different kinds of multivariate analysis7,29. However, visual identification of such zones can be problematic. Firstly, there may be no visible features to base the selection on. For instance, the different cell wall layers of woody material cannot be distinguished in the white light image with the commonly used setups for vibrational microspectroscopy. Secondly, the chemically distinct zones may not (yet) be correlated to features visible in the white light image. For instance, at the starting phase of an infection process, healthy and newly infected cells may appear morphologically identical, whereas their chemical composition may already differ. And finally, large sets of spectra from different samples may need to be compared, which requires a reliable (semi)automated and objective approach. This is the case when comparisons of genotypes, cell types, or treatments need to be performed with many biological and technical replicates. Therefore, different methods have been developed to identify chemically different zones based on the spectral information of the hyperspectral images, in addition to the white light image.

### *Band intensity heat maps*

The quickest, easiest and thereby most commonly applied strategy for vibrational microspectroscopic data analysis is the mapping of integrated band areas (intensity heat maps)8,18,30. This method assumes that the intensity variation of the band used for the mapping is only determined by the (relative) concentration variation of the compound(s) associated with that band. Thus, it can give erroneous results when bands overlap or shift, and it is very sensitive to baseline correction errors and spectral artifacts (dispersive line shapes, noise, etc.). Moreover, heat maps based on single bands do not consider the entire spectral information, which is a significant drawback if the specificity of the mapped band is low (non-diagnostic band). Even in the case of perfect diagnostic bands, no information is obtained regarding the ratio of the mapped compound in relation to other compounds, which could better describe the chemical variation over the sample than the distribution of a single compound. Finally, heat maps may show gradients within the sample but no clear boundaries.

### *Principal component analysis (PCA)*

Principal component analysis (PCA) is a multivariate analysis method that takes the entire spectrum of the pixel into account to describe the variation within the image using a small number of basic contributions (principal components) related to spectral and spatial behavior31. It overcomes the above-mentioned problems of band intensity heat maps and is better suited to handle biological variation. Thus, PCA (alongside segmentation/cluster analysis) has gained popularity in the analysis of vibrational microspectroscopic images of biological materials14,32,33. However, interpreting the results of PCA in terms of chemistry remains difficult, since the scores (abstract distribution maps) and the loadings (abstract spectra) of principal components are estimated to be completely uncorrelated to each other, which is not the case for distribution maps and spectra of real chemical compounds. As a consequence, principal components cannot be generally associated with single chemical compounds and may even contain information unrelated to the chemical compounds (e.g. scattering, fluorescence, etc.) if no proper pre-processing has been carried out. Additionally, several principal components may be required to differentiate distinct chemical zones. This further complicates data interpretation and visualization, and makes it problematic to uniformly differentiate the same kinds of zones in several images. Finally, PCA loadings are generally hard to compare to reference spectra, making it difficult to evaluate the contribution of different chemical compounds to each principal component.

### *Multivariate curve resolution – alternating least squares (MCR-ALS) analysis*

Another approach that shares the underlying bilinear mathematical model of PCA but considerably improves the interpretation of the results is multivariate curve resolution using alternating least squares (MCR-ALS), followed by cluster analysis34. Essentially, MCR-ALS assumes that the complex spectrum in every pixel of an image can be described as a linear combination of the signal of a set of pure component spectra, conveniently weighted according to their abundance in each pixel. Under this assumption, MCR works by un-mixing the original complex measurement (image data set) into the contributions of each of the pure components providing a signal. This makes it particularly suitable for the analysis of hyperspectral images. In mathematical terms, the MCR-ALS model is described as **D** = **CST**, where **D** is a matrix containing the spectra of all pixels of the image, and **ST** and **C** are matrices of the pure spectral signatures and the related distribution maps (concentration profiles) of the image constituents, respectively (Figure_5). Each one of the resolved (un-mixed) pure contributions is thus represented by a pure spectral signature (a row in **ST**) and a related distribution map, showing its abundance in each pixel of the image (a column in **C**). The main difference with respect to PCA is that, instead of imposing orthogonality (i.e. lack of correlation) among components, MCR models the shapes of distribution maps and spectra according to natural chemical, spectroscopic and mathematical properties of the image data (e.g. nonnegativity). Optimization of the pure component spectra and the distribution maps is iterative until convergence is achieved. A “pure component” in the context of MCR image analysis can be a pure chemical compound, or a part of the sample with a consistent spectral signature (i.e. a homogenous mixture of compounds). “Pure” in the latter case means that it cannot be un-mixed (resolved) further.

The pure component distribution maps provided by MCR-ALS can be used as starting (input) information for subsequent cluster analysis, since these maps are compressed representations of the information in the original image data set (Figures 1 and 5). MCR followed by clustering is a powerful tool to identify chemically distinct zones of the biological sample, either based on pure chemical compounds or on mixtures of pure compounds in similar proportion. It works well even for small hyperspectral image maps (Figure 6), which is of great advantage when the

scanned tissue zones need to be kept small to enable screening many samples in reasonable time. Different software packages and detailed descriptions are available for applying MCR-ALS (and subsequent clustering)34,35. These are extremely powerful and versatile, but the variety of options and settings requires experience in chemometrics or spectroscopy for proper use. In particular, MCR-ALS optimizes the spectral and concentration profiles (distribution maps) of the image components under certain constraints. Constraints are criteria of biochemical or mathematical origin that the sought profiles must fulfill. We can divide the large variety of constraints into two main categories: those that are applicable for images and those that are designed for other kinds of data sets (e.g. processes). The latter includes constraints such as unimodality, closure, hard-modelling, etc.36,37. Since these do not apply to the present protocol, they will not be addressed here any further. Among constrains applicable for image analysis, the most commonly used is non-negativity. Concentration values in the distribution maps naturally fulfill this criterion, and pure spectra of compounds in vibrational spectroscopy also contain zero (baseline) or positive values (bands) only, unless specific operations (spectral subtraction or derivation) have been performed. To avoid intensity fluctuations in the recovered pure spectral signatures, normalization of the pure spectra in **ST** is also applied. These basic constraints can be applied to all vibrational image data sets and are pre-fixed in the present script for ease-of-use. It is important to note that MCR-ALS is an iterative method, and as such the correct progress of the optimization depends on several aspects. The first is the starting point for the optimization, determined by the number of components and initial pure spectral estimates. The second is the endpoint of the optimization, which is determined by the number of iterations and the convergence limit and by the subsequent quality assessment of the results obtained. We provide guidelines for selecting the correct values for these factors, together with notes on how to evaluate whether the optimum results have been obtained by MCR-ALS or the analysis needs to be repeated to test a different set of parameters.

### Determining the number of components in the image data set

The number of components must be determined before MCR-ALS can be applied. In some cases, this may be known already beforehand (e.g. mixtures with known components). However, when no *a priori* information is available, singular value decomposition (SVD) can be used to estimate the number of components. SVD is an algorithm that describes the data set by an abstract model of bilinear contributions, formally analogous to the MCR-ALS model (see above). Assuming correct pre-processing (i.e. elimination of intense baseline contributions or other artifacts), the relevance of each component is defined by the magnitude of its singular value: high singular values relating to biochemical contributions and small values to noise. The number of components that have significant singular values is therefore chosen as total number of components used in the MCR-ALS model. However, in biological images the exact number of components is often not evident from SVD values alone (Figure 7). It is therefore recommended to test several models using different numbers of components and evaluate the results in terms of model fit quality and interpretability of the final maps and spectra obtained.

### Initial estimates

To start the optimization process by alternating least squares, an initial estimate of the pure spectral signatures (the matrix **ST**, see above) should be available. In the context of image analysis, the best option is to use a method that is based on the selection of the purest spectra in the image data set. For the present work, we chose a method based on SIMPLISMA38 to determine the initial pure spectra estimates. SIMPLISMA (SIMPLe-to-use Inter- active Self-modeling Mixture Analysis) is a method based on the sequential choice of pixels with the purest spectral signatures within the raw image data set. It works well even on hyperspectral

images of any size, and has been used successfully in various Raman and FTIR microspectroscopic image analyses38-40.

## Evaluation of the results

The main criteria for the quality assessment of the MCR-ALS results are satisfactory model fit (mathematical evaluation) and meaningful distribution maps and spectra (biological, chemical and spectroscopic evaluation). We provide the basic information for evaluating each of these criteria. It has to be noted that no priority among these criteria can be set, and thus they should always be evaluated together.

### *Input data visualization (spectroscopic evaluation)*

While chronologically this step precedes MCR-ALS (see PROCEDURE), it is part of the evaluation tools. Total (non-normalized) intensity plots that are generated before and after preprocessing can be used to determine whether the chemical image overlaps perfectly with the white light image or whether there are other distortions (out of focus effects, fluorescence problems, etc.) that could affect component mapping and segmentation. It is an important tool for ensuring that data quality after pre-processing is good (i.e. artifacts are removed), and that signal intensities match anatomical features (see Figure 6).

### *Model fit (mathematical validation)*

The lack of fit at the end of the MCR-ALS process is the first parameter to indicate whether the model describes the data well. A satisfactory result should be reached with the selected number of components, i.e. the uncertainty of the model (lack of fit) should be in agreement with the experimental noise in the data. There is no default optimum value, since it strongly depends on the quality of the original data. When the lack of fit is unexpectedly high, several options are available. The best general strategy is to test models with an increased number of components and see how this affects the model fit. If the fit improves, the original model did not include sufficient information to describe the system. If the model fit worsens or does not improve significantly after adding a new component, it is likely to be an unnecessary compound, unless the biological, chemical or spectroscopic evaluation finds it necessary and meaningful (see below).

Lengthening the iteration process either by increasing the number of iterations or by decreasing the convergence criterion is only helpful if very few iterations were initially used. In other GUIs that use the MCR-ALS algorithm with more constraint options35, a high lack of fit can also be associated with constraints that are improper or too strict. However, this is not the case in the simplified GUI presented in this work, since the non-negativity constraint should already be met after pre-processing.

### *Component profiles (biological, chemical and spectroscopic evaluation of concentration maps and spectral signatures)*

After performing MCR-ALS, the concentrations of the resolved pure components should be mapped. These distribution (concentration/component) maps should be evaluated to find matches to features observable in the visible image, global intensity plots or with any *a priori* biological knowledge (biological evaluation). In the best case, these component maps can already visualize zones of distinct chemical composition. In this respect, they provide an analytical endpoint, agreeing with or providing complementary information to segmentation maps.

However, random maps or maps that do not reflect the biological features of the sample can be artifacts or products of inadequate pre-processing or data quality/quantity (Figure 6). Together with the component distribution maps, the spectra of the corresponding resolved components should also be inspected (chemical and spectroscopic evaluation). Spectral shapes should be generally meaningful and never completely different from the features of the raw spectra. Ideally, the spectra should contain characteristic bands of certain compounds in image areas covered by sample, and only background or signal unrelated to the biological material in areas that are free from the sample (e.g. glass or embedding media in cell lumen or at tissue boundaries, Figure 2). While artifacts in theory could also be resolved as components, these should be eliminated by the pre-processing step (see the strong fluorescence signal in the beginning of the recording, which diminishes during the scan in Figure 6). Such artifacts are easily detected in both the component maps and the corresponding spectra, clearly indicating improper pre-processing.

### *Reference spectra matching (chemical evaluation)*

Optionally, the spectra of the resolved components can be compared to reference spectra for identification (chemical evaluation). Although the pure components identified by MCR-ALS are often pure compounds (or very good approximations of them), some can still have a mixed character. This can be due to the absence of more powerful (e.g. local rank41,42) constraints in the data analysis pipeline, or simply because certain conditions of compound overlap are not fulfilled by the sample. The reference spectra to be matched to the resolved pure components should be recorded separately by using well-defined chemicals (e.g. different wood biopolymers extracted by wet chemical methods, relevant model compounds or mixtures with known compositions), or extracted from welldefined regions of other images for investigating similarities. One of the most common methods for spectral matching is based on Euclidean distances (dot products), because it is computationally undemanding yet robust43-45. In a simplified view, comparison is based on determining the distance (difference) between each data point of the component and the reference spectra. The closer they are to each other, the higher the match is. Such point-by-point matching requires the reference spectra to have the exact same number of data points as the sample it is matched to. In spectroscopic terminology, it translates to covering the same spectral region with the same spectral resolution.

Additionally, spectra should be as similar as possible in terms of minimum and maximum intensity, so data points are not distant simply because of offset or intensity differences. For this reason, resolved pure component spectra and loaded reference spectra are automatically offset corrected and area normalized by the script presented in this protocol. To further facilitate matching, the loaded reference spectra can be pre-treated in the same way as the raw input data was, using the same parameters for baseline correction and smoothing. It has to be pointed out that reference matching is entirely optional and the results must always be carefully evaluated. On one hand, Euclidean distances can produce relatively high matches even in cases where spectra are obviously different (false positive match). Consider the following hypothetical scenario: both the component and the reference spectra have a single characteristic band each, albeit in different positions, otherwise only small uncharacteristic bands or large regions with only baseline. In this case, the two spectra will be rather closely matching at every data point except for the two small zones where their respective characteristic bands are located. Such false positive matches are easily identified by visual inspection (Supplementary Figure 3). On the other hand, significant portions of the spectra can be different, resulting in low Euclidean matches. However, this may simply be due to the fact that the pure components are not pure compounds but unresolved mixtures. In this case, a low match indicates a low content of the reference compound in the mixture, rather than a

poor match (Supplementary Figure 3). While in a sense, it provides a false negative match, it is harder to detect than the false positive match in the hypothetical example above. Imperfect matches can also originate from reference spectra that do not mimic exactly the pure spectrum of a compound in the particular biological sample analyzed. For further notes on reference matching, consult the ADDITIONAL COMMENTS AND LIMITATIONS section.

### *Segmentation (biological and chemical evaluation)*

One of the challenges in hyperspectral imaging is the identification of equivalent pixels between independent images so that appropriate comparisons can be made in statistical tests. Component maps often show gradients of their respective components, and thus cannot always be used to find clear boundaries between chemically different areas. In addition, zones based on multiple component profiles cannot easily be identified between independent images. While manual categorization on a pixel-by-pixel basis is time-consuming and subjective, image segmentation can be performed automatically and objectively by k-means clustering of the MCR-ALS concentration profiles34. The key parameter for k-means clustering is the number of clusters, which can be determined manually or iteratively using silhouette values. The silhouette value of each data point (pixel spectrum) in the image describes how similar that pixel is to other pixels in its own cluster compared to pixels belonging to other clusters. The similarity is determined based on the average difference from pixels in the same cluster versus the minimum average difference to pixels in a different cluster28. There are numerous measures of such differences28, including Euclidean distances. Nevertheless, the number of clusters determined by silhouette values can always be manually overruled in both directions, i.e. fewer or more clusters can be selected for k-means clustering based on *a priori* knowledge of the sample or biologically justified expectations. In general, it is recommended to test different numbers of clusters and compare the results. The resulting segmentation maps, for instance, should be compared to the white light image to evaluate whether they match anatomical features or sample areas of interest to the biologist (biological evaluation). In addition, the contribution of each pure component to a particular cluster (image segment) should be investigated by means of the corresponding centroid plot (chemical evaluation). This strategy is useful to clearly identify chemically different regions with distinct boundaries in a spatial and compositional manner and can be more powerful than heat maps of single band intensities or pure component maps. Spectra from equivalent regions of multiple images can thus be compared directly, or by using multivariate methods, such as orthogonal projections to latent structures – discriminant analysis (OPLS-DA7,29).

## ADDITIONAL COMMENTS AND LIMITATIONS

Our GUI is based on modules that were primarily designed to handle a broad set of analytical challenges (multiple methods from chromatography to spectroscopy, multiple series of spectra, multiple images, etc.) (http://www.science.uva.nl/~hboelens/publications/draftpub/Eilers_2005.pdf and reference 35). To enhance ease-of-use and to streamline the analysis for a specific task only (namely vibrational microspectroscopic image analysis of biological materials), certain parameters have been fixed, preventing end-user modifications and excluding certain analytical options altogether. Here, we outline the key limitations of this protocol. Consultation with spectroscopy and chemometrics experts using Table 1 may be necessary to determine if these limitations make the protocol unsuitable for a specific task. In particular, non-negativity constraints in the present script are fixed for both spectra and concentrations in MCR-ALS. The selection of the number of components is based on singular value decomposition or manual input only, and starting conditions are locked to pure spectral estimates using a method based

on SIMPLISMA38. The spectra in each pixel must have equal spectral range and resolution. Moreover, the presented script does not allow simultaneous multitechnique or multi-image processing. However, it does allow for spectral preprocessing, including asymmetric least squares (AsLS) baseline correction, optional smoothing and area normalization. AsLS baseline correction is limited to spectra with non-negative peaks and thus cannot handle derivatives.Smoothing is only available by the Savitzky-Golay method.

In addition to the practical constraints of MCR-ALS listed in Table 1, some aspects should be taken into account. MCR-ALS is often applied in a dynamic way to test different sets of starting parameters until satisfactory final results are obtained. It is important to note that MCR-ALS is not a nested method, i.e. increasing the model size does not imply adding new components to the existing ones. Instead, it means that the full system is re-defined with a new set of components. Therefore, it is recommended to test different number of components (Figure 7) on the same image to ensure optimum un-mixing. Another way to influence the optimization is by using local rank constraints35,41,42. While these constraints are highly valuable in cases of extreme compound overlap (to better define the presence and absence of compounds in pixels), they are very difficult to automate. In the case of biological materials, where the different elements are reasonably well compartmented (tissues, cells, subcellular zones), MCR-ALS using only non-negativity constraints often provide very satisfactory results without the need of local-rank constrains. Since MCR-ALS is not used with all its capabilities in the present script, unmixing may not be perfect in some cases. Thus, the resulting pure components may be good approximations of pure compounds, but still mixtures. Reference matching is helpful in this respect, since it can show the dominating compounds in the pure component spectra, but it has its own limitations. Most importantly, the separately recorded reference compounds may not be chemically identical to the native, *in situ* compounds, integrated into biological systems and linked to other biomolecules (differences in e.g. cellulose crystallinity, protein structure/folding, polymer orientation, etc.). This is often reflected in their spectra and results in less than perfect matches. It is important to note that reference spectra must be recorded with similar settings to the samples, having identical spectral region and resolution. If needed, reference spectra should be pre-processed in the same way as the image (i.e. baseline correction and smoothing).

When these limitations render the present protocol unsuitable for a particular dataset, we refer to a more complete MCR-ALS GUI with additional MATLAB scripts35, focusing on the resolution data analysis step. It offers more versatility and can be used with a wider range of constraints to a broad variety of data analysis problems (including simultaneous multi-technique and multi-image processing), but it also requires a better understanding of chemometrics. Finally, as a practical limitation, the input data for our GUI must be either an ASCII file (e.g. common tab or space delimited .txt file) or a Mathworks MATLAB .mat file. For requirements regarding files, folders and formatting, please consult the PROCEDURE section.

## EXPERIMENTAL DETAILS

### Biological material

The materials we chose to illustrate the different steps of our method are cryosections from xylem (wood) of deep-frozen hybrid aspen (*Populus tremula x Populus tremuloides*) stems. Raman image data was recorded by available standard procedures14. Small maps (< 200 pixels) were used to demonstrate the suitability of our method for screening large sample numbers where recording times of individual images need to be kept short (and therefore pixel numbers low). We illustrate this by showing how our pipeline can be used to distinguish the chemically distinct cell wall layers in such images. However, our protocol is versatile and well suited to different kinds of vibrational (FTIR or Raman) microspectroscopic images of diverse sample types and image sizes (pixel numbers). To demonstrate this versatility, we

provide a further example using an FTIR microspectroscopic image of a biomedical mammalian tissue (the pancreas of 6-weeks-old C57BL/6 mouse) (Figure 2).

### Input data

The following input data files were used in the protocol (and are available as supplementary material, including the corresponding white light images in .jpg format):

BadExample_Aspen_Raman.txt is a Raman microspectroscopic image with significant fluorescence problems and limited number of pixels. It was recorded over normal wood in hybrid poplar (see Biological material above), using a Renishaw inVia microscope, 100x magnification, 514 nm Ar+ laser excitation, 1 µm laser spot and step size, 1 second exposure time, in static mode.

GoodExample_Aspen_Raman.txt was recorded with the exact same settings as BadExample_Aspen_Raman.txt, over the tension wood area of hybrid poplar. The corresponding white light images of the Aspen examples contain overlaid single band intensity heat map, marking the area scanned by Raman microspectroscopy. These heat maps were created using the built in functions of Renishaw's WiRE software, calculating the band total band area in the 1560-1650 cm-1 region, corresponding to aromatic -C=C- vibrations (lignin). The reference compound spectra (cellulose.txt, lignin.txt and Dglucuronicacid.txt) were also recorded with the exact same settings as BadExample_Aspen_Raman.txt, using the same chemicals as described in reference 7.

064x064_MousePancreas_FTIR.mat is an FTIR microspectroscopic image, recorded over the wax embedded pancreas section of a 6-weeks-old C57BL/6 mouse, using a Bruker Tensor 27 FTIR spectrometer with an attached Hyperion 3000 microscopy accessory and a 64 x 64 liquid nitrogen cooled focal plane array (FPA) detector. 32 interferograms were co-added for improved signal-tonoise ratio. Spectral resolution of 4 cm-1 was used and a zero filling factor of 2 was applied.

## MATERIALS

### Equipment Hardware

- Standard equipped PC or Mac with minimum system requirements to run the software (see below) and enough free disk space for saving the results. For the TIMING sections in the protocol, the following architectures were used:
1) Apple MacBook Pro 15'' Retina, 2.4 GHz Intel Core i7 CPU, 8 GB 1600 MHz DDR3 RAM, Intel HD Graphics 4000 1024 MB, 256 GB SDD, OS X 10.9.1
2) Fujitsu Siemens Celsius desktop, 2.67 GHz Intel Xeon CPU, 8 GB RAM, NVidia GeForce 6600 256 MB, 1TB HDD, Windows 7 Professional (64-bit)
3) Apple MacBook Air 13'', 2.13 GHz Intel Core 2 Duo, 2 GB 1067 MHz DDR3, NVIDIA GeForce Graphics 9400M 256 MB, OS X 10.9.1
4) Apple iMac 22", 2.7 GHz Core i5 20GB 1333 MHz DDR3, AMD Radeon HD 6770M 512 MB, OS X 10.9

### Software

- MathWorks MATLAB version R2012a or newer recommended. The GUI has been successfully tested on MATLAB version R2009b with no problems. However, older versions are not supported.

- MCR_ALS_PlantImaging_v1.m and .fig files (available for download, free of charge at www.kbc.umu.se/vibrationaldownload.html)

For the TIMING sections in the protocol, the following MATLAB versions have been used:
1) Version 8.0.0.783 (R2012b), Mac (64bit)
2) Version 8.2.0.701 (2013b), Win (64 bit)
3) Version 8.2.0.701 (2013b), Mac (64 bit)

## PROCEDURE

### Input data

CAUTION: In order for the GUI to work, the input files must be either tab or space delimited ASCII files (e.g. .txt files), or MATLAB .mat files. Other file types or files with a different data structure than outlined below need to be reformatted to match the input file requirements of the GUI.

ASCII files must have the following format
1 1 1800 4528
1 1 1799 4321
1 1 1798 4413
. . …. ….
1 2 1800 4619
1 2 1799 4721
1 2 1798 4812
. . …. ….
1 3 1800 4311
1 3 1799 4289
1 3 1798 4156
. . …. ….
. . …. ….
2 1 1800 4555
2 1 1799 4369
2 1 1798 4611
. . …. ….

In this case the first column contains the Y coordinate of the pixel, the second column contains the X coordinate of the pixel, the third column contains the wavenumbers and the fourth column contains the spectral intensities. This is the default format when Raman image maps are recorded by Renishaw's WiRE software and exported as .txt files.

MATLAB .mat files must have the following structure:
4000 0.0473 0.0498 0.0953 ... 0.0622
3998 0.0423 0.0123 0.1022 ... 0.0817
3996 0.0494 0.0331 0.1169 ... 0.0724
…. ……. ……. ……. … ……
400 0.0512 0.0678 0.0744 ... 0.0688

In this case, the first column contains the wavenumbers, thereafter each column contains the intensities of one pixel. The first pixel is the one located in the first row of the first column of the image; the second pixel is the first row of the second column, etc. This is the default format with Bruker's OPUS 7 export into MATLAB.

CAUTION: The .mat file does not contain information regarding X and Y coordinates, only the total number of pixels (i.e. the .mat file of an image containing 4x20 pixels has the same dimensions as the .mat file of an image containing 8x10 pixels). Thus, the number of X and Y pixels needs to be supplied separately when .mat files are used as input. Alternatively, the filename can contain this information in the following format:
"AAAxBBB_examplefilename.mat". In this case, AAA and BBB denote the number of pixels in the X and Y dimensions, respectively. For instance 064x032_examplefilename.mat indicates an image with 64x32 pixels.

CAUTION: Since .mat input files are processed faster, they are recommended for larger data sets (> 200 pixels).

CAUTION: The optional reference spectra input files must be tab or space separated ASCII files (e.g. .txt files), with the first column containing wavenumbers and the second column containing the corresponding intensities. No other formats are accepted.

CAUTION: Folder and filenames must contain standard alphanumeric characters only (i.e. unaccented Latin letters, numbers and underscore). No special characters are allowed. Do not use capitals, only small letters for file extensions (i.e. .jpg and not .JPG, .txt and not .TXT, .mat and not .MAT, etc), as MATLAB is case sensitive.
For better organization, keep files in their dedicated folder without un-necessary nested subfolders to facilitate navigation.
The image files used as examples in the present protocol are provided as supplementary material, together with three reference spectra files. These can be used for testing the protocol and as formatting guides for own data.

**Starting steps** TIMING: ca. **30** seconds - **5** minutes
**1)** Start MATLAB.
**2)** Navigate to the folder containing the "MCR_ALS_PlantImaging_v1.m" script, using the "Current Folder" panel in MATLAB
**3)** Double-click on the MCR_ALS_PlantImaging_v1.m file in the "Current Folder" panel to load the script to the main MATLAB interface.
**4)** Run the script. This step can be performed using option A or B.
A) In the MATLAB interface, under the EDITOR tab, click on the "Run" button
B) Right-click on the file MCR_ALS_PlantImaging_v1.m and select "run" in the opening context sensitive menu

CAUTION: If the MCR_ALS_PlantImaging_v1.m file is not in the default MATLAB folder but it is automatically loaded at startup (i.e. it was left in the Editor window last time MATLAB was shut down), MATLAB will ask whether to change to that folder or add that folder to the path. In this case, select changing to that folder.

TROUBLESHOOTING
**5)** In the "Open File" dialog box, navigate to the hyperspectral image data file to be processed, select it and click "Open".

OPTIONAL: If a .mat file was selected with no pixel dimension defined in the filename, an additional dialog box will open, where the correct X and Y pixel numbers need to be supplied manually.

CAUTION: Data needs to comply with the formatting rules outlined in the Input data section of this PROCEDURE.

TROUBLESHOOTING

**6)** The data is loaded and the relevant parts of the GUI are automatically populated.

a. The name of the loaded data file is displayed in red in the bottom right part of the interface below the “Segmentation Map (Clusters)” and “Centroid Profiles” plots.

b. The top left plot in the interface labeled “Original Spectra” (in the “Pre-processing” frame) shows the loaded spectra in a rainbow set of colors and without modification.

CAUTION: To speed up the displaying step on slower computers, this plot contains maximum 100 spectra. If the original dataset contains more spectra, 100 of them are randomly selected for display.

c. The “Selected Spectrum” plot displays in red the currently selected spectrum in the dataset (by default the first one or the first randomly selected one if there are more than 100 spectra in total), the AsLS calculated baseline in blue (using the default settings, step 10) and the resulting corrected spectrum in black. The dropdown menu above the plot shows the number of the currently selected spectrum.

TROUBLESHOOTING

d. The “Corrected Spectra” plot in the bottom left of the interface shows the result of the pre-processing with the set parameters. It displays the same set of spectra as shown in the “Original Spectra” plot, using the exact same color for each spectrum.

e. The “Total Intensity Map” in the top right of the interface, under the “Visualization” frame shows a total intensity map created by determining the area under all bands in the entire spectral region. The colors range from dark blue (lowest intensity) to dark red (highest intensity).

CAUTION: This plot uses the raw data input with no baseline correction at startup. The plot automatically updates once the preprocessed data is submitted for MCR-ALS analysis (step 17)

f. A white light image is automatically loaded as long as it is located in the same folder and has the exact same filename as the data file loaded, with .jpg as extension. It is shown in the “White Light Image” plot in the top far right of the interface, in the “Visualization” frame. Certain button texts in the GUI are color coded to differ from the default black. Red text means the button starts an important stage of the analysis, while blue text means the button saves the results of an important stage of the analysis.

CAUTION: At initial startup, displaying the “Corrected Spectra” plot may be slow, requiring up to several minutes to load large data files on slow computers.

TROUBLESHOOTING

**7)** Optional: Save any or all of the three plots “Original Spectra”, “Selected Spectrum” and “Corrected Spectra” by clicking on their respective “Save Plot” button, located on the top right of each plot. Suggestions for file type (.pdf by default), filename and folder are pre-filled in the opening save dialog boxes, but can be changed.

TROUBLESHOOTING

**8)** Optional: Save the total intensity plot by clicking on the “Save Intensity Map” button, located to the left of the plot. Suggestions for file type (.pdf by default), filename and folder are pre-filled in the opening dialog box, but can be changed.

TROUBLESHOOTING

**9)** Optional: Manually load a white light image by clicking on the "Load White Light Image" button in the "Visualization" frame. Use the opening dialog box to navigate to the correct folder and select the file containing the white light image.

TROUBLESHOOTING

**Pre-processing steps** TIMING: ca. **1**-**20** minutes

**10)** Perform baseline correction using Asymmetrical Least Squares (AsLS) fitting (http://www.science.uva.nl/~hboelens/publications/draftpub/Eilers_2005.pdf). Adjust the baseline using either the sliders or the direct input text boxes for lambda and p values in the "Baseline" frame of the interface.

CRITICAL STEP: The aim is to completely remove the baseline while retaining all spectral information (Figure 3a). Generally, the best strategy is to keep the p value to the minimum, since there should be no negative peaks, and only vary the lambda value. Testing lambda values is recommended in the range 102 to 109, varying one order of magnitude among successive testing steps to clearly see the effect on the fitted baseline shape (http://www.science.uva.nl/~hboelens/publications/draftpub/Eilers_2005.pdf).

CAUTION: Too high lambda values result in underfitting (i.e. not all of the baseline is removed, Figure 3c), while too low lambda values result in overfitting (i.e. spectral bands are removed as well, not only baseline, Figure 3b)

**11)** Confirm that the baseline correction performs equally well on different parts (pixels) of the image by selecting different spectra from the drop-down menu above the "Selected Spectrum" plot.

TROUBLESHOOTING

**12)** Optional: Tick the checkbox "S-G filtering" in the "Smoothing" frame of the interface to perform smoothing of the spectra using Savitzky- Golay filtering27. The "Selected Spectrum" plot automatically updates and shows the smoothed spectra in green.

CAUTION: The polynomial order (in the box labeled "Order") must be lower than the frame size (in the box labeled "Frame"), which must be an odd number. The larger the difference between Order and Frame is, the more smoothing is applied. If Order=Frame-1, no smoothing is performed28.

CAUTION: Using too much smoothing results in the loss of spectral resolution, distortions and merging of bands (Figure 4).

TROUBLESHOOTING

**13)** Optional: Tick the checkbox labeled "Area Norm" to perform area (total intensity) normalization over the entire spectral region (see the BACKGROUND section and Supplementary Figure 2).

**14)** Click on the "Update Plot" button above the "Corrected Spectra" plot to check the combined outcome of all spectral pre-processing steps (steps 10-13) on the spectra. If suboptimal, return to Step 10.

CRITICAL STEP: Spectra need to be free from physical and optical artifacts as much as possible before proceeding to the following steps. Otherwise, these artifacts can be resolved as pure components by MCRALS and distort the results (Figure 6)

**15)** Optional: Save the pre-processed data in the same format as the original input file (ASCII or .mat) by clicking on the "Save Corrected" button in the bottom left of the interface.

CAUTION: Saving in ASCII .txt format can be slow for large data sets (large spectral range and many pixels). In those cases, .mat formats are recommended as input.
**16)** Optional: Pass the input data to multivariate curve resolution without performing any pre-processing (i.e. no baseline correction, smoothing or area normalization) by clicking on the "Untreated for MCR" button.

CAUTION: Only use this step if the data has already been pre-processed previously and is just re-loaded (Figure 1).
**17)** Click on the "Pre-Treated for MCR" button to conclude all preprocessing steps and send the data for multivariate curve resolution – alternating least squares (MCR-ALS) analysis. Clicking this button automatically updates the Total Intensity Map in the "Visualization" frame on the top right of the interface (step 6e) and populates the Singular Value Decomposition plot in top middle of the interface (step 18) in the "Multivariate Curve Resolution" frame.

CAUTION: The updated Total Intensity Map considers baseline changes and smoothing, but it does NOT consider area normalization, since that would result in the loss of features due to all intensities being equal after normalization.

**Data analysis steps** TIMING: ca. **5** - **10** minutes

**18)** Select the number of components. The choice needs to be based on the results of singular value decomposition (top plot in the "Multivariate Curve Resolution" frame in the central part of the interface, Figure 7), or on *a priori* knowledge and expectations. This step can be performed using two options
A) Select the number of components using the drop down-menu of Eigen values, listed in ascending order of component number (lower Eigen value for higher component number). The textbox below automatically updates to show the selected number of components in red.
B) Type the number of components into the text box with the same name and hit enter. The drop-down menu above automatically updates to show the corresponding singular value.
After selecting the number of components, the "Singular value decomposition" plot automatically zooms in to show the selected number of components +5. In addition, selecting the number of components automatically populates the "Pure Spectra Estimation (Initial Values)" plot with initial pure spectral estimates; calculated using a SIMPLISMA based method (see BACKGROUND section). The "Purest Pixels" list automatically updates to show the pixel number of the pure spectral estimates (step 19).

CRITICAL STEP: Selecting the correct number of components ultimately determines the outcome of the analysis. Different number of components should be tested and the results evaluated to see which gives the best result in terms of chemistry and biology (see the BACKGROUND section).

CAUTION: Adding a new component is not an additive process. In other words, it does NOT leave the original components intact, but recalculates ALL components (see the BACKGROUND section).

TROUBLESHOOTING
**19)** Inspect the "Pure Spectra Estimation (Initial Values)" plot to see whether spectral profiles are reasonable or not.

CRITICAL STEP: It is important to see whether the pure spectral estimates are meaningful or not, since this can help in selecting the correct number of components (Figure 7, and BACKGROUND section). If adding a new component (step 18) does not result in a significantly

different new spectral estimate, it is likely that the new component is not required and will not be well resolved. (BACKGROUND section)

CAUTION: The colors used for each component in the "Pure Spectra Estimation (Initial Values)" plot are kept constant throughout the interface. Thus, the legend of this plot can always be consulted to determine which color represents which component in subsequent plots. The only exception is the Reference Spectra Matching plot (step 37).

TROUBLESHOOTING

**20)** Optional: Change the noise allowed (in percentage) for the calculation of initial estimates in the textbox labeled "Noise" and hit enter. While the default 10% is generally safe, different values can be tested and their effect inspected in the "Pure Spectra Estimation (Initial Values)" plot.

TROUBLESHOOTING

**21)** Optional: Mark the location of the purest pixels on the Total Intensity Map by ticking the checkbox labeled "Mark Purest" in the Visualization frame of the interface, to the left of the Total Intensity Map.

CAUTION: Use the location of these pixels in the image together with the "Pure Spectra Estimates (Initial Values)" plot (step 19, Figure 7) to evaluate whether the correct number of components has been selected or not.

CAUTION: Once this step has been reached, the locations of the purest pixels are NOT updated automatically after returning to step 18 to select a different number of components. To force updating the locations of the purest pixels, untick and tick the "Mark Purest" checkbox. However, it only works if a higher number of components is selected then previously in step 18. In case a lower number of components is selected, the "Mark Purest" checkbox is unable to properly display the fewer components and the interface needs to be restarted by completing steps 44-47 and returning to step 4.

TROUBLESHOOTING

**22)** Click on the "Show Eigen Vectors" button to the right of the "Singular value decomposition" plot to show the Eigen vectors of the singular value decomposition in a separate window. Eigenvector profiles showing a noisy random pattern refer to non-relevant contributions. This option is especially valuable for people with more thorough background in chemometrics or when more insight is required, since distinction between noisy pattern and minor spectral features is not always obvious. As such, it can be safely sidestepped in general practice.

**23)** Optional: Before starting the MCR analysis, a quick look at the description of the data set by bilinear models based on components can be obtained. Clicking on the "Data Preview" button to perform a PCA analysis opens a separate window showing four different plots. The first and second plot refer to an approximate description of the data set with a bilinear model of meaningful contributions obtained with the "Initial Pure Spectra Estimates" (first plot), and the related concentration profiles that are estimated using the original data set and the spectral estimates by least squares calculation without any constraints (second plot). The third and fourth plots show the bilinear model obtained by PCA, using the same number of components as set for the MCR analysis. In addition, the main Command Window of MATLAB displays the PCA analysis parameters (number of components, lack of fit and CPU time). These plots provide a first insight into the behavior of the data set that will be improved with MCR analysis. Although informative, this step is not strictly necessary to perform the MCR analysis and can be optionally sidestepped.

**24)** Set the parameters for MCR-ALS. Since most constrains are already pre-set (see Table 1), the two main parameters to be adjusted at this stage are the number of iterations and the convergence limit. The convergence limit determines the sigma change between consecutive iterations (improvement of fit), below which the solution is considered to be optimal and is therefore not refined further. The number of iterations sets the maximum number of iterations performed, unless convergence is achieved before. In practice, it is difficult to know the optimum values for these parameters beforehand, and therefore the default values can be used at start.

CAUTION: High convergence limits or low number of iterations will result in quicker (and perhaps suboptimal) analysis, which can be preferred for a quick overview.

CAUTION: The live updating MCR Optimized Spectra and Concentration plots (step 25) are the most visual indicators to determine whether the convergence limit and the number of iterations are set properly or not. If either of these plots changes significantly even at the last iteration, lower the convergence limits and / or increase the number of iterations. Similarly, if the "MCR Results" displays anything else than "CONVERGENCE ACHIEVED" (step 25) adjusting the convergence limit and / or the number of iterations may be need. If a warning of divergence appears, the way in which the MCR analysis has been set (number of components, initial estimates, etc.) and the initial submitted data set (preprocessing options) should be reconsidered (return to step 18 or step 10, respectively).

**25)** Perform MCR-ALS analysis by clicking on the "Perform MCR" button. The analysis is performed with the pre-set constrains (see Table 1 and the BACKGROUND section) and parameters set in step 24. The MATLAB Command Window and the "MCR Results" frame in the center of the interface both display the results at every step of iteration. The first line of the display shows the current number of iterations ("Iterations:"). Below this is the status of the analysis, which can be "FIT IS IMPROVING" or "FIT IS NOT IMPROVING", resulting in "CONVERGENCE ACHIEVED" (when the change in sigma values is below the convergence limit set in step 24), "FIT NOT IMPROVED 20 TIMES. STOP" (when the fit has not improved for 20 consecutive iterations, indicating divergence), or "MAX NR OF ITERATIONS" (when the iteration count reaches the maximum allowed in step 24, without reaching convergence or stopping for divergence). The lines following the status display show the percentage change in sigma values, the fitting errors expressed as percentage lack of fit for PCA and the experimental variation, and finally the percentage of variation explained. In addition to these parameters, the MATLAB Command Window also lists the sum of squares in PCA reproduction, the sigma with respect to the experimental data, and finally gives a summary line at the end of the iteration. Of these parameters, the percentage sigma change can be useful for adjusting the convergence limit in step 24 for subsequent analysis, while the lack of fit and variance explained are indicative of how well the current model describes the data (see the BACKGROUND section).

CRITICAL STEP: A model with low fit and low explained variance cannot be considered a good representation of the data. If these values cannot be improved by changing key pre-processing parameters (return to step 10) or MCR-ALS parameters (such as the number of components (return to step 18), noise in initial estimates (return to step 20), or maximum number of iterations and convergence limits (return to step 24) (Figure 1)), consultation with chemometrics and spectroscopy experts is necessary.

CRITICAL STEP: During the MCR-ALS analysis, the plots "MCR Optimized Spectra" and "MCR Optimized Concentration" automatically update with each iteration. These plots (especially the spectral output) are important for evaluating the results and to see whether there is any

significant change during the iterations of MCR-ALS. If changes are still significant at the last iteration, the convergence limit and / or the maximum number of iterations need to be adjusted (return to step 24).

**26)** Optional: Save the "MCR Optimized Spectra" and "MCR Optimized Concentration" plots using their respective "Save Plot" button, located above each plot. Suggestions for file type (.pdf by default), filename and folder are pre-filled in the opening save dialog boxes, but can be changed.

**27)** Optional: Save the "MCR Optimized Spectra" and "MCR Optimized Concentrations" matrices using their respective "Save Matrix" button, located above each plot. Suggestions for filename and folder are pre-filled in the opening save dialog boxes, but can be changed. The resulting .mat files contain the variables "SOpt" and "COpt", respectively. The variable SOpt stores the optimized spectral profiles of the pure resolved components in N x W format, where N is the number of components and W is the number of wavenumbers (i.e. each row represents one spectrum, see **ST** in Figure 5). The variable "COpt" stores the optimized concentration profiles of the pure resolved components in M x N format, where M is the number of total pixels and N is the number of components (i.e. each column contains the concentration of a single pure component in all pixels, see **C** in Figure 5).

**Visualization steps (evaluation)** TIMING: ca. **10** seconds – **1** minute

**28)** Optional: If no white light image is loaded by default, or if a new one needs to be loaded instead of the default one, click on the "Load White Light Image" button in the "Visualization" frame in the upper right of the main interface. A standard load dialog box opens, allowing for navigation and single image selection, listing only .jpg files by default.

CAUTION: multiple images cannot be selected and loaded

TROUBLESHOOTING

**29)** Optional: Tick / untick the "Mark Purest" checkbox to show / hide the purest pixels in both the Total Intensity Map and in the Component Maps (step 30). This can be useful to see which regions in the image are associated with each pure component.

CAUTION: The marks only show locations, but they do not display which pure component is associated with a certain pixel. They do not show pixel numbers either.

TROUBLESHOOTING

**30)** Click on the "Show Component Maps" button in the "Visualization" frame in the upper right section of the interface. This opens a new window, which contains two sets of plots. On the left, distribution maps (concentration profiles) are shown for each pure component as intensity heat maps. Colors range from dark blue (lowest intensity) to dark red (highest intensity). On the right, the spectral profiles for the corresponding pure component are shown, using the same colors as in the "Pure Spectra Estimation (Initial Values)" plot (step 19).

CRITICAL STEP: Component maps should match anatomical features of interest and should be in agreement with features observable in the visible image, the Total Intensity Plot or with *a priori* biological knowledge. Random maps or maps that do not reflect the biological features of the sample are likely to be artifacts or products of a bad preprocessing. In such cases, a new model should be tested by changing key MCR parameters, such as the number of components (return to step 18), noise in initial estimates (return to step 20), or maximum number of iterations and convergence limits (return to step 24) (Figure 1)). Alternatively, pre-processing parameters can also be adjusted (return to step 10). If none of the above results in an

improved match of the Component Maps to biological features, consultation with chemometrics and spectroscopy experts is necessary.

CRITICAL STEP: Spectra of the resolved pure components should be generally meaningful and never completely different from the features of the raw spectra. If only artifacts are resolved in ALL components, consultation with chemometrics and spectroscopy experts is necessary.

CAUTION: If the "Mark Purest" checkbox is checked (step 29) before the "Show Component Maps" button is pressed, the purest pixel for each component is marked in their respective Component Map plot.

TROUBLESHOOTING

**31)** Optional: Save the "Total Intensity Map" plot by clicking on the "Save Intensity Map" button of the main interface. A save dialog box opens, with pre-filled values for filename, format and location, which can be changed.

**32)** Save the most important MCR-ALS results by clicking on the "Save MCR Results" button. A save dialog box opens, with pre-filled values for filename, format and location. While filename and location can be changed, the format needs to remain .mat. The saved .mat file contains four variables. The variable "COpt" stores the optimized concentration profiles of the pure resolved components in M x N format, where M is the total number of pixels and N is the number of components (i.e. each column contains the concentration of a single pure component in all pixels, see **C** in Figure 5). The variable "SOpt" stores the optimized spectral profiles of the pure resolved components in N x W format, where N is the number of components and W is the number of wavenumbers (i.e. each row represents one spectrum, see **ST** in Figure 5). The variable "R2Opt" contains the variance explained at the last iteration, expressed in values of percentage divided by 100 (i.e. R2Opt = 0.98 means 98 percent variation explained). The variable "SDOpt" contains two numbers: the percentage lack of fit in terms of PCA and in terms of the experimental data)

**Optional: Reference spectra matching steps (evaluation)** TIMING: ca. **30** seconds – **5** minutes

**33)** Load reference spectra by clicking on the "Load Reference Spectra" button in the "Match Components to Reference Spectra" frame in the middle right of the interface. A load dialog box opens, where multiple files can be selected.

CAUTION: Spectra must be in tab or space separated ASCII format (e.g. standard .txt files), with the first column containing wavenumbers and the second column containing the corresponding intensities. No other formats are accepted.

CAUTION: Each file must contain only one spectrum

CAUTION: All reference spectra must be in the same folder and must have unique alphanumeric filenames (i.e. only unaccented Latin letters and numbers and underscore are allowed, no special characters). It is also important to have filenames that are representative for the reference compound, since these names will be displayed in legends and tables.

**34)** Optional: Tick the "Pre-treat References" checkbox to pre-process the loaded reference spectra, using the same parameters as for image preprocessing (lambda and p values for the baseline correction, step 10, and polynomial order and frame size for the Savitzky-Golay smoothing, step 12).

CAUTION: Reference spectra and the pure resolved component spectra are ALWAYS area normalized before reference matching, irrespective of the "Pre-treat References" checkbox.

**35)** Optional: Show the loaded reference spectra in a separate window by clicking on the "Show Reference Spectra" button. This can be used to evaluate the quality of the reference spectra and to determine whether pre-processing is necessary or not.

CAUTION: The reference spectra shown in this plot are always area normalized.

TROUBLESHOOTING

**36)** Perform reference matching based on Euclidean distances by clicking on the "Perform Reference Matching (Dot Product)" button. The results are automatically displayed in the table below this button, showing the percentage match of each reference spectrum to each pure component.

CRITICAL STEP: The resulting percentage matches should always be considered as indicative only and should always be evaluated by visual inspection of the matches (step 37 below). See the Reference matching section of the Introduction regarding false positive and negative matches.

CAUTION: The percentage matches are all individuals and do not add up to 100% total.

CAUTION: Depending on the number of components and reference spectra, scrolling may be needed to display all values.

**37)** Always inspect the matching results of step 36 by clicking on the "Show Matches" button. A new window opens, containing one plot for each pure resolved component.

CRITICAL STEP: Critically examine matching results to exclude false positive / negative matches (see the Reference matching section in the Introduction). In case of uncertainties, consult a spectroscopy expert to avoid over-interpretation of the results.

CAUTION: This is the only plot where the pure component colors are not maintained. Instead, each component has its own separate plot in which it is displayed in solid thick black lines, while all reference spectra are displayed in dashed thin colored lines, in accordance with the figure legend.

**38)** Save the reference matching results by clicking the "Save Match Results" button. A save dialog box opens, with pre-filled values for filename, format and location. While filename and location can be changed, the format needs to remain .mat. The saved .mat file contains the same table as displayed in the main interface, including row and column headings (i.e. component and reference names, respectively)

**Optional (recommended): Image segmentation steps (evaluation)** TIMING: ca. **10** seconds – **5** minutes

**39)** Determine the number of clusters for segmentation maps. This step can be performed using option A or B
A) Click on the "Silhouette Clusters" button. This performs silhouette clustering to determine the number of clusters for segmentation maps. A message box appears to prompt the user to wait while silhouette clustering is in progress. When the process finishes, the message box

automatically closes and the “Number of clusters” textbox updates. While silhouette clustering is entirely optional, it is recommended to get an initial overview of the data.
B) Enter the number of clusters directly in the “Number of clusters” textbox. This kind of manual input is needed when segmentation maps with different numbers of clusters need to be tested (step 40), or when the biological question at hand demands a certain number of clusters. In these cases, silhouette clustering (option A) is not used or its results need to be overruled.
**40)** Click on the “K-Means Clustering” button to perform k-means clustering using the number of clusters determined in step 39. This automatically updates the “Segmentation Map (Clusters)” and “Centroid Profiles” plots. A legend is automatically created for the “Segmentation Map (Clusters)” plot to show the colors representing each cluster. The colors used in the “Centroid Profiles” plot refer to the resolved pure components and are the same as throughout the interface. Therefore, the “Centroid Profiles” plot has no separate legend. Instead, the legend of the “Pure Spectra Estimation (Initial Values)” plot (step 19) should be consulted.

CRITICAL STEP: If segmentation maps are expected to provide detailed information regarding chemically distinct zones in the sample, different numbers of clusters may need to be tested. In that case, return to step 39, option B.

CAUTION: The cluster’s number to which a particular pixel in the image belongs is randomly determined by k-means clustering. Therefore, rerunning k-means clustering can result in different coloring. However, the clustering results (boundaries of clusters) do not change, unless a different number of clusters is selected. In other words, the same pixels will belong to the same cluster, only the coloring of the “Segmentation Map (Clusters)” plot is altered, together with the “Centroid Profiles” plot to reflect the change in cluster order.

**41)** Optional: Save the “Segmentation Map (Clusters)” plot and legend by clicking on the “Save Segmentation Plot” button. A save dialog box opens, with pre-filled values for filename, format and location, which can be changed.
**42)** Optional: Save the “Centroid Profiles” plot by clicking on the “Save Centroid Plot” button. A save dialog box opens, with pre-filled values for filename, format and location, which can be changed.
**43)** Optional: Save the results of the k-means clustering by clicking on the “Save Segment Results” button. A save dialog box opens, with prefilled values for filename, format and location. While filename and location can be changed, the format needs to remain .mat. The saved .mat file contains the variables “IDX" and “Centr”. The variable “IDX” stores the cluster number each pixel of the image belongs to. The variable “Centr” is a K x N matrix (where K is the number of clusters and N is the number of components), describing the contribution of each component to each cluster.

### Finishing steps

**44)** Close the interface window to finish the analysis

CAUTION: Unsaved plots of the main interface window cannot be recovered after this step!

CAUTION: The most important variables remain in the MATLAB basic workspace and can still be saved, until step 46.

**45)** Close all additional open figure windows (Eigen Vectors, Pre-MCR PCA results, Component Maps, Reference Spectra, Reference Matches) individually.

CAUTION: Unsaved plots cannot be recovered after closing their respective figure windows!

**46)** Type “clear all” at the MATLAB Command Window prompt to clear the MATLAB workspace and memory from all variables.

CAUTION: No data can be recovered after this step!

**47)** Type “clc” at the MATLAB Command Window prompt to clear the MATLAB Command Window.
**48)** If new image data needs to be processed, return to step 4.
**49)** After processing the final data, close MATLAB. The next analysis will have to start from step 1.

## TIMING

Starting steps: ca. 30 seconds - 5 minutes
Pre-processing steps: ca. 1 - 20 minutes
Data analysis steps: ca. 2 - 30 minutes
Visualization steps: ca. 10 seconds - 1 minute
Reference spectra matching steps: ca. 30 seconds - 5 minutes
Image segmentation steps: ca. 10 seconds - 5 minutes
Finishing steps: ca. 10 - 20 seconds

## ANTICIPATED RESULTS

Using vibrational microscpectroscopic (hyperspectral) images of a biological sample the described procedure is able to identify a) the number of distinct chemical components that can be differentiated in the sample based on their spectral profiles; b) the pure spectral profiles of each component; c) the relative concentration of each component in each pixel of the image (plotted as a distribution map); d) the number and distribution of zones with distinct chemical characteristics (segmentation). Additionally, the spectral profiles of the resolved unique chemical components can be matched to the spectral profiles of reference compounds identification. The primary application areas benefiting from these results are the biological and medical sciences, where changes in chemical composition need to be interpreted in a spatial context. This includes mapping the effects of disease or pathogens, genetic modifications, environmental factors or inherent biochemical differences between cell or tissue types. Below, we illustrate the power of the method by finding chemically distinct zones within cell wall layers of woody plant tissues (tension wood of hybrid aspen, Figure 6, right panel) and within a larger mammalian tissue sample (mouse pancreas, Figure 2) in the absence of clear visible boundaries.

### Resolving cell wall layers

Plant cell walls are heterogeneous mixtures of mainly four major types of biopolymers: cellulose, hemicelluloses, lignins and pectins. The structure, composition and relative proportion of these biopolymers can vary not only between tissues and cell types, but also along a developmental gradient and even within a single cell wall. Cell walls of woody tissues in plants are characterized by their layered structure, with different layers having distinct proportions of the above mentioned polymers and distinct ultrastructural features46. While different cell wall layers may not be visible in white light images, they can in theory be differentiated based on their spectral profiles, i.e. chemical composition. This is of great importance when the effects of e.g. genetic modification or environmental signals on cell wall

layer development and composition need to be investigated. In such studies, it is imperative to compare the same cell wall layers in wild type and transgenic or untreated and treated plants, without the influence from the neighboring cell wall layers or cells. As visible features often do not exist for differentiating the cell wall layers, and simple distance measurements from easily recognized features (such as the lumen) cannot be used due to cell wall thickness variations, spectral information need to be used. Heat maps (based either on single band intensity or on entire spectral profiles) are of limited use in this respect, since they are often unspecific or only provide concentration gradients, but no distinct boundaries among zones (Figure 6d, right panel). Using our protocol, however, cell wall layers can be clearly defined in segmentation schemes. The example shown in Figure 6, right column is the hyperspectral image of the cross-section of a hybrid aspen stem, containing tension wood fibres. We chose this example for demonstration due to clear and characteristic chemical composition differences of tension wood fibers, which allow easy validation of the results. In addition to the cell wall layers of normal wood fibers46 tension wood fibers have a thick cell wall layer on the lumen side of the cell wall. This additional layer (G-layer) is extremely rich in cellulose but mostly free from lignin that is present in the underlying layers (reference 47, and references therein). Since cellulose and lignin have clearly distinguishable Raman spectroscopic fingerprints, and since the G-layer is rather thick, it is expected to be easily distinguishable from the underlying cell wall layers. Accordingly, the pure components clearly resolve both lignin and cellulose spectral profiles and the corresponding distribution maps highlight areas that are rich in each (see Figure 6d and e, right column: C1 and C2 spectral profiles and maps for lignin and cellulose, respectively). In particular, the distribution map for component 2 (C2) can be used to select clear G-layer pixels in the image. However, to separate the underlying, thinner and chemically more similar cell wall layers, further analysis (image segmentation) is required. Four clusters were selected for k-means clustering based on Silhouette values and *a priori* knowledge of the sample, as we expected to resolve the lumen and three cell wall layers: 1) the cellulose-rich G-layer on the lumen side, 2) the lignin-rich middle lamella separating adjacent cells, and 3) the underlying S-layer, which is a mixture of different thin and chemically similar layers that cannot be distinguished at the spatial resolution of the experiment (see below). The segmentation map (Figure 6f, right column, segmentation map) clearly defines these zones. Cluster 2 (light blue) is the G-layer, and it has the highest contribution of component 2 (cellulose, green contribution in the Centroid Profile plot, Figure 6g, right column). Cluster 1 (dark blue) represents the S-layer with high contributions from both lignin (component 1, blue line) and cellulose (component 2, green line). Cluster 4 (brown) is almost exclusively lignin (component 1, blue line) and can thus be identified as the middle lamella. Anatomically, the dark blue zone on the right side of the segmentation map, wedged between two G-layers (light blue), should be expected to contain pixels describing the middle lamella, and thus should have a brown line in the center (indicating the presence of cluster 4). There are two main reasons for its absence. Firstly, the lateral resolution of the image is too low to resolve zones thinner than 1 µm. Secondly, even if the middle lamella is almost 1 µm thick, it may not coincide with the pixel boundaries, i.e. parts of this cell wall layer can belong to different pixels. This decreases its contribution to the spectrum of those pixels, which have significant signals from neighboring zones. In short, no clear pixel for the middle lamella can be found in this part of the image, which is also slightly out of focus, further decreasing detection limits of smaller contributions. Nevertheless, the middle lamella could be detected in the cell walls that appear to left in the image, since it was better resolved in that location: it was thicker, due to the cell corner covered in this part of the image, coincided with pixel boundaries more and was more in focus. Finally, cluster 3 (yellow) represents the lumen, having the highest contributions from components 3 and 4, dominated by noise (red and cyan, respectively in Figure 6e, right column, C3 and C4). Based on these results, representative pixels for not only the G-layer, but also for other cell wall layers can be identified, their representative spectra extracted and compared to spectra of the corresponding cell wall layers of other genotypes or trees growing

in the absence of tension wood formation. This example clearly illustrates the power of MCR-ALS in component identification, both for spectral profiles and component distribution maps. It also highlights where component maps may not be sufficient for the identification of chemically distinct zones with clear boundaries. In those cases, segmentation maps help to achieve clear results. It is important to note that the clear interpretation of the clusters is only possible because the image segmentation is based on the resolved concentration profiles by MCR-ALS. This does not only speed up the segmentation process compared to classical approaches using the full pixel spectra, but also allows the straightforward interpretation of centroid profiles as mixtures of pure components in different proportions.

### Resolving chemically distinct zones in a mammalian tissue

Hyperspectral imaging for the medical sciences is challenging, since clear identification of chemical differences is required in order to assess pathological conditions with certainty. For instance, the monitoring of disease progression requires a set of quantitative markers that are clearly associated with (the different stages of) the disease and can be accurately followed. In the case of type I diabetes, for instance, the loss of pancreatic beta cells has been followed by 3D optical tomography for this purpose48-50. The beta cells are responsible for insulin production and are organized into the islets of Lagerhans, scattered throughout the exocrine tissue. Visually differentiating islet cells in the surrounding exocrine tissue without staining is difficult, and thus this serves as an ideal example to demonstrate the applicability of the present procedure to biomedical challenges.

Following the steps of the present protocol, the FTIR microspectroscopic image of a mouse pancreas section was resolved into 5 pure components (Figure 2). None of the pure components are pure chemical compounds, but they have contributions from various proteins (bands around 1550 and 1650 cm-1), as well as lipids and carbohydrates (bands between 1000-1200 cm-1), and the embedding medium (mostly in C2 and 5, green and violet respectively, and to a lesser extent in C1 (blue) due to smearing during sectioning, Figure 2c, spectral profiles). While it can be important to know the exact chemical composition of different zones within the tissue (or even within cells, as in the cell wall layer example above), here we exclusively focus on using our procedure to detect a known feature in the image (the islet cells embedded in the exocrine tissue), based on its unique spectral fingerprint. Thus, interpreting the spectra in terms of exact chemical compounds is beyond the scope of the protocol, and the emphasis is on the different spatial maps. The pure component maps (Figure 2c, concentration profiles) are unable to detect islet cells directly, but they already highlight problematic areas (see the component maps of C2 and 5 in particular, but also of C1 to a lesser extent, since according to their spectral profiles, the embedding medium is clearly contributing to the signal in these zones. Figure 2c). This is also clear in the segmentation map (cluster 3, brown, Figure 2d). However, the segmentation map also resolves chemically different zones within the pancreatic tissue, with cluster 1 (dark blue) associated to the islet cells, and cluster 2 (light green) to the exocrine tissue. Note also that these zones have the least contribution from embedding medium (contributions of C2 and C5, green and violet, respectively, in the Centroid Plot, Figure 2e) and thus provide the most reliable spectral profiles of the tissue. Extracting spectra from these chemically different zones at different stages of the disease development provides valuable information regarding the biochemical changes associated with the pathological conditions. In general, such information can be used for assessing risk factors, monitoring disease progression or even for developing targeted therapies for selected cell types or tissue zones.

## ACKNOWLEDGEMENTS

The authors thank the Vibrational Spectroscopy Core Facility of the Chemical Biological Centre at Umeå University for full Raman and FTIR microspectroscopy instrumentation access and resources dedicated to method development. We thank Prof. Björn Sundberg at the Umeå Plant Science Centre for initializing the project using hybrid aspen and continued support. Prof. Ulf Ahlgren and Christoffer Nord at the Umeå Centre for Molecular Medicine are acknowledged for providing the mouse pancreas sample. Anna de Juan and Romà Tauler acknowledge financial support of the European Union project CHEMAGEB and Joaquim Jaumot from the Spanish government (grant CTQ2012-11572).

## AUTHOR CONTRIBUTIONS

The project was initiated by J.F. and A.G.; plant samples were prepared by J. F., Raman hyperspectral images were recorded by J. F. and A. G., FTIR hyperspectral image were recorded by A. G.; Raman reference compounds were prepared and recorded by J. F. and A. G.; MCR-ALS was developed by R. T., J.J. and A. J.; the present MATLAB GUI was designed, written and tested by A. G., J. F., H. H. and J. J.; the manuscript was written by J. F., H. H., R. T., A. J. and A. G.; the project was coordinated and supervised by A. G.

## COMPETING FINANCIAL INTEREST

The authors declare no competing financial interests.

## TABLES

Table 1 Constraints for single image and single method datasets. For detailed descriptions of the constraints, consult the Introduction, reference 35, and references therein

## References

1 Geladi, P., Grahn, H. & Burger, J. Multivariate Images, Hyperspectral Imaging: Background and Equipment. in *Techniques and Applications of Hyperspectral Image Analysis* (eds HF. Grahn & PLM. Geladi) (John Wiley & Sons Ltd, 2007).

2 Hall, H., Cheung, J. & Ellis, B. Immunoprofiling reveals unique cell-specific patterns of wall epitopes in the expanding Arabidopsis stem. *Plant J.* **74**, 134-147, (2013).

3 Wilson, S. & Bacic, A. Preparation of plant cells for transmission electron microscopy to optimize immunogold labeling of carbohydrate and protein epitopes. *Nat. Protoc.* **7**, 1716-1727, (2012).

4 Fabian, H. *et al.* Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy. *Biochim. Biophys. Acta* **1758**, 874-882, (2006).

5 Nijssen, A. *et al.* Discriminating basal cell carcinoma from its surrounding tissue by Raman spectroscopy. *J. Invest. Dermatol.* **119**, 64-69, (2002).

6 Sobottka, S., Geiger, K., Salzer, R., Schackert, G. & Krafft, C. Suitability of infrared spectroscopic imaging as an intraoperative tool in cerebral glioma surgery. *Anal. Bioanal. Chem.* **393**, 187-195, (2009).

7 Gorzsás, A., Stenlund, H., Persson, P., Trygg, J. & Sundberg, B. Cell-specific chemotyping and multivariate imaging by combined FT-IR microspectroscopy and orthogonal projections to latent structures (OPLS) analysis reveals the chemical landscape of secondary xylem. *Plant J.* **66**, 903-914, (2011).

8 Gierlinger, N. & Schwanninger, M. Chemical imaging of poplar wood cell walls by confocal Raman microscopy. *Plant Physiol.* **140**, 1246-1254, (2006).

9 Chang, S.-S., Salmén, L., Olsson, A.-M. & Clair, B. Deposition and organisation of cell wall polymers during maturation of poplar tension wood by FTIR microspectroscopy. *Planta*, (2013).
10 Pesquet, E. *et al.* Non-cell-autonomous postmortem lignification of tracheary elements in Zinnia elegans. *The Plant Cell* **25**, 1314-1328, (2013).
11 Tsai, A. *et al.* Constitutive expression of a fungal glucuronoyl esterase in Arabidopsis reveals altered cell wall composition and structure. *Plant Biotechnol. J.* **10**, 1077-1087, (2012).
12 Horvath, L. *et al.* Distribution of wood polymers within the cell wall of transgenic aspen imaged by Raman microscopy. *Holzforschung* **66**, 717-725, (2012).
13 Schmidt, M. *et al.* Label-free in situ imaging of lignification in the cell wall of low lignin transgenic Populus trichocarpa. *Planta* **230**, 589-597, (2009).
14 Gierlinger, N., Keplinger, T. & Harrington, M. Imaging of plant cell walls by confocal Raman microscopy. *Nat. Protoc.* **7**, 1694-1708, (2012).
15 Richter, S., Müssig, J. & Gierlinger, N. Functional plant cell wall design revealed by the Raman imaging approach. *Planta* **233**, 763-772, (2011).
16 Gierlinger, N. *et al.* Cellulose microfibril orientation of Picea abies and its variability at the micron-level determined by Raman imaging. *J. Exp. Bot.* **61**, 587-595, (2010).
17 Gierlinger, N., Schwanninger, M., Reinecke, A. & Burgert, I. Molecular changes during tensile deformation of single wood fibers followed by Raman microscopy. *Biomacromolecules* **7**, 2077-2081, (2006).
18 Naumann, A., Navarro-Gonzalez, M., Peddireddi, S., Kues, U. & Polle, A. Fourier transform infrared microscopy and imaging: Detection of fungi in wood. *Fungal Genet. Biol.* **42**, 829-835, (2005).
19 Wilson, R. H. *et al.* The mechanical properties and molecular dynamics of plant cell wall polysaccharides studied by Fourier-transform infrared spectroscopy. *Plant Physiol.* **124**, 397-405, (2000).
20 Faix, O. Classification of Lignins from Different Botanical Origins by Ft-Ir Spectroscopy. *Holzforschung* **45**, 21-27, (1991).
21 Kataoka, Y. & Kondo, T. Quantitative analysis for the cellulose I alpha crystalline phase in developing wood cell walls. *Int. J. Biol. Macromol.* **24**, 37-41, (1999).
22 Akerholm, M., Hinterstoisser, B. & Salmen, L. Characterization of the crystalline structure of cellulose using static and dynamic FT-IR spectroscopy. *Carbohydr. Res.* **339**, 569-578, (2004).
23 Wetzel, D. FT-IR Microspectroscopy Imaging of Plant Materal. in *Infrared and Raman Spectroscopic Imaging* (eds R. Salzer & HW. Siesler) (Wiley- VCH Verlag GmbH & Co. KGaA, 2009).
24 Gorzsás, A. & Sundberg, B. Chemical fingerprinting of Arabidopsis using Fourier transform infrared (FT-IR) spectroscopic approaches. *Methods in molecular biology (Clifton, N.J.)* **1062**, 317-352, (2014).
25 Jirasek, A., Schulze, G., Yu, M. M. L., Blades, M. W. & Turner, R. F. B. Accuracy and precision of manual baseline determination. *Appl. Spectrosc.* **58**, 1488-1499, (2004).
26 Eilers, P. H. C. Parametric time warping. *Anal. Chem.* **76**, 404-411, (2004). 27 Savitzky, A. & Golay, M. J. E. Smoothing + Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **36**, 1627-&, (1964).
28 MatLab computer program (MathWorks, 2013R).
29 Stenlund, H., Gorzsás, A., Persson, P., Sundberg, B. & Trygg, J. Orthogonal projections to latent structures discriminant analysis modeling on in situ FT-IR spectral imaging of liver tissue for identifying sources of variability. *Anal. Chem.* **80**, 6898-6906, (2008).
30 Baranska, M., Schulz, H., Rosch, P., Strehle, M. A. & Popp, J. Identification of secondary metabolites in medicinal and spice plants by NIR-FT-Raman microspectroscopic mapping. *Analyst* **129**, 926-930, (2004).

31 de Juan, A., Maeder, M., Hancewicz, T., Duponchel, L. & Tauler, R. Chemometric Tools for Image Analysis. in *Infrared and Raman Spectroscopic Imaging* (eds R. Salzer & HW. Siesler) Ch. 2, 65-106 (Wiley-VCH Verlag GmbH & Co. KGaA, 2009).
32 Bonnier, F. & Byrne, H. J. Understanding the molecular information contained in principal component analysis of vibrational spectra of biological systems. *Analyst* **137**, 322-332, (2012).
33 Tran, T. N., Wehrens, R. & Buydens, L. M. C. Clustering multispectral images: a tutorial. *Chemometrics Intellig. Lab. Syst.* **77**, 3-17, (2005). 52
34 Piqueras, S., Duponchel, L., Tauler, R. & de Juan, A. Resolution and segmentation of hyperspectral biomedical images by multivariate curve resolution-alternating least squares. *Anal. Chim. Acta* **705**, 182-192, (2011).
35 Jaumot, J., Gargallo, R., de Juan, A. & Tauler, R. A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. *Chemometrics Intellig. Lab. Syst.* **76**, 101-110, (2005).
36 de Juan, A., Rutan, S. C. & Tauler, R. Two-way data analysis: Multivariate Curve Resolution: Iterative resolution methods. in *Comprehensive Chemometrics* (eds S. Brown, R. Tauler, & R. Walczak) 325-344 (Elsevier B. V., 2009).
37 Tauler, R., Smilde, A. & Kowalski, B. Selectivity, Local Rank, 3-Way Data- Analysis and Ambiguity in Multivariate Curve Resolution. *J. Chemom.* **9**, 31-58, (1995).
38 Windig, W. & Guilment, J. Interactive Self-Modeling Mixture Analysis. *Anal. Chem.* **63**, 1425-1432, (1991).
39 Windig, W. Spectral data files for self-modeling curve resolution with examples using the Simplisma approach. *Chemometrics Intellig. Lab. Syst.* **36**, 3-16, (1997).
40 Batonneau, Y., Laureyns, J., Merlin, J. C. & Bremard, C. Self-modeling mixture analysis of Raman microspectrometric investigations of dust emitted by lead and zinc smelters. *Anal. Chim. Acta* **446**, 23-37, (2001).
41 de Juan, A., Maeder, M., Hancewicz, T. & Tauler, R. Local rank analysis for exploratory spectroscopic image analysis. Fixed Size Image Window- Evolving Factor Analysis. *Chemometrics Intellig. Lab. Syst.* **77**, 64-74, (2005).
42 de Juan, A., Maeder, M., Hancewicz, T. & Tauler, R. Use of local rank-based spatial information for resolution of spectroscopic images. *J. Chemom.* **22**, 291-298, (2008).
43 Li, J. F., Hibbert, D. B., Fuller, S., Cattle, J. & Way, C. P. Comparison of spectra using a Bayesian approach. An argument using oil spills as an example. *Anal. Chem.* **77**, 639-644, (2005).
44 Mark, H. & Workman, J. *Chemometrics in Spectroscopy*. (Elsevier, 2007).
45 Linusson, A., Wold, S. & Norden, B. Fuzzy clustering of 627 alcohols, guided by a strategy for cluster analysis of chemical compounds for combinatorial chemistry. *Chemometrics Intellig. Lab. Syst.* **44**, 213-227, (1998).
46 Plomion, C., Leprovost, G. & Stokes, A. Wood formation in trees. *Plant Physiol.* **127**, 1513-1523, (2001).
47 Felten, J. & Sundberg, B. Biology, Chemistry and Structure of Tension Wood. in *Cellular Aspects of Wood Formation Plant Cell Monographs* (ed J. Fromm) 203-224 (Springer Verlag, 2013).
48 Alanentalo, T. *et al.* Tomographic molecular imaging and 3D quantification within adult mouse organs. *Nat. Methods* **4**, 31-33, (2007).
49 Alanentalo, T. *et al.* Quantification and Three-Dimensional Imaging of the Insulitis-Induced Destruction of beta-Cells in Murine Type 1 Diabetes. *Diabetes* **59**, 1756-1764, (2010).
50 Hornblad, A., Cheddad, A. & Ahlgren, U. An improved protocol for optical projection tomography imaging reveals lobular heterogeneities in pancreatic islet and beta-cell mass distribution. *Islets* **3**, 204-208, (2011).

## Figure Captions

Figure1
Overview flowchart listing the steps of the data analysis protocol The data analysis can be broken down into five major parts (gray shaded areas, which also reflects the layout of the GUI (Supplementary Figure 1): Preprocessing, Multivaraite curve resolution, Visualization, Reference matching and Image segmentation, each having a number of steps and options. The major stages are labeled with gray ellipsoids and numbers that correspond to the steps in the PROCEDURE section. Data analysis starts with Sample Input and follows the arrows, although some steps are optional. For details, consult the running text. Paralelograms indicate data matrices, rectangles indicate processing steps, diamonds indicate key conditional choices, and bubbles denote plots.

Figure 2
Demonstrating the use of the PROCEDURE on an FTIR hyperspectral image of a biomedical mammalian sample (mouse pancreas) For sample description and recording parameters, see the MATERIALS section. a) White light image. The tissue boundary in the bottom left corner is clearly visible. b) Total intensity map, following pre-processing (AsLS baseline correction, lambda = 100, p = 0.001; Savitzky-Golay smoothing, order = 1, frame = 3). The tissue boundaries are more visible (low intensity regions shown in blue) and indicate a crack in the tissue section, as result of drying. When comparing this heat map to the white light image, a region in the sample crack can be noticed, where the embedding material is clearly visible. c) Pure component (C1-5) distribution maps (concentration profiles, left) and the corresponding spectral profiles (right), following MCR-ALS (5 components, iteration limit: 50, convergence limit: 1). The white diamonds mark the location of the purest pixel for each component. Based on the spectral profiles and comparing the distribution maps of each component, it is clear that C2 and 5 (and to a lesser extent C1) contain contributions from the embedding medium.
d and e) Segmentation map and the corresponding centroid profile, respectively, following k-means clustering (3 clusters), based on the concentration profiles in c). Three clusters were selected, since the image is expected to contain empty areas (covered by the embedding medium), cells of the exocrine tissue, cells of the islets of Lagerhans. The analysis is clearly able to differentiate distinct zones in the image. Cluster 1 (dark blue) corresponds to the islets of Lagerhans, having a large contribution from C3 (red). Cluster 2 (light green) corresponds to cells of the exocrine tissue, with an increased contribution from C1 (blue) and C4 (cyan). Finally, cluster 3 mostly corresponds to sample-free areas of the image, with high contribution from signals of the embedding medium.

Figure 3
The effect of different lambda values on the results of asymmetrical least squares (AsLS) baseline correction. The Raman spectrum of a poplar fiber cell wall is used as an example, showing the original spectrum in red, and the calculated baseline in blue, using a) the optimal lambda value, resulting in a spectrum that does not contain any broad features of leftover baseline, yet retains even small intensity bands (black). b) overly low lambda values, which generate baselines that follow the data too closely, resulting in over-fitting. As a result, spectral band intensities diminish, or disappear entirely (black) c) overly high lambda values, which generate a very linear baseline, resulting in under-fitting. As a result, the broad baseline features are not removed (black) The p value was kept at the value of 0.001 (default GUI minimum) for all plots.

Figure 4

The effect of Savitzky-Golay (S-G) smoothing. The Raman spectrum of a poplar fiber cell wall is used as an example. The original (raw) spectrum without S-G smoothing is shown in solid black lines. The solid black line shows the raw spectrum with no smoothing. Using a mild smoothing (dashed green line, first order polynomial, frame size = 3) reduces noise while at the same time retains band intensity, position and shape. The dashed red line shows the result of oversmoothing (first order polynomial, frame size 29). While noise is undoubtedly reduced, band intensity, position and shape are all compromised. The inset shows a magnified spectral region for clarity.

Figure 5
Schematic illustration of multivariate curve resolution – alternating least squares (MCR-ALS) analysis on a hypothetical example. The hyperspectral image is described as a 3-dimensional datacube**,** with x * y = M spatial datapoints (number of image pixels) and W spectral datapoints (wavenumbers). The image is first unfolded (dashed arrow) to form the matrix **D**, which is the input for MCR-ALS. **D** is unmixed by MCR-ALS, using N number of pure components (N=4 is used as an example in the figure), resulting in two matrices: **C** and **ST**. **C** is an M * N matrix, every column of which contains the concentration profile of a pure component (the relative concentrations of the component in each pixel). **ST** is an N * W matrix, every row of which contains the spectra of a pure component (illustrated as blue, green, red and cyan lines). Since **C** is a reduced, noise-free representation of the original data (**D**), it can be used to construct segmentation maps and the corresponding centroid profiles. The segmentation maps use K number of clusters to describe K chemically unique zones in the sample (K = 4 is used as an example in the figure). Each cluster has different contributions from each pure component, as represented by the centroid profiles. The colors of the bars are the same blue, green, red and cyan as used for the pure components for easy identification. For example, cluster 1 (brown in the segmentation map) has the highest contribution from the pure component represented by the green spectrum. Similarly, cluster 3 (green in the segmentation map) has the highest contribution from the "blue" pure component, etc.

Figure 6
Demonstrating the importance of data amount and quality, using Raman microspectroscopic images of cross-sections of hybrid aspen wood fibers as examples. The left column contains a hyperspectral image with low data quality and amount ("bad example", too few pixels, significant fluorescence problems). The right column contains a hyperspectral image with just acceptable data quality and amount ("good example", enough pixels, fluorescence problems can be removed by pre-processing). The corresponding data files (BadExample_Aspen_Raman.txt and GoodExample_Aspen_Raman.txt, respectively) are provided as Supplementary Material for download. Recording parameters are described in the Materials section. a) White light images, with overlaid band intensity heat maps based on the integral value of the 1560-1650 cm-1 band (aromatic -C=C- vibration) to highlight the area used for Raman microspectroscopy. Note that the single band intensity map in the right column is non-informative, due to a uniformly low distribution of lignin in most of the image area (tension wood). In the left column, it indicates that the Raman image is offset to the left compared to the white light image. b) Total intensity maps before pre-processing. Note that the first row of the "bad example" image has much higher (gradually decreasing) intensity, due to the fluorescence problems caused by the presence of lignin. The corresponding map of the "good example" image is free from this distortion. c) Total intensity maps after pre-processing, using AsLS baseline correction (lambda = 30,000, p = 0.001) and Savitzky-Golay smoothing (order = 1, frame = 3). The quality is considerably improved, matching anatomical features in the white light image (a). d) Pure component distribution maps following MCR-ALS using five components (C1-5), an iteration limit of 50 and a convergence limit of 0.1 (default values). The maps of the "bad example" image only indicate the presence or absence of cell walls (c.f. C1

and C2 / C3). Moreover, the first and second row in these maps are still visibly different, indicating that the MCR-ALS results still contain contribution from fluorescence artifacts. On the other hand, the “good example” maps clearly indicate resolved cell wall layers (c.f. C1 and C2) e) The corresponding pure component spectral profiles. In the “bad example” most component spectra have high noise level (except C1, corresponding to the sample area in the image, see d) above). The lignin band (around 1600 cm-1) is detectable in most component spectra. In addition, the spectra of C2 and C3 have high intensity bands around 800 cm-1, which is unrelated to the sample and coincides with the lumen (see d) above). In the “good example”, the spectral profiles of C1 and C2 can be clearly associated to lignin and cellulose, respectively, while C5 is a mixture of both. The spectra of C3 and C4 have much worse signal to noise ratios and contain bands unrelated to the sample (around 800 cm-1) f and g) Segmentation maps and the corresponding centroid profiles, respectively, following k-means clustering, based on the concentration profiles in d). For the “bad example”, silhouette clustering returned 2 clusters, associated to sample and lumen (data not shown). In this case, the cluster number was manually set to 4, to be able to compare to the “good example”. It is clear from the segmentation map of the “bad example” that cluster 1 describes the sample, while the other 3 clusters only describe different parts of the lumen, with more or less contribution from fluorescence or stray signal close to the sample boundaries (cluster 2). It is also confirmed by the centroid profiles (g). In the case of the “good example”, the clusters can be associated to distinct cell wall layers. Cluster 1 is the secondary cell wall, with a high contribution from both cellulose and lignin (C2, green, and C1, blue, respectively). Cluster 2 is the Glayer, having the highest contribution of cellulose (C2, green). Cluster 4 is the middle lamella, having by far the largest contribution from lignin (C1, blue). Cluster 3 is the lumen, with high contributions from components unrelated to the sample (C3, red and C4, cyan).

Figure 7
Selecting the number of components for MCR-ALS, based on singular value decomposition (SVD) a) The “Singular Value Decomposition” plot of the main interface, showing the singular values in a typical Raman hyperspectral image of wood fibers in the cross-section of a hybrid aspen stem. The largest drop is seen between the first and second singular values (marked b). Singular values level off somewhat at the fourth component (mark c), with a minor drop between the fifth (mark d) and sixth (mark e) component. Thereafter, singular values change only marginally. Based on these values, two, four, five or six components can be selected for MCRALS (marks b, c, d and e, respectively). The corresponding “Pure Spectra Estimate (Initial Values)” plots are shown in figures b-e, and are used to help deciding the number of components for MCR-ALS b) Two components model: the spectrum for component 1 (blue) shows clear bands, while the spectrum for component 2 (green) only contains contribution from fluorescence baseline and noise. This indicates that using only two components would likely resolve only cell wall and lumen differences (i.e. showing the presence or absence of sample) c) Four components model: All four spectral profiles are unique, and all but one (for component 2, green) also contain clear bands. This indicates that all components are unique, and describe different chemical components in the cell wall (and the lumen, component 2, green). d) Five components model: Two of the spectra are practically identical (components 2 and 5, green and purple, respectively) and only represent fluorescence baseline and noise. This indicates that there is likely no need for the fifth component. e) Six components model: While the spectrum of component 6 (yellow) contains real bands, it is essentially identical to the spectrum of component 1 (blue) except for generally lower intensity. Thus it is not unique. As seen in the five components model already, the spectrum of component 5 is identical to that of component 2 and only describes fluorescence baseline and noise. This indicates that there is likely no need for the sixth component either. Based on the above evaluations, the best option is to use four components in the subsequent MCR-ALS analysis.

Supplementary Figure 1
Screenshot of the main window of the graphical user interface (GUI) The important steps are labeled with gray ellipsoids and numbers that correspond to the steps in the PROCEDURE section and match the ones of Figure 1

Supplementary Figure 2
Demonstrating the effects of area normalization, using a Raman hyperspectral image of wood fibers in the cross-section of a hybrid aspen stem as an example The panels show representative steps of the workflow, using non-normalized (left column) and area normalized (right column) data. Other pre-processing steps were applied uniformly (AsLS baseline correction, lambda = 30,000; p=0.001; Savitzky-Golay smoothing, order = 1, frame = 3) a and b) Pure component (C1-C5) distribution maps (concentration profiles, left), following MCR-ALS (5 components, iteration limit: 50, convergence limit: 0.1). The white diamonds mark the location of the purest pixel for each component. These locations do not change, but the distribution maps are markedly different in intensity. The non-normalized maps contain larger intensity fluctuations due to fluorescence (especially noticeable for C3 and C4) c and d) The corresponding pure spectral profiles (C1-5). Notice that the spectral profiles are identical, i.e. the normalization had no effects on the spectral profiles (no distortions). e and f) Segmentation maps, following k-means clustering with 4 clusters, based on the concentration profiles in a and b, respectively. Please note that k-means clustering arbitrarily assigns the cluster numbers, thus they do not match in e and f. This results in a coloring difference between the two plots. Nevertheless, the different zones are almost identical, except for a few borderline pixels (pixels 1, 12, 17 and 21). g and h) The corresponding centroid profiles, showing the contribution of different components to each cluster. The effect of normalization is very obvious, with more absolute amounts in the non-normalized centroid plots, and more relative (proportional) amounts in the normalized centroid plots. When signal intensity is comparable (cluster 2 for the non-normalized data and cluster 4 for the normalized data), the component profiles are very similar. In other cases, the normalization “stretches” the contributions and discrepancies occur (c.f. cluster 1 in both cases).

Supplementary Figure 3
Demonstrating the importance of visual inspection of reference spectra matching results. MCR-ALS resolved pure component spectra from Raman microspectroscopic images of cross-sections of hybrid aspen wood fibers are used as examples (see “good example” in Figure 6), with spectra of pure cellulose, lignin and D-glucoronic acid used as references (see the Materials section) a) The percentage matches based on Euclidean distances (dot products) b-f) Area normalized spectra plots showing each component (thick black line), cellulose (blue), lignin (red) and D-glucoronic acid (green). As can be seen, all matches to D-glucoronic acid can be discarded despite the high percentage hits, since none its characteristic bands match the component spectra (most notably the band at 1750 cm-1, resulting from -C=O stretching, is absent). Component 1 matching lignin and component 2 matching cellulose can be confirmed as legitimate matches. Component 3 and 4 can have contributions from cellulose too, while component 5 is clearly an unresolved mixture, containing both lignin and cellulose.

Figure 1

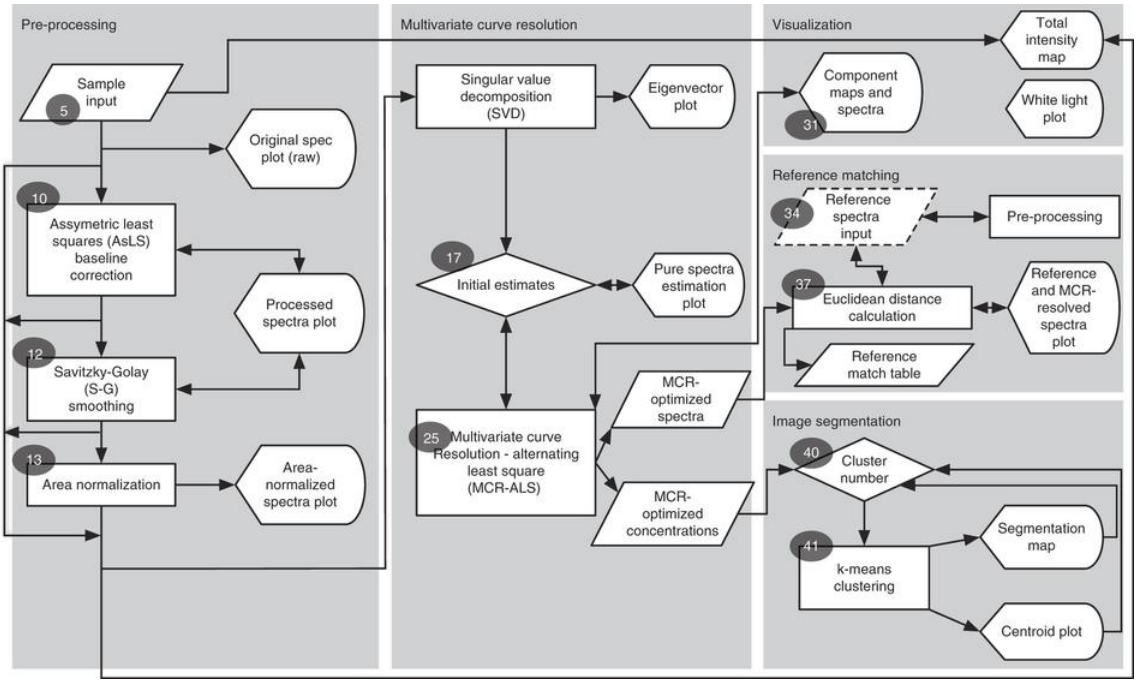
Pre-processing
Sample input
5
Original spec plot (raw)
10
Assymetric least squares (AsLS) baseline correction
Processed spectra plot
12
Savitzky-Golay (S-G) smoothing
13
Area normalization
Area-normalized spectra plot
Multivariate curve resolution
Singular value decomposition (SVD)
Eigenvector plot
17
Initial estimates
Pure spectra estimation plot
25
Multivariate curve Resolution - alternating least square (MCR-ALS)
MCR-optimized spectra
MCR-optimized concentrations
Visualization
Component maps and spectra
31
Total intensity map
White light plot
Reference matching
34
Reference spectra input
Pre-processing
37
Euclidean distance calculation
Reference and MCR-resolved spectra plot
Reference match table
Image segmentation
40
Cluster number
41
k-means clustering
Segmentation map
Centroid plot

Figure 2

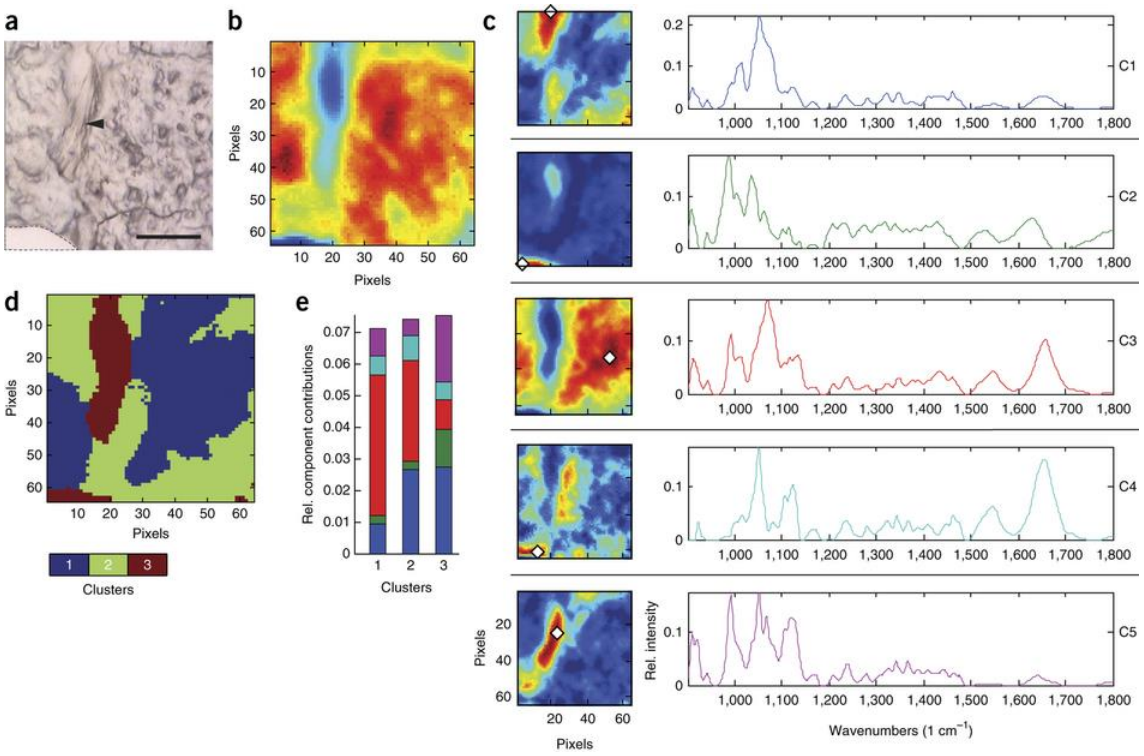
a
b
c
d
e
Pixels
Clusters
Rel. component contributions
Rel. intensity
Wavenumbers (1 cm$^{-1}$)
C1
C2
C3
C4
C5

Figure 3

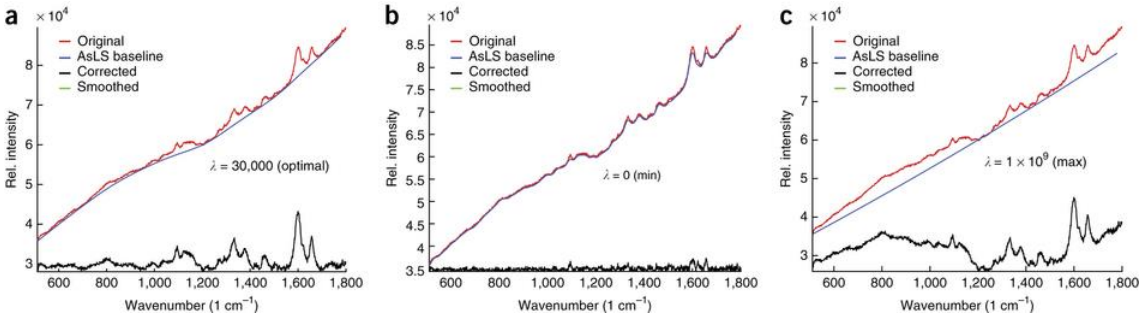
a
b
c
× 10^4
Original
AsLS baseline
Corrected
Smoothed
Rel. intensity
λ = 30,000 (optimal)
λ = 0 (min)
λ = 1 × 10^9 (max)
Wavenumber (1 cm^−1)
600
800
1,000
1,200
1,400
1,600
1,800

Figure 4

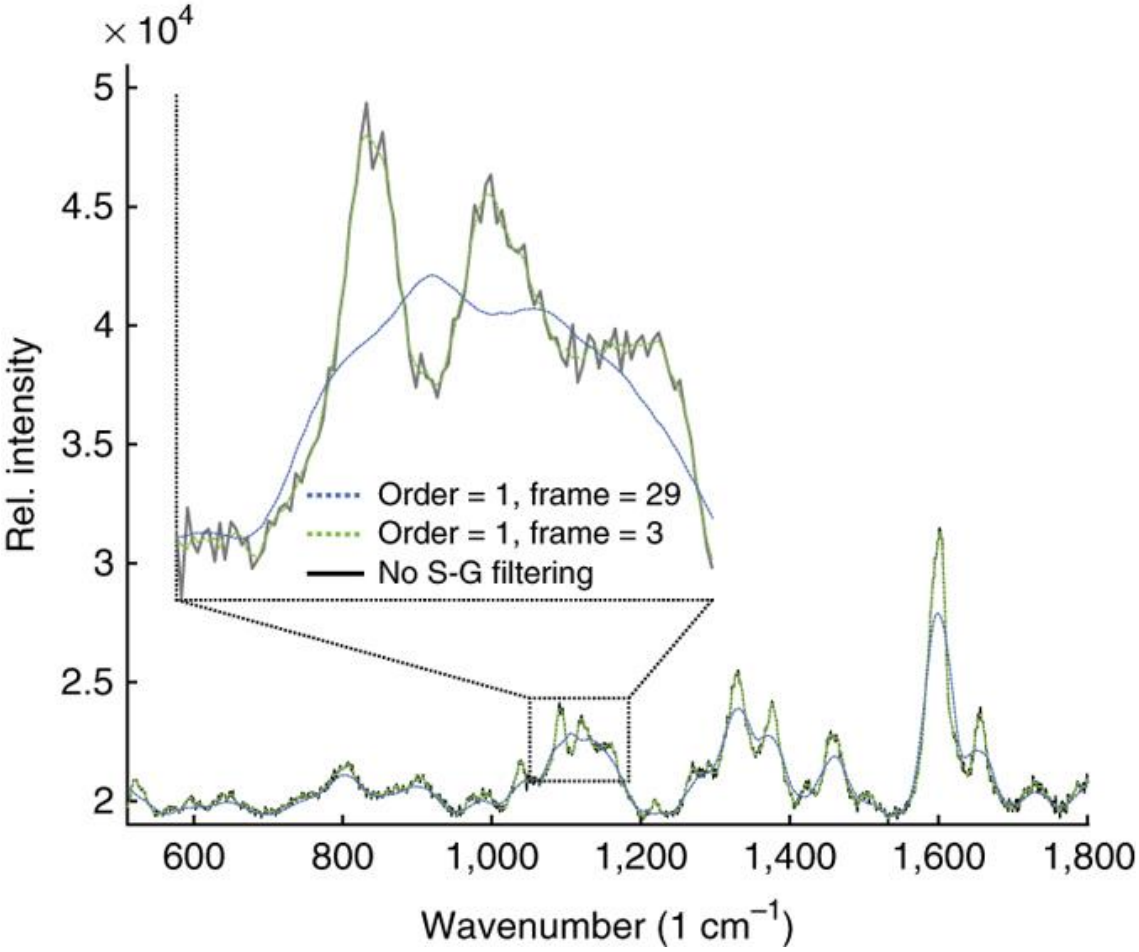
× 10^4
5
4.5
4
3.5
3
2.5
2
Rel. intensity
Order = 1, frame = 29
Order = 1, frame = 3
No S-G filtering
600
800
1,000
1,200
1,400
1,600
1,800
Wavenumber (1 cm^−1)

Figure 5

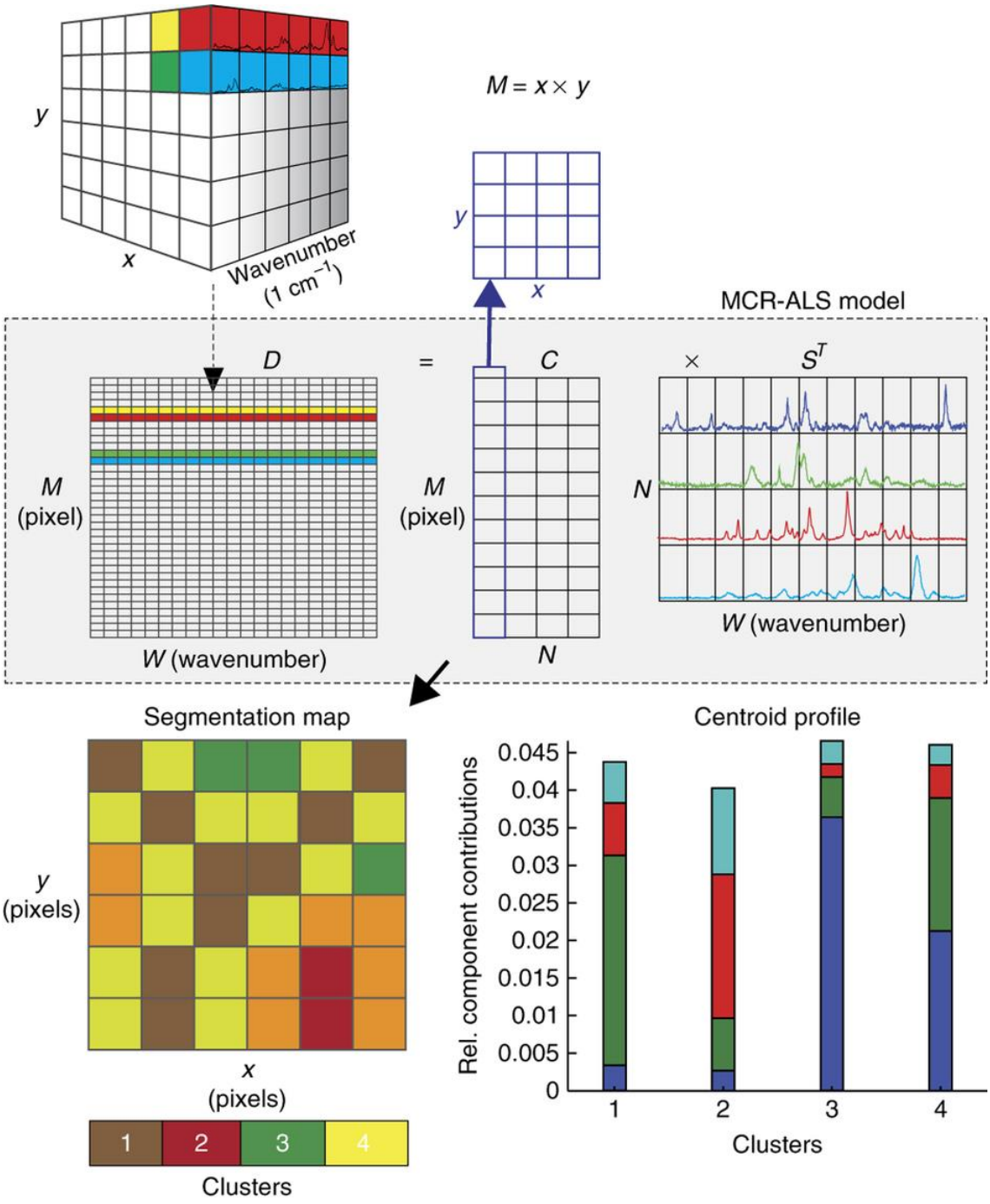
y
x
Wavenumber
(1 cm−1)
M = x × y
y
x
MCR-ALS model
D
=
C
×
ST
M
(pixel)
M
(pixel)
N
W (wavenumber)
N
W (wavenumber)
Segmentation map
y
(pixels)
x
(pixels)
1
2
3
4
Clusters
Centroid profile
Rel. component contributions
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
1
2
3
4
Clusters

Figure 6

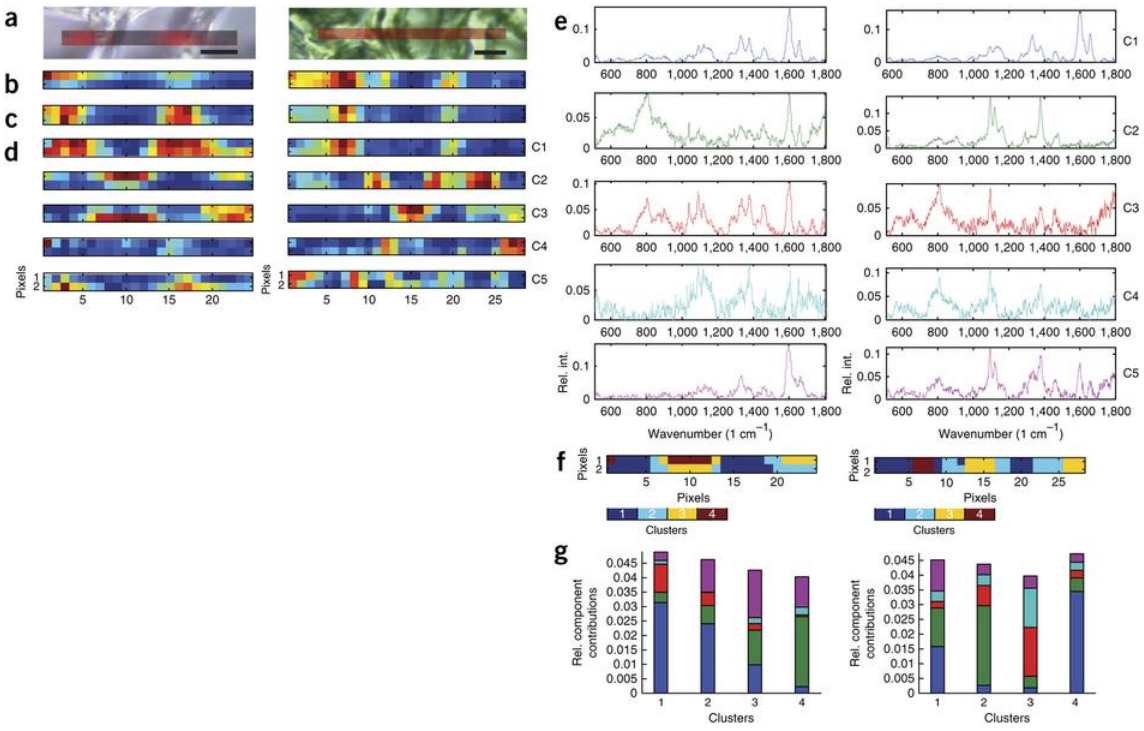
a
b
c
d
C1
C2
C3
C4
C5
Pixels
e
Rel. int.
Wavenumber (1 cm$^{-1}$)
f
Pixels
Clusters
g
Rel. component contributions
Clusters

Figure 7

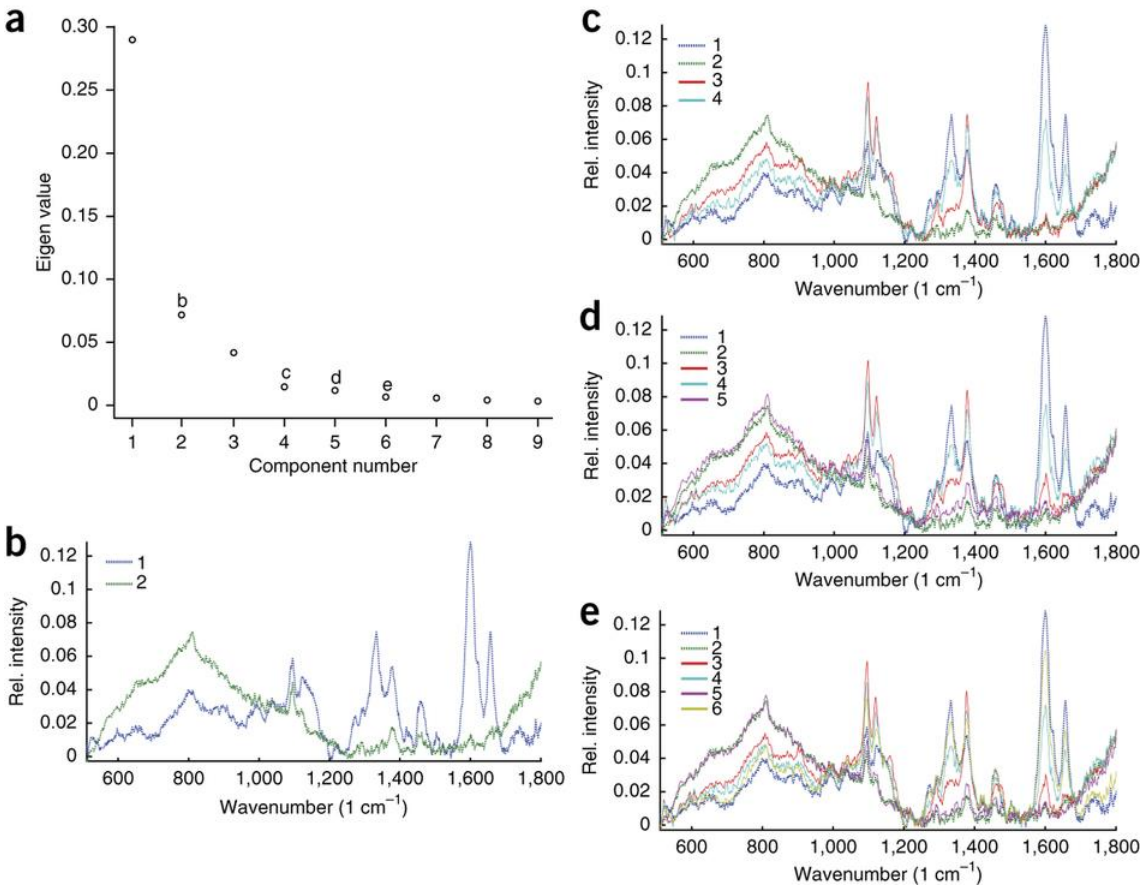
a
Eigen value
Component number
b
c
d
e
b
Rel. intensity
Wavenumber (1 cm−1)
c
d
e

Table 1

| Constraint | Possible choice in standalone scripts[40] | Possible choice in the present script |
|---|---|---|
| For detailed descriptions of the constraints, refer to the INTRODUCTION and Jaumot *et al*.[40] and references therein. | | |
| AsLS baseline correction | No baseline correction is possible | Value: *P* value adjustable between 0.001 and 1; λ value adjustable between 1 and 1,000,000,000 |
| Non-negativity for concentration profiles | Value: can be set to yes or no individually for each component<br>Method: can be set to non-negative least squares, fast non-negative least squares and forced to zero | Value: fixed to yes for all components<br>Method: fixed to using fast non-negative least squares |
| Non-negativity for spectral profiles | Value: can be set to yes or no individually for each component<br>Method: can be set to non-negative least squares, fast non-negative least squares and forced to zero | Value: fixed to yes for all components<br>Method: Fixed to using fast non-negative least squares |
| Unimodality for concentration profiles | Value: can be set to yes or no individually for each component<br>Method: vertical, horizontal or average. Tolerance level can be adjusted | Value: fixed to no for all components |
| Unimodality for spectral profiles | Value: can be set to yes or no individually for each component<br>Method: vertical, horizontal or average. Tolerance level can be adjusted | Value: fixed to no for all components |
| Closure for concentration profiles | Value: can be set to yes or no individually for each component<br>Method: fixed to a single value or allowing changing values. Equality or nonequality (equal or less than a preselected value) can be chosen | Value: fixed to no for all components |
| Closure for | Value: can be set to yes or no individually for | Value: fixed to spectra |

| Constraint | Possible choice in standalone scripts[40] | Possible choice in the present script |
|---|---|---|
| spectral profiles | each component<br>Method: fixed to a single value or allowing changing values. Equality or nonequality (equal or less than a preselected value) can be chosen | equal length |
| Equality for concentration profiles | Value: can be set to yes or no individually for each component<br>Method: selectivity or local rank information via an auxiliary matrix. Equality and nonequality (equal or less than the predefined value in the auxiliary matrix) can be chosen for the constrained concentrations | Value: fixed to no for all components |
| Equality for spectral profiles | Value: can be set to yes or no individually for each component<br>Method: known compound spectra via an auxiliary matrix. Equality and nonequality (equal or less than the predefined value in the auxiliary matrix) can be chosen for the constrained spectra | Value: fixed to no for all componen |

Table 2

| Step | Problem | Possible reason | Solution |
|---|---|---|---|
| 4 | No EDITOR tab in MATLAB | The 'Editor' panel is not active | Click on the filename in the 'Editor' panel, which will bring up the EDITOR tab |
| 5, 9, 29, 34 | No files are selectable | File extension does not match | Change the drop-down menu of the open dialog box to show all file types Note that MATLAB is case-sensitive (i.e., .jpg and .JPG are not equivalent) |
| | | Finder does not update (Mac users only) | Change the drop-down menu of the open dialog box to show all file types and select the file even if it is grayed out |
| | File does not load | Improper file formatting or name | Make sure that the file is in the correct format (see the Input data section of thisPROCEDURE), and that the folder names do not contain special characters |
| 6 | Script is extremely slow to start | Too large data files in .txt format | Save the data in .mat format |
| | | Insufficient computing power | Run MATLAB in native environments and not through hosted environments, e.g., parallels and so on<br>Shut down unnecessary processes |
| 7, 8, 27, 32, 42, 43 | Saved plots are of suboptimal quality | File format is suboptimal | Change the file type in the save dialog box |
| | Text is cropped in the GUI | Low-resolution computer screen | If possible, increase the screen resolution. Maximize the GUI window. Refer toSupplementary Figure 1 for the full version of texts in |

| Step | Problem | Possible reason | Solution |
|---|---|---|---|
| | | | the GUI |
| 10, 11, 19, Box 1, step 3 | Legend obscures the 'Selected Spectrum' plot | High-intensity data points in the low wavenumber region of the spectrum | Click and drag on the legend to move it to another location |
| Box 1, step 6 | No white light image is loaded at the start | No exact match of the data filename with lowercase .jpg extension for the figure was found | Manually load a white light image figure by the 'Load White Light Image' button in the 'Visualization' frame (Step 9 in the PROCEDURE) |
| 12 | The smoothed spectrum in the 'Selected Spectrum' plot does not update after changing S-G smoothing parameters | Variables are not updated in the MATLAB script | Press 'Enter'/'Return' after typing the value in the 'Order' and 'Frame' textboxes. Unmark and mark the 'S-G filtering' checkbox |
| 18 | Changing the number of components does not refresh the 'Pure Spectra Estimates (Initial Values)' | Variables are not cleared and updated in the MATLAB script, especially when changing to lower number of components | Restart the GUI and the entire analysis by completing Steps 44–47 and by returning to Step 4. CAUTION This will result in the loss of all unsaved data |
| 20 | The 'Pure Spectra Estimation (Initial Values)' plot does not refresh after changing the noise level | Variables are not updated in the MATLAB script | Force update by changing the number of components in Step 18 to a lower value and then back to the desired value |
| 21, 30 | Changing the number of components does not refresh the Pure Pixel markings in the 'Total | Variables are not cleared and updated in the MATLAB script, | Force updating the locations of the purest pixels, untick and tick the 'Mark Purest' checkbox. Please note that this only works if a higher |

| Step | Problem | Possible reason | Solution |
|---|---|---|---|
| | Intensity' plot and in the 'Component Maps' | especially when changing to lower number of components | number of components are selected than previously. In case a lower number of components are selected, the 'Mark Purest' checkbox is permanently unable to properly display the fewer components and the interface needs to be restarted by completing Steps 44–47 and returning to Step 4. Note that his will result in the complete loss of unsaved data |
| 26 | MCR-ALS text results are illegible or cropped in the GUI | Low-resolution computer screen | Increase the screen resolution<br>Maximize the GUI window<br>Read the MCR-ALS results in full text version in the MATLAB Command window |
| 30, 31 | The separate window containing the pure component maps does not show/hide the purest pixel after ticking/unticking the 'Mark Purest' checkbox of the main GUI window | Once the separate figure window is open, it does not refresh automatically | First close the separate window showing the component maps, tick or untick the 'Mark Purest' checkbox on the main interface window and then click on the 'Show Component Maps' button to reopen the updated component maps figure window |
| 36 | The separate window containing the reference spectra does not update when ticking or unticking the 'Pre-treat References' checkbox in the main interface window | Once the separate figure window is open, it does not refresh automatically | First close the separate window showing the reference spectra plot, tick or untick the 'Pre-treat References' checkbox on the main interface window and then click on the 'Show Reference Spectra' button to reopen the updated reference spectra plot window |