TUTORIAL

COCHINEAL CHARACTERIZATION IN HISTORICAL TEXTILES WITH MODELS FROM PARTIAL-LEAST SQUARES DISCRIMINANT ANALYSIS

By Ana Serrano and Andre van den Doel

In this tutorial, it is described step-by-step how to characterize chromatographic results of cochineal dyes found in fibre samples from historical textiles, through models developed with partial-least squares discriminant analysis (PLS-DA). This tutorial has been developed in the framework of the doctoral project "The Red Road of the Iberian Expansion: Cochineal and the Global Dye Trade".

Before reading this document any further, it is highly advisable that a careful reading of the following paper is undertaken: Serrano, A., Doel, A. van den, Bommel, M. van, Hallett, J., Joosten, I., Berg, K. J. van den, 2015, 'Investigation of crimson-dyed fibres for a new approach on the characterization of cochineal and kermes dyes in historical textiles', *Analytica Chimica Acta*, 897, pp. 116-127. This paper lays the basis of the current tutorial and any information related to the development of the PLS-DA models can be found there.

Please note that this tutorial is only applicable for the characterization of chromatographic results, which have been acquired by faithfully following the sample preparation and analytical conditions described in the above-mentioned paper (see also below). This is due to the fact that all reference samples (silk and wool samples dyed with cochineal dyes), which comprise the PLS-DA models, were prepared and analysed under those conditions. Therefore, any other samples that are submitted to different sample preparation and/or analytical conditions, should present non-equivalent chromatographic profiles, in relation to those in the PLS-DA models.

DISCLAIMER: The authors of this tutorial do not take any responsibility for the results obtained by users in further applications of the PLS-DA models, let alone if these are not applied according to the exact specifications of this tutorial. Moreover, these **are free of charge models, generously provided by the authors, for researchers to use at no additional expense**. Hence, the authors are not responsible for any scams or inappropriate usage given to the models. Contact from third-parties, for seeking out any type of profit-making deals, are completely discouraged. The authors are the sole responsible parties for these models, and any other claims on their appropriation are false.

If the users ought to apply the PLS-DA models and further publish their results, **the authors much appreciate if their work is acknowledged**. Please provide the following reference when necessary:

Serrano, A., Doel, A. van den, "Cochineal characterization in historical textiles with models from partial-least squares discriminant analysis models", launched online on 1 October 2016, at https://github.com/CochinealDyes/PLS-DA-models.

CHROMATOGRAMS ACQUISITION AND DATA PRE-TREATMENT

STEP 1: Ensure to follow the same sample preparation and analytical conditions.

Sample preparation (Fig. 1). Historical textile samples are accurately weighted (0.1-0.3 mg). A two-step extraction method is used (Fig. 1), using dimethyl sulfoxide (DMSO), prior to an acidic solution of hydrochloric acid (HCl) 37%: methanol (MeOH): H_2O (2:1:1, v/v/v).

In this way, information on the potential presence of different dyestuffs (in unknown historical textile samples) can be obtained, as DMSO is able to extract vat and direct dyes, whereas the acidic solution, mordant dyes:

1) 50 µL DMSO is added to the fibres and heated up to 80 °C in a water bath for 10 min, after which, the extract is transferred to another vial;

2) 50 µL of the acidic solution is added to the fibres and heated up to 100 °C in a water bath for 10 min.

After the extraction, the dye extracts are evaporated to dryness under gentle nitrogen flow, and the resulting dry residues are reconstituted with the DMSO extracts, thus combining the two steps. These are then centrifuged for 10 min at 7000 rpm, and part of the resulting supernatant is transferred to new vails. These vails can be centrifuged once more prior analyses.

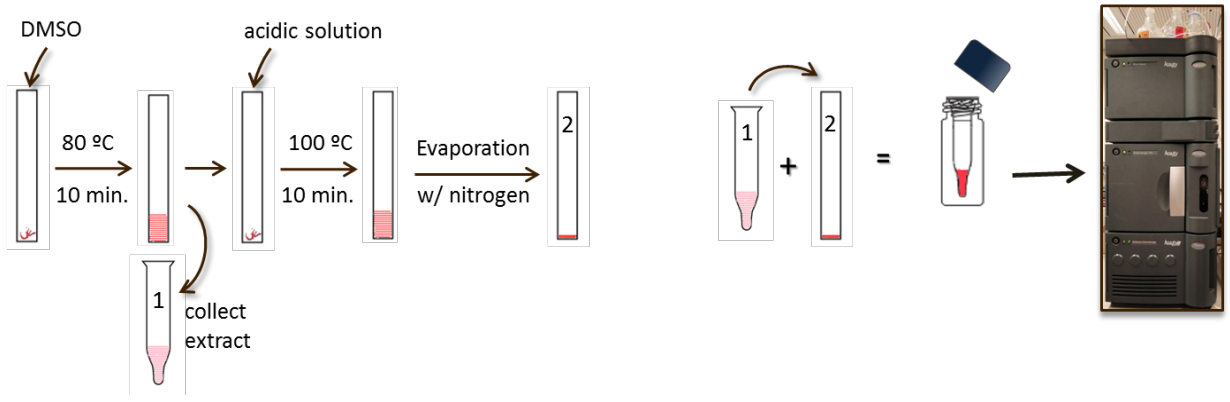


Fig. 1: Scheme of historical textile samples preparation using a two-step extraction method.

Analytical conditions. Analyses are carried out using ultra-high performance liquid chromatography (UHPLC). The authors used a Waters AcquityTM H-class UHPLC system (Waters Corporation, Milford, MA, U.S.A.) equipped with a quaternary solvent delivery system, a column oven, an autosampler and a photo-diode array (PDA) detector.

PDA data was recorded from 200 to 800 nm with a resolution of 1.2 nm (2 scan/s), and the analyses monitoring was settled at a detection wavelength of 254 nm. The equipment was controlled by Empower 3.0 Chromatography Data Software from Waters Corporation.

Separation is performed using a method published by Serrano et al. [1], which has demonstrated to deliver suitable chromatographic detection and resolution for the analysis of mixtures of insect dyes.

Thereby, analytical conditions are carried out using a Waters Acquity® UHPLC BEH Shield RP18 1.7 µm of 2.1 x 150 mm column, protected by a filter unit (0.2 µm), with 2 µL injection volume, a flow rate of 0.2 ml min-1 and a constant temperature of 40 °C.

The mobile phase comprises 10% aqueous methanol (v/v) (solvent A), pure methanol (solvent B) and 1% aqueous formic acid (v/v) (solvent C) in a gradient elution program scheduled for a 40 min run: 0–1.33 min, isocratic gradient of 80A:10B:10C (v/v/v); 1.33–2.33, linear gradient to 74A:16B:10C (v/v/v); 2.33–5.33, linear gradient to 55A:35B:10C (v/v/v), kept in isocratic gradient until 9 min; 9–14 min, linear gradient to 30A:60B:10C (v/v/v); 14–25 min, linear gradient to 5A:85B:10C (v/v/v); 25–26 min, linear gradient to 100B, kept for 4 min; and 30–32 min, linear gradient to 80A:10B:10C (v/v/v), kept for 8 min.

[1] Serrano, A., Bommel, M. van, Hallett, J., 2013, 'Evaluation between ultrahigh pressure liquid chromatography and high-performance liquid chromatography analytical methods for characterizing natural dyestuffs', *Journal of Chromatography A*, 1318, pp. 102-111.

STEP 2: Be sure to make a good qualitative interpretation of the chromatographic results.

For the qualitative interpretation of the chromatographic results, chromatograms are examined at 275 nm, as the major and minor compounds from the insect dyes (Table 1), can be detected at this wavelength.

Table 1. Representative compounds in cochineal and kermes dye extracts [2].

Insect Dyes	**Representative compounds**
Kermes	Flavokermesic acid (fk), kermesic acid (ka)
Polish cochineal *(P. polonica*)	ppI, carminic acid (ca), dcIV, dcVII, fk, ka
American cochineal (*D. coccus*)	dcII, ca, dcIV, dcVII, fk, ka
Armenian cochineal (*P. hamelli*)	dcII, ca, dcIV, dcVII, fk, ka

Moreover, characterization of the compounds and minor compounds should always be possible, as long as the respective PDA spectra can still be recognized. This step is essential to recognize whether minor compounds should be considered noise, something that is particularly helpful when characterizing species of cochineal in unknown historical samples.

Characterization of cochineal dyes with PLS-DA only considers two chromatographic regions (absorbance always at 275 nm): 14.5–17.5 and 20.5–24 min (retention time), because this includes relevant dye compounds (dcIV, dcVII and/or fk and ka), as shown in Fig. 2.

Given the usual high amount of carminic acid (ca) compound in cochineal-dyed fibres, it is expected that this is always detected in historical cochineal-dyed samples, along with the minor compounds. Therefore, PLS-DA models are only suitable for chromatographic results of historical textile samples that exhibit a characteristic chromatographic profile (Fig. 3) of cochineal or, possibly, a mixture with kermes.

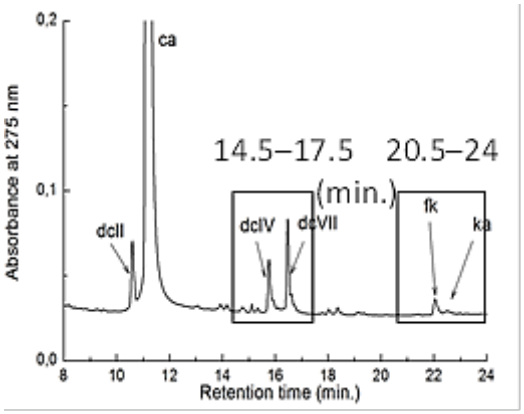


Fig. 2: Chromatographic regions used to perform PLS-DA.

[2] J. Wouters, A. Verhecken, 1989, "The coccid insect dyes: HPLC and computerized diode-array analysis of dyed yarns", *Studies in Conservation*, 34, pp. 189–200.

However, if the response of ca is too small, the minor compounds should not be detected and, for this reason, such chromatographic results cannot be considered in further PLS-DA models (Fig. 4). Also, if a sample is characterized with a complex mixture of dyestuffs (Fig.5), chromatographic regions may be too different from those in the reference samples and, therefore, PLS-DA results may not be accurate.

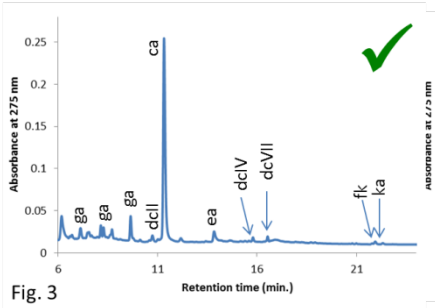


Fig. 3

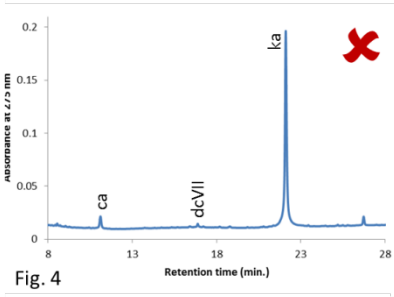


Fig. 4

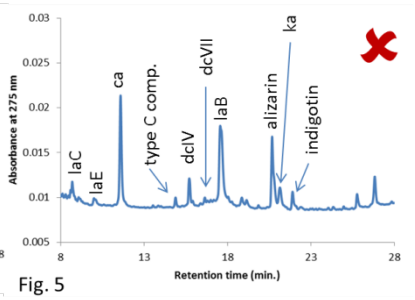


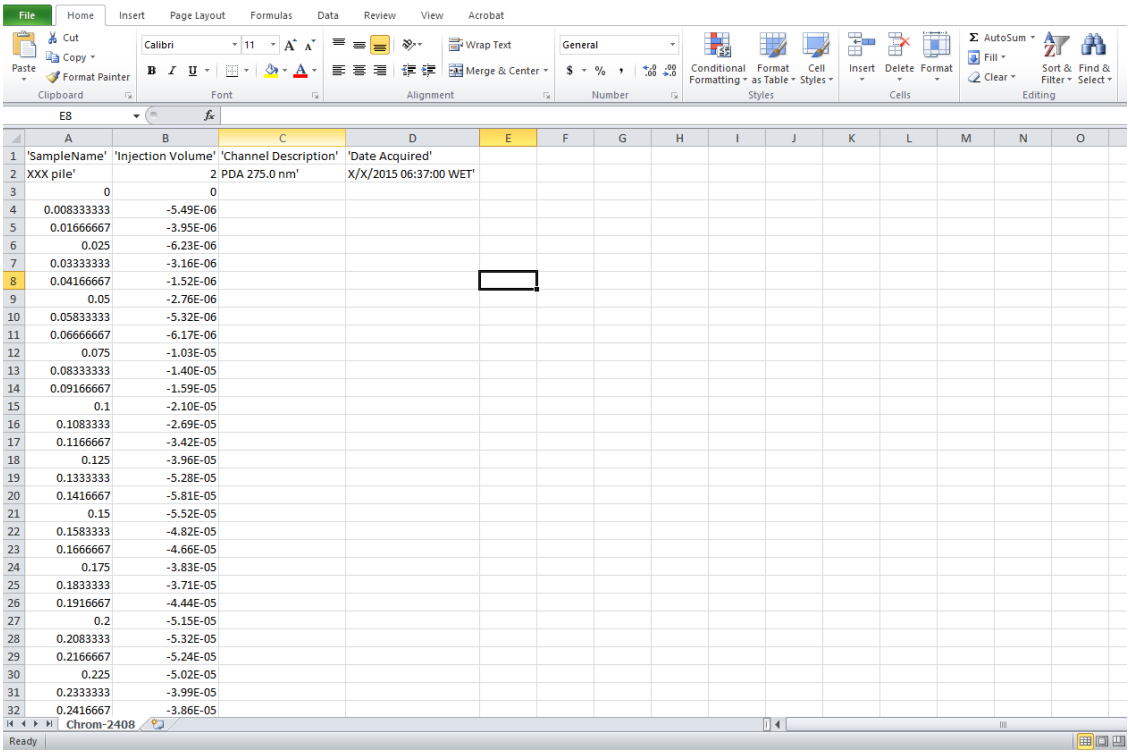
Fig. 5

Fig. 3: Chromatogram characterized by a mixture of cochineal and tannins (gallic acid – ga, and ellagic acid - ea); Fig. 4: Chromatogram characterized by a mixture of kermes (ka) and a small amount of cochineal (ca and dcVII); Fig. 5: Chromatogram characterized by a complex mixture of dyestuffs – cochineal (ca, dcIV, dcVII, ka), lac dye (la's), brazilwood (type c comp.), madder (alizarin) and indigo/woad (indigotin).

Ideally, chromatographic results should present only the relevant dye compounds (dcIV, dcVII and/or fk and ka) in the regions submitted to PLS-DA.

STEP 3: Prepare the chromatograms for the PLS-DA models.

Once it is decided which chromatograms from historical samples can be submitted to PLS-DA, process them at 275 nm and export them from the UHPLC software (authors have used Empower 3), as Excel files. When opening each Excel file (named in this text as FILE 1), this should have a similar appearance to Fig. 6.



	A	B	C	D
1	'SampleName'	'Injection Volume'	'Channel Description'	'Date Acquired'
2	XXX pile'	2	PDA 275.0 nm'	X/X/2015 06:37:00 WET'
3	0	0		
4	0.008333333	-5.49E-06		
5	0.01666667	-3.95E-06		
6	0.025	-6.23E-06		
7	0.03333333	-3.16E-06		
8	0.04166667	-1.52E-06		
9	0.05	-2.76E-06		
10	0.05833333	-5.32E-06		
11	0.06666667	-6.17E-06		
12	0.075	-1.03E-05		
13	0.08333333	-1.40E-05		
14	0.09166667	-1.59E-05		
15	0.1	-2.10E-05		
16	0.1083333	-2.69E-05		
17	0.1166667	-3.42E-05		
18	0.125	-3.96E-05		
19	0.1333333	-5.28E-05		
20	0.1416667	-5.81E-05		
21	0.15	-5.52E-05		
22	0.1583333	-4.82E-05		
23	0.1666667	-4.66E-05		
24	0.175	-3.83E-05		
25	0.1833333	-3.71E-05		
26	0.1916667	-4.44E-05		
27	0.2	-5.15E-05		
28	0.2083333	-5.32E-05		
29	0.2166667	-5.24E-05		
30	0.225	-5.02E-05		
31	0.2333333	-3.99E-05		
32	0.2416667	-3.86E-05		

Fig. 6: Typical Excel file from a chromatogram exported from UHPLC software (FILE 1).

Here, the first column (left) corresponds to the retention time, whereas the second column (right), to the absorbance at 275 nm.

When building the chromatogram in the Excel file, the absorbance must be corrected (add +0.01 to all data points of the right column) for a better representation (Fig. 7).

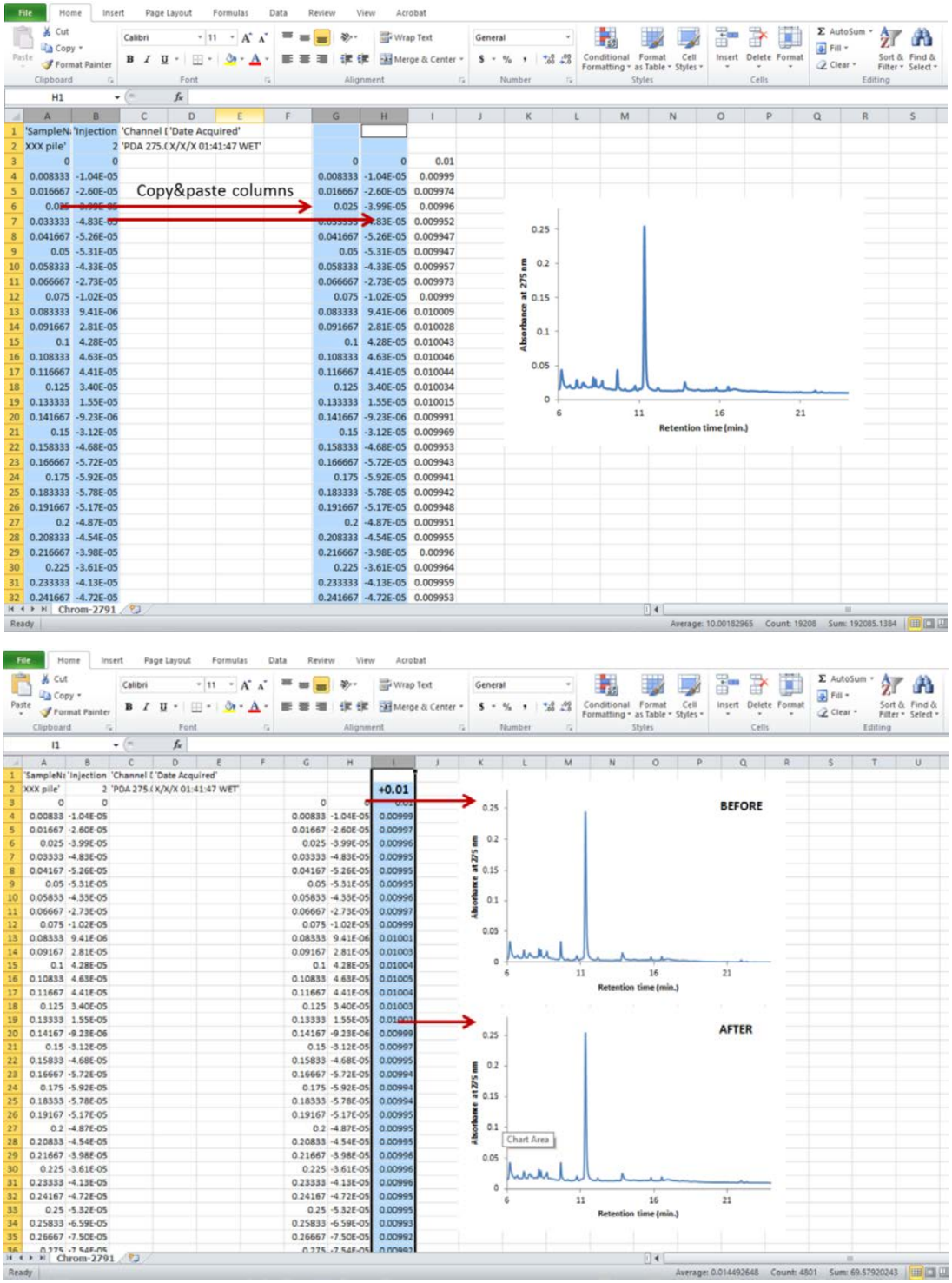


Fig. 7: Correction of chromatogram for a better representation.

Although this correction seems an unnecessary step, it is required to show consistency with the reference samples of the PLS-DA models – these too have been corrected. Also, building the chromatograms is not essential here, but it helps to verify if the exported data is correct, and, especially, to verify the retention time of carminic acid, which will be soon used as a reference (see below).

After all chromatograms have been corrected, open the Excel file "Historical samples" (named in this text as FILE 2), found in "Matlab" folder. The first worksheet "Labels" (Fig. 8) helps organizing the information available for the analysed historical samples, as well as respective qualitative interpretations (detection of dye compounds) and further PLS-DA attributions.

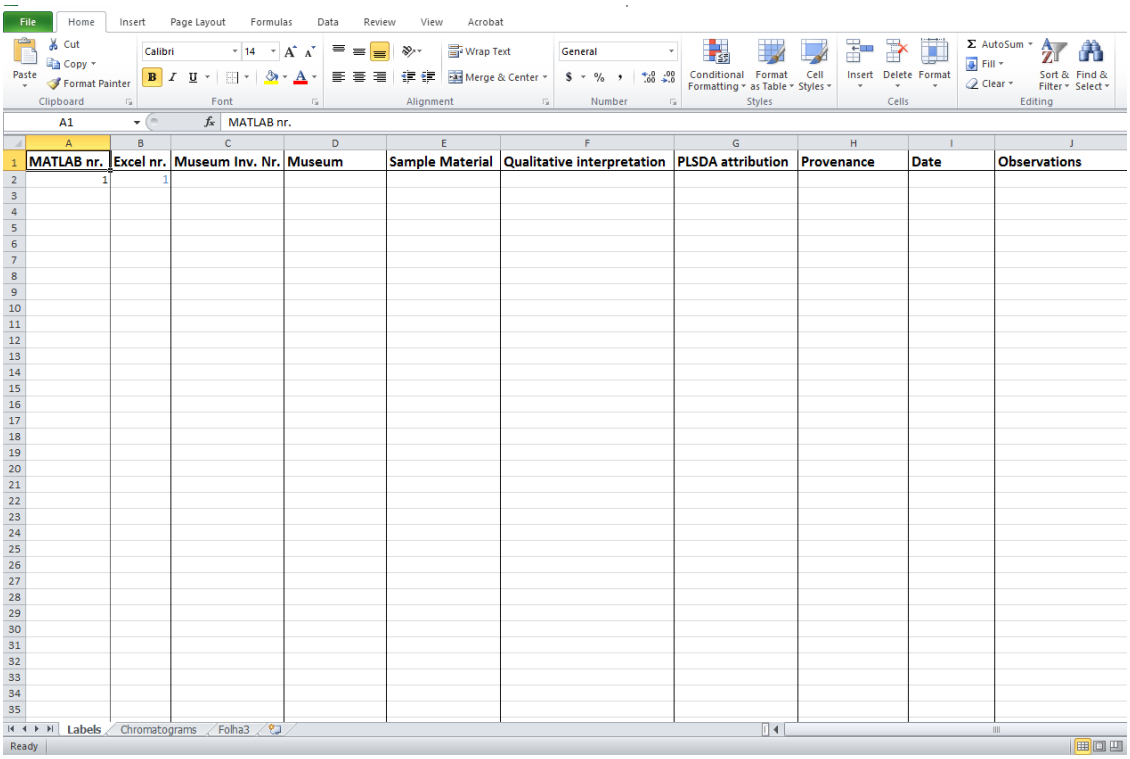


Fig. 8: "Labels" worksheet in Excel file "Historical samples" (FILE 2).

Ultimately, this table can be very useful to compare all results and give the best cochineal dye attribution. **PLS-DA results should not be considered alone, but always be compared with the qualitative interpretations, as well as with the provenance and the date of the textiles, if available.** For specific examples, read the paper suggested at the beginning of this tutorial.

A second worksheet “Chromatograms” is then available in FILE 2, and this is where the corrected absorbance from FILE 1 is added. Hence, copy the right column from FILE 1, and then, past it in the worksheet “Chromatograms” from FILE 2 – make sure to paste the data as values (Fig. 9).

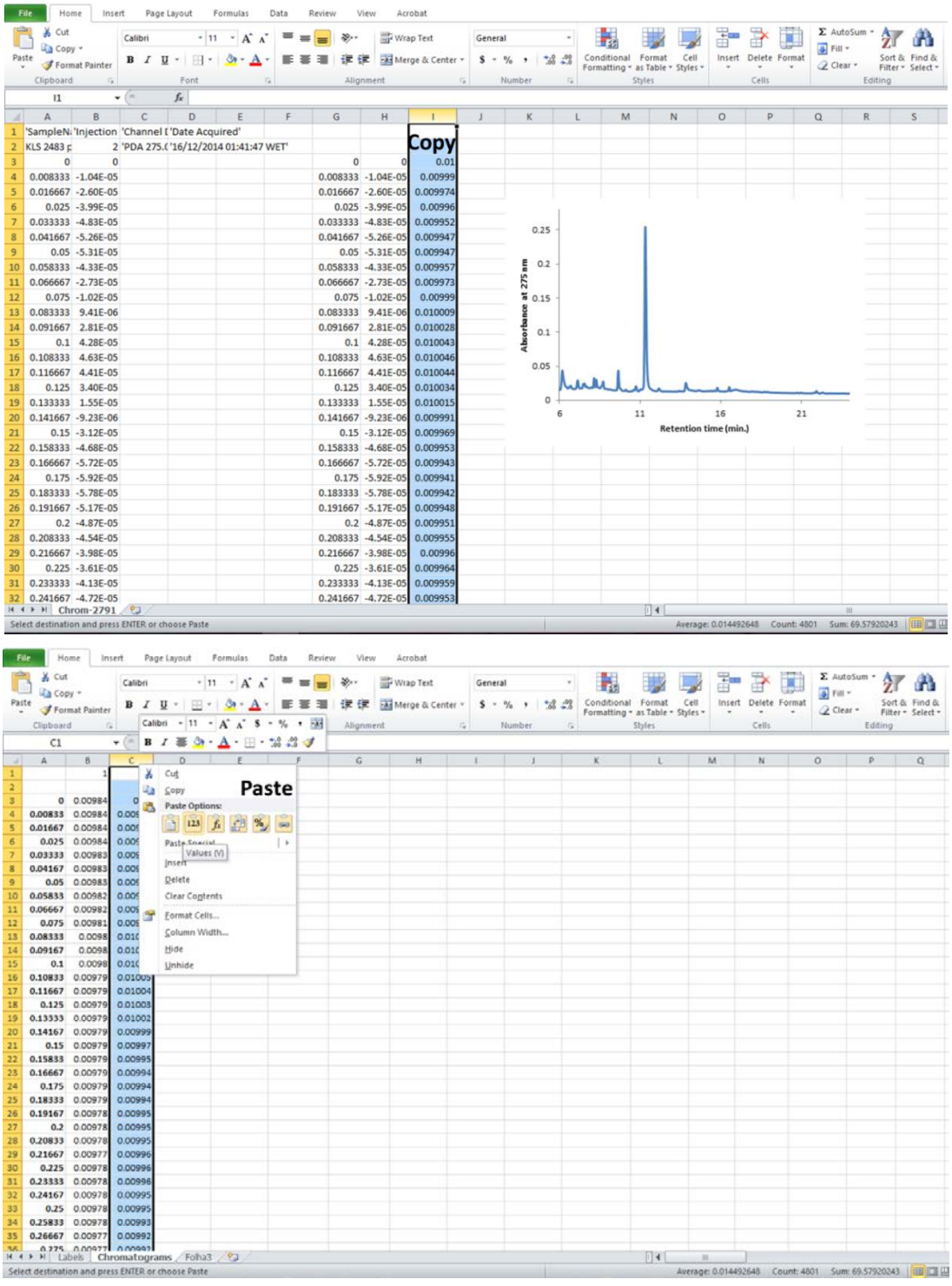


Fig. 9: Import of chromatograms from FILE 1 to FILE 2.

In this “Chromatograms” worksheet, two more columns are presented for reference. The first (most left) corresponds to the retention time, whereas the second (right next), to the absorbance of a chromatogram that belongs to a historical sample (used as example here).

It is important that the rows (or data points) of the retention time correspond to those of the retention time column of the FILE 1. If not, the retention time of FILE 1 needs to be added as well, along with the absorbance column. These two new columns should then be compared with the two columns that are already present in FILE 2.

In order to standardize all chromatograms for the PLS-DA models, it is of upmost importance to centre all by the carminic acid peak. For this, select all data points from the absorbance and find the maximum peak (Fig. 10) - usually corresponds to carminic acid.

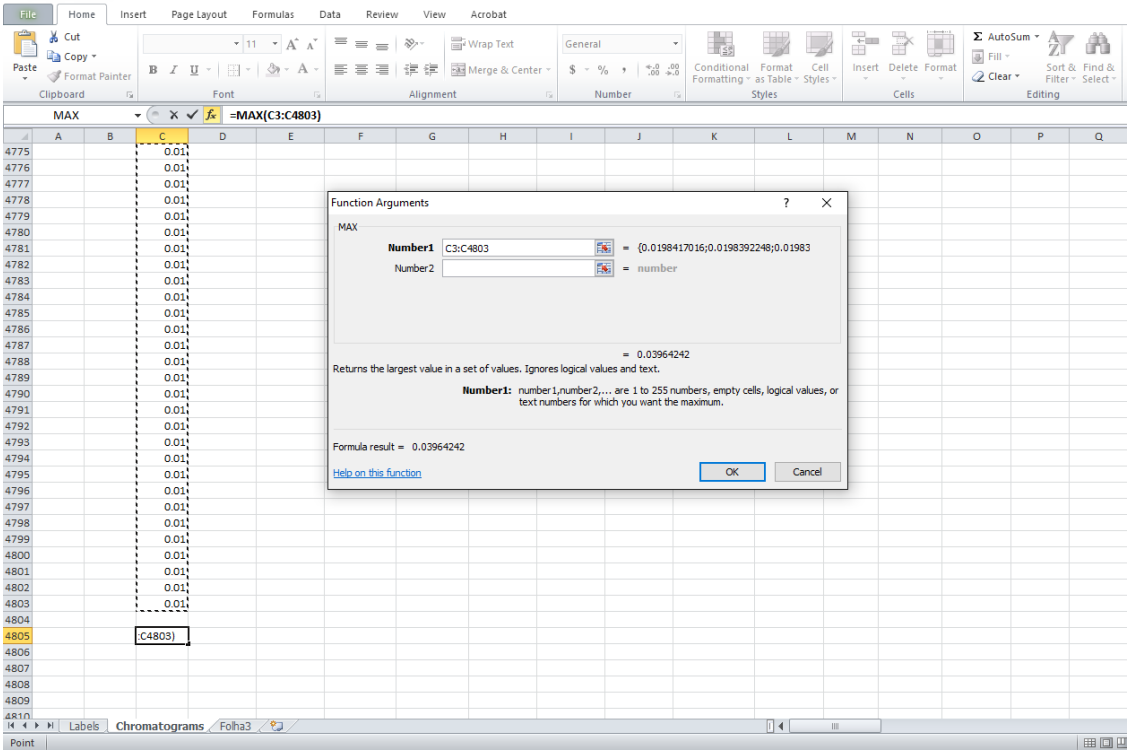


Fig. 10: Finding the maximum absorbance peak in the chromatogram.

If the maximum peak is not carminic acid, the chromatogram representation in FILE 1 must be examined, in order to check roughly what is the retention time and absorbance of carminic acid. This helps to find the peak, when scrolling throughout all data points in FILE 2. In the case of the chromatogram (used as example) already available in FILE 2, carminic acid eluted around minute 11.32.

However, it is clear that the data point corresponding to carminic acid is not centered with the previous chromatogram (on the left – in red). In this case, the data points from the right column need to be moved up (or down, depending on the situation), until they match those from the left column (Fig. 11).

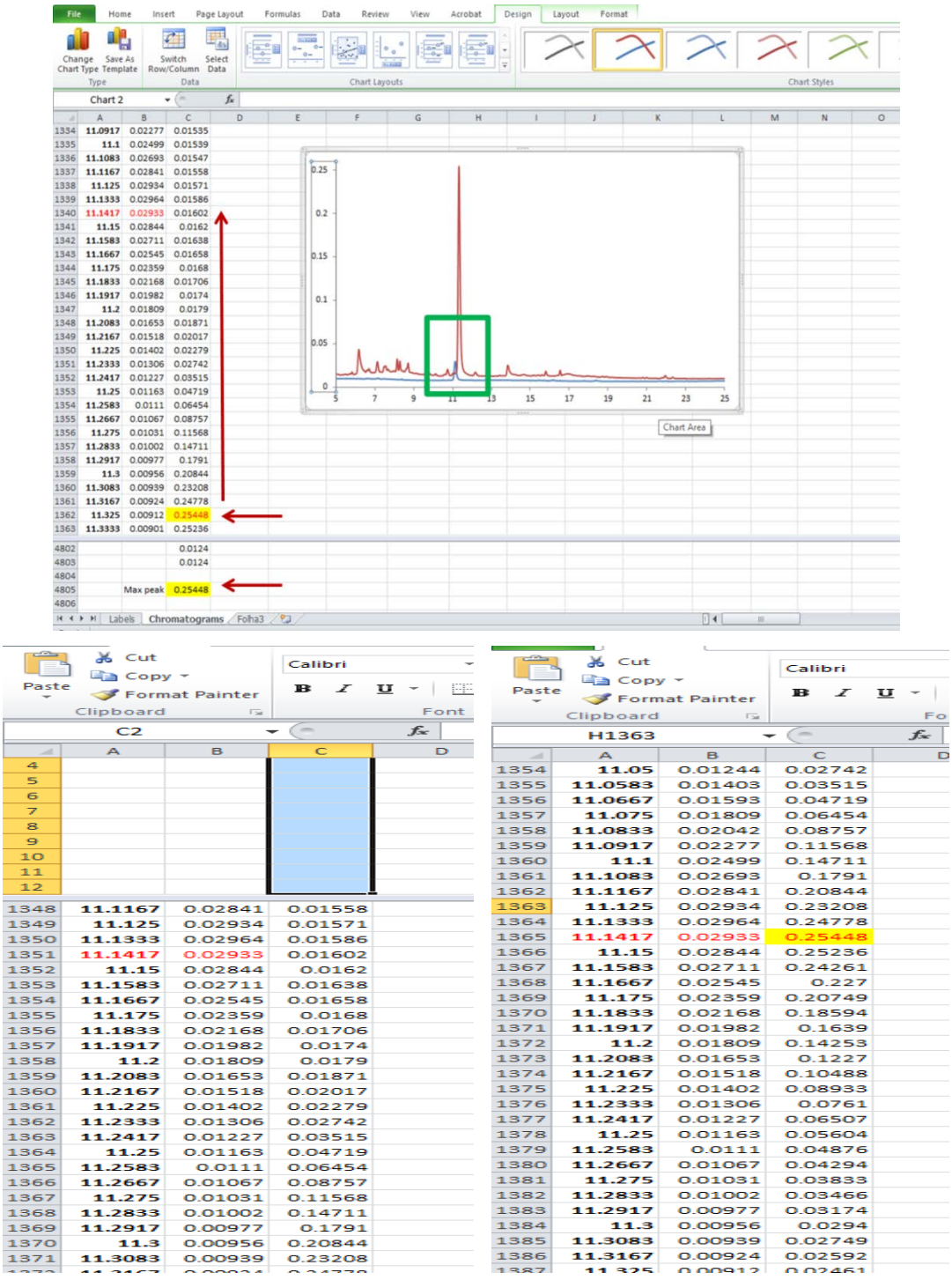


Fig. 11: Centering chromatogram with the reference chromatogram, through carminic acid.

If any other chromatograms ought to be added, these should always be added next to the first columns and their main carminic peak aligned with the reference column. When this is done, the first reference absorbance column can be deleted, so that it will not appear in the PLS-DA models.

If a new retention time column is added (different from the one originally present in FILE 2), then this should be aligned first with the other retention time column and the old must be removed. The reference for alignment should be the same (or very similar) retention time as before: minute 11.147. This number should never change, as it helps to align all historical samples with the reference samples in the PLS-DA models.

Once all alignment is completed, all data falling outside the boundaries of the retention time can be removed (Fig. 12). Anything after minute 30 is also irrelevant and can be removed, because the last ten minutes of the analytical run are mainly for stabilization, and no compounds are expected to elute.

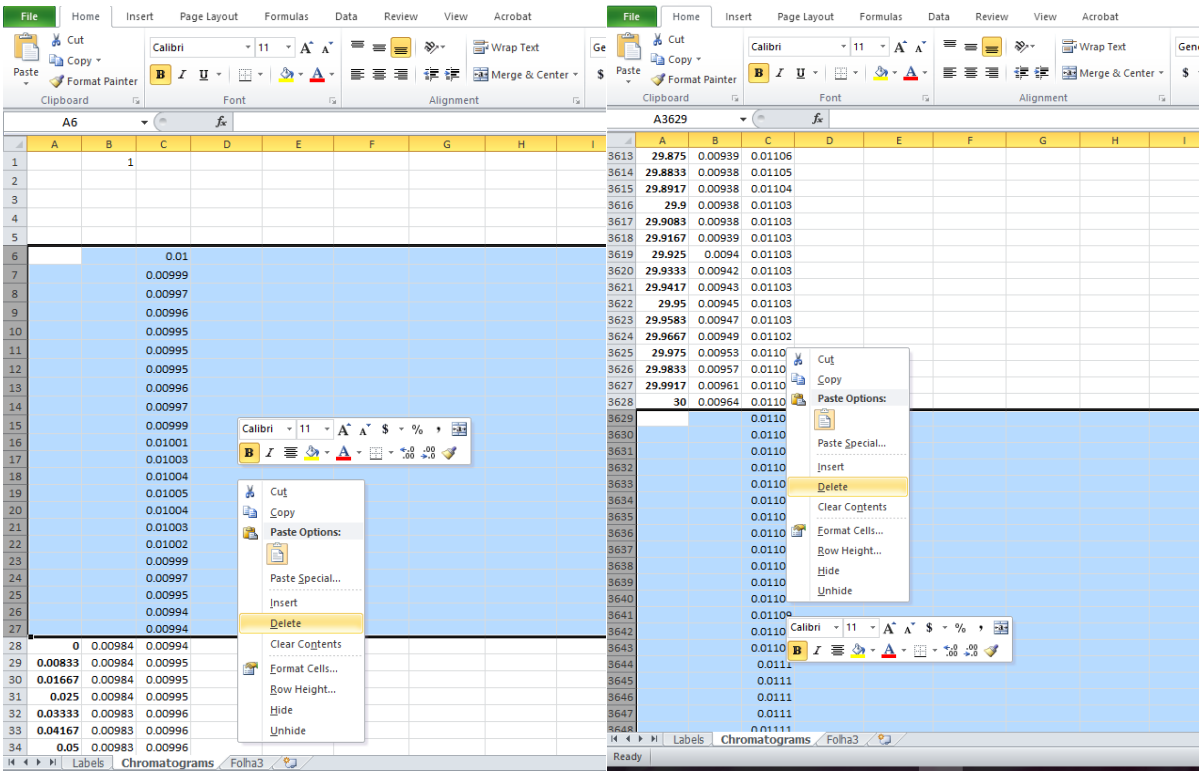


Fig. 12: Removal of additional rows falling outside the retention time (0-30 min.).

If, by chance, the alignment requires a lot of moving of data points (up or down), and the first rows of the absorbance column become empty in relation to the data points of the retention time, this may create an error when performing PLS-DA. Therefore, in any empty rows found, it should always be added 0 (zero).

Also, to ensure comparison with the reference chromatograms, check if all chromatograms start on row 4, i.e., if **min. 0 corresponds to row 4**.

IMPORTANT NOTE: If chromatograms of historical **silk and wool** samples are added to FILE 2, they actually must be separated into two different files. This is because cochineal dyes have been reported to behave differently in both types of fibres (for more details, see paper mentioned in the beginning of this tutorial) and, therefore, they need to be projected onto different PLS-DA models (silk or wool). If this is the case, create a new FILE 2 and copy/paste only the chromatograms belonging to cochineal-dyed silk **OR** wool. In the end, there should be two "Historical samples" files: one with silk historical samples, and another with wool historical samples.

CHROMATOGRAMS PROJECTION ONTO PLS-DA MODELS

STEP 1: Preparing the software and pre-processing the chromatograms.

In order to perform PLS-DA, a relatively recent version of Matlab software is necessary to be installed, as well as the add-on PLS Toolbox. The authors used Matlab R2015a (Mathworks, Natick, MA, USA) and 8.2 PLS Toolbox (Eigenvector Research, Manson, WA, USA).

DISCLAIMER: This tutorial was made with PLS toolbox version 8.2 and other versions may be slightly different.

Once Matlab and PLS Toolbox are properly installed, it is possible to proceed with the application of this tutorial.

Prior to the submission of FILE 2 containing the historical samples, it is necessary to build first the PLS-DA models. These are prepared with samples of known identity, i.e., artificially aged and non-aged silk OR wool fibres, experimentally-dyed with insect dyes. These can be found in the Excel files "Dyed fibres - silk", "Dyed fibres - wool", "Photo-degraded fibres - silk" and "Photo-degraded fibres - wool". By opening any of these files (found in the "Matlab" folder), the first worksheet "Labels" displays the samples of known identity, grouped into four classes of insect species:

- American cochineal (black – class 1 for dyed fibres and class 5 for aged fibres);
- Armenian cochineal (blue – class 2 for dyed fibres and class 6 for aged fibres);
- Polish cochineal (red – class 3 for dyed fibres and class 7 for aged fibres);
- Mixture of American cochineal and kermes (green – class 4 for dyed fibres and class 8 for aged fibres).

Acknowledging each of these classes is essential to understand the results obtained with the PLS-DA models.

These four Excel files should NEVER be modified at any time. Any small change in their layout may cause interference on the Matlab scripts and, hence, cause errors in the loading and preprocessing of the chromatograms.

Loading data in Matlab and Preprocessing. Open Matlab and set the "Matlab" folder (in the "Tutorial" folder) as the current folder: at the upper left corner of the current folder window, use the folder icon with the green arrow. In this folder there are all Matlab scripts and calibration data (Excel files) required to build de PLS-DA models. Then, add the warping folder to the path: Right click on the "Warping" folder, hover over "Add to path" and choose either one of the two options (Fig. 13). Now the scripts (.m files) can access the functions in the warping folder.

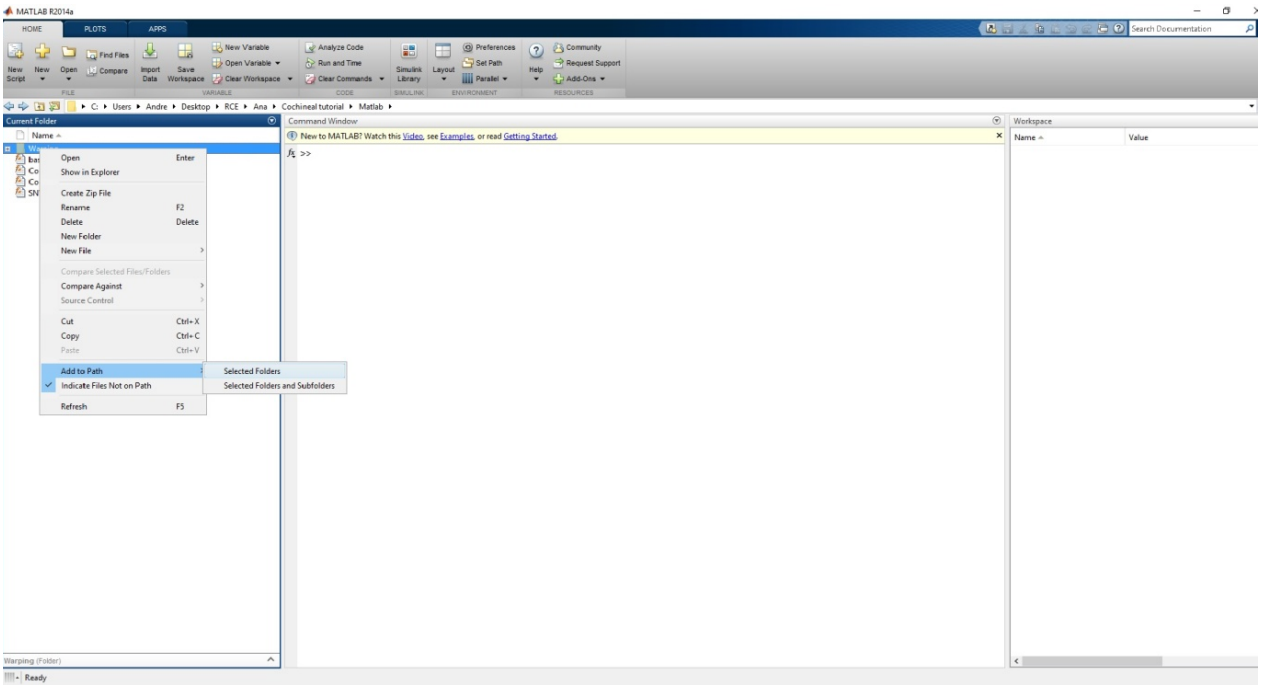


Fig. 13: Adding "Warping" folder to current Matlab folder.

Run the file cochinealPLS.m by typing "[Xcal Xtest] = CochinealPLS();" in the Matlab command window (Fig. 14). You will be prompted to "Select file(s) containing calibration samples" and "Select file(s) containing test samples".

Calibration samples are samples of known origin that are used to define the model – those from Excel files "Dyed fibres - silk", "Dyed fibres - wool", "Photo-degraded fibres - silk" and "Photo-degraded fibres - wool". Depending on which type of historical samples that will be characterized (silk OR wool), select the corresponding Excel files as calibration samples (Dyed fibres and Photo-degraded fibres).

Test samples are samples of unknown origin, which class of insect species needs to be predicted (American, Armenian, Polish and American and kermes). Therefore, the "Historical samples" Excel file (or FILE 2), which was built in previous steps of this tutorial, must be selected. In accordance with the calibration samples, this file must only contain silk OR wool historical samples.

Note that it is possible to select multiple files with calibration or test samples, as long as they are in the same folder.

Running this script can take several minutes to hours, depending on the size of the data and the speed of the computer. It is also possible that this script will stop running in computers with limited memory (RAM). In this case, the error "Out of memory" will appear. For example, this occurred on a computer with 3 Gb of usable RAM. Therefore, a computer with higher memory will be required to build the models.

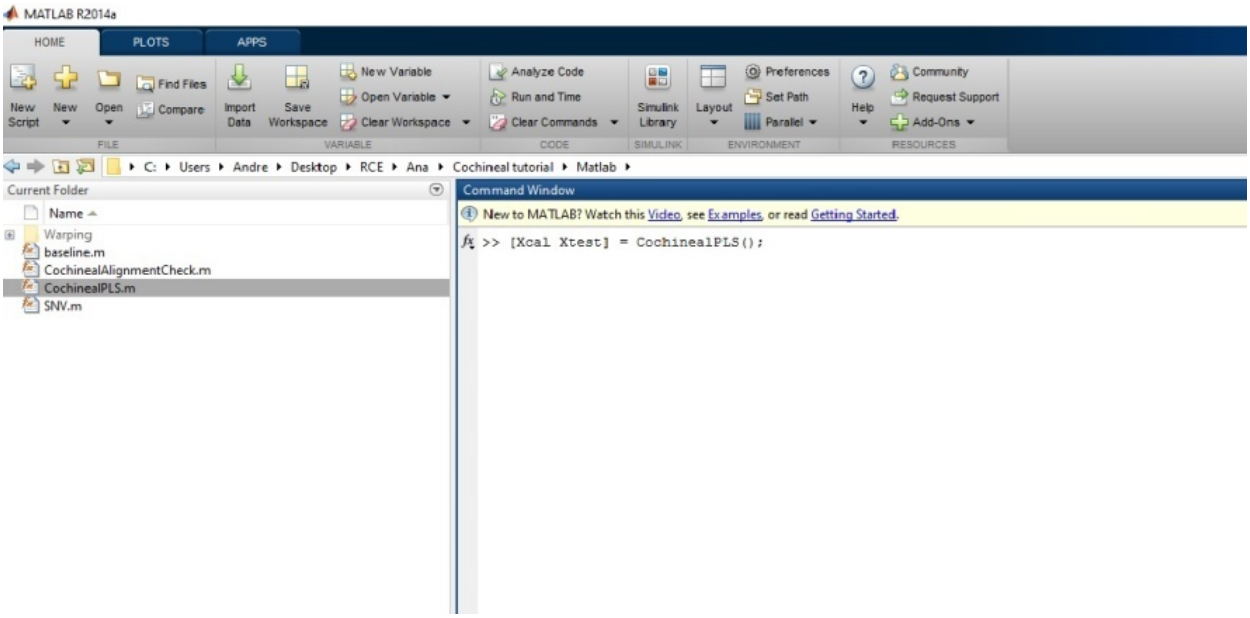


Fig. 14. Running "CochinealPLS" in the Matlab command window (in the centre).

The “CochinealPLS” file automatically does a few preprocessing steps. Some of these steps require parameters that can be set manually (see below for details). To change the parameters, type “edit cochinealPLS” in the Matlab command window and change the parameters in the top section of the script file in the editor (Fig. 15). Then save the changes before running again the script.

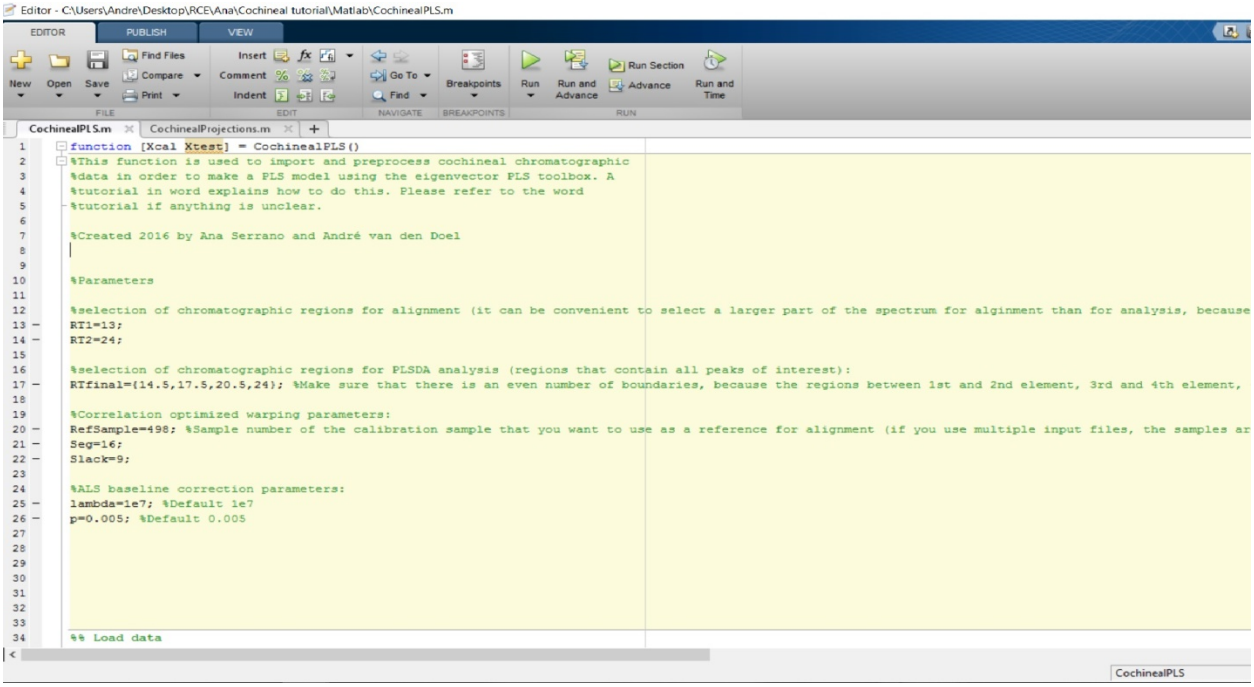


Fig. 15: Changing parameters in CochinealPLS script.

[Note that standard normal variate (SNV) scaling and mean centering are also applied to the chromatograms when running the script, but these do not require input parameters.]

Change of Parameters (more information in the script)

Selection of chromatographic region for alignment: This is the part of the chromatogram that is used for alignment. It must be one continuous region and must at least contain the regions for PLS-DA analysis. It may be advisable to select a slightly larger region for alignment, so that the peaks at the edges of the analysis region are also properly aligned.

Selection of chromatographic regions for PLS-DA analysis: This selects which regions of the chromatogram (in terms of retention time) are used for analysis. These regions must include the peaks (absorbance) that are suspected to be different between samples of different classes (insect species). Peaks that are expected to be the same for all classes can be excluded because they only add unnecessary noise to the model.

Alignment: Chromatograms are aligned using correlation optimised warping (COW). This divides a chromatogram in segments which are stretched or compressed to align it to a reference spectrum. Therefore, it is best to select a reference spectrum that contains all (or most) peaks that are found in all chromatograms. In this case, two parameters can be optimized: segment length and slack. The slack determines how much each segment can be stretched or compressed. The shorter the segment length, or the larger the slack, the more severely a chromatogram can be warped. Too much warping may wrongly align peaks of different compounds.
Depending on which fibres are characterized (silk or wool), be aware to select the proper reference spectra:

- For silk samples, set RefSample=**498** – sample 63.6 in Excel file ("Dyed fibres – silk");
- For wool samples, set RefSample=**270** – sample 64.6 in Excel file ("Dyed fibres – wool").

Baseline correction: Asymmetric least squares is used for baseline correction, and it has two parameters: p for asymmetry and λ (lambda) for smoothness. A higher λ results in a smoother baseline estimate. In this case, it is suggested to use $0.001 \leq p \leq 0.1$ and $10^2 \leq \lambda \leq 10^9$. For more information, follow the link:
https://zanran_storage.s3.amazonaws.com/www.science.uva.nl/ContentPages/443199618.pdf

During the run of the script, several figures will open to show the raw chromatograms, the results of alignment and baseline correction. Pay attention to the alignment (Fig. 16) of the black chromatograms (test samples), in relation to the colourful ones (calibration samples). Principally, the dcIV, dcVII and flavokermesic and kermesic acid peaks of the black chromatograms should correspond to those of the colourful. Also, pay attention to the baseline correction of the chromatograms (Fig. 17).

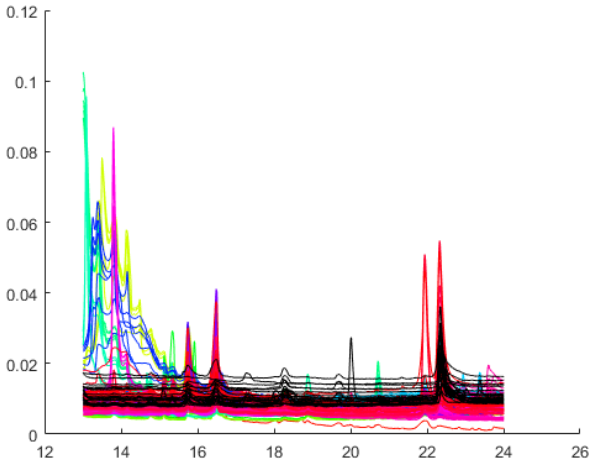


Fig. 16. Results of alignment.

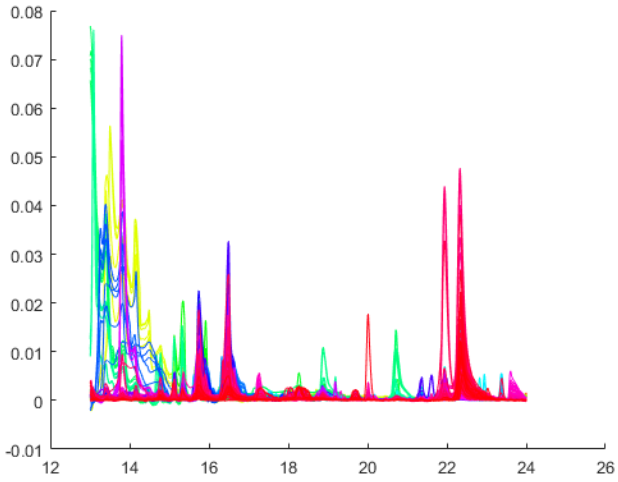


Fig. 17. Results of baseline correction.

Although the above-mentioned parameters are currently set to optimally preprocess the chromatograms, it is possible that some need to be slightly adjusted in order to match the calibration samples with test samples. Hence, if the results are not satisfactory, change some parameters and run the script again to improve the results. For instance, if the alignment of the peaks in the chromatograms is not correct, adjust the parameters in the alignment using COW, until a better alignment of the peaks is achieved.

STEP 2: Preparing the calibration model.

Start the PLS toolbox by typing “browse” in the MATLAB command line. Then, in the Analysis Tools pane, expand CLASSIFICATION and double click on PLSDA to make a PLSDA model (Fig. 18).

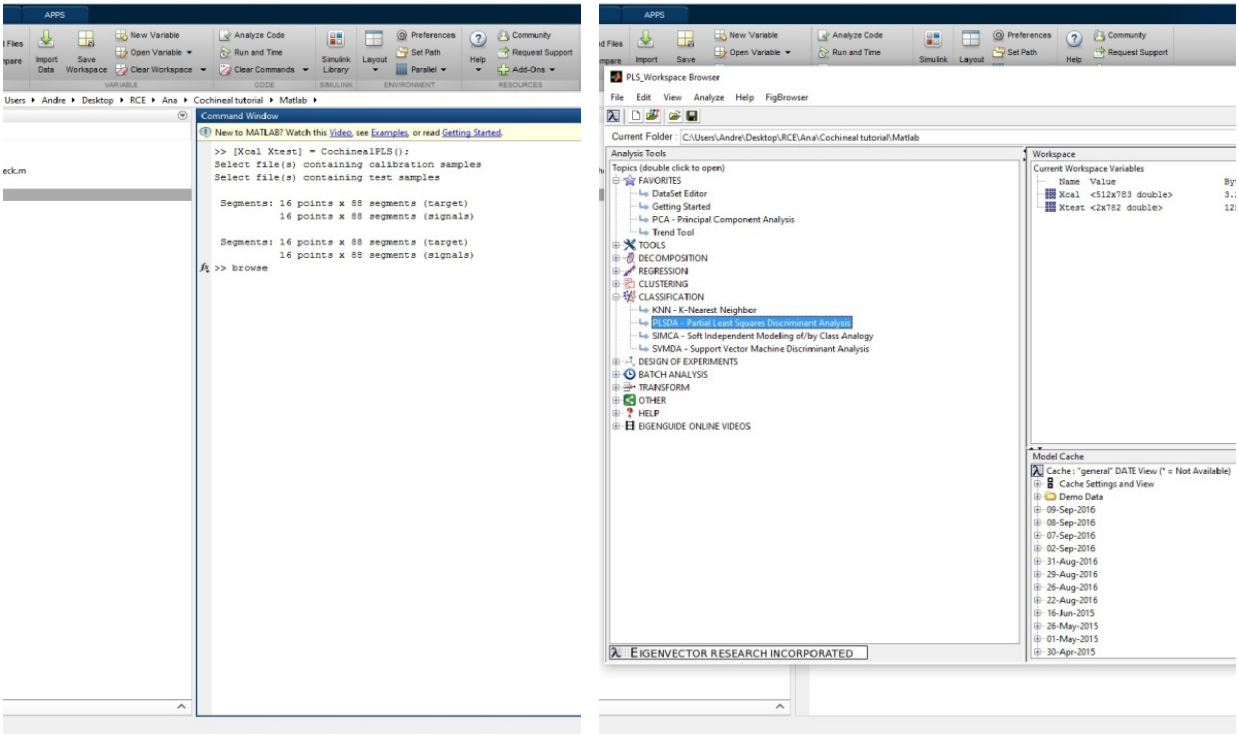


Fig. 18: Typing “browse” in the Matlab command window to start the PLS toolbox (at the left) and clicking on PLSDA in the Analysis Tools pane (at the right).

In the PLSDA window right click on the blue X button, hover over “Import data” and then select “Workspace/MAT file”. Select Xcal and click on “Load” (Fig. 19).

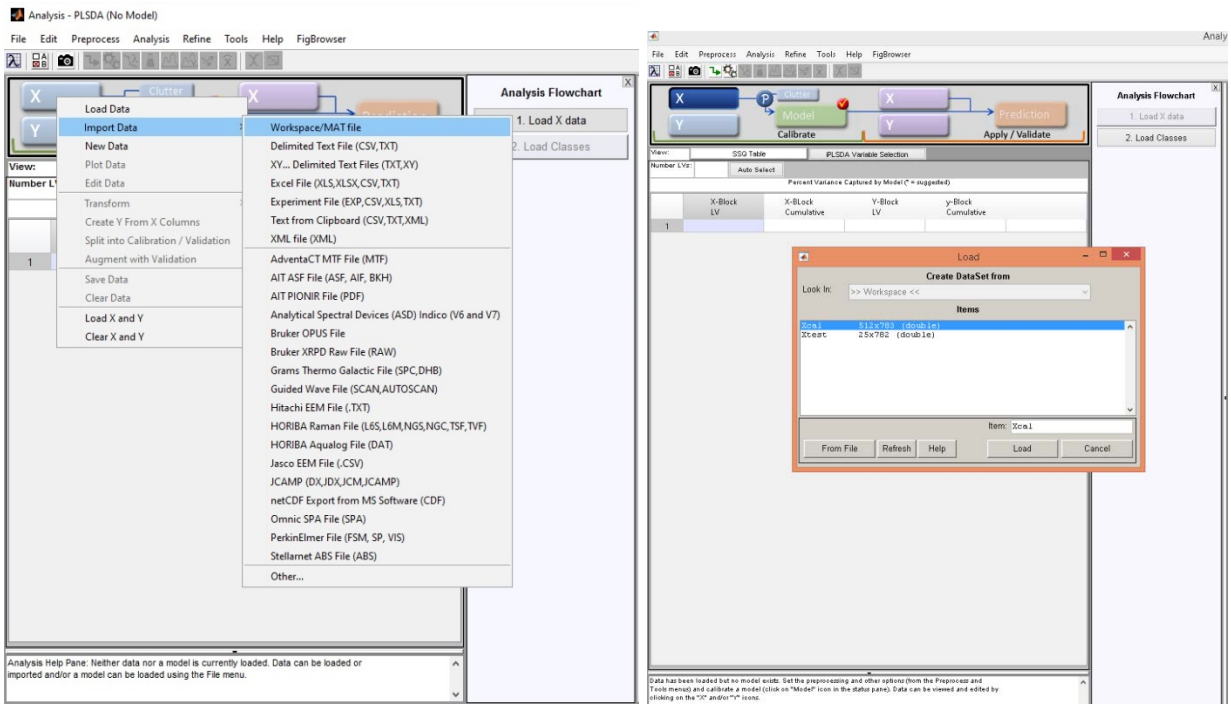


Fig. 19: Importing and loading the calibration data.

The calibration data has now been loaded, but it is still needed to assign classes to the samples. Right click on the blue *X* again and select “Edit data”. A window (DataSet Editor) appears. Right click on the ‘1’ above the first column and select “Use as class”. Click OK on the window that pops up and says “Moved to Class set 1 for mode 1”. The class labels should now be shown in red (Fig. 20). Close this window.

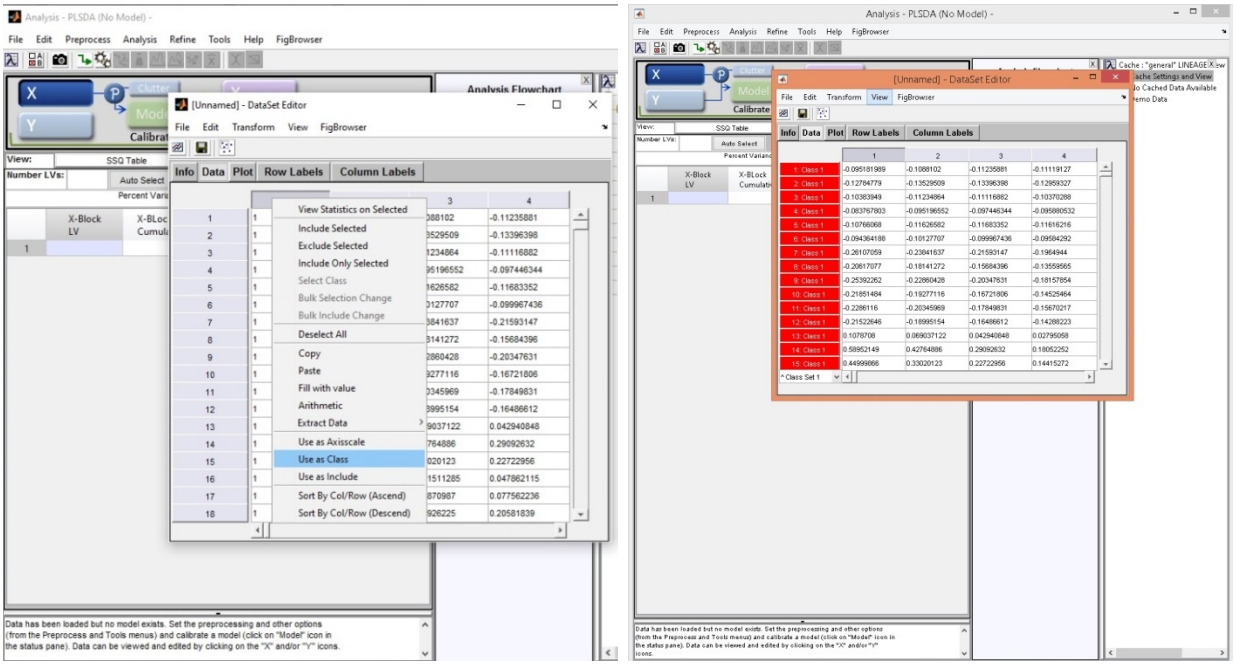


Fig. 20: Setting the class labels.

As previously explained in this tutorial (p. 14), the calibration samples have been assigned with a class number, depending on the cochineal species:

- American cochineal (black – class 1 for dyed fibres and class 5 for aged fibres);
- Armenian cochineal (blue – class 2 for dyed fibres and class 6 for aged fibres);
- Polish cochineal (red – class 3 for dyed fibres and class 7 for aged fibres);
- Mixture of American cochineal and kermes (green – class 4 for dyed fibres and class 8 for aged fibres).

Samples of different classes can be modelled as if they are the same class. For instance, American cochineal-dyed and aged fibres (classes 1 and 5, respectively) can be selected together in one same class. This can be useful if there is no interest in showing separation between these two types of samples – especially when cochineal species are characterized in unknown historical samples. Therefore, the model will focus only on the difference between cochineal species; and not on the additional difference between dyed fibres and aged fibres.

To make class groups, click on "3.Select Class Groups" (right side pane "Analysis Flowchart") and click *OK* in the next popup window. Then, select the classes to be modelled as a group (press CTRL button to select two classes), and click on "Model As Group": Class 1 with 5, Class 2 with 6, Class 3 with 7, and Class 4 with 8, as shown in Fig. 21 and described above. Click *OK* after all the classes have been grouped.

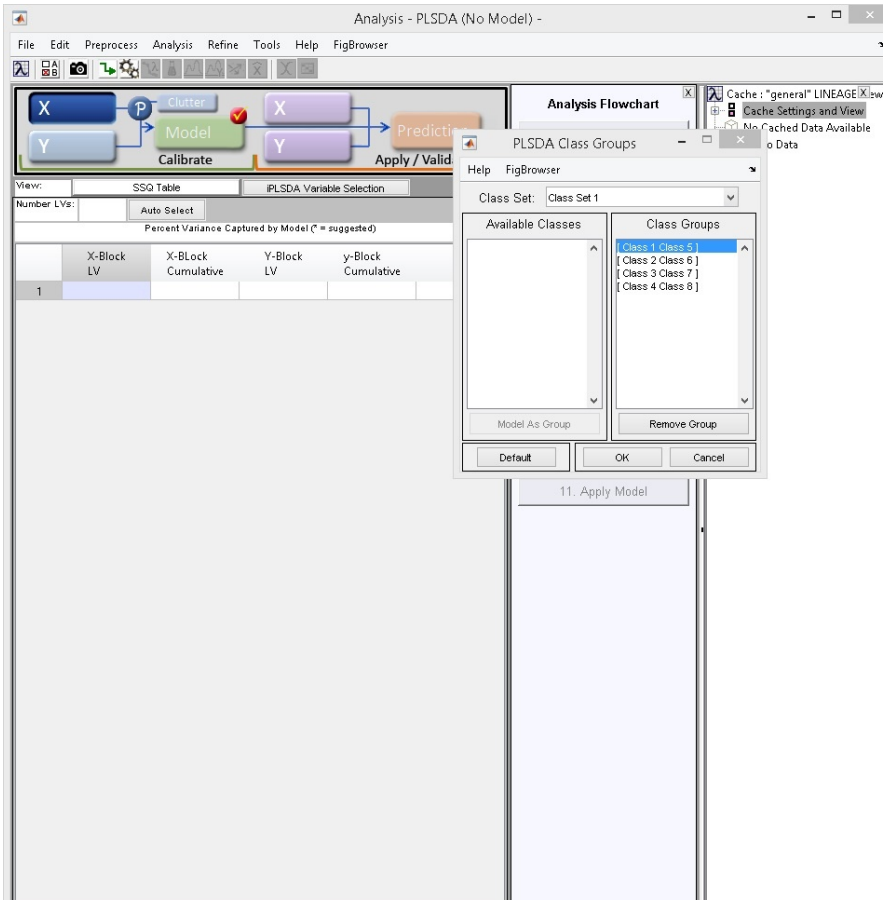


Fig. 21: Modelling the class labels.

Click on “4.Choose Preprocessing” (in pane “Analysis Flowchart”) and remove “Autoscale” (Fig. 22). Because the data is already preprocessed (SNV scaled and mean centered – see above), no more preprocessing is required. Click OK after removing “Autoscale”.

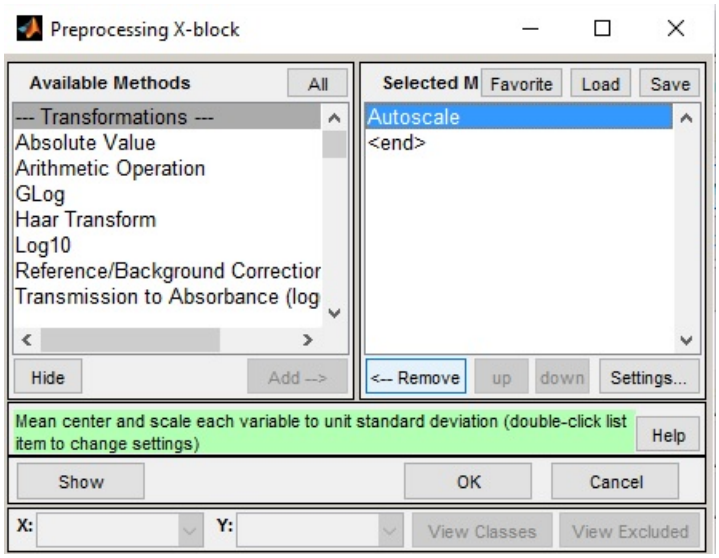


Fig. 22. Removing “Autoscale” from the preprocessing steps.

After, click on choose “5.Cross-Validation”. This should be already set to venetian blinds and, thus, the maximum number of latent variables (LVs) and number of data splits should be at their default values (Fig. 23). Change the “Samples per Blind (Thickness)” to **6**. This number corresponds to the number of replicates undertaken for each sample, i.e, the number of times each sample was analysed with UHPLC. Click *OK* or *Close*.

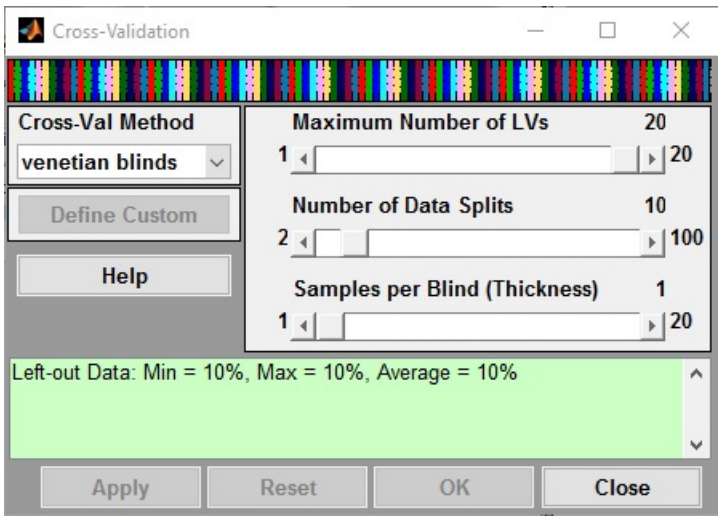


Fig. 23: Cross-validation settings.

Now everything is set and it is possible to build the calibration model.

Building the calibration model. Click on “6.Build Model” to build the calibration model. After some calculation, the model is ready and some statistics are shown, namely the variance explained (LV and Cumulative) and the prediction errors (CV Class Errors 1 to 4) (Fig. 24). These results in Fig. 24 suggest that the optimal number of latent variables is 4, as indicated with an asterisk on the right side of the table (“current*”).

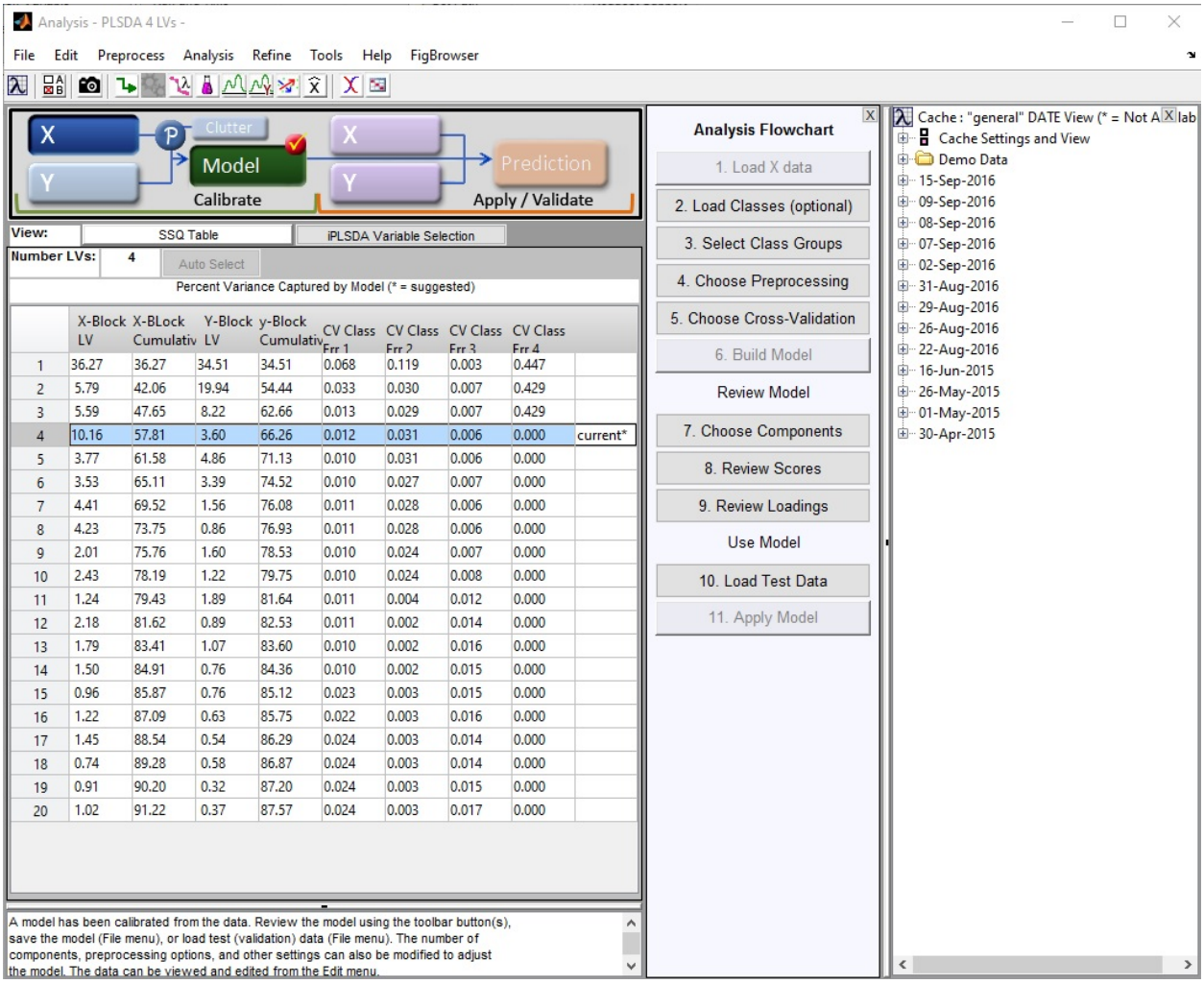


Fig. 24: Calibration model with an optimal number of 4 LVs.

To further investigate the model, click on the purple Erlenmeyer () at the top of the screen. A window with several figures will open (Fig. 25).

The plot on the top left shows a plot of Q-residuals against Hotelling T^2. Double click on it to open a larger version of this plot in a separate window. Use the "Make Selection" tool (), in the toolbar above the figure, to select all samples in the upper right quadrant (Fig. 26). These samples are not well fitted by the model but they do have a large influence on it. Hence, right click on one of the samples and select "Exclude Selection" (click *OK* if a popup window appears). Removing these samples leads to a model that better fits the majority of the samples, as well as to more accurate predictions.

Click away the figure windows.

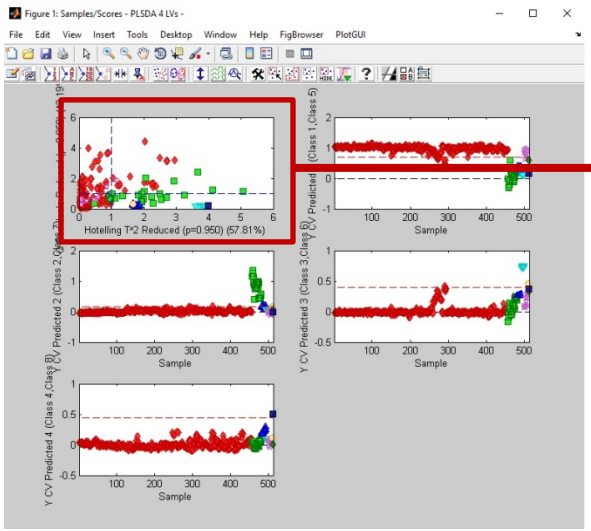


Figure 25. Plots of PLSDA model.

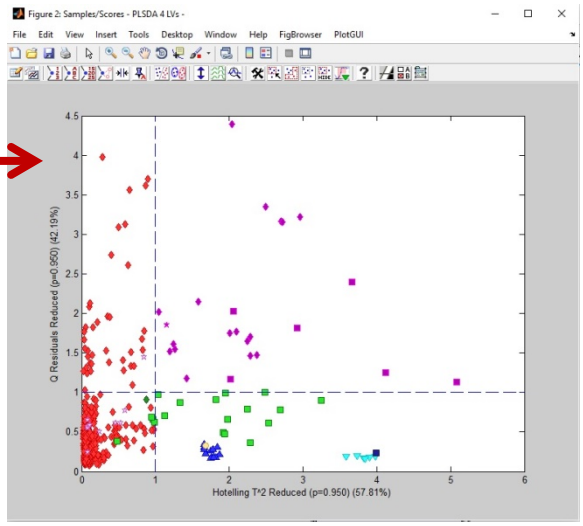
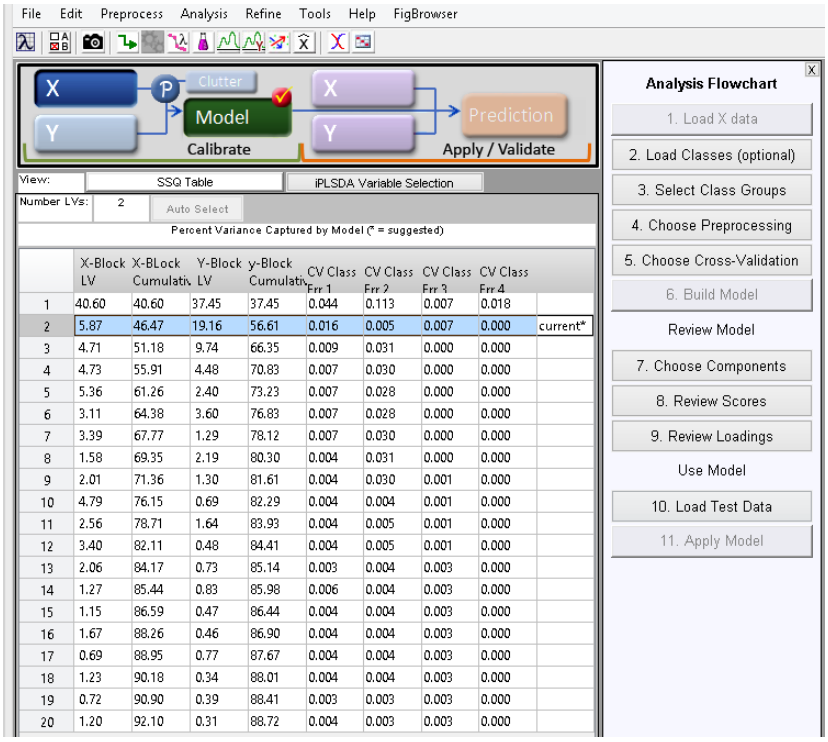


Figure 26. Samples in the upper right quadrant that need to be removed.

Since the samples were removed, now the table with the variance explained and prediction errors have disappeared. Therefore, the model needs to be recalculated. Click on “Build Model” again. For the model in figure 27 the optimal number of latent variables is 2, as indicated with an asterisk on the right side of the table (“current*”) – Fig. 27. Click on the purple Erlenmeyer to investigate the final calibration model.



	X-Block LV	X-BLock Cumulati	Y-Block LV	y-Block Cumulati	CV Class Err 1	CV Class Err 2	CV Class Err 3	CV Class Err 4	
1	40.60	40.60	37.45	37.45	0.044	0.113	0.007	0.018	
2	5.87	46.47	19.16	56.61	0.016	0.005	0.007	0.000	current*
3	4.71	51.18	9.74	66.35	0.009	0.031	0.000	0.000	
4	4.73	55.91	4.48	70.83	0.007	0.030	0.000	0.000	
5	5.36	61.26	2.40	73.23	0.007	0.028	0.000	0.000	
6	3.11	64.38	3.60	76.83	0.007	0.028	0.000	0.000	
7	3.39	67.77	1.29	78.12	0.007	0.030	0.000	0.000	
8	1.58	69.35	2.19	80.30	0.004	0.031	0.000	0.000	
9	2.01	71.36	1.30	81.61	0.004	0.030	0.001	0.000	
10	4.79	76.15	0.69	82.29	0.004	0.004	0.001	0.000	
11	2.56	78.71	1.64	83.93	0.004	0.005	0.001	0.000	
12	3.40	82.11	0.48	84.41	0.004	0.005	0.001	0.000	
13	2.06	84.17	0.73	85.14	0.003	0.004	0.003	0.000	
14	1.27	85.44	0.83	85.98	0.006	0.004	0.003	0.000	
15	1.15	86.59	0.47	86.44	0.004	0.004	0.003	0.000	
16	1.67	88.26	0.46	86.90	0.004	0.004	0.003	0.000	
17	0.69	88.95	0.77	87.67	0.004	0.004	0.003	0.000	
18	1.23	90.18	0.34	88.01	0.004	0.004	0.003	0.000	
19	0.72	90.90	0.39	88.41	0.003	0.003	0.003	0.000	
20	1.20	92.10	0.31	88.72	0.004	0.003	0.003	0.000	

Figure 27. Calibration model with an optimal number of 2 LVs.

The “Plot Controls” box can be used to change the plots. The most common plot is the scores of LV1 on the x-axis and the scores of LV2 on the y-axis (Fig. 29, **A**) - it is also possible to plot other latent variables when the model has more than 2.

By default, the other plots show the predicted Y values for different classes (or class groups). If the predicted Y value is above a threshold value (indicated by the dashed red line) a sample is predicted to belong to that class.

Therefore, on plot **B**, the red and pink samples are predicted to belong to Class 1, Class 5 (American cochineal); on plot **C**, the green samples are predicted to belong to Class 2, Class 6 (Armenian cochineal); on plot **D**, the blue triangle samples, to Class 3, Class 7 (Polish cochineal); and on plot **E**, the blue square samples, to Class 4, Class 8 (mixture of American cochineal and kermes).

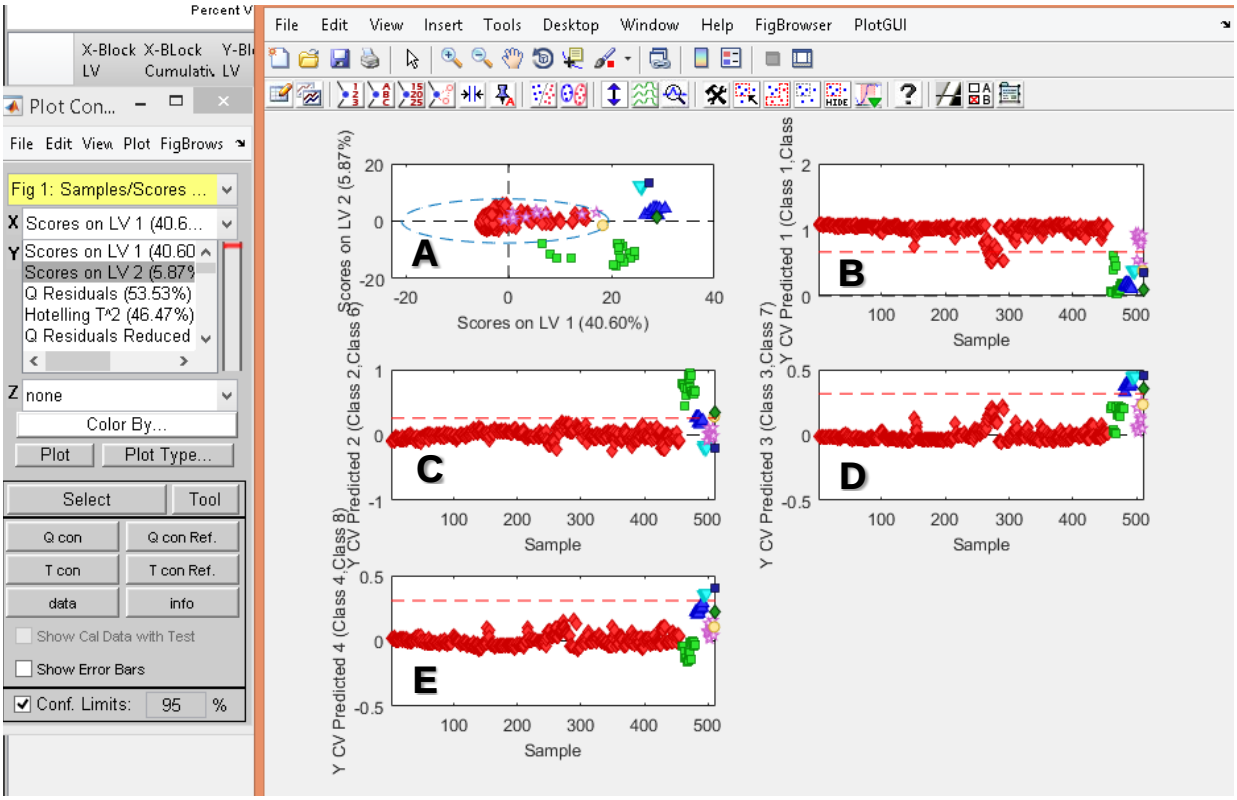


Fig. 29: Plots of the calibration model.

Another way to investigate the model is the confusion table (Fig. 30). This can be found on the upper toolbar of the PLS toolbox analysis window ( ).

The confusion table shows the actual class against the predicted class; how many misclassifications there are in the calibration set; and which classes get "confused". For instance, from the actual 447 samples from Class*1, 439 are actually predicted as "Class 1, Class 5", whereas 8 are wrongly predicted as "Class 2, Class 6". Therefore, these 7 samples are said to be "confused".

Always check the confusion table after cross-validation, as it is easier to interpret than the plots.

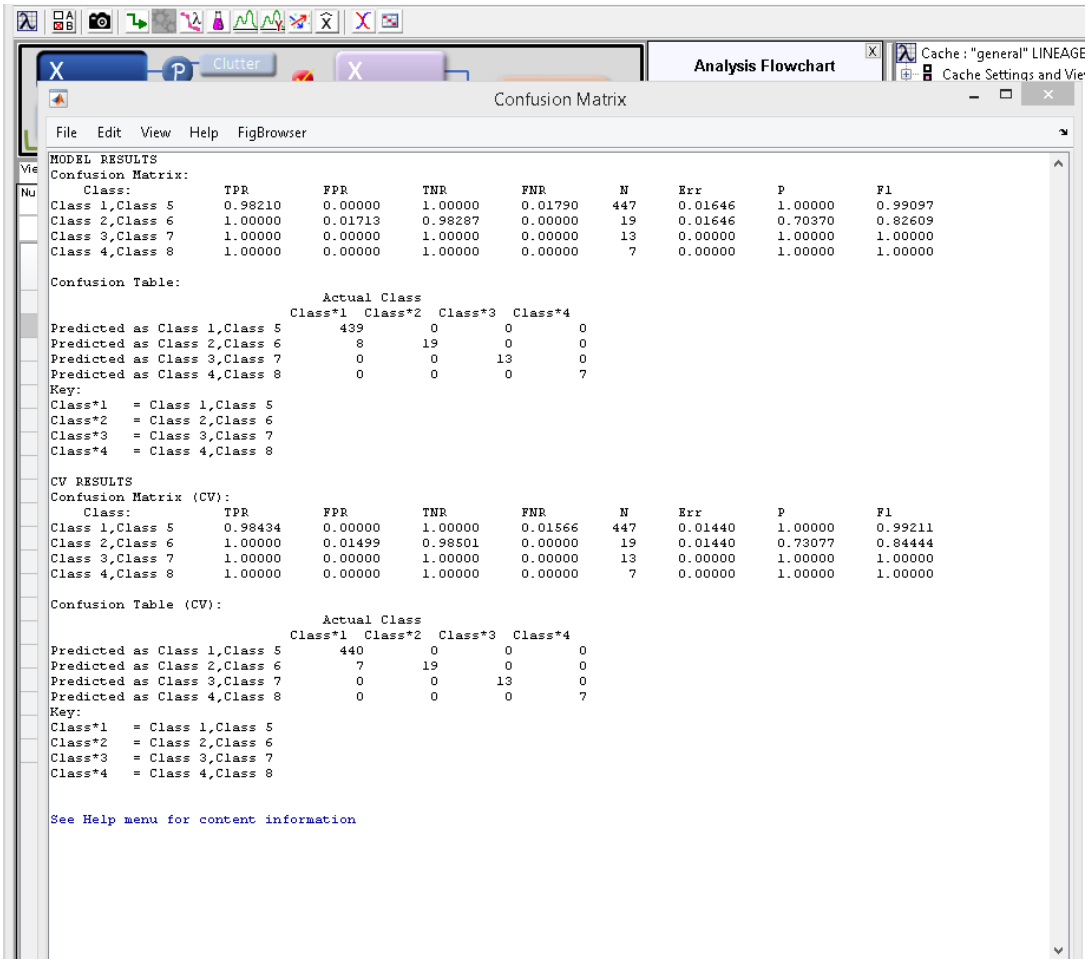


Fig. 30: Depiction of confusion table.

Building the model with calibration and test samples. To import the test data, right click on the purple *X* button, hover over import data and select “Workspace/MAT” file. Then, select *Xtest* and click “Load” (Fig. 31).

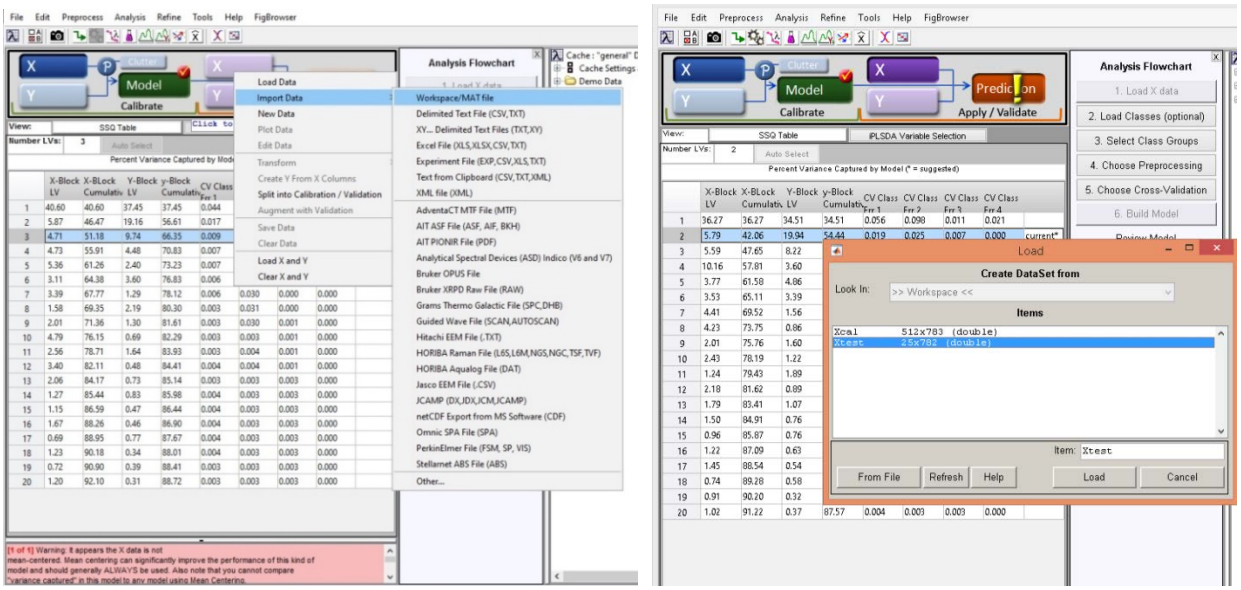


Fig. 31: Importing the test samples.

Now the “Prediction” button will light up and a yellow exclamation mark appears (Fig. 32). Click on this button to apply the model to the test set and, thus, predict the classes of the unknown test samples.

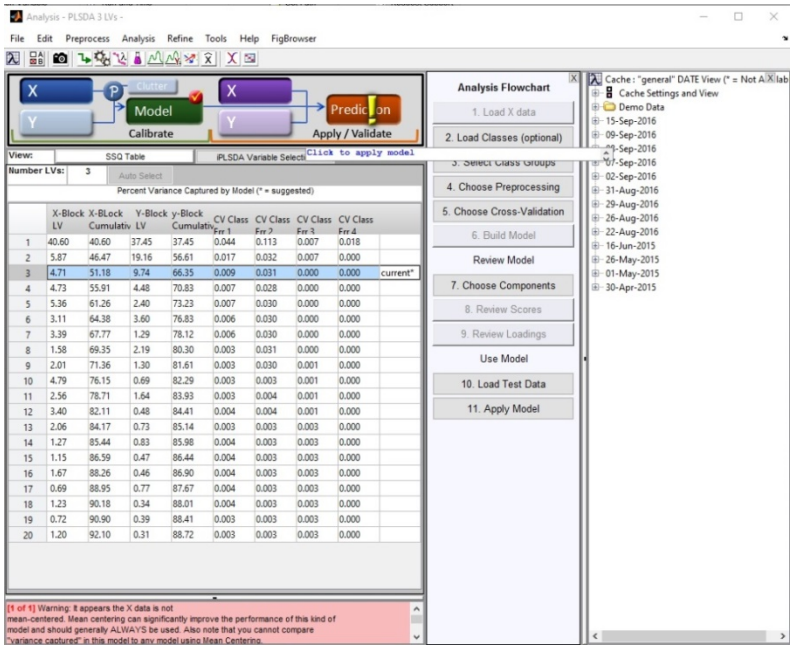


Fig. 32. Prediction button lights with the yellow exclamation mark.

Click on the purple Erlenmeyer () again to view scores and sample statistics of the test samples. It is, for example, very useful to look at the class predictions. A set of test samples is used here as example in Fig. 33. Here it is possible to see that the majority of the samples are strictly predicted (only two samples are not) (**A**); and these samples are strictly predicted as American cochineal (Class 1, Class 5 - **C**). If considering most probable class predictions, all samples are considered as American cochineal (**B**), and none as Polish cochineal (**E**), or a mixture of American cochineal and kermes (**D**).

The difference between strict and most probable class predictions is explained below in the interpretation section.

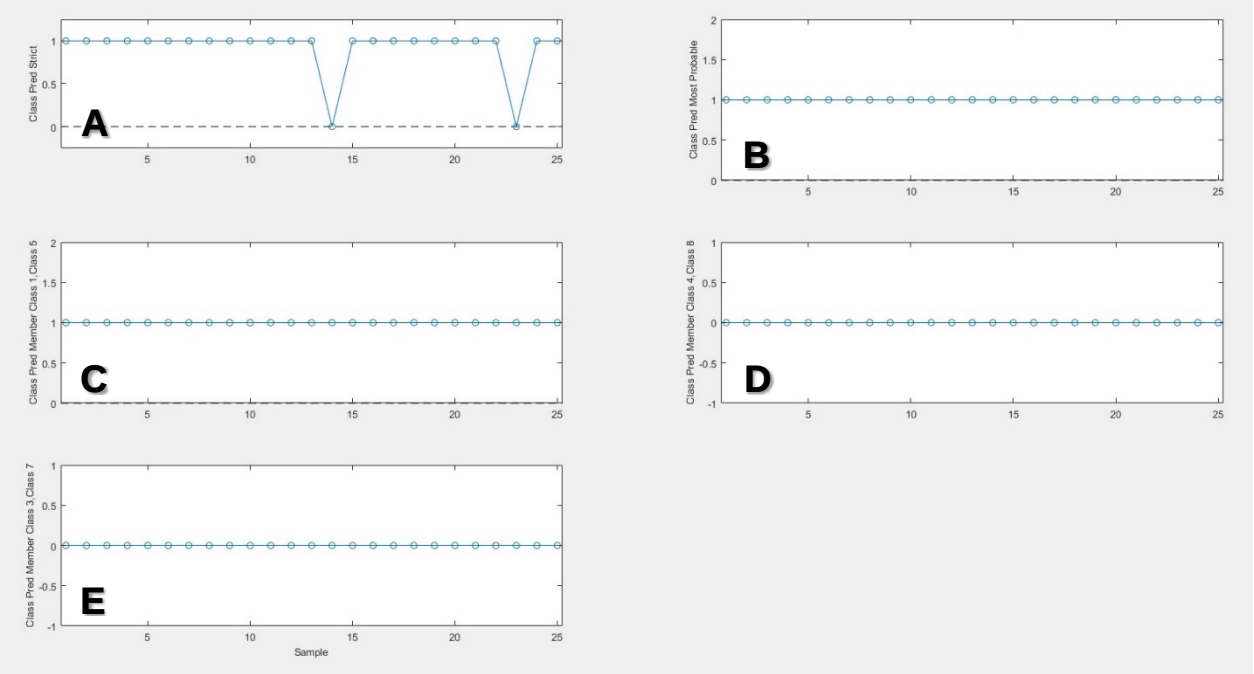


Fig. 33: Class predictions of the test samples.

To plot together the calibration samples and the test samples, along with their predictions, they need to be firstly exported as .csv files. For this, click “File” (at the upper toolbar of the PLS toolbox), “Export Predictions”, and then, select “Calibration” and “Test/Validation” (Fig. 34). Save these files (name of the files or the location where they are saved is not specific).

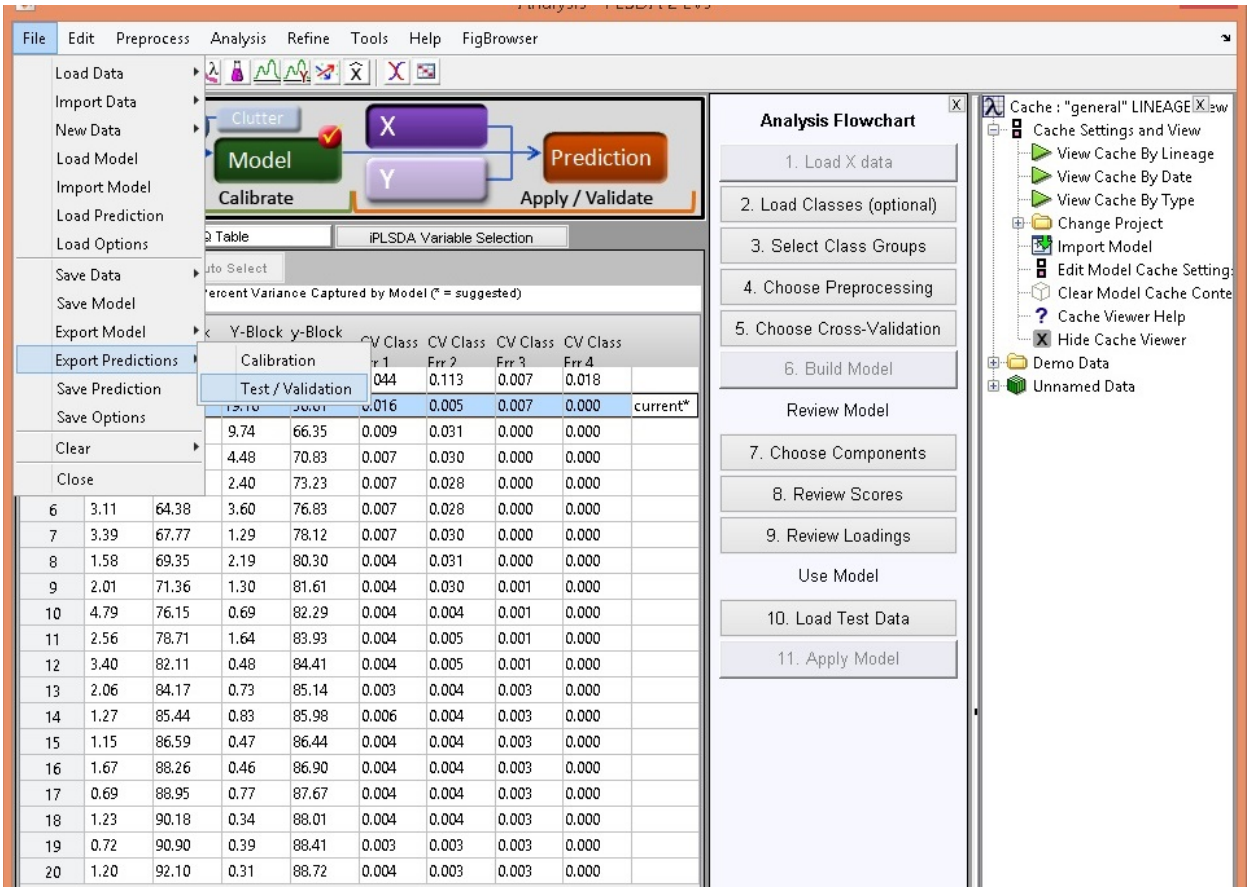


Fig. 34: Exporting the calibration and test samples with respective predictions.

In Matlab, type “CochinealProjections” in the Matlab command window, in order to plot the projections of the test samples together with the calibration samples. A window will be prompted to select a file containing the calibration scores and predictions, and a file containing the test scores and predictions (Fig. 35). These are the files that have just been exported from the PLS toolbox.

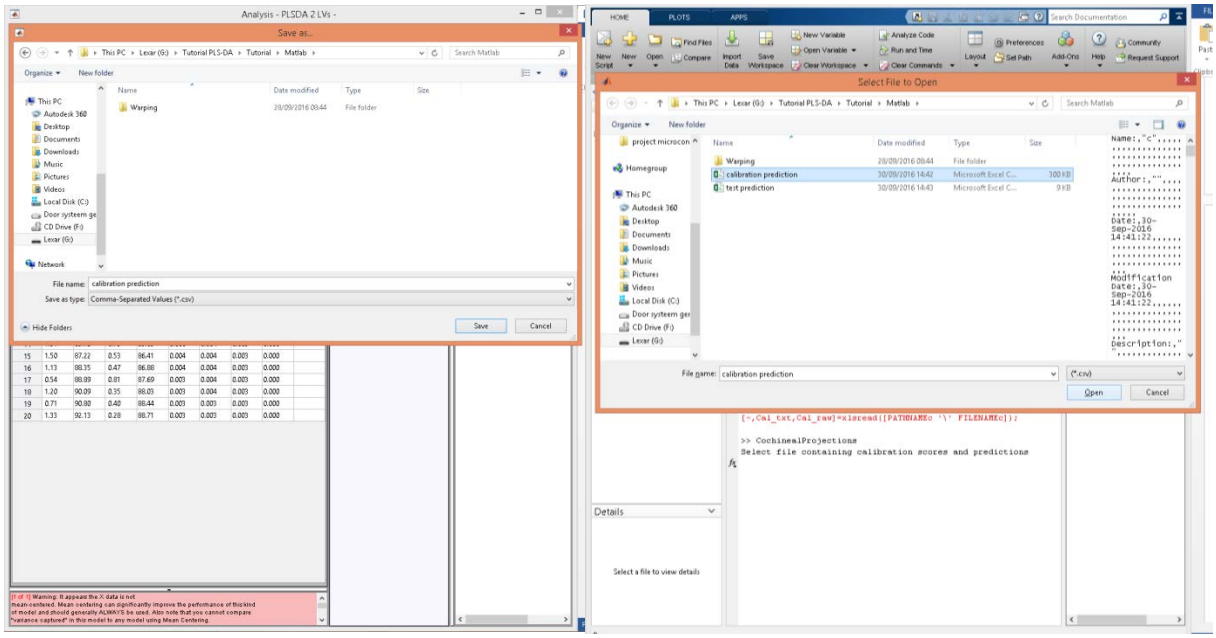


Fig. 35: Exporting the calibration scores, test scores and predictions from the PLS toolbox and importing them into Matlab.

If an error occurs, related to the excel reading of the .csv files, it is possible that the data in these files is not properly divided by columns. If this is the case, try to find an option for changing the “List separator”. This option can either be found in the Excel software or in the additional settings of Windows language settings (the authors did not try other operating systems). In these settings, if there is a semi-colon (;), change it to a colon (,), as in Fig. 36.

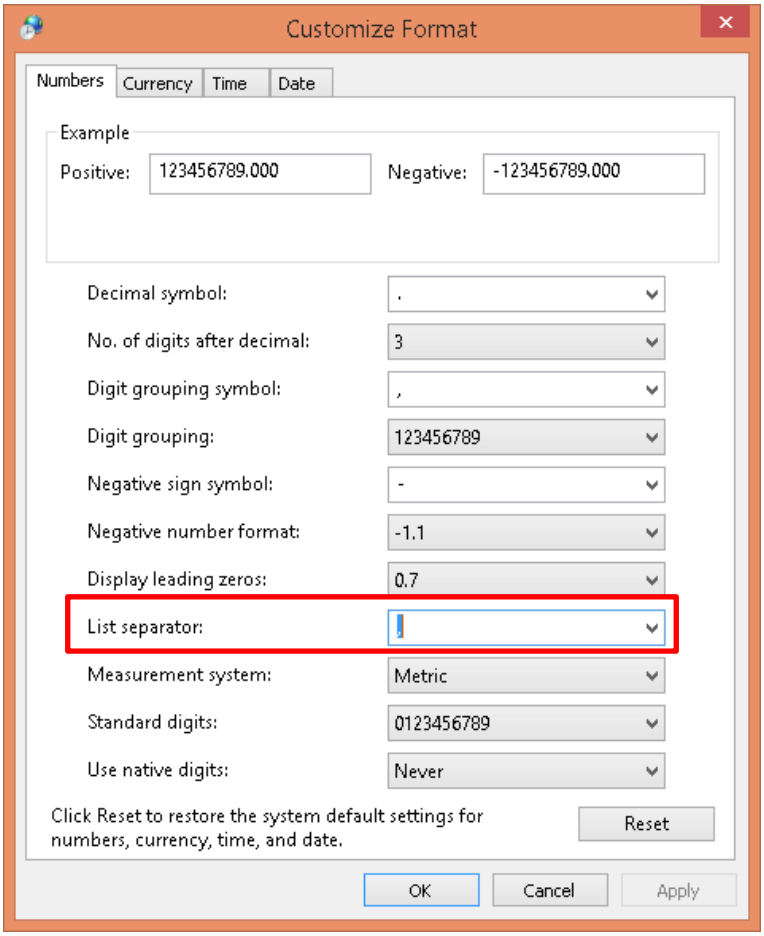


Fig. 36: Change of list separator, in case of .csv reading error.

Several figures will open in Matlab. The number of figures depends on the number of latent variables of the model, as all possible combinations of latent variables are plotted against each other. Usually (and in this case), LV1 vs LV2 is the most informative plot (Fig. 37), because these latent variables explain most of the variance – however, in other cases, it is possible that other projections can be useful as well.

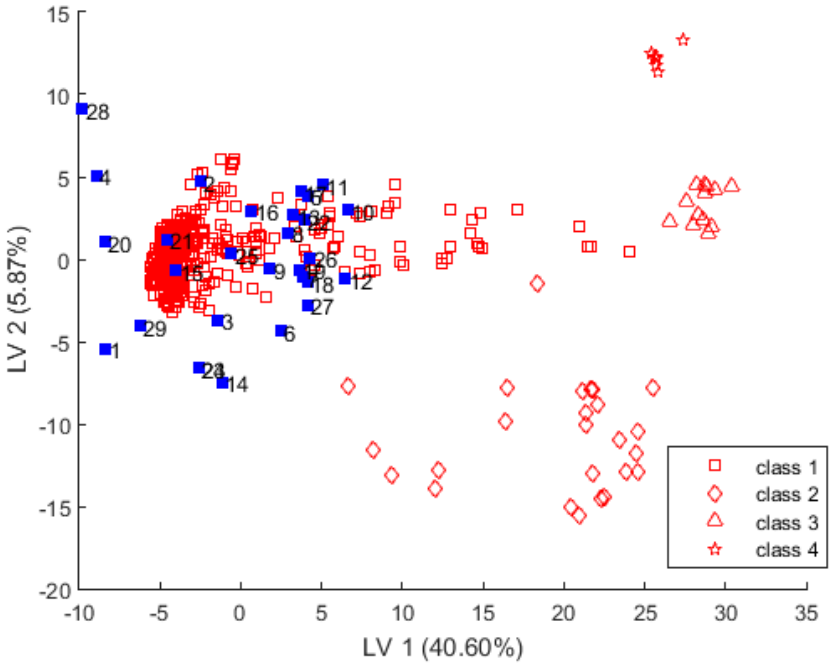


Fig. 37: PLS-DA plot of calibration and test scores, corresponding to dyed and aged silk fibres (red scores) and to silk historical fibres (blue scores).

In Fig. 37, the calibration scores (red triangles, squares, lozenges and stars) correspond to both aged and non-aged samples. As previously specified in "3.Select Class Groups" of the PLS toolbox, the aged and non-aged samples are plotted as part of the same group (cannot be distinguished). Thus, class 1 corresponds to American cochineal (squares), class 2 to Armenian cochineal (lozenges), class 3 to Polish cochineal (triangles) and class 4 to the mixture of American cochineal and kermes (stars).

It is possible that only test samples (blue scores) with strict predictions are shown. In the case of Fig. 37, all historical samples are strictly predicted as American cochineal. The numbers next to each score correspond to every historical sample labelled originally in the Excel file "Historical samples" (FILE 2).

To depict the most probable predictions of all samples, type “edit CochinealProjections” in the Matlab command window - or double click on the script, shown on the left tab “Current folder”. Once it is opened, change the prediction setting in the top section of the script, from “strict” to “prob” (Fig. 38).

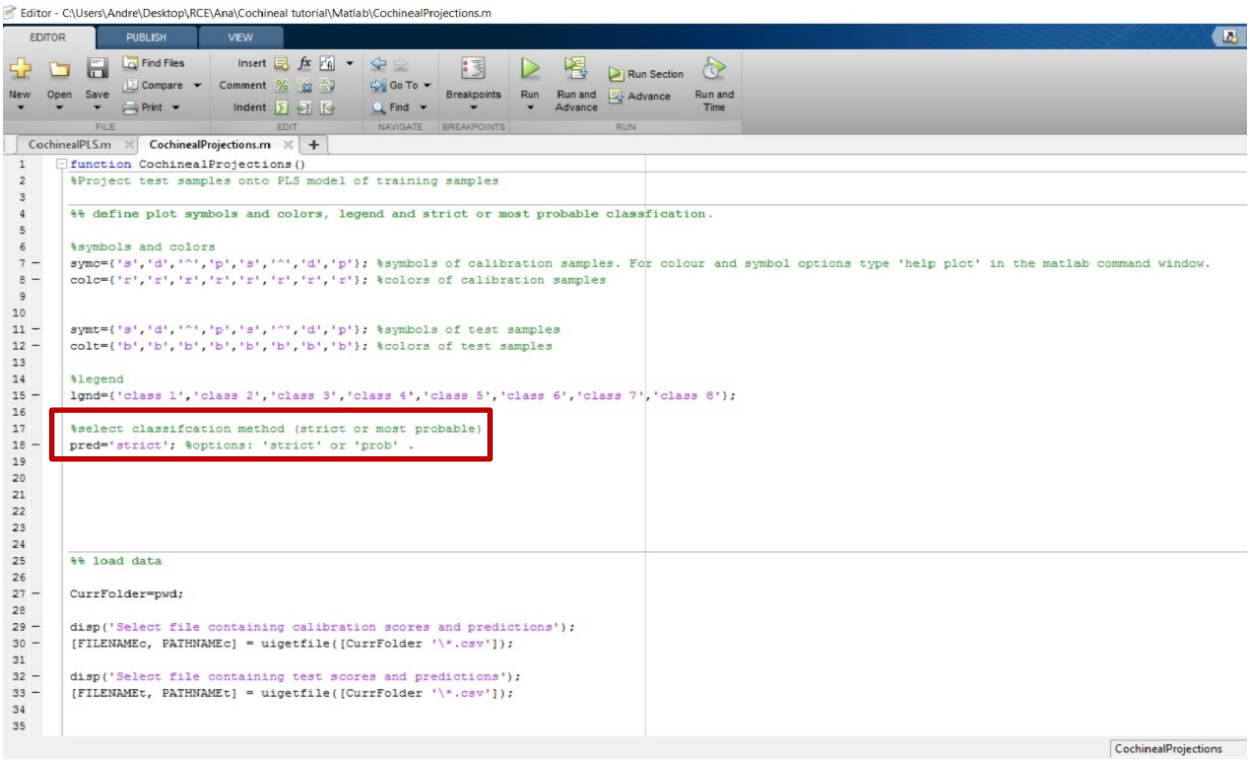


Fig. 38: Changing the prediction setting.

In this script, it is also possible to change the legend and the plot colours and symbols.

However, this is also possible to be undertaken directly on the figure. For this, go back to the screen of the figure and click on the cursor symbol at the top ( ). Then, double click on the figure to open the Property Editor (Fig. 39). Here, it is possible to change the font of the axes and legend, as well as the colour and shape of the plots – double click on the plots and a new tab will open on the bottom of the screen (Fig. 40).

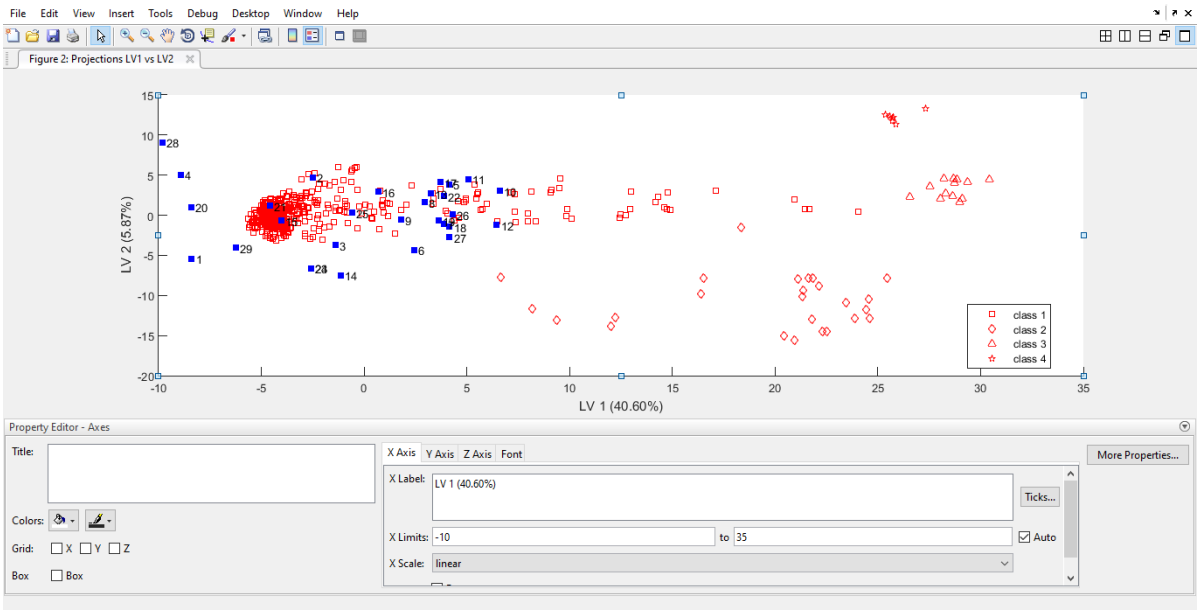


Fig. 39: Editing tab for the font of the axes and legend of the figure.

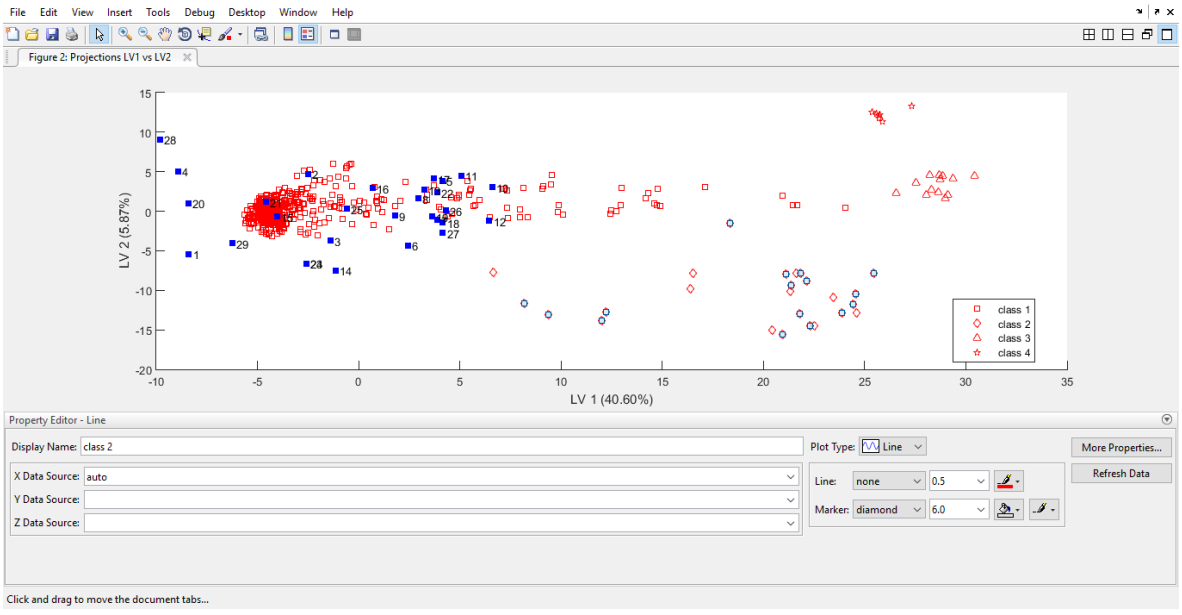
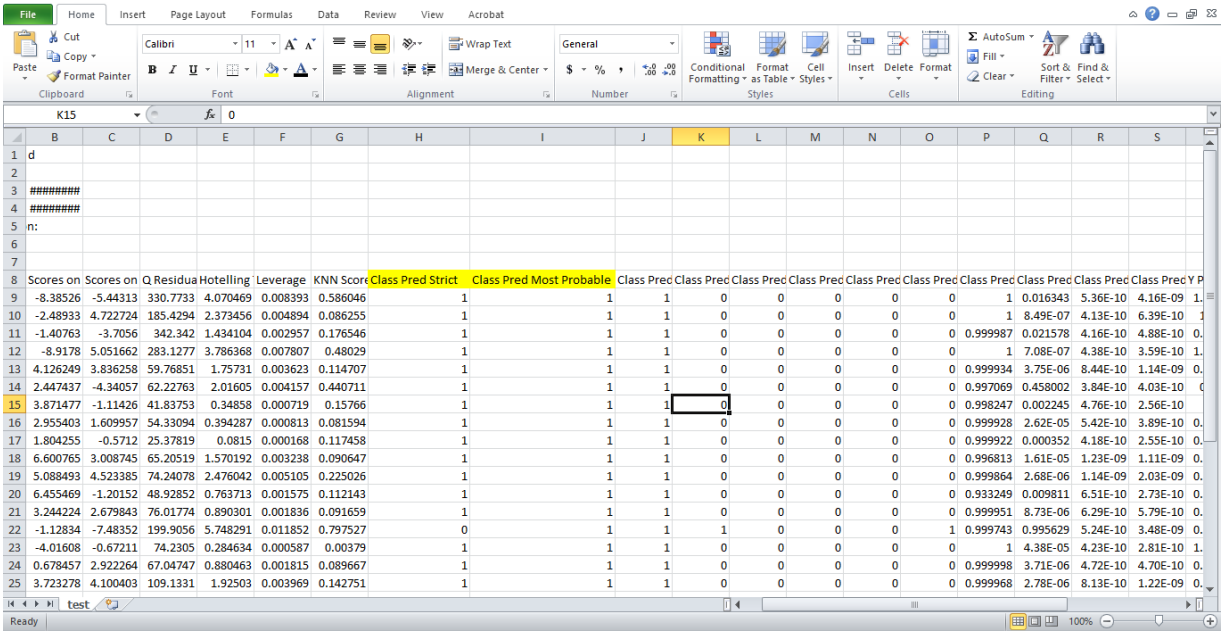


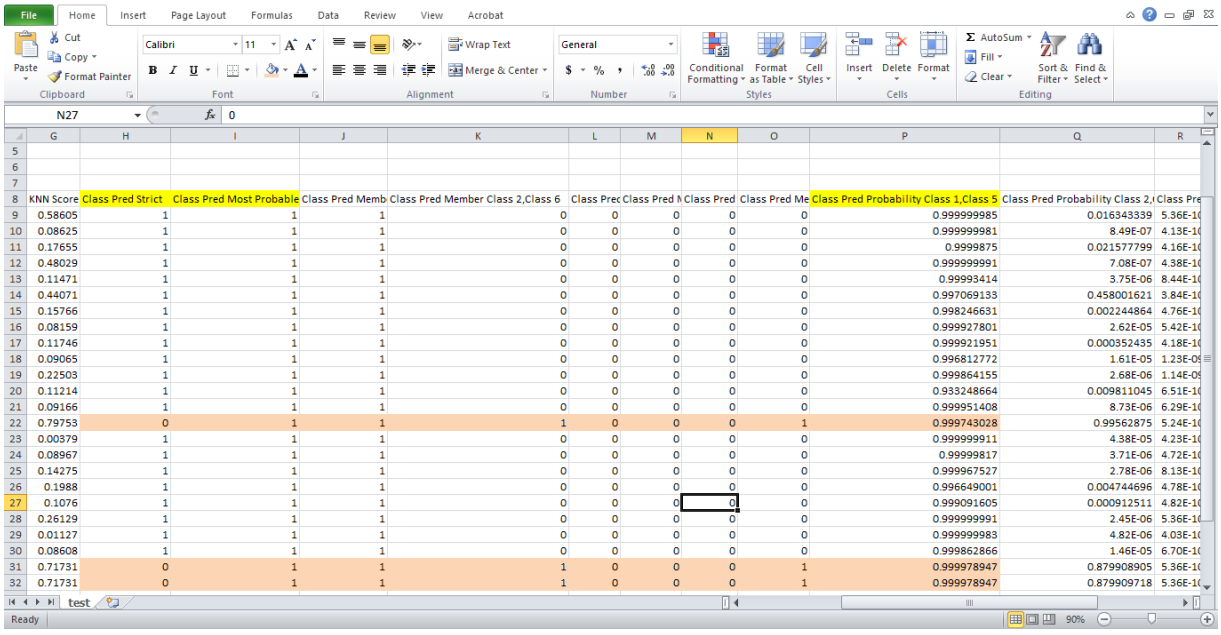
Fig. 40: Editing tab for colours and shapes in the figure.

To better understand the strict and most probable predictions of the characterized historical samples, open the exported .csv file of test scores with Excel (Fig. 41). In the first column (A), add the labels of each of the historical samples, originally described in the Excel file "Historical samples" (FILE 2). Follow the exact order of samples given in FILE 2; the number of rows should then correspond in both files. The following columns describe each sample, namely in terms of their scores on LV1 and LV2, for instance.



	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	d																	
2																		
3	########																	
4	########																	
5	n:																	
6																		
7																		
8	Scores on	Scores on	Q Residua	Hotelling	Leverage	KNN Score	Class Pred Strict	Class Pred Most Probable	Class Pred	Class Pred	Class Pred	Class Pred	Class Pred	Class Pred	Class Pred	Class Pred	Class Pred	Y P
9	-8.38526	-5.44313	330.7733	4.070469	0.008393	0.586046	1	1	1	0	0	0	0	0	1	0.016343	5.36E-10	4.16E-09	1.
10	-2.48933	4.722724	185.4294	2.373456	0.004894	0.086255	1	1	1	0	0	0	0	0	1	8.49E-07	4.13E-10	6.39E-10	1
11	-1.40763	-3.7056	342.342	1.434104	0.002957	0.176546	1	1	1	0	0	0	0	0	0.999987	0.021578	4.16E-10	4.88E-10	0.
12	-8.9178	5.051662	283.1277	3.786368	0.007807	0.48029	1	1	1	0	0	0	0	0	1	7.08E-07	4.38E-10	3.59E-10	1.
13	4.126249	3.836258	59.76851	1.75731	0.003623	0.114707	1	1	1	0	0	0	0	0	0.999934	3.75E-06	8.44E-10	1.14E-09	0.
14	2.447437	-4.34057	62.22763	2.01605	0.004157	0.440711	1	1	1	0	0	0	0	0	0.997069	0.458002	3.84E-10	4.03E-10	0
15	3.871477	-1.11426	41.83753	0.34858	0.000719	0.15766	1	1	1	0	0	0	0	0	0.998247	0.002245	4.76E-10	2.56E-10	
16	2.955403	1.609957	54.33094	0.394287	0.000813	0.081594	1	1	1	0	0	0	0	0	0.999928	2.62E-05	5.42E-10	3.89E-10	0.
17	1.804255	-0.5712	25.37819	0.0815	0.000168	0.117458	1	1	1	0	0	0	0	0	0.999922	0.000352	4.18E-10	2.55E-10	0.
18	6.600765	3.008745	65.20519	1.570192	0.003238	0.090647	1	1	1	0	0	0	0	0	0.996813	1.61E-05	1.23E-09	1.11E-09	0.
19	5.088493	4.523385	74.24078	2.476042	0.005105	0.225026	1	1	1	0	0	0	0	0	0.999864	2.68E-06	1.14E-09	2.03E-09	0.
20	6.455469	-1.20152	48.92852	0.763713	0.001575	0.112143	1	1	1	0	0	0	0	0	0.933249	0.009811	6.51E-10	2.73E-10	0.
21	3.244224	2.679843	76.01774	0.890301	0.001836	0.091659	1	1	1	0	0	0	0	0	0.999951	8.73E-06	6.29E-10	5.79E-10	0.
22	-1.12834	-7.48352	199.9056	5.748291	0.011852	0.797527	0	1	1	1	0	0	0	1	0.999743	0.995629	5.24E-10	3.48E-09	0.
23	-4.01608	-0.67211	74.2305	0.284634	0.000587	0.00379	1	1	1	0	0	0	0	0	1	4.38E-05	4.23E-10	2.81E-10	1.
24	0.678457	2.922264	67.04747	0.880463	0.001815	0.089667	1	1	1	0	0	0	0	0	0.999998	3.71E-06	4.72E-10	4.70E-10	0.
25	3.723278	4.100403	109.1331	1.92503	0.003969	0.142751	1	1	1	0	0	0	0	0	0.999968	2.78E-06	8.13E-10	1.22E-09	0.

Fig. 41: Exported .csv file with strict and most probable predictions of historical samples.



	G	H	I	J	K	L	M	N	O	P	Q	R
5												
6												
7												
8	KNN Score	Class Pred Strict	Class Pred Most Probable	Class Pred Memb	Class Pred Member Class 2,Class 6	Class Prec	Class Pred M	Class Pred	Class Pred Me	Class Pred Probability Class 1,Class 5	Class Pred Probability Class 2,(	Class Pre
9	0.58605	1	1	1	0	0	0	0	0	0.999999985	0.016343339	5.36E-10
10	0.08625	1	1	1	0	0	0	0	0	0.999999981	8.49E-07	4.13E-10
11	0.17655	1	1	1	0	0	0	0	0	0.9999875	0.021577799	4.16E-10
12	0.48029	1	1	1	0	0	0	0	0	0.999999991	7.08E-07	4.38E-10
13	0.11471	1	1	1	0	0	0	0	0	0.99993414	3.75E-06	8.44E-10
14	0.44071	1	1	1	0	0	0	0	0	0.997069133	0.458001621	3.84E-10
15	0.15766	1	1	1	0	0	0	0	0	0.998246631	0.002244864	4.76E-10
16	0.08159	1	1	1	0	0	0	0	0	0.999927801	2.62E-05	5.42E-10
17	0.11746	1	1	1	0	0	0	0	0	0.999921951	0.000352435	4.18E-10
18	0.09065	1	1	1	0	0	0	0	0	0.996812772	1.61E-05	1.23E-09
19	0.22503	1	1	1	0	0	0	0	0	0.999864155	2.68E-06	1.14E-09
20	0.11214	1	1	1	0	0	0	0	0	0.933248664	0.009811045	6.51E-10
21	0.09166	1	1	1	0	0	0	0	0	0.999951408	8.73E-06	6.29E-10
22	0.79753	0	1	1	1	0	0	0	1	0.999743028	0.99562875	5.24E-10
23	0.00379	1	1	1	0	0	0	0	0	0.999999911	4.38E-05	4.23E-10
24	0.08967	1	1	1	0	0	0	0	0	0.99999817	3.71E-06	4.72E-10
25	0.14275	1	1	1	0	0	0	0	0	0.999967527	2.78E-06	8.13E-10
26	0.1988	1	1	1	0	0	0	0	0	0.996649001	0.004744696	4.78E-10
27	0.1076	1	1	1	0	0	0	0	0	0.999091605	0.000912511	4.82E-10
28	0.26129	1	1	1	0	0	0	0	0	0.999999991	2.45E-06	5.36E-10
29	0.01127	1	1	1	0	0	0	0	0	0.999999983	4.82E-06	4.03E-10
30	0.08608	1	1	1	0	0	0	0	0	0.999862866	1.46E-05	6.70E-10
31	0.71731	0	1	1	1	0	0	0	1	0.999978947	0.879908905	5.36E-10
32	0.71731	0	1	1	1	0	0	0	1	0.999978947	0.879909718	5.36E-10

Fig. 42: Comparison between strictly and most probable predicted samples.

To understand which samples are strictly predicted, examine the column “Class Pred Strict”. In the case of Figs. 41 and 42, all samples displaying “1” are strictly predicted as American cochineal (Class 1), but samples displaying “0” are not strictly predicted. The next column “Class Pred Most Probable” shows that, despite not strictly predicted, these samples are still predicted as American cochineal (Class 1 – “1”). The probability of this attribution is expressed in a scale from 0 to 1, in the column “Class Pred Probability Class 1,Class 5”. Here, it is possible to observe that these most probable predictions have a probability of almost 1 or, in other words, of almost 100%.

However, it is worthwhile noticing that, although these samples are attributed as American cochineal (also in the column “Class Pred Member Class 1,Class 5” – “1”), they are also attributed as Armenian cochineal in “Class Pred Member Class 2,Class 6” (“1”). Indeed, despite the fact that the “Class Pred Probability Class 1,Class 5” column gives almost 100% probability of American cochineal attribution, a very high percentage is also given in “Class Pred Probability Class 2,Class 6”. This means that the model is not able to strictly predict these samples either into the American or the Armenian cochineal classes.

While in many cases the percentage of probable predictions can help interpreting the results of non-strictly predicted samples, it is always important to compare these results with the qualitative UHPLC-PDA interpretations, as well as with the provenance and the date of the textiles, if available.

Further notes for the interpretation of PLS-DA models

PLS is a projection method and its correspondent PLS model is defined by scores and loadings (similar to PCA). The scores indicate the position of samples on latent variables and the scoreplot shows the relationship between samples. The loadings indicate how the latent variables are constructed from the original variables. The loadings reveal which chromatographic regions are important for a specific latent variable (every latent variable has its own loadings). Samples with a high positive score on a latent variable have relatively large peaks in chromatographic areas that have big positive loadings and relatively small peaks in chromatographic areas that have big negative loadings. When a sample has a high negative score, it is the other way around.

The PLS algorithm finds a projection of the data which is most suited to predict a Y variable. In the PLSDA model described above, the Y variable contains class labels and this ensures an optimal separation between the classes in the scoreplot. A class label 1 indicates that the sample belongs to the class and a 0 indicates that the sample does not belong to the class. When the predicted Y value is above a certain threshold (between 0 and 1) the sample is predicted to belong to the class. When a data set contains multiple classes, Y is a matrix with a column for each class and membership is predicted to each class separately.

Because separate predictions are made for each class, it is possible that predictions are ambiguous. It could happen that a sample is predicted to belong to no class at all, or to multiple classes. Therefore the PLS toolbox outputs both *most probable* and *strict* predictions. The most probable prediction is the class for which the sample has the highest probability of belonging to that class. It may be interesting to see which class is most probable, but there is a high risk of misclassification (both when several classes are probable and when no class is probable). Therefore a more conservative approach is to use strict predictions. A strict prediction only classifies a sample when it belongs to exactly one class.